New Paradigms and Optimality Guarantees in Statistical Learning and Estimation

Yu-Xiang Wang

CMU-ML-17-105
December 2017

Machine Learning Department
School of Computer Science
& Department of Statistics and Data Science
Dietrich College of Humanities and Social Sciences
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Thesis Committee:
Ryan J. Tibshirani, Chair
Stephen E. Fienberg
Jing Lei
Alexander J. Smola
Adam D. Smith, Boston University

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research was supported by NSF Grant BCS-0941518, DMS-1309174, DMS-1554123, a grant by Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, a grant from Adobe, a gift from Google and a Baidu Scholarship.

Keywords: machine Learning, statistics, differential privacy, nonparametric regression, trend filtering, contextual bandits, off-policy evaluation, adaptive data analysis, sequential selective estimation

To my family and to Steve.

Abstract

Machine learning (ML) has become one of the most powerful classes of tools for artificial intelligence, personalized web services and data science problems across fields. Within the field of machine learning itself, there had been quite a number of paradigm shifts caused by the explosion of data size, computing power, modeling tools, and the new ways people collect, share, and make use of data sets.

Data privacy, for instance, was much less of a problem before the availability of personal information online that could be used to identify users in anonymized data sets. Images, videos, as well as observations generated over a social networks, often have highly localized structures, that cannot be captured by standard nonparametric models. Moreover, the "common task framework" that is adopted by many subdisciplines of AI has made it possible for many people to collaboratively and repeated work on the same data set, leading to implicit overfitting on public benchmarks. In addition, data collected in many internet services, e.g., web search and targeted ads, are not iid, but rather feedbacks specific to the deployed algorithm.

This thesis presents technical contributions under a number of new mathematical frameworks that are designed to partially address these new paradigms.

- Firstly, we consider the problem of statistical learning with privacy constraints. Under Vapnik's general learning setting and the formalism of differential privacy (DP), we establish simple conditions that characterizes the private learnability, which reveals a mixture of positive and negative insight. We then identify generic methods that reuses existing randomness to effectively solve private learning in practice; and discuss weaker notions of privacy that allows for more favorable privacy-utility tradeoff.
- Secondly, we develop a few generalizations of trend filtering, a locally-adaptive nonparametric regression technique that is minimax in 1D, to the multivariate setting and to graphs. We also study specific instances of the problems, e.g., total variation denoising on d-dimensional grids more closely and the results reveal interesting statistical computational trade-offs.
- Thirdly, we investigate two problems in sequential interactive learning: a) off-policy evaluation in contextual bandits, that aims to use data collected from one algorithm to evaluate the performance of a different algorithm; b) the problem of adaptive data analysis, that uses randomization to prevent adversarial data analysts from a form of "p-hacking" through multiple steps of sequential data access.

In the above problems, we will provide not only performance guarantees of algorithms but also certain notions of optimality. Whenever applicable, careful empirical studies on synthetic and real data are also included.

Acknowledgments

The work presented in this document could not have been possible without the help of my amazing group of advisors, Steve Fienberg, Ryan Tibshirani, Alex Smola and Jing Lei. I feel extremely privileged to have also worked closely with each of them during my years in Carnegie Mellon. I have learned so much from them to the point that the mixture distribution of their way of thinking and their taste of research are now part of who I am.

I am especially grateful for Steve Fienberg's wisdom and perspective as both a research advisor and a mentor. Despite his excruciating fight with cancer that lasted my entire PhD candidature, Steve had been able to take very good care of me both academically and personally. He gave me freedom to pursue my broad interest, connected me to the best people in each specific research topic, made himself available for technical discussions and even coached me in writing and presentation. Steve is everything a PhD student could have asked for in an advisor. I will always remember Steve fondly for every moment that I had the fortune to share with him and I will always miss his iconic heart-warming wink!

After Steve passed away in Dec 2016, Ryan Tibshirani and Jing Lei took over as co-advisors and carried me through graduation. As a matter of fact, I had been receiving a tremendous amount of training in theoretical statistical research by working closely with these two young and shining stars in statistics since 2013. Their acute instinct in research, deep theoretical insight and the meticulous pursuit of mathematical rigor have repeatedly led research projects to those "Hooray!" moments.

Alex Smola had been my de facto co-advisor in the machine learning side since the very beginning. He is clearly one of the sharpest and most knowledgeable person I have ever met. Discussions with him always keep me inspired and a little overwhelmed at the same time. It happened more than once that only after several days did I truly appreciate Alex's point and exclaimed "Oh I see! That is brilliant, Alex!" Alex also challenged me to "get out of the comfort zone" and to read more broadly so as to "connect the dots". These are things that I appreciated even more as I get further along in my career.

Besides my advisors, the thesis also contains collaborative work with my peers Veeru Sadhanala, James Sharpnack (Chapter 6 and Chapter 7), as well as Alekh Agarwal and Miro Dudik from Microsoft Research NYC (Chapter 9). I could not have completed these work without them. I would like to express my special thanks to Adam Smith. Discussion with Adam at various privacy-related meetings allowed us to significantly improve the learnability work, presented in Chapter 2, as well as the sequential selective estimation work presented in Chapter 10.

During my four years at CMU, I had the pleasure of collaborating with a large group of brilliant researchers and writing papers on a variety of topics that are not directly related to this thesis. These awesome folks are Yining Wang, Ziqi Liu, Aarti Singh, Kyle Soska, Mu Li, Seth Flaxman, Dougal Sutherland, David Wei Dai, Willie Neiswanger, Eric Xing, Suvrit Sra and Xi Chen. I really enjoyed working with each of them and I have learned a whole lot out of each paper we wrote.

I would like to thank the faculty members who prepared excellent course materials and taught with excitement and insight. I particularly enjoyed Ryan Tibshirani's convex optimization, Gary Miller's spectral graph theory, Alex Smola and Suvrit Sra's Advanced Optimization and Randomized algorithm, Eric Xing's probabilistic graphical models, Jing Lei's Intermediate Statistics, Steve Fienberg's Data Privacy, Jiashun Jin's Advanced statistics theory, Larry Wasserman and Ryan Tibshirani's Statistical Machine Learning and the Advanced Data Analysis courses ran by Valerie Ventura.

I also would like to thank other faculty members who I had memorable and helpful discussions with. These include Geoff Gordon, Christos Faloutsos and Manuela Veloso from machine learning, Avrim Blum and Gary Miller from the computer science department, Larry Wasserman and Edward Kennedy from the statistics department, Kun Zhang from the philosophy department, David Krackhardt, Beibei Li and Siyuan Liu from Heinz college and so on. I appreciate their time and effort in sharing their expertise.

Special thanks to the staff members of the university for creating a supportive and efficient work environment. In particular, I owe my gratitude to Diane Stidle who acted as the caring "mom" to me (and everyone) in the MLD family, to Sharon Cavlovich who helped to send Steve's recommendation letter out countless number of times during my job market search as well as to Mallory Deptola and Barbara Diecks who helped with all my reimbursement for conference trips.

Over the years, I have had numerous ad hoc discussions with my fellow PhD students and resident postdocs. In particular, I probably have bugged Yining Wang, Veeru Sadhanala, Yaoliang Yu, Andrew Wilson, Yifei Ma, Zichao Yang, Alnur Ali, Manzil Zaheer, Kirthevansan Kandasamy, Aaditya Ramdas, Ahmed Hefny, Sashank Reddi, Jisu Kim and Ina Fiterau a lot more than others. I personally find these discussions very enlightening. I feel lucky to be part of a talented and lively community.

I am also grateful for all the mentorship and career tips that I received from people not in CMU. The list includes Huan Xu, Suvrit Sra, Miro Dudik, Alekh Agarwal, Lihong Li, Ju Sun, John Wright, Cynthia Dwork, Adam Smith, Sesa Slakovic, Beka Steorts, Theo Damoulas, Yuchen Zhang, Xiaolin Yang, Anima Anandkumar, Edo Liberty, William Wang, Xifeng Yan and many others. Their timely advice helped me to make informed decisions in troubled times.

I would like to thank my office mates, Micol, Alnur, Hyun Ah, Kenny and my friends Qiong Zhang, Nick Kim, Ying Yang, Yifei Ma, Gus Xia, Jing Xiang, Carlton Downey, Adams Wei Yu, Diyi Yang, Justin Hyun, Yan Xia, Bin Deng, Zhong Zhou, Han Zhao, Yu Liu, Yuling Shi and many others who I cannot enumerate for supporting me on this fun journey. Many of them will surely remember our epic bbq parties.

Last but not least, I would like to thank my parents for their unconditional love and support, my wife Su for staying married with me despite many of my late night deadline rushes, and my son Joshua Jia-Shu Wang for descending upon us and being the cutest baby ever born.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Machine learning (ML) is an interdisciplinary field that studies how computers can learn to perform tasks, e.g., prediction, decision making, by looking at examples, and keep improving as it sees more examples. The idea of machine learning dates back to Alan Turing [227] and his seminal 1950 article on “computing machinery and intelligence”, where he describes emphatically a “learning machine” that can “learn from experience”. The first such machine that was designed is arguably Rosenblatt’s perceptron algorithm [179] in 1958. However, limited by the computational resources, systematic studies and practical use of machine learning have not really started until late 1980s.

Classical applications of machine learning include recognizing spoken words [141] and handwritten digits [139], detecting fraudulent activities in credit card transactions [52], indexing and summarizing complex collection of textual documents [34, 63], finding hidden group structures in social networks [5], recommending new movies to users based on their past viewing/rating history [132], labeling sentences with part-of-speech tags [135] and so on.

These applications are supported by decades of theoretical research in machine learning that provides concrete theoretical guarantees for many learning algorithms. In particular, the field of statistical learning focuses on understanding learning algorithms statistically and characterizing the relationship among the number of training examples, prediction accuracy and computational cost [229, 234].

Thanks to the explosion of data size and the availability of high performance computing lately, machine learning has become one of the most powerful classes of tools for artificial intelligence, personalized web services and data science problems across many fields. Machine learning has been increasingly recognized and deployed as the state-of-the-art methods for industry-scale computer vision problems such as object recognition/segmentation with a thousand classes, natural language processing problems such as parsing, sequence labeling and machine translation (see [140] and the references therein). As of today, technologies empowered by machine learning have already revolutionized many aspects of our daily life.

The wide applicability of machine learning also poses new theoretical challenges to the field itself. Many such challenges require us to incorporate new considerations or revisit the underlying assumptions of statistical learning. Here is a partial list of them:

- The use of ML in medical, financial, political and many other sensitive areas is greatly limited due to privacy concerns. Formal tools for data sharing, data processing and running machine learning without breaching the privacy of individual data record are much needed but lack mainstream attention.
- New data formats such as images, videos, geographic regions, social networks are often modeled as noisy observations of underlying signals with certain regularity structures. Standard denoising techniques based on kernels and splines often require the underlying signals to be homogeneous smooth, therefore are unable to effectively denoise and preserve fine-grained localized structures at the same time. Can we weaken such assumptions, e.g., to allowing heterogeneity in smoothness without losing much in statistical and computational complexity? Or is there a price to pay?
- Learning systems are often interactive. In many ML systems, e.g., web search, recommendation system, the algorithms being used will largely affect the data being collected. In this case, how can we evaluate new algorithms and make sure they outperform old ones without actually running them on real users?
- Real-life applications of ML are often sequential and with humans in the loop. The choices of models for your next ImageNet entry are likely to be based on previously reported results. Similarly, the scientific questions to investigate and results to include into your next publication on genome-wise association studies will most likely be an outcome of an exploratory data analysis and some manual eye-balling of visualized results. Such selections can implicitly cause overfitting, which is largely the root of false discoveries in science [86, 138] and overly optimistic performance numbers in computer vision benchmarks [226].

This thesis focuses on formally addressing these questions under a range of practically-applicable mathematical frameworks, which I refer to as “new paradigms”. For instance, in private machine learning, the new paradigm is to only privately release the recommendation engine as web services rather than trying to privatize the raw data set. Similarly, in “adaptive data analysis”, we consider the case when each model is chosen adaptively based on the results of previous models.

The remainder of the thesis will be broken into three parts.

The first part describes a sequence of investigations on the theory and algorithms of private statistical learning — a setting that aims at designing algorithms to learn from a data set as a whole while protecting individual records in the data set from being identified. We will be working with differential privacy [84] and its variants and establish intriguing connections between the task of privacy and learning-theoretic quantities.

The second part will be devoted to efforts in generalizing “trend filtering” [129, 221], a recent development in locally adaptive nonparametric regression, to graphs and higher dimensions, such that it is more broadly applicable to a wide range of practical problems.

The third part contains new technical results on two modern sequential learning models: contextual bandits and adaptive data analysis that are designed to address respectively the third and fourth problem listed above.

The main contribution of these thesis will be about theoretical analysis that yields upper bounds for learning performance guarantees and information-theoretical lower bounds for certifying optimality and/or revealing (sometimes surprising) fundamental limits. From my humble point of view, theoretical guarantees are still highly relevant for ML methods, because one can never enumerate enough empirical evidence to conclude that the method will work on new problems, or to be confident about whether a particular approach cannot be improved further for solving a class of problems.

1.1 Organization and notation

As we explained earlier, the thesis will be organized in to three parts, with each part consisting a few chapters, presenting a sequence of work under that category. Each chapter is based on a previously published article (or one that is under review), and will be made as independent as possible. They will have their own introduction and conclusion as well as self-contained proofs of technical results. In this way, readers can pick up any chapter and understand the context of the results without needing to keep referring to other parts of the thesis.

We will however provide a synopsis at the beginning of each part to sketch the "big picture", which motivates the general line of research and explains the connections and dependence of each chapters within that part. At the end of every part, we will briefly describe the subsequent work in the line of research and applications including both our own work that did not make to the thesis and follow-up work in the literature.

As we will be presenting a very heterogeneous list of topics, no effort has been made to ensure all chapters use the same notations. However, standard mathematical notations will remain the same and consistent throughout the thesis. We will define the specific notations in each chapter separately.

Part I

Towards practical machine learning with privacy guarantee

Motivation and overview

Increasing public concerns regarding data privacy have posed obstacles in the development and application of new machine learning methods in medical, financial, political and many other sensitive areas as data collectors and curators may no longer be able to share data for research purposes. Here is an example of the many requests from practitioners that our group has received (with direct identifiers removed):

> “Please let me introduce myself, my name is ******, and I am a researcher at the Office of Financial Research (OFR), at the ******. The OFR is charged at providing analytics and data driven research for monitoring and managing financial stability, across the regulatory community. One major issue we have been facing is how to share confidential data and analysis across regulatory agencies, or when collaborating with academics outside of the federal government.”

In today’s big data era, stories about privacy breaches from even anonymized data appear regularly (e.g., medical records,[1] Netflix,[2] NYC Taxi,[3]) and the general public is becoming increasingly informed and also alarmed about how their data are being used. We know that the potential harm is not limited to the data release itself but also inferences that can be made about individuals with the potential aid of auxiliary information. Therefore, it is not surprising that individuals and data collectors are reluctant to share their data, even when there is a strong scientific incentive to do so.

We attribute such conservativeness in data sharing in part to the lack of *practical* statistical machine learning tools for conducting private data analysis. Existing techniques include statistical confidentiality and disclosure control [76, 117] developed in the statistics community beginning in the 1970s and differential privacy (see, e.g. [80]) popularized by the theoretical computer science and cryptography communities since 2006 [77, 84]. The traditional statistical approaches focus on the macroscopic statistical tradeoff between disclosure risk and data utility but are

[1] “Anonymized” data really isn’t and here’s why not," by Nate Anderson `http://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/`

[2] “How To Break Anonymity of the Netflix Prize Dataset"`http://arxiv.org/abs/cs/0610105`

[3] “Lessons from improperly anonymized taxi logs," by Nathan Yau, `http://flowingdata.com/2014/06/23/lessons-from-improperly-anonymized-taxi-logs/`

fragile under auxiliary information. Differential privacy has three main advantages over other approaches:

(1) it rigorously quantifies the privacy property of any data analysis mechanism;

(2) it controls the amount of privacy leakage regardless of the attacker's resource or knowledge;

(3) it has useful interpretations from the perspectives of Bayesian inference and statistical hypothesis testing, and hence fits naturally in the general framework of statistical machine learning, e.g., see [78, 144, 202, 245, 249], as well as applications involving regression [55, 217] and GWAS data [254], etc.

However there is still a huge gap between theory and its wide application to practical problems. The key drawback of DP is precisely its strength: it is a worst-case guarantee that protects the privacy of an arbitrary data point in an arbitrary dataset, which does not take into account properties of a real dataset. DP also does not easily scale to the kinds of big data setting of greatest interest to the machine learning community.

This part of the thesis presents a sequence of work that attempts to understand the pros and cons of differential privacy in statistical learning and to make DP more practical. This involves incorporating the notion of DP as a key resource in the standard theory of statistics and machine learning, and developing generic and easy-to-use tools for learning with DP. The challenge is often not about privacy alone but rather how we can maximize the utility of the learned model from a fixed dataset under privacy constraint, or equivalently, how we can minimize the sample complexity under privacy constraint. DP is intricately connected to notions such as stability, generalization and robustness [78, 87, 246, 247] in learning theory; also, randomization for various other purposes can potentially be exploited for privacy [245]. Therefore, instead of considering differential privacy directly in its own right, it makes sense to formally put it as a new resource in the statistical machine learning framework and to consider it jointly with all other desiderata.

We will also consider designing general-purposed and practical algorithms for differentially private statistical learning. In particular, the algorithm that samples from a Gibbs distribution induced by the loss function to be minimized, or equivalently the posterior distribution seems to be a natural candidate, and is used quite extensively for other purposes that is not related to privacy at all. This is first investigated in Chapter 3 and subsequently additional properties of this algorithm was studied in Chapter 4, and its guarantees substantially improved for specific models in Chapter 5.

Another underlying thread of the chapters involves (weakened or more adaptive) notions of differential privacy that allows for weaker assumptions, more favorable privacy-utility trade-offs, closer connections to learning tasks and ability to take advantage of structures in the data set. In Chapter 4 we will propose On-Average KL Privacy, which measures the average privacy cost of an data distribution when running an algorithm on data set is drawn from the same data distribution. It turns out that On-Average KL Privacy is equivalent to generalization for the posterior sampling algorithm. In Chapter 5, we will unify differential privacy and On-Average KL Privacy by a more low-level definition called per-instance differential privacy (pDP). This can be used to calculate the heterogeneous privacy loss incurred for each individual (in the data set or outside the data set) separately, and also be used as a low-level tool to provide tighter privacy analysis of a given algorithm.

Chapter 2

Characterizing learnability under a differential privacy constraint

While machine learning has proven to be a powerful data-driven solution to many real-life problems, its use in sensitive domains has been limited due to privacy concerns. A popular approach known as *differential privacy* offers provable privacy guarantees, but it is often observed in practice that it could substantially hamper learning accuracy. In this chapter we study the learnability (whether a problem can be learned by any algorithm) under Vapnik's general learning setting with differential privacy constraint, and reveal some intricate relationships between privacy, stability and learnability. In particular, we show that a problem is privately learnable *if an only if* there is a private algorithm that asymptotically minimizes the empirical risk (AERM). In contrast, for non-private learning AERM alone is not sufficient for learnability. This result suggests that when searching for private learning algorithms, we can restrict the search to algorithms that are AERM. In light of this, we propose a conceptual procedure that always finds a universally consistent algorithm whenever the problem is learnable under privacy constraint. We also propose a generic and practical algorithm and show that under very general conditions it privately learns a wide class of learning problems. Lastly, we extend some of the results to the more practical (ϵ, δ)-differential privacy and establish the existence of a phase-transition on the class of problems that are approximately privately learnable with respect to how small δ needs to be.

2.1 Introduction

In this chapter we focus on the following fundamental question about differential privacy and machine learning:

What problems can we learn with differential privacy?

Most literature focuses on designing differentially private extensions of various learning algorithms, where the methods depend crucially on the specific context and differ vastly in nature. But with the privacy constraint, we have less choice in developing learning and data analysis algorithms. It remains unclear how such a constraint affects our ability to learn, and if it is possible to design a generic privacy-preserving analysis mechanism that is applicable to a wide class of learning problems.

Our Contributions We provide a general answer to the relationship between learnability and differential privacy under Vapnik's General Learning Setting [235] in four aspects.

1. We characterize the subset of problems in the General Learning Setting that can be learned under differential privacy. Specifically, we show that a sufficient and necessary condition for a problem to be privately learnable is the existence of an algorithm that is differentially private and asymptotically minimizes the empirical risk. This characterization generalizes previous studies of the subject [19, 124] that focus on binary classification in discrete domain under the PAC learning model. Technically, the result relies on the now well-known intuitive observation that "privacy implies algorithmic stability" and the argument in Shalev-Shwartz et al. [191] that shows a variant of algorithmic stability is necessary for learnability.

2. We also introduce a weaker notion of learnability, which only requires consistency for a class of distributions $\mathfrak{D}$. Problems that are not privately learnable (a surprisingly large class that includes simple problems such as 0-1 loss binary classification in continuous feature domain [54]) are usually private $\mathfrak{D}$-learnable for some "nice" distribution class $\mathfrak{D}$. We characterize the subset of private $\mathfrak{D}$-learnable problems that are also (non-privately) learnable using conditions analogous to those in distribution-free private learning.

3. Inspired by the equivalence between privacy learnability and private AERM, we propose a generic (but impractical) procedure that always finds a consistent and private algorithm for any privately learnable (or $\mathfrak{D}$-learnable) problems. We also study a specific algorithm that aims at minimizing the empirical risk while preserving the privacy. We show that under a sufficient condition that relies on the geometry of the hypothesis space and the data distribution, this algorithm is able to privately learn (or $\mathfrak{D}$-learn) a large range of learning problems including classification, regression, clustering, density estimation and etc, and it is computationally efficient when the problem is convex. In fact, this generic learning algorithm learns any privately learnable problems in the PAC learning setting [19]. It remains an open problem whether the second algorithm also learns any privately learnable problem in the General Learning Setting.

4. Lastly, we provide a preliminary study of learnability under the more practical (ϵ, δ)-differential privacy. Our results reveal that whether there is separation between learnability and approximate private learnability depends on how fast δ is required to go to 0 with respect to the size of the data. Finding where the exact phase transition occurs is an open problem of future interest.

Our primary objective is to understand the conceptual impact of differential privacy and learnability under a general framework and the rates of convergence obtained in the analysis may be suboptimal. Although we do provide some discussion on polynomial time approximations to the proposed algorithm, learnability under computational constraints is beyond the scope of this chapter.

Related work While a large amount of work has been devoted to finding consistent (and rate optimal) differentially private learning algorithms in various settings [e.g., 17, 55, 121, 128], the characterization of privately learnable problems were only studied in a few special cases.

Kasiviswanathan et al. [124] showed that, for binary classification with a finite discrete hypothesis space, anything that is non-privately learnable is privately learnable under the agnostic Probably Approximately Correct (PAC) learning framework, therefore "finite VC-dimension" characterizes the set of private learnable problems in this setting. Beimel et al. [19] extends Kasiviswanathan et al. [124] by characterizing the sample complexity of the same class of problems, but the result only applies to the realizable (non-agnostic) case. Chaudhuri and Hsu [54] provided a counter-example showing that for continuous hypothesis space and data space, there is a gap between learnability and learnability under privacy constraint. They proposed to fix this issue by either weakening the privacy requirement to labels only or by restricting the class of potential distribution. While meaningful in some cases, these approaches do not resolve the learnability problem in general.

A key difference of our work from Beimel et al. [19], Chaudhuri and Hsu [54], Kasiviswanathan et al. [124] is that we consider a more general class of learning problems and provide a proper treatment in a statistical learning framework. This allows us to capture a wider collection of important learning problems (see Figure 2.1a and Table 2.1).

It is important to note that despite its generality, Vapnik's general learning setting still does not nearly cover the full spectrum of private learning. In particular, our results do not apply to improper learning (learning using a different hypothesis class) as considered in Beimel et al. [19] or to structural loss minimization (the loss function jointly take all data points as input) considered in Beimel et al. [20]. Also, our results do not address the sample complexity problem, which remains open in the general learning setting even for learning without privacy constraints.

Our characterization of private learnability (and private $\mathfrak{D}$-learnability) in Section 2.3 uses a recent advance in the characterization of general learnability given by Shalev-Shwartz et al. [191]. Roughly speaking, they showed that a problem is learnable if and only if there exists an algorithm that (i) is stable under small perturbation of training data, and (ii) behaves like empirical risk minimization (ERM) asymptotically. We also makes use of a folklore observation that "Privacy $\Rightarrow$ Stability $\Rightarrow$ Generalization". The connection of privacy and stability appeared as early as 2008 in a conference version of Kasiviswanathan et al. [124]. Further connection to "generalization" recently appeared in blog posts[1], stated as a theorem in Appendix F of Bassily et al. [17], and was shown to hold with strong concentration in Dwork et al. [87].

Dwork et al. [87] is part of an independent line of work [18, 36, 86, 109] on adaptive data analysis, which also stems from the observation that privacy implies stability and generalization. Comparing to adaptive data analysis works, our focus is quite different. Adaptive data analysis

[1] For instance, Frank McSherry described in a blog post an example of exploiting differential privacy for measure concentration `http://windowsontheory.org/2014/02/04/differential-privacy-for-measure-concentration/`; Moritz Hardt discussed the connection of differential privacy to stability and generalization in his blog post `http://blog.mrtz.org/2014/01/13/false-discovery`.

work focus on the impact of k on how fast the maximum absolute error of k-adaptively chosen queries goes to 0 as a function of n, while this chapter is concerned with whether the error can go to 0 at all for each learning problem when we require the learning algorithm be differentially private with $\epsilon < \infty$. Nonetheless, we acknowledge that Theorem 7 in Dwork et al. [87] provides an interesting alternative proof for "differentially private learners have small generalization error", when choosing the statistical query as evaluating a loss function at a privately learned hypothesis. The connection is not quite obvious and we provide a more detailed explanation in Section 2.9.

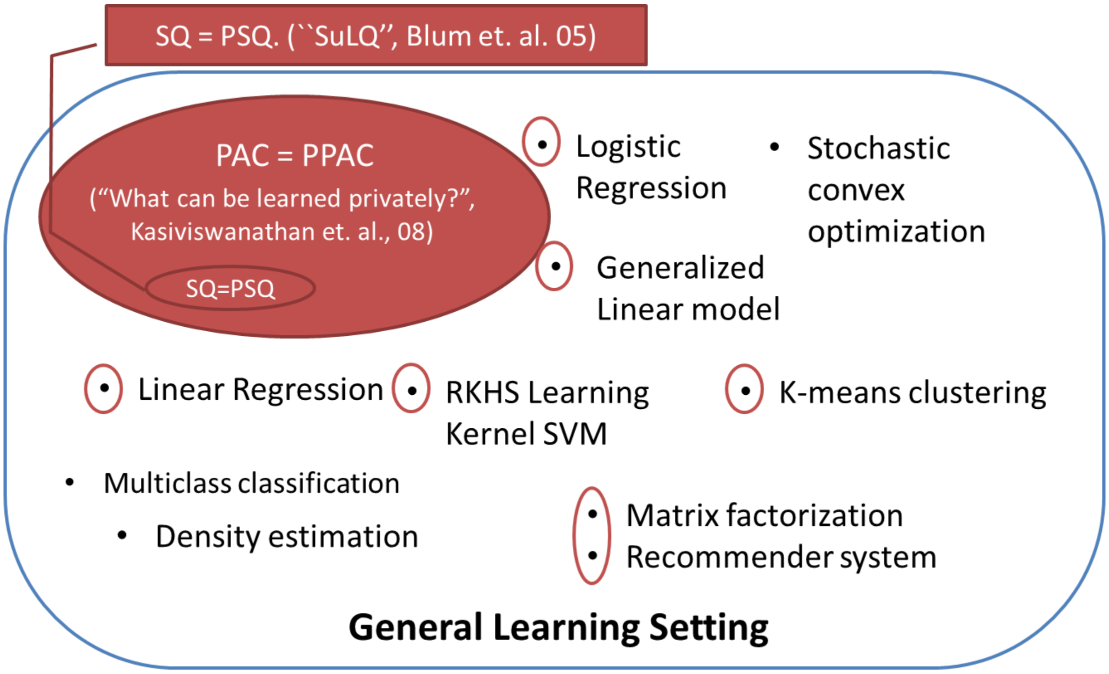
The main tool used in the construction of our generic private learning algorithm in Section 2.4 is the Exponential Mechanism [156], which provides a simple and differentially-private approximation to the maximizer of a score function among a candidate set. In the general learning context, we use the negative empirical risk as the utility function, and apply the exponential mechanism to a possibly pre-discretized hypothesis space. This exponential mechanism approach was used in Bassily et al. [17] for minimizing convex and Lipschitz functions. The sample discretization procedure has been considered in Chaudhuri and Hsu [54] and Beimel et al. [19]. Our scope and proof techniques are different. Our strategy is to show that, under some general regularity conditions, the exponential mechanism is stable and behaves like ERM. Our sublevel set condition has the same flavor as that in the proof of Bassily et al. [17, Theorem 3.2], although we do not need the loss function to be convex or Lipschitz.

Stability, privacy and generalization were also studied in Thakurta and Smith [217] with different notions of stability. More importantly, their stability is used as an assumption rather than a consequence, so their result is not directly comparable to ours.

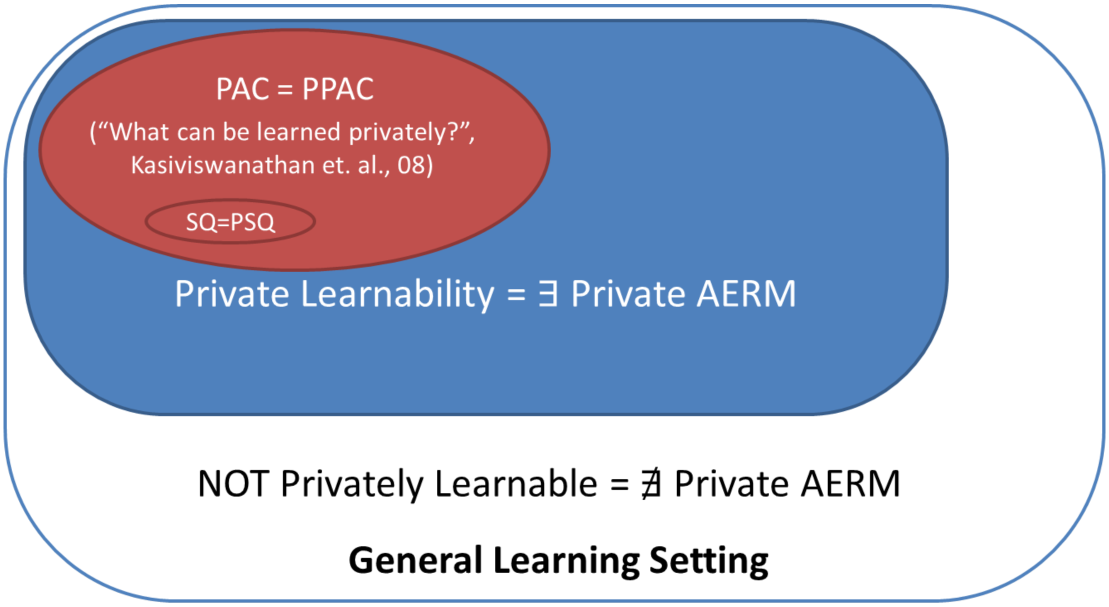
2.2 Background

2.2.1 Learnability under the General Learning Setting

In the General Learning Setting of Vapnik [235], a learning problem is characterized by a triplet $(\mathcal{Z}, \mathcal{H}, \ell)$. Here $\mathcal{Z}$ is the sample space (with a σ-algebra). The hypothesis space $\mathcal{H}$ is a collection of models such that each $h \in \mathcal{H}$ describes some structures of the data. The loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ measures how well the hypothesis h explains the data instance $z \in \mathcal{Z}$. For example, in supervised learning problems $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is the feature space and $\mathcal{Y}$ is the label space; $\mathcal{H}$ defines a collection of mapping $h : \mathcal{X} \to \mathcal{Y}$; and $\ell(h, z)$ measures how well h predicts the feature-label relationship $z = (x, y) \in \mathcal{Z}$. This setting includes problems with continuous input/output in potentially infinite dimensional spaces (e.g. RKHS methods), hence is much more general than PAC learning. In addition, the general learning setting also covers a variety of unsupervised learning problems, including clustering, density estimation, principal component analysis (PCA) and variants (e.g., Sparse PCA, Robust PCA), dictionary learning, matrix factorization and even Latent Dirichlet Allocation (LDA). Details of these examples are given in Table 2.1 (the first few are extracted from Shalev-Shwartz et al. [191]).



(a) Illustration of general learning setting. Examples of known DP extensions are circled in **maroon**.


PAC = PPAC
("What can be learned privately?", Kasiviswanathan et. al., 08)
SQ=PSQ
Private Learnability = ∃ Private AERM
NOT Privately Learnable = ∄ Private AERM
General Learning Setting

(b) Our characterization of private learnable problems in the general learning setting (in **blue**).

Figure 2.1: Illustration of general learning setting and our contribution in understanding differentially private learnability.

To account for the randomness in the data, we are primarily interested in the case where the data $Z = \{z_1, ..., z_n\} \in \mathcal{Z}^n$ are independent samples drawn from an unknown probability distribution $\mathcal{D}$ on $\mathcal{Z}$. We denote such a random sample by $Z \sim \mathcal{D}^n$. For a given distribution $\mathcal{D}$, let $R(h)$ be the expected loss of hypothesis h and $\hat{R}(h, Z)$ the empirical risk from a sample $Z \in \mathcal{Z}^n$:

$$R(h) = \mathbb{E}_{z\sim\mathcal{D}}\ell(h, z), \qquad \hat{R}(h, Z) = \frac{1}{n}\sum_{i=1}^{n}\ell(h, z_i)\,.$$

The optimal risk $R^* = \inf_{h\in\mathcal{H}} R(h)$ and we assume that it is achieved by an optimal $h^* \in \mathcal{H}$. Similarly, the minimal empirical risk $\hat{R}^*(Z) = \inf_{h\in\mathcal{H}} \hat{R}(h, Z)$ is achieved by $\hat{h}^*(Z) \in \mathcal{H}$. For a possibly randomized algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{H}$ that learns some hypothesis $\mathcal{A}(Z) \in \mathcal{H}$ given data sample Z, we say $\mathcal{A}$ is *consistent* if

$$\lim_{n\to\infty} \mathbb{E}_{Z\sim\mathcal{D}^n}\left(\mathbb{E}_{h\sim\mathcal{A}(Z)}R(h) - R^*\right) = 0. \tag{2.1}$$

In addition, we say $\mathcal{A}$ is consistent with rate $\xi(n)$ if

$$\mathbb{E}_{Z\sim\mathcal{D}^n}\left(\mathbb{E}_{h\sim\mathcal{A}(Z)}R(h) - R^*\right) \le \xi(n), \text{ where } \lim_{n\to\infty}\xi(n) \to 0. \tag{2.2}$$

Since the distribution $\mathcal{D}$ is unknown, we cannot adapt the algorithm $\mathcal{A}$ to $\mathcal{D}$, especially when privacy is a concern. Also, even if $\mathcal{A}$ is pointwise consistent for any distribution $\mathcal{D}$, it may have different rates for different $\mathcal{D}$ and potentially be arbitrarily slow for some $\mathcal{D}$. This makes it hard to evaluate whether $\mathcal{A}$ indeed learns the learning problem and forbids the study of the learnability problem. In this study, we adopt the stronger notion of learnability considered in Shalev-Shwartz et al. [191], which is a direct generalization of PAC-learnability [229] and agnostic PAC-learnability [126] to the General Learning Setting as studied by Haussler [111].

Definition 2.1 (Learnability, [191])**.** *A learning problem is learnable if there exists an algorithm $\mathcal{A}$ and rate $\xi(n)$, such that $\mathcal{A}$ is consistent with rate $\xi(n)$ for any distribution $\mathcal{D}$ defined on $\mathcal{Z}$.*

This definition requires consistency to hold universally for any distribution $\mathcal{D}$ with a uniform (distribution-independent) rate $\xi(n)$. This type of problem is often called *distribution-free learning* [229], and an algorithm is said to be *universally consistent* with rate $\xi(n)$ if it realizes the criterion.

2.2.2 Differential privacy

Differential privacy requires that if we arbitrarily perturb a database by only one data point, the output should not differ much. Therefore, if one conducts a statistical test for whether any individual is in the database or not, the false positive and false negative probabilities cannot both be small [249]. Formally, define “Hamming distance”

$$d(Z, Z') := \#\{i = 1, ..., n : z_i \neq z_i'\}\,. \tag{2.3}$$

Problem	Hypothesis class $\mathcal{H}$	$\mathcal{Z}$ or $\mathcal{X} \times \mathcal{Y}$	Loss function ℓ
Binary classification	$\mathcal{H} \subset \{f : \{0,1\}^d \to \{0,1\}\}$	$\{0,1\}^d \times \{0,1\}$	$1(h(x) \neq y)$
Regression	$\mathcal{H} \subset \{f : [0,1]^d \to \mathbb{R}\}$	$[0,1]^d \times \mathbb{R}$	$\vert h(x) - y\vert^2$
Density Estimation	Bounded distributions on $\mathcal{Z}$	$\mathcal{Z} \subset \mathbb{R}^d$	$-\log(h(z))$
K-means Clustering	$\{S \subset \mathbb{R}^d : \vert S\vert = k\}$	$\mathcal{Z} \subset \mathbb{R}^d$	$\min_{c\in h} \Vert c - z\Vert^2$
RKHS classification	Bounded RKHS	RKHS$\times\{0,1\}$	$\max\{0, 1 - y\langle x, h\rangle\}$
RKHS regression	Bounded RKHS	RKHS$\times\mathbb{R}$	$\vert\langle x, h\rangle - y\vert^2$
Sparse PCA	Rank-r projection matrices	$\mathbb{R}^d$	$\Vert hz - z\Vert^2 + \lambda\Vert h\Vert_1$
Robust PCA	All subspaces in $\mathbb{R}^d$	$\mathbb{R}^d$	$\Vert\mathcal{P}_h(z) - z\Vert_1 + \lambda\text{rank}(h)$
Matrix Completion	All subspaces in $\mathbb{R}^d$	$\mathbb{R}^d \times \{1,0\}^d$	$\min_{b\in h} \Vert y \circ (b - x)\Vert^2 + \lambda\text{rank}(h)$
Dictionary Learning	All dictionaries $\in \mathbb{R}^{d\times r}$	$\mathbb{R}^d$	$\min_{b\in\mathbb{R}^r} \Vert hb - z\Vert^2 + \lambda\Vert b\Vert_1$
Non-negative MF	All dictionaries $\in \mathbb{R}^{d\times r}_+$	$\mathbb{R}^d$	$\min_{b\in\mathbb{R}^r_+} \Vert hb - z\Vert^2$
Subspace Clustering	A set of k rank-r subspaces	$\mathbb{R}^d$	$\min_{b\in h} \Vert\mathcal{P}_b(z) - z\Vert^2$
Topic models (LDA)	$\{\mathbb{P}(\text{word}\vert\text{topic})\}$	Documents	$-\max_{b\in\{\mathbb{P}(\text{Topic})\}} \sum_{w\in z} \log \mathbb{P}_{b,h}(w)$

Table 2.1: An illustration of problems in the General Learning setting.

Definition 2.2 (ϵ-Differential Privacy, [77]). *An algorithm $\mathcal{A}$ is ϵ-differentially private, if*

$$\mathbb{P}(\mathcal{A}(Z) \in H) \leq \exp(\epsilon)\mathbb{P}(\mathcal{A}(Z') \in H)$$

for $\forall\ Z,\ Z'$ obeying $d(Z, Z') = 1$ and any measurable subset $H \subseteq \mathcal{H}$.

There are weaker notions of differential privacy. For example (ϵ, δ)-differential privacy allows for a small probability δ where the privacy guarantee does not hold. In this chapter, we will mainly work with the stronger ϵ-differential privacy. In Section 2.6 we discuss the problem of (ϵ, δ)-differential privacy and extend some of the results to this setting.

Our objective is to understand whether there is a gap between learnable problems and privately learnable problems in the general learning setting, and to quantify the tradeoff required to protect privacy. To achieve this objective, we need to show the existence of an algorithm that learns a class of problems while preserving differential privacy. More formally, we define

Definition 2.3 (Private learnability). *A learning problem is privately learnable with rate $\xi(n)$ if there exists an algorithm $\mathcal{A}$ that satisfies both universal consistency (as in Definition 2.1) with rate $\xi(n)$ and ϵ-differential privacy with privacy parameter $\epsilon < \infty$.*

We can view the consistency requirement Definition 2.3 as a measure of utility. This utility is not a function of the observed data, however, but rather how the results generalize to unseen data.

The following lemma shows that the above definition of private learnability is actually equivalent to a seemingly much stronger condition with a vanishing privacy loss ϵ.

Lemma 2.4. *If there is an ϵ-DP algorithm that is consistent with rate $\xi(n)$ for some constant $0 < \epsilon < \infty$, then there is a $\frac{2}{\sqrt{n}}\left(e^{\epsilon} - e^{-\epsilon}\right)$-DP algorithm that is consistent with rate $\xi(\sqrt{n})$.*

The proof, given in Section 2.8.1, uses a subsampling theorem adapted from Beimel et al. [21, Lemma 4.4]. The intuition behind the lemma is that when we subsample randomly with probability $1/\sqrt{n}$, then with only probability $1/\sqrt{n}$ will the two data sets be different and that reduces the privacy cost by a factor that is proportional to $1/\sqrt{n}$. The same result also holds for (ϵ, δ)-DP, in Section 2.6, we will prove a subsampling lemma for (ϵ, δ)-DP.

There are many approaches to design differentially private algorithms, such as noise perturbation using Laplace noise [77, 84] and the Exponential Mechanism [156]. Our construction of generic differentially private learning algorithms applies the Exponential Mechanism to penalized empirical risk minimization. Our argument will make use of a general characterization of learnability described below.

2.2.3 Stability and Asymptotic ERM

An important breakthrough in learning theory is a full characterization of all learnable problems in the General Learning Setting in terms of stability and empirical risk minimization [191]. Without assuming uniform convergence of empirical risk, Shalev-Shwartz et al. showed that a problem is learnable if and only if there exists a “strongly uniform-RO stable” and “always asymptotically empirical risk minimization” (Always AERM) randomized algorithm that learns the problem. Here “RO” stands for “replace one”. Also, any strongly uniform-RO stable and “universally” AERM (weaker than “always” AERM) learning rule learns the problem consistently. Here we give detailed definitions.

Definition 2.5 (Universally/Always AERM, [191])**.** *A (possibly randomized) learning rule $\mathcal{A}$ is Universally AERM if for any distribution $\mathcal{D}$ defined on domain $\mathcal{Z}$*

$$\mathbb{E}_{Z\sim\mathcal{D}^n}\left[\mathbb{E}_{h\sim\mathcal{A}(Z)}\hat{R}(h,Z)-\hat{R}^*(Z)\right]\to 0,\ \ \textit{as}\ \ n\to\infty$$

where $\hat{R}^*(Z)$ is the minimum empirical risk for data set Z. We say $\mathcal{A}$ is Always AERM, if in addition,

$$\sup_{Z\in\mathcal{Z}^n}\mathbb{E}_{h\sim\mathcal{A}(Z)}\hat{R}(h,Z)-\hat{R}^*(Z)\to 0,\ \ \textit{as}\ \ n\to\infty\,.$$

Definition 2.6 (Strongly Uniform RO-Stability, [191])**.** *An algorithm $\mathcal{A}$ is strongly uniform RO-stable if*

$$\sup_{z\in\mathcal{Z}}\ \sup_{\substack{Z,Z'\in\mathcal{Z}^n,\\ d(Z,Z')=1}}\left|\mathbb{E}_{h\sim\mathcal{A}(Z)}\ell(h,z)-\mathbb{E}_{h\sim\mathcal{A}(Z')}\ell(h,z)\right|\to 0\ \textit{as}\ n\to\infty.$$

where $d(Z,Z')$ is defined in (4.1)*, in other word, Z and Z' can differ by at most one data point.*

Since we will not deal with other variants of algorithmic stability in this chapter (e.g., hypothesis stability [125], uniform stability [38] and leave-one-out (LOO) stability in Mukherjee et al. [162]), we simply call Definition 2.6 stability or uniform stability. Likewise, we will refer to ϵ-differential privacy as just “privacy” although there are several other notions of privacy in the literature.

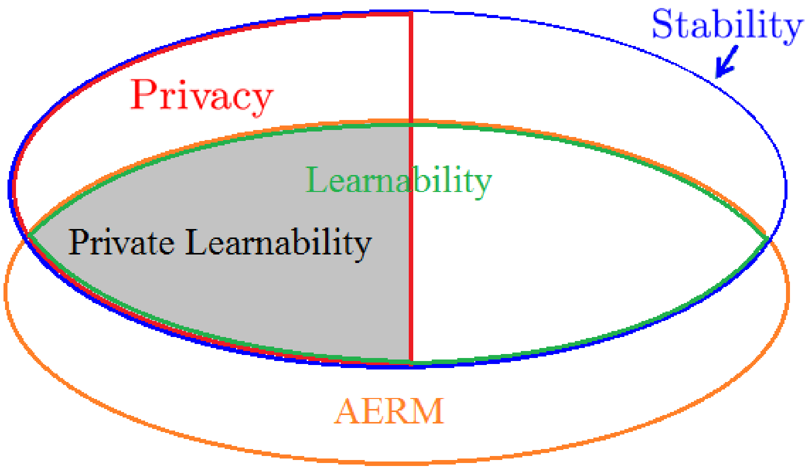


Figure 2.2: A summary of the relationships of various notions revealed by our analysis.

2.3 Characterization of private learnability

We are now ready to state our main result. The only assumption we make is the uniform boundedness of the loss function. This is also assumed in Shalev-Shwartz et al. [191] for the learnability problem without privacy constraints. Without loss of generality, we can assume $0 \leq \ell(h, z) \leq 1$.

Theorem 2.7. *Given a learning problem $(\mathcal{Z}, \mathcal{H}, \ell)$, the following statements are equivalent.*

1. *The problem is privately learnable.*
2. *There exists a differentially private universally AERM algorithm.*
3. *There exists a differentially private always AERM algorithm.*

The proof is simple yet revealing, we will present the arguments for $2 \Rightarrow 1$ (sufficiency of AERM) in Section 2.3.1 and $1 \Rightarrow 3$ (necessity of AERM) in Section 2.3.2. $3 \Rightarrow 2$ follows trivially from the definition of "always" and "universal" AERM.

The theorem says that we can stick to ERM-like algorithms for private learning, despite that ERM may fail for some problems in the (non-private) general learning setting [191]. Thus a standard procedure for finding universally consistent and differentially private algorithms would be to approximately minimize the empirical risk using some differentially private procedures [17, 55, 128]. If the utility analysis reveals that the method is AERM, we do not need to worry about generalization as it is guaranteed by privacy. This consistency analysis is considerably simpler than non-private learning problems where one typically needs to control generalization error either via uniform convergence (VC-dimension, Rademacher complexity, metric entropy, etc) or to adopt the stability argument [191].

This result does not imply that privacy is helping the algorithm to learn in any sense, as the simplicity is achieved at the cost of having a smaller class of learnable problems. A concrete example of a problem being learnable but not privately learnable is given in [54] and we will revisit it in Section 2.3.3. For some problems where ERM fails, it may not be possible to make it AERM while preserving privacy. In particular, we were not able to privatize the problem in Section 4.1 of Shalev-Shwartz et al. [191].

To avoid any potential misunderstanding, we stress that Theorem 3 is a characterization of learnability, *not* learning algorithms. It does not prevent the existence of a universally consistent learning algorithm that is private but not AERM. Also, the characterization given in Theorem 3 is about consistency, and it does not claim anything on sample complexity. An algorithm that is AERM may be suboptimal in terms of convergence rate.

2.3.1 Sufficiency: Privacy implies stability

A key ingredient in the proof of sufficiency is a well-known heuristic observation that differential privacy by definition implies uniform stability, which is useful in its own right.

Lemma 2.8 (Privacy ⇒ Stability)**.** *Assume $0 \leq \ell(h, z) \leq 1$, any ϵ-differentially private algorithm satisfies $(e^{\epsilon} - 1)$-stability. Moreover if $\epsilon \leq 1$ it satisfies 2ϵ-stability.*

The proof of this lemma comes directly from the definition of differential privacy so it is algorithm independent. The converse, however, is not true in general (e.g., a non-trivial deterministic algorithm can be stable, but not differentially private.)

Corollary 2.9 (Privacy + Universal AERM ⇒ Consistency)**.** *If a learning algorithm $\mathcal{A}$ is $\epsilon(n)$-differentially private and $\mathcal{A}$ is universally AERM with rate $\xi(n)$, then $\mathcal{A}$ is universally consistent with rate $\xi(n) + e^{\epsilon(n)} - 1 = O(\xi(n) + \epsilon(n))$.*

The proof of Corollary 2.9, provided in the Section 9.7, combines Lemma 2.8 and the fact that consistency is implied by stability and AERM (Theorem 2.35). Our Theorem 2.35 is based on minor modifications of Theorem 8 in Shalev-Shwartz et al. [191]. In fact, Corollary 2.9 can be stated in a stronger per distribution form, since universality is not used in the proof. We will revisit this point when we discuss a weaker notion of private learnability below.

Lemma 2.4 and Corollary 2.9 together establishes $2 \Rightarrow 1$ in Theorem 3.

If for a problem privacy and always AERM cannot coexist, then the problem is not privately learnable. This is what we will show next.

2.3.2 Necessity: Consistency implies Always AERM

To prove that the existence of an always AERM learning algorithm is necessary for any private learnable problems, it suffices to construct such a learning algorithm for each learnable problem.

Lemma 2.10 (Consistency + Privacy ⇒ Private Always AERM)**.** *If $\mathcal{A}$ is a universally consistent learning algorithm satisfying ϵ-DP with any $\epsilon > 0$ and consistent with rate $\xi(n)$, then there is* *another universally consistent learning algorithm $\mathcal{A}'$ that is always AERM with rate $\xi(\sqrt{n})$ and satisfies $\frac{2}{\sqrt{n}}(e^{\epsilon} - e^{-\epsilon})$-DP.*

Lemma 2.10 is proved in Section 2.8.2. The proof idea is to run $\mathcal{A}$ on a size $O(\sqrt{n})$ random subsample of Z, which will be universally consistent with a slower rate, differentially private with $\epsilon(n) \to 0$ (Lemma 2.34), and at the same time always AERM. The last part uses an argument in Lemma 24 of Shalev-Shwartz et al. [191] which appeals to the universality of $\mathcal{A}$'s consistency on a specific discrete distribution supported on the given data set Z.

As pointed out by an anonymous reviewer, there is a simpler proof by invoking Theorem 10 of Shalev-Shwartz et al. [191] that says any consistent and generalizing algorithm must be AERM and a result [e.g., 17, Appendix F] that says "privacy ⇒ generalization". This is a valid observation. But their Theorem 10 is proven using a detour through "generalization", which leads to a slower rate than what we are able to obtain in Lemma 2.10 using a more direct argument.

2.3.3 Private Learnability vs. Non-private Learnability

Now we have a characterization of all privately learnable problems, a natural question to ask is that whether any learnable problem is also privately learnable. The answer is "yes" for learning in Statistical Query (SQ)-model and PAC Learning model (binary classification) with finite hypothesis space, and is "no" for continuous hypothesis space [54].

By definition, all privately learnable problems are learnable. But now that we know that privacy implies generalization, it is tempting to hope that privacy can help at least some problem to learn better than any non-private algorithm. In terms of learnability, the question becomes: Could there be a (learnable) problem that is *exclusively* learnable through private algorithms? We now show that such a problem does not exist.

Proposition 2.11. *If a learning problem is learnable by an ϵ-DP algorithm $\mathcal{A}$, then it is also learnable by a non-private algorithm.*

The proof is given in Section 2.8.3. The idea is that $\mathcal{A}(Z)$ defines a distribution over $\mathcal{H}$. Pick an $z \in \mathcal{Z}$. If $z \notin Z$, algorithm $\mathcal{A}' = \mathcal{A}$. Otherwise, $\mathcal{A}'(Z)$ samples from a slightly different distribution than $\mathcal{A}(Z)$ that does not affect the expectation much.

On the other hand, not all learnable problems are privately learnable. This can already be seen from Chaudhuri and Hsu [54], where the gap between learning and private learning is established. We revisit Chaudhuri and Hsu's example in our notation under the general learning setting and produce an alternative proof by showing that differential privacy contradicts *always AERM*, then invoking Theorem 3 to show the problem is not privately learnable.

Proposition 2.12 ([54, Theorem 5])**.** *There exists a problem that is learnable by a non-private algorithm, but not privately learnable. In particular, any private algorithm cannot be* always AERM *in this problem.*

We describe the counterexample and re-establish the impossibility of private learning for this problem using the contrapositive of Theorem 3, which suggests that if privacy and always AERM algorithm cannot coexist for some problem, then the problem is not privately learnable.

Consider the binary classification problem with $\mathcal{X} = [0, 1]$, $\mathcal{Y} = \{0, 1\}$ and 0-1 loss function. Let $\mathcal{H}$ be the collection of threshold functions that output $h(x) = 1$ if $x > h$ and $h(x) = 0$ otherwise. This class has VC-dimension 1, and hence the problem is learnable.

Next we will construct $K = \lceil \exp(\epsilon_n n) \rceil$ data sets such that if $K - 1$ of them obey AERM, the remaining one cannot be. Let $\eta = 1/\exp(\epsilon n)$, $K := \lceil 1/\eta \rceil$. Let $h_1, h_2, ..., h_K$ be a disjoint thresholds such that they are at least η apart and $[h_i - \eta/3, h_i + \eta/3]$ are disjoint intervals.

If we take $Z_i \subseteq [h_i - \eta/3, h_i + \eta/3]$ with half of the points in $[h_i - \eta/3, h_i)$ and the other half in $(h_i, h_i + \eta/3]$ and we label each data point in it with $\mathbf{1}(z > h_i)$, then empirical risk $\hat{R}(h_i, Z_i) = 0$ $\forall i = 1, ..., K$. So for any AERM learning rule, $\mathbb{E}_{h \sim \mathcal{A}(Z_i)} \hat{R}(h, Z_i) \to 0$ for all i. For some sufficiently large n, $\mathbb{E}_{h \sim \mathcal{A}(Z_i)} \hat{R}(h, Z_i) < 0.1$.

Now consider Z_1,

$$\mathbb{P}(\mathcal{A}(Z_1) \notin [h_1 - \eta/3, h_1 + \eta/3]) \geq \sum_{i=2}^{K} \mathbb{P}(\mathcal{A}(Z_1) \in [h_i - \eta/3, h_i + \eta/3]),$$

since these intervals are disjoint. Then by the definition of ϵ-DP,

$$\mathbb{P}(\mathcal{A}(Z_1) \in [h_i - \eta/3, h_i + \eta/3]) \geq \exp(-\epsilon n) \mathbb{P}(\mathcal{A}(Z_i) \in [h_i - \eta/3, h_i + \eta/3]). \tag{2.4}$$

It follows that $\mathbb{P}(\mathcal{A}(Z_i) \in [h_i - \eta/3, h_i + \eta/3]) > 0.9$ otherwise $\mathbb{E}_{h \sim \mathcal{A}(Z_i)} \hat{R}(h, Z_i) \geq 0.1$, therefore

$$\mathbb{P}(\mathcal{A}(Z_1) \notin [h_1 - \eta/3, h_1 + \eta/3]) \geq K \exp(-\epsilon n) 0.9 \geq 0.9, \tag{2.5}$$

and $\mathbb{E}_{h \sim \mathcal{A}(Z_i)} \hat{R}(h, Z_i) \geq 0.9 \times 1 = 0.9$, which violates the "always AERM" condition that requires $\mathbb{E}_{h \sim \mathcal{A}(Z_1)} \hat{R}(h, Z_1) < 0.1$. Therefore, the problem is not privately learnable.

As is pointed out by an anonymous reviewer, the same conclusion of this impossibility result of privately learning thresholds on $[0, 1]$ can be drawn numerically through the characterization of the sample complexity [19], via the bound that depends logarithmically on the $\log(|\mathcal{H}|)$ and on $[0, 1]$ this number is infinite. The above analysis provides different insights about the problem. We will be using it again for understanding the separation of learnability and learnability under (ϵ, δ)-Differential Privacy later in Section 2.6.

2.3.4 Private $\mathfrak{D}$-learnability

The above example implies that even very simple learning problems may not be privately learnable. What this suggests is that often one cannot hope to obtain the same kind of uniform learnability results typical in nonprivate learning. However, most data sets of practical interest have nice distributions. Therefore, it makes sense to consider a smaller class of distributions, e.g., distributions with smooth densities that have bounded kth order derivative, or those having bounded total variation. These are common assumptions in non-parametric statistics, such as kernel density estimation, smoothing spline regression and mode clustering. Similarly, in high dimensional statistics, there are often assumptions on the structures of the underlying distribution, such as sparsity, smoothness, and low-rank conditions.

Definition 2.13 ((Private) $\mathfrak{D}$-learnability)**.** *We say a learning problem $(\mathcal{Z}, \mathcal{H}, \ell)$ is $\mathfrak{D}$-learnable if there exists a learning algorithm $\mathcal{A}$ that is consistent for every unknown distribution $\mathcal{D} \in \mathfrak{D}$. If in addition, the problem is $\mathfrak{D}$-learnable under ϵ-differential privacy for some $0 \leq \epsilon < \infty$, then we say the problem is privately $\mathfrak{D}$-learnable.*

Almost all of our arguments hold in a per distribution fashion, therefore they also hold for any such subclass $\mathfrak{D}$. The only exception is the necessity of “always AERM” (Lemma 2.10), where we used the universal consistency on an arbitrary discrete uniform distribution in the proof. The characterization still holds if the class $\mathfrak{D}$ contains all finite discrete uniform distributions. For general distribution classes, we characterize private $\mathfrak{D}$-learnability using a weaker “universally AERM” (instead of “always AERM”) under the assumption that the problem itself is learnable in a distribution-free setting without privacy constraints.

Lemma 2.14 (private $\mathfrak{D}$-learnability $\Rightarrow$ private $\mathfrak{D}$-universal AERM)**.** *If an ϵ-DP algorithm $\mathcal{A}$ is $\mathfrak{D}$-universally consistent with rate $\xi(n)$ and the problem itself is learnable in a distribution-free sense with rate $\xi'(n)$, then there exists a $\mathfrak{D}$-universally consistent learning algorithm $\mathcal{A}'$ that is $\mathfrak{D}$-universally AERM with rate $12\xi'(n^{1/4}) + \frac{37}{\sqrt{n}} + \xi(\sqrt{n})$ and satisfies $\frac{2}{\sqrt{n}}(e^{\epsilon} - e^{-\epsilon})$-DP.*

The proof, given in Section 2.8.4, shows that the algorithm $\mathcal{A}'$ that applies $\mathcal{A}$ to a random subsample of size $\lfloor \sqrt{n} \rfloor$ is AERM for any distribution in the class $\mathfrak{D}$.

Theorem 2.15 (Characterization of private $\mathfrak{D}$-learnability)**.** *A problem is privately $\mathfrak{D}$-learnable* if *there exists an algorithm that is $\mathfrak{D}$-universally AERM and differentially private with privacy loss $\epsilon(n) \to 0$. If in addition, the problem is (distribution-free and non-privately) learnable, then the converse is also true.*

Proof. The “if” part is exactly the same as the argument in Section 2.3.1, since both Lemma 2.8 and Lemma 2.9 holds for each distribution independently. Under the additional assumption that the problem itself is learnable (distribution-free and non-privately), the “only if” part is given by Lemma 2.14. □

This result may appear to be unsatisfactory due to the additional assumption of learnability. It is clearly a strong assumption because many problems that are $\mathfrak{D}$-learnable for a practically meaningful $\mathfrak{D}$ are not actually learnable. We provide one such example here.

Example 2.16. *Let the data space be $[0, 1]$, the hypothesis space be the class of all* finite *subset of $[0, 1]$ and the loss function $\ell(h, z) = 1_{z \notin h}$. This problem is not learnable, and not even $\mathfrak{D}$-learnable when $\mathfrak{D}$ is the class of all discrete distributions with finite number of possible values. But it is $\mathfrak{D}$-learnable when $\mathfrak{D}$ is further restricted with an upper bound on the total number of possible values.*

Proof. For any discrete distribution with a finite support set, there is an $h \in \mathcal{H}$ such that the optimal risk is 0. Assume the problem is learnable with rate $\xi(n)$, then for some n $\xi(n) < 0.5$.

However, we can always construct a uniform distribution over $3n$ elements and it is information-theoretically impossible for any estimators based on n samples from the distribution to achieve a risk better than $2/3$. The problem is therefore not learnable. When we assume an upper bound N on the maximum number of bins of the underlying distribution, then the ERM which outputs just the support of all observed data will be universally consistent with rate $\xi(n) = N/n$. □

It turns out that we cannot hope to *completely* remove the assumption from Theorem 2.15. The following example illustrates that some form of qualification (implied by the learnability assumption) is necessary for the converse statement to be true.

Example 2.17. *Consider the learning problem in Example 2.16. Let $\mathfrak{D}$ be the class of all continuous distributions. There is a learning problem that is s privately $\mathfrak{D}$-learnable but no private AERM algorithm exists.*

Proof. Let the learning problem be that in Example 2.16 and $\mathfrak{D}$ be the class of all continuous distributions defined on $[0, 1]$. Consider The learning algorithm $\mathcal{A}(Z)$ always returns $h = \emptyset$.
The optimal risk for any continuous distribution is 1 because any finite subset is of measure 0, output $\emptyset$ is 0-consistent and 0-generalizing, but not AERM, since the minimum empirical risk is 0. $\mathcal{A}$ is also 0-differentially private, therefore the problem is privately $\mathfrak{D}$-learnable for $\mathfrak{D}$ being the set of all continuous distributions.
However, it is not privately $\mathfrak{D}$-learnable via an AERM, i.e., no private AERM algorithm exists for this problem. We prove this by contradiction. Assume an ϵ-DP AERM algorithm exists, the subsampling lemma ensures the existence of an $\epsilon(n)$-DP AERM algorithm $\mathcal{A}'$ with $\epsilon(n) \to 0$. $\mathcal{A}'$ is therefore generalizing by stability, and it follows that the $\mathcal{A}'$ has risk $\mathbb{E}_{h\sim\mathcal{A}'(Z)}R(h)$ converging to 0. But there is no $h \in \mathcal{H}$ such that $R(h) < 1$, giving the contradiction. □

Interestingly, this problem is $\mathfrak{D}$-learnable via a non-private AERM algorithm, which always outputs $h = Z$. This is 0-consistent, 0-AERM but not generalizing. This example suggests that $\mathfrak{D}$-learnability and learnability are quite different because for learnable problems, if an algorithm is consistent and AERM, then it must also be generalizing [191, Theorem 10].

2.3.5 A generic learning algorithm

The characterization of private learnability suggests a generic (but impractical) procedure that learns all privately learnable problems (in the same flavor as the generic algorithm in Shalev-Shwartz et al. [191] that learns all learnable problems). This is to solve

$$\operatorname*{argmin}_{\substack{(\mathcal{A},\epsilon):\\ \mathcal{A}:\mathcal{Z}^n\to\mathcal{H},\\ \mathcal{A}\text{ is }\epsilon\text{-DP}}} \left[\epsilon + \sup_{Z\in\mathcal{Z}^n}\left(\mathbb{E}_{h\sim\mathcal{A}(Z)}\hat{R}(h,Z) - \inf_{h\in\mathcal{H}}\hat{R}(h,Z)\right)\right], \tag{2.6}$$

or to privately $\mathfrak{D}$-learn the problem when (2.6) is not feasible

$$\underset{\substack{(\mathcal{A},\epsilon):\\ \mathcal{A}:\mathcal{Z}^n\to\mathcal{H},\\ \mathcal{A}\text{ is }\epsilon\text{-DP}}}{\operatorname{argmin}} \left[\epsilon + \sup_{\mathcal{D}\in\mathfrak{D}} \mathbb{E}_{Z\sim\mathcal{D}^n}\left(\mathbb{E}_{h\sim\mathcal{A}(Z)}\hat{R}(h,Z) - \inf_{h\in\mathcal{H}}\hat{R}(h,Z)\right)\right]. \tag{2.7}$$

Theorem 2.18. *Assume the problem is learnable. If the problem is private learnable,* (2.6) *will always output a universally consistent private learning algorithm. If the problem is private $\mathfrak{D}$-learnable,* (2.7) *will always output a $\mathfrak{D}$-universally consistent private learning algorithm.*

Proof. If the problem is private learnable, by Theorem 3 there exists an algorithm $\mathcal{A}$ that is $\epsilon(n)$-DP and always AERM with rate $\xi(n)$ and $\epsilon(n)+\xi(n)\to 0$. This $\mathcal{A}$ is a witness in the optimization so we know that any minimizer of (2.6) will have a objective value that is no greater than $\epsilon(n)+\xi(n)$ for any n. Corollary 2.9 concludes its universal consistency. The second claim follows from the characterization of private $\mathfrak{D}$-learnability in Theorem 2.15. □

It is of course impossible to minimize the supremum over any data Z, nor is it possible to efficiently search over the space of all algorithms, let alone DP algorithms. But conceptually, this formulation may be of interest to theoretical questions related to the search of private learning algorithms and the fundamental limit of machine learning under privacy constraints.

2.4 Private learning for penalized ERM

Now we describe a generic and practical class of private learning algorithms, based on the idea of minimizing the empirical risk under privacy constraint:

$$\underset{h\in\mathcal{H}}{\text{minimize}}\, F(Z,h) = \frac{1}{n}\sum_{i=1}^{n}\ell(h,z_i) + g_n(h). \tag{2.8}$$

The first term is empirical risk and the second term vanishes as n increases so that this estimator is asymptotically ERM. The same formulation has been studied before in the context of differentially private machine learning [55, 128], but our focus is more generic and does not require the objective function to be convex, differentiable, continuous, or even have a finite dimensional Euclidean space embedding, hence covers a larger class of learning problems.

Our generic algorithm for differentially private learning is summarized in Algorithm 1. It applies the exponential mechanism [156] to penalized ERM. We note that this algorithm implicitly requires that $\int_{\mathcal{H}}\exp(\frac{\epsilon(n)}{2\Delta q}q(h,Z))dh<\infty$, otherwise the distribution is not well-defined and it does not make sense to talk about differential privacy. In general, if $\mathcal{H}$ is a compact set with a finite volume (with respect to a base measure, such as the Lebesgue measure or counting measure), then such a distribution always exists. We will revisit this point and discuss the practicality of this assumption in the Section 2.5.3.

Algorithm 1 Exponential Mechanism for regularized ERM

```
**Input:** Data points $Z = \{z_1, ..., z_n\} \in \mathcal{Z}^n$, loss function $\ell$, regularizer $g_n$, privacy parameter $\epsilon(n)$ and a hypothesis space $\mathcal{H}$.
1. Construct utility function $q(h, Z) := -\frac{1}{n}\sum_{i=1}^{n} \ell(h, z_i) - g_n(h)$, and its sensitivity $\Delta q := \sup_{h\in\mathcal{H}, d(Z,Z')=1} |q(h, Z) - q(h, Z')| \leq \frac{2}{n} \sup_{h\in\mathcal{H}, z\in\mathcal{Z}} |\ell(h, z)|$.
2. Sample $h \in \mathcal{H}$ with probability $\mathbb{P}(h) \propto \exp(\frac{\epsilon(n)}{2\Delta q} q(h, Z))$.
**Output:** $h$.
```

Using the characterization results developed so far, we are able to give sufficient conditions for consistency of private learning algorithms without having to establish uniform convergence. Define the sublevel set as

$$\mathcal{S}_{Z,t} = \{h \in \mathcal{H} \mid F(Z, h) \leq t + \inf_{h\in\mathcal{H}} F(Z, h)\}, \tag{2.9}$$

where $F(h, Z)$ is the regularized empirical risk function defined in (2.8). In particular, we assume the following conditions:

A1. Bounded loss function: $0 \leq \ell(h, z) \leq 1$ for any $h \in \mathcal{H}, z \in \mathcal{Z}$.

A2. Sublevel set condition: There exist constant positive integer n_0, positive real number t_0, and a sequence of regularizer g_n satisfying $\sup_{h\in\mathcal{H}} |g_n(h)| = o(n)$, such that for any $0 < t < t_0$, $n > n_0$

$$\mathbb{E}_{Z\sim\mathcal{D}^n}\left(\frac{\mu(\mathcal{H})}{\mu(\mathcal{S}_{Z,t})}\right) \leq K\left(\frac{1}{t}\right)^{\rho}, \tag{2.10}$$

where $K = K(n), \rho = \rho(n)$ satisfy $\log K + \rho \log n = o(n)$. Here the measure μ may depend on context, such as Lebesgue measure ($\mathcal{H}$ is continuous) or counting measure ($\mathcal{H}$ is discrete).

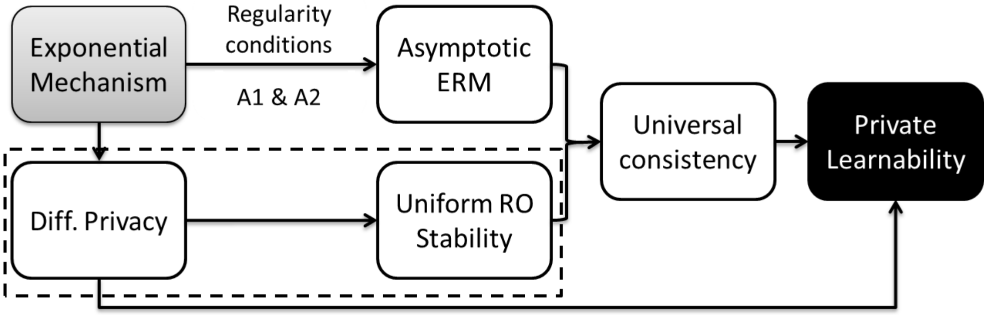
The first condition of boundedness is common. It is assumed in Vapnik's characterization for ERM learnability and Shalev-Shwartz et al.'s general characterization of all learnable problems. In fact, we can always consider $\mathcal{H}$ to be a sublevel set such that the boundedness condition holds. For the second condition, the intuition is that we require the sublevel set to be large enough such that the sampling procedure will return a good hypothesis with large probability. $\mu(\mathcal{S}_t)$ is a critical parameter in the utility guarantee for the exponential mechanism [156]. Also, it is worth pointing out that A2 implies that the exponential distribution is well-defined.

Theorem 2.19 (General private learning)**.** *Let $(\mathcal{Z}, \mathcal{H}, \ell)$ be any problem in the general learning setting. Suppose we can choose g_n such that A.1 and A.2 are satisfied with (ρ, K, g_n, n_0, t_0) for a distribution $\mathcal{D}$, then Algorithm 1 satisfies $\epsilon(n)$-privacy and is consistent with rate*

$$\xi(n) = \frac{9[\log K + (\rho + 2)\log n]}{n\epsilon(n)} + 2\epsilon(n) + \sup_{h\in\mathcal{H}} |g_n(h)|. \tag{2.11}$$

In particular, if $\epsilon(n) = o(1)$, $\sup_{h\in\mathcal{H}} |g_n(h)| = o(1)$ and $\log K + \rho \log n = o(n\epsilon(n))$ for all $\mathcal{D}$ (in $\mathfrak{D}$) Algorithm 1 privately learns ($\mathfrak{D}$-learns) the problem.

We give an illustration of the proof in Figure 2.3. The detailed proof, based on the stability argument [191], is deferred to Section 2.8.5.



*Dashed box works for any DP algorithms (including Exp. Mech.)

Figure 2.3: Illustration of Theorem 2.19: conditions for private learnability in general learning setting.

To see that Theorem 2.19 actually contains a large number of problems in the general learning setting. We provide concrete examples that satisfy A1 and A2 below for both privately learnable and privately $\mathfrak{D}$-learnable problems that can be learned using Algorithm 1.

2.4.1 Examples of privately learnable problems

We start from a few cases where Algorithm 1 is universally consistent for all distributions.

Example 2.20 (Finite discrete $\mathcal{H}$). *Suppose $\mathcal{H}$ can be fully encoded by M-bits, then*

$$\mu(\mathcal{S}_t)/\mu(\mathcal{H}) \geq |\mathcal{H}|^{-1} = 2^{-M}\,,$$

since there are at least 1 optimal hypothesis for each function and now μ is the counting measure. In other word, we can take $K = 2^M$ and $\rho = 0$ in the (2.11)*. Plug this into the expression and take $g_n \equiv 0$, $\epsilon(n) = \sqrt{(M + \log n)/n}$, we get a rate of consistency $\xi(n) = O(\frac{M+\log n}{\sqrt{n}})$. In addition, if we can find a data-independent covering set for a continuous space, then we can discretize the space and the result same results follow. This observation will be used in the construction of many private learning algorithms below.*

Example 2.21 (Lipschitz functions/Hölder class). *Let $\mathcal{H}$ be a compact, β_p-regular subset of $\mathbb{R}^d$ satisfying $\mu(B \cap \mathcal{H}) \geq \beta_p \mu(B)$ for any ℓ_p ball $B \subset \mathbb{R}^d$ that is small enough. Assume that $F(Z, \cdot)$ is L-Lipschitz on $\mathcal{H}$: for any $h, h' \in \mathcal{H}$,*

$$|F(Z,h) - F(Z,h')| \leq L\|h - h'\|_p\,.$$

Then for sufficiently small t, we have Lebesgue measure

$$\mu(\mathcal{S}_t) \geq \beta_p\,(t/L)^d$$

and Condition A.2 holds with $K = \mu(\mathcal{H})\beta_p^{-1}L^d$, $\rho = d$. Furthermore, if we take $\epsilon(n) = \sqrt{\frac{d(\log L+\log n)+\log(\mu(\mathcal{H})/\beta_p)}{n}}$, the algorithm is $O\left(\sqrt{\frac{d(\log L+\log n)+\log(\mu(\mathcal{H})/\beta_p)}{n}} + \sup_{h\in\mathcal{H}}|g_n(h)|\right)$-consistent.

This shows that condition A2 holds for a large class of low-dimensional problems of interest in machine learning and one can learn the problem privately without actually needing to find a covering set algorithmically. Specifically, the example includes many practically used methods such as logistic regression, linear SVM, ridge regression, even multi-layer neural networks, since the loss functions in these methods are jointly bounded in (Z, h) and Lipschitz in h.

The example also raises an interesting observation that while differentially private classification is not possible in a distribution-free setting for 0-1 loss function [54], it is learnable under smoother surrogate loss, e.g., logistic loss or hinge loss. In other words, private learnability and computational tractability both benefit from the same relaxation.

The Lipschitz condition still requires the dimension of the hypothesis space to be $o(n)$. Thus it does not cover high-dimensional machine learning problems where $d \gg n$, nor does it contain the example of Shalev-Shwartz et al. [191] that ERM fails.

For high dimensional problems where d grows with n, typically some assumptions or restrictions need to be made either on the data or on the hypothesis space (so that it becomes essentially low-dimensional). We give one example here for the problem of sparse regression.

Example 2.22 (Best subset selection)**.** *Consider $\mathcal{H} = \{h \in \mathbb{R}^d : \|h\|_0 < s, \|h\|_2 \leq 1\}$ and let $\ell(h, z)$ be an L-Lipschitz loss function. The solution can only be chosen from $\binom{d}{s} < d^s$ different s-dimensional subspaces. We can apply Algorithm 1 twice to first sample a support set S with utility function being the $-\min_{h\in\mathcal{H}_S} F(Z, h)$, and then sample a solution in the chosen s-dimensional subspace. By the composition theorem this two-stage procedure is differentially private. Moreover, by the arguments in Example 2.20 and Example 2.21 respectively, we have an $\mu(\mathcal{S}_t) \geq \left(\frac{1}{d}\right)^s$ for the subset selection and $\mu(\mathcal{S}_t) \geq (\frac{t}{L})^s$ for the low-dimensional regression. Note that $\rho = 0$ in both cases and the dependency on the ambient dimension d is on the logarithm. The first stage ensures that for the chosen support set $\mathcal{S}$, $\min_{h\in\mathcal{H}_S} F(Z, h)$ is close to $\min_{h\in\mathcal{H}} F(Z, h)$ by $O(\frac{s\log d+\log n}{n\epsilon(n)})$ in expectation and (the second stage ensures that the sampled hypothesis from $\mathcal{H}_S$ would have objective function close to $\min_{h\in\mathcal{H}_S} F(Z, h)$ by $O(\frac{s\log L+s\log n+\log(\mu(\mathcal{H}_S)/\beta_p)}{n\epsilon(n)})$. This leads to an overall rate of consistency (they simply add up) of $O(\frac{s(\log d+\log n+L)+\log(\mu(\mathcal{H}_S)/\beta_p)}{\sqrt{n}})$ if we choose $\epsilon(n) = 1/\sqrt{n}$.*

2.4.2 Examples of privately $\mathfrak{D}$-learnable problems.

For problems where private learnability is impossible to achieve, we may still apply Theorem 2.19 to prove the weaker private $\mathfrak{D}$-learnability for some specific class of distributions.

Example 2.23 (Finite Representation Dimension in the General Learning Setting)**.** *For binary classification problems with 0-1 loss (PAC learning), this has been well-studied. In particular, Beimel et al. [19] characterized the sample complexity of privately learnable problems using a combinatorial condition they call a "Probabilistic Representation", which basically involves finding a finite, data-independent set of hypotheses to approximate any hypothesis in the class. Their claim is that if the "representation dimension" is finite, then the problem is privately learnable, otherwise it is not. We can extend the notion of probabilistic representation beyond the finite discrete and countably infinite hypothesis class considered in Beimel et al. [19] to cases* *when the problem is not privately learnable (e.g, learning threshold functions on* $[0,1]$*). The existence of probabilistic representation for all distributions in* $\mathfrak{D}$ *would lead to a* $\mathfrak{D}$*-universally private learning algorithm.*

Another way to define a class of distribution $\mathfrak{D}$ is to assume the existence of a reference distribution that is close to any distribution of interest as in [54].

Example 2.24 (Existence of a public reference distribution)**.** *To deal with the* 0-1 *loss classification problems on a continuous hypothesis domain, Chaudhuri and Hsu [54] assume that there exists a data-independent reference distribution* $\mathcal{D}^*$*, which by multiplying a fixed constant on its density, uniformly dominates any distributtion of interest. This essentially produces a subset of distributions* $\mathfrak{D}$*. The consequence is that one can build an* ϵ*-net of* $\mathcal{H}$ *with metric defined on the risk under* $\mathcal{D}^*$ *and this will also be a (looser) covering set of any distribution* $\mathcal{D} \in \mathfrak{D}$*, thereby learning the problem for any distribution in the set.*

The same idea can be applied to the general learning setting. For any fixed reference distribution $\mathcal{D}^*$ *defined on* $\mathcal{Z}$ *and constant* c*,*

$$\mathfrak{D} = \{\mathcal{D} = (\mathcal{Z}, \mathcal{F}, \mathbb{P}) \mid \mathbb{P}_{\mathcal{D}}(z \in A) \leq c\mathbb{P}_{\mathcal{D}^*}(z \in A) \textit{ for } \forall A \in \mathcal{F}\}$$

is a valid set of distributions and we are able to $\mathfrak{D}$*-privately learn this problem whenever we can construct a sufficiently small cover set with respect to* $\mathcal{D}^*$ *and reduce the problem to Example 2.20. This class of problems includes high-dimensional and infinity dimensional problems such as density estimation, nonparametric regression, kernel methods and essentially any other problems that are strictly learnable [234], since they are characterized by one-sided uniform convergence (and the corresponding entropy condition).*

2.4.3 Discussion on uniform convergence and private learnability

Uniform convergence requires that $\mathbb{E}_{Z\sim\mathcal{D}^n} \sup_{h\in\mathcal{H}} |\hat{R}(h,Z) - R(h)| \to 0$ for any distribution $\mathcal{D}$ with a distribution independent rate. Most machine learning algorithms rely on uniform convergence to establish consistency result (e.g., through complexity measure such as VC-dimension, Rademacher Complexity, covering and bracketing numbers and so on). In fact, the learnability of ERM algorithm is characterized by the one-sided uniform convergence [234], which is only slightly weaker than requiring uniform convergence on both sides.

A key point in Shalev-Shwartz et al. [191] is that the learnability (by any algorithm) in general learning setting is no longer characterized by variants of uniform convergence. However, the class of privately learnable problems is much smaller. Clearly, uniform convergence is not sufficient for a problem to be privately learnable (see Section 2.3.3), but is it necessary?

In binary classification with discrete domain (agnostic PAC Learning), since VC-dimension being finite characterizes the class of privately PAC learnable problems, the necessity of uniform convergence is clear. This could also be more explicitly seen from Beimel et al. [19] where the *probabilistic representation dimension* is a form of uniform convergence on its own.

In the general learning setting, the problem is still open. We were not able to prove that private learnability implies uniform convergence, but we could not construct a counter example either. All our examples in this section do implicitly or explicitly uses uniform convergence, which seems to hint at a positive answer.

2.5 Practical concerns

2.5.1 High confidence private learning via boosting

We have stated all results so far in expectation. We can easily convert these to the high-confidence learning paradigm by applying Markov's inequality, since convergence in expectation to the minimum risk implies convergence in probability to the minimum risk. While the $1/\delta$ dependence on the failure probability δ is not ideal, we can apply a similar meta-algorithm "boosting"[188] as in Shalev-Shwartz et al. [191, Section 7] to get a $\log(1/\delta)$ rate. The approach is similar to cross-validation. Given a pre-chosen positive integer a, the original boosting algorithm randomly partitions the data into $(a+1)$ subsamples of size $n/(a+1)$, and applies Algorithm 1 on the first a partitions, obtaining a candidate hypotheses. The method then returns the one hypothesis with smallest validation error, calculated using the remaining subsample. To ensure differential privacy, our method instead uses the exponential mechanism to sample the best candidate hypothesis, where the logarithm of sampling probability is proportional to the negative validation error.

Theorem 2.25 (High-confidence private learning)**.** *If an algorithm $\mathcal{A}$ privately learns a problem with rate $\xi(n)$ and privacy parameter $\epsilon(n)$, then the boosting algorithm $\mathcal{A}'$ with $a = \log\frac{3}{\delta}$ is $\max\left\{\epsilon\left(\frac{n}{\log(3/n)+1}\right), \frac{\log(3/\delta)+1}{\sqrt{n}}\right\}$-differentially private, its output h obeys*

$$R(h) - R^* \le e\xi\left(\frac{n}{\log(3/\delta)+1}\right) + C\sqrt{\frac{\log(3/\delta)}{n}}$$

for an absolute constant C with probability at least $1-\delta$.

2.5.2 Efficient sampling algorithm for convex problems

Our proposed exponential sampling based algorithm is to establish a more explicit geometric condition upon which AERM holds, hence the algorithm may not be computationally tractable. Ignoring the difficulty of constructing the ϵ-covering set of an exponential number of elements, sampling from the set alone is not a polynomial time algorithm. But we can solve a subset of the continuous version of our Algorithm 1 described in Theorem 2.19 in polynomial time to arbitrary accuracy (see also Bassily et al. [17, Theorem 3.4]).

Proposition 2.26. *If $n^{-1}\sum_{i=1}^{n}\ell(h,z_i) + g_n(h)$ is convex in h and $\mathcal{H}$ is a convex set, then the sampling procedure in Algorithm 1 can be solved in polynomial time.*

Proof. When $n^{-1}\sum_{i=1}^{n}\ell(h,z_i)+g_n(h)$ is convex, the utility function $q(h,Z)$ is concave in h. The density to be sampled from in Algorithm 1 is proportional to $\exp(\frac{\epsilon nq(h,Z)}{B})$ and is log-concave. The Markov chain sampling algorithm in Applegate and Kannan [11] is guaranteed to produce a sample from a distribution that is arbitrarily close to the target distribution (in the total variation sense) in polynomial time. □

2.5.3 Exponential mechanism in infinite domain

As we mention earlier, the results in Section 2.4 based on the exponential mechanism implicitly assumes certain regularity conditions that ensures the existence of a probability distribution.

When $\mathcal{H}$ is finite, the existence is trivial. On the other hand, an infinite set $\mathcal{H}$ is tricky in that there may not exist a proper distribution that satisfies $\mathbb{P}(h)\propto e^{\frac{\epsilon}{2\Delta_q}q(Z,h)}$ for at least some $q(Z,h)$. For instance, if $\mathcal{H}=\mathbb{R}$ and $q(Z,h)\equiv 1$ then $\int_{\mathbb{R}} e^{\frac{\epsilon}{2\Delta_q}q(Z,h)}dh=\infty$. Such distributions that are only defined up to scale with no finite normalization constants are called improper distributions. In case of finite dimensional non-compact set, this translates into an additional assumption on the loss function and the regularization term.

Things get even trickier when $\mathcal{H}$ is an infinite dimensional space, such as a subset of a Hilbert space. While probability measures can still be defined, no density function can be defined on such spaces. Therefore, we cannot use exponential mechanism to define a valid probability distribution.

The practical implication is that exponential mechanism is really only applicable to cases when the hypothesis space $\mathcal{H}$ allows for definitions of densities in the usual sense, or then $\mathcal{H}$ can be approximated by such a space. For example, a separable Hilbert space can be studied by finite-dimensional projections. Also, we can approximate RKHS induced by translation invariant kernels via random Fourier features [174].

2.6 Results for learnability under (ϵ,δ)-differential privacy

Another way to weaken the definition of private learnability is through (ϵ,δ)-approximate differential privacy.

Definition 2.27 ([83])**.** *An algorithm $\mathcal{A}$ obeys (ϵ,δ)-differential privacy if for any Z,Z' such that $d(Z,Z')\leq 1$, and for any measurable set $\mathcal{S}\subset\mathcal{H}$*

$$\mathbb{P}_{h\sim\mathcal{A}(Z)}(h\in\mathcal{S})\leq e^{\epsilon}\mathbb{P}_{h\sim\mathcal{A}(Z')}(h\in\mathcal{S})+\delta.$$

We define a version of the problem to be

Definition 2.28 (Approximately Private Learnability)**.** *We say a learning problem is $\Delta(n)$-approximately privately learnable for some pre-specified family of rate $\Delta(n)$ if for some $\epsilon<\infty$, $\delta(n)\in\Delta(n)$, there exists a universally consistent algorithm that is $(\epsilon,\delta(n))$-DP.*

This is a completely different subject to study and the class of approximately privately learnable problems could be substantially larger than the pure privately learnable problems. Moreover, the picture may vary with respect to how small $\delta(n)$ is required to be. In this section, we present our preliminary investigation on this problem.

Specifically, we will consider two questions:

1. Does the existence of an (ϵ, δ)-DP always AERM algorithm characterize the class of approximately private learnable problems?
2. Are all learnable problems approximately privately learnable for different choices of $\Delta(n)$?

The minimal requirement in the same flavor of Definition 2.3 would be to require $\Delta(n) = \{\delta(n) | \delta(n) \to 0\}$. The learnability problem turns out to be trivial under this definition due to the following observation.

Lemma 2.29. *For any algorithm $\mathcal{A}$ that acts on Z, $\mathcal{A}'$ that runs $\mathcal{A}$ on a randomly chosen subset of Z of size $\sqrt{n}$ is $(0, \frac{1}{\sqrt{n}})$-DP.*

Proof. Let Z and Z' be adjacent datasets that differs only in data point i. For any i and any $S \in \sigma(\mathcal{H})$.

$$\begin{aligned}
\mathbb{P}(\mathcal{A}'(Z) \in S) &= \mathbb{P}_I(\mathcal{A}(Z_I) \in S | i \in I)\mathbb{P}(i \in I) + \mathbb{P}_I(\mathcal{A}(Z_I) \in S | i \notin I)\mathbb{P}(i \notin I) \\
&= \mathbb{P}_I(\mathcal{A}(Z_I) \in S | i \in I)\mathbb{P}(i \in I) + \mathbb{P}_I(\mathcal{A}(Z_I') \in S | i \notin I)\mathbb{P}(i \notin I) \\
&= \mathbb{P}(\mathcal{A}'(Z') \in S) + [\mathbb{P}_I(\mathcal{A}(Z_I) \in S | i \in I) - \mathbb{P}_I(\mathcal{A}'(Z_I) \in S | i \in I)]\mathbb{P}(i \in I) \\
&\leq \mathbb{P}(\mathcal{A}'(Z') \in S) + \mathbb{P}(i \in I) \\
&= e^0 \mathbb{P}(\mathcal{A}'(Z') \in S) + \frac{1}{\sqrt{n}}.
\end{aligned}$$

This verifies the $(0, 1/\sqrt{n})$-DP of algorithm $\mathcal{A}'$. □

The above lemma suggests that if $\delta(n) = o(1)$ is all we need for the *approximately private learnability*, then any consistent learning algorithm can be made approximately DP by simply subsampling. In other words, any learnable problem is also learnable under approximate differential privacy.

To get around this triviality, we need to specify a sufficiently fast rate of $\delta(n)$ going to 0. While it is common to require that $\delta(n) = o(1/\text{poly}(n))$ [2] for cryptographically strong privacy protection, requiring $\delta(n) = o(1/n)$ is already enough to invalidate the above subsampling argument and makes the problem of learnability a non-trivial one.

Again, the question is whether AERM characterizes approximately private learnability and whether there is a gap between the class of learnable and approximately privately learnable problems.

Here we show that the "folklore" Lemma 2.8 and subsampling lemma (Lemma 2.34) can be extended to work with (ϵ, δ)-DP and then we provide a positive answer to the first question.

[2] Here the notation "$o(1/\text{poly}(n))$" means "decays faster than any polynomial of n". A sequence $a(n) = o(1/\text{poly}(n))$ if and only if $a(n) = o(n^{-r})$ for any $r > 0$.

Lemma 2.30 (Stability of (ϵ,δ)-DP)**.** *If $\mathcal{A}$ is (ϵ,δ)-DP, and $0 \le \ell(h,z) \le 1$, then $\mathcal{A}$ is $(e^\epsilon - 1 + \delta)$-Strongly Uniform RO-stable.*

Proof. For any Z, Z' such that $d(Z,Z') \le 1$ and for any $z \in \mathcal{Z}$. Let the event $E = \{h | p(h) \ge p'(h)\}$,

$$\begin{aligned}
&\left|\mathbb{E}_{h\sim\mathcal{A}(Z)}\ell(h,z) - \mathbb{E}_{h\sim\mathcal{A}(Z')}\ell(h,z)\right| = \left|\int_h \ell(h,z)p(h)dh - \int_h \ell(h,z)p'(h)dh\right| \\
\le&\sup_{h,z}\ell(h,z)\int_E p(h) - p'(h)dh \le \int_E p(h) - p'(h)dh = \mathbb{P}_{h\sim\mathcal{A}(Z)}(h\in E) - \mathbb{P}_{h\sim\mathcal{A}(Z')}(h \in E) \\
\le&(e^\epsilon - 1)\mathbb{P}_{h\sim\mathcal{A}(Z')}(h\in E) + \delta \le e^\epsilon - 1 + \delta.
\end{aligned}$$

The last line applies the definition of (ϵ,δ)-DP. □

Lemma 2.31 (Subsampling Lemma of (ϵ,δ)-DP)**.** *If $\mathcal{A}$ is (ϵ,δ)-DP, then $\mathcal{A}'$ that acts on a random subsample of Z of size γn obeys (ϵ',δ')-DP with $\epsilon' = \log(1+\gamma e^\epsilon(e^\epsilon - 1))$ and $\delta' = \gamma e^\epsilon \delta$.*

Proof. For any event $E \in \sigma(\mathcal{H})$, let i be the coordinate where Z and Z' differs

$$\begin{aligned}
&\mathbb{P}_{h\sim\mathcal{A}'(Z)}(h\in E) = \gamma\mathbb{P}_{h\sim A(Z_I)}(h\sim E|i\in I) + (1-\gamma)\mathbb{P}_{h\sim A(Z_I)}(h\sim E|i\notin I) \\
=&\gamma\mathbb{P}_{h\sim A(Z_I)}(h\sim E|i\in I) + (1-\gamma)\mathbb{P}_{h\sim A(Z_I')}(h\sim E|i\notin I) \\
=&\gamma\mathbb{P}_{h\sim A(Z_I)}(h\sim E|i\in I) - \gamma\mathbb{P}_{h\sim A(Z_I')}(h\sim E|i\in I) + \gamma\mathbb{P}_{h\sim A(Z_I')}(h\sim E|i\in I) \\
&+ (1-\gamma)\mathbb{P}_{h\sim A(Z_I')}(h\sim E|i\notin I) \\
=&\mathbb{P}_{h\sim\mathcal{A}'(Z')}(h\in E) + \gamma[\mathbb{P}_{h\sim A(Z_I)}(h\sim E|i\in I) - \mathbb{P}_{h\sim A(Z_I')}(h\sim E|i\in I)] \\
\le&\mathbb{P}_{h\sim\mathcal{A}'(Z')}(h\in E) + \gamma(e^\epsilon - 1)\mathbb{P}_{h\sim A(Z_I')}(h\sim E|i\in I) + \gamma\delta, \qquad (2.12)
\end{aligned}$$

where in last line, we apply (ϵ,δ)-DP of $\mathcal{A}$.

It remains to show that $\mathbb{P}_{h\sim A(Z_I')}(h\sim E|i\in I)$ is similar to $\mathbb{P}_{h\sim\mathcal{A}'(Z')}(h\in E)$. First,

$$\mathbb{P}_{h\sim\mathcal{A}'(Z')}(h\in E) = \gamma\mathbb{P}_{h\sim\mathcal{A}(Z_I')}(h\in E|i\in I) + (1-\gamma)\mathbb{P}_{h\sim\mathcal{A}(Z_I')}(h\in E|i\notin I). \qquad (2.13)$$

Denote $\mathcal{I}_1 = \{I|i\in I\}$, $\mathcal{I}_2 = \{I|i\notin I\}$. We known $|\mathcal{I}_1| = \binom{n-1}{\gamma n-1}$, and $|\mathcal{I}_2| = \binom{n-1}{\gamma n}$ and $|\mathcal{I}_1|/|\mathcal{I}_2| = \gamma n/(n-\gamma n)$. For every $I\in\mathcal{I}_2$ there are precisely γn elements $J\in\mathcal{I}_1$ such that $d(I,J)=1$. Likewise, for every $J\in\mathcal{I}_1$, there are $n-\gamma n$ elements $I\in\mathcal{I}_2$ such that $d(I,J)=1$. It follows by symmetry that if we apply (ϵ,δ)-DP to $1/\gamma n$ of each $I\in\mathcal{I}_2$ and change I to their corresponding $J\in\mathcal{I}_1$, then each $J\in\mathcal{I}_1$ will receive $(n-\gamma n)/\gamma n$ “contribution” in total from the sum over all $I \in \mathcal{I}_2$.

$$\begin{aligned}
&\mathbb{P}_{h\sim\mathcal{A}(Z'_I)}(h \in E|i \notin I) = \frac{1}{|\mathcal{I}_2|}\sum_{I\in\mathcal{I}_2}\mathbb{P}_{h\sim\mathcal{A}(Z'_I)}(h\in E)\\
=&\frac{1}{|\mathcal{I}_2|}\sum_{I\in\mathcal{I}_2}\sum_{j=1}^{\gamma n}\frac{1}{\gamma n}\mathbb{P}_{h\sim\mathcal{A}(Z'_I)}(h\in E)\\
\geq&\frac{|\mathcal{I}_1|}{|\mathcal{I}_2|}\frac{1}{|\mathcal{I}_1|}\sum_{J\in\mathcal{I}_1}\frac{n-\gamma n}{\gamma n}e^{-\epsilon}(\mathbb{P}_{h\sim\mathcal{A}(Z'_J)}(h\in E)-\delta)\\
=&\frac{1}{|\mathcal{I}_1|}\sum_{J\in\mathcal{I}_1}e^{-\epsilon}(\mathbb{P}_{h\sim\mathcal{A}(Z'_J)}(h\in E)-\delta) = e^{-\epsilon}\mathbb{P}_{h\sim\mathcal{A}(Z'_I)}(h\in E|i\in I) - e^{-\epsilon}\delta
\end{aligned}$$

Substitute into (2.13), we get

$$\mathbb{P}_{h\sim\mathcal{A}(Z'_I)}(h\in E|i\in I) \leq \frac{1}{\gamma+(1-\gamma)e^{-\epsilon}}\mathbb{P}_{h\sim\mathcal{A}'(Z')}(h\in E) + \frac{(1-\gamma)e^{-\epsilon}}{\gamma+(1-\gamma)e^{-\epsilon}}\delta.$$

We further relax the upper bound to a simple form $e^{\epsilon}\mathbb{P}_{h\sim\mathcal{A}'(Z')}(h\in E)+\delta$ and substitute into (2.12), we have

$$\mathbb{P}_{h\sim\mathcal{A}'(Z)}(h\in E) \leq (1+\gamma e^{\epsilon}(e^{\epsilon}-1))\mathbb{P}_{h\sim\mathcal{A}'(Z')}(h\in E) + \gamma\delta + \gamma(e^{\epsilon}-1)\delta,$$

which concludes the proof. □

Using the above two lemmas, we are able to establish the same result which says that AERM characterizes the approximate private learnability for certain classes of $\Delta(n)$.

Theorem 2.32. *A problem is $\Delta(n)$-approximately privately learnable implies that there exists an always AERM algorithm that is $(\epsilon(n), n^{-1/2}e^{\epsilon}\delta(\sqrt{n}))$-DP for some $\epsilon(n)\to 0$ and $\delta(\sqrt{n})\in\Delta(n)$. The converse is also true if $n^{-1/2}e^{\epsilon}\delta(\sqrt{n})\in\Delta(n)$.*

Proof. If we have an always AERM algorithm with $\xi_{erm}(n)$ that is $(\epsilon(n),\delta(n))$-DP for $\delta(n)\in\Delta(n)$. Then by Lemma 2.30, this algorithm is strongly uniform RO-stable with rate $e^{\epsilon(n)}-1+\delta(n)$. By Theorem 2.35, the algorithm is universally consistent with rate $\xi_{erm}(n)+e^{\epsilon(n)}-1+\delta(n)$. This establishes the "if" part.

To see the "only if" part, by definition if a problem is $\Delta(n)$-approximately privately learnable with ϵ and $\delta(n)\in\Delta(n)$. Then by Lemma 2.31 with $\gamma=1/\sqrt{n}$, we get an algorithm that obeys the privacy condition. It remains to prove always AERM, which requires exactly the same arguments in the proof of Lemma 2.10. Details are omitted. □

Note that the results above suggest that in the two canonical settings $\Delta(n)=o(1/n)$ or $\Delta(n)=o(1/\text{poly}(n))$, existence of a private AERM algorithm that satisfies the stronger constraint $\epsilon(n)=o(1)$ characterizes the learnability.

The next question that whether any learnable problems are also approximately privately learnable would depend on how fast $\delta(n)$ is required to decay. We know that when we only have $\Delta(n) =$

$o(1)$, all learnable problems are approximately privately learnable, and when we have $\Delta(n) = \{0\}$, only a strict subset of these problems is privately learnable. The following result establishes that when $\delta(n)$ needs to go to 0 with a sufficiently fast rate, there is separation between learnability and approximately private learnability.

Proposition 2.33. *Let $\Delta(n) = \{\delta(n)|\delta(n) \leq \tilde{\delta}(n)\}$ for some sequence $\tilde{\delta}(n) \to 0$. The following statements are true.*

- *All learnable problems are $\Delta(n)$-approximately privately learnable, if $\tilde{\delta}(n) = \omega(1/n)$.*
- *There exists a problem that is learnable but not $\Delta(n)$-approximately privately learnable, if $\tilde{\delta}(n) \leq \frac{\exp(-\epsilon(n^2)n^2)}{n}$*

Proof. The first claim follows from the same argument in Lemma 2.29. If a problem is learnable, there exists a universally consistent learning algorithm $\mathcal{A}$. The algorithm that applies $\mathcal{A}$ on a $\tilde{\delta}(n)$-fraction random subsample of the dataset is $(0, \tilde{\delta}(n))$-DP and universally consistent with rate $\xi(n\tilde{\delta}(n))$. Since $\tilde{\delta}(n) = \omega(1/n)$, $n\tilde{\delta}(n) \to \infty$.

We now show that when we require a fast decaying $\delta(n)$, then suddenly the example in Section 2.3.3 due to Chaudhuri and Hsu [54] becomes not approximately privately learnable even for (ϵ, δ)-DP. Let Z, Z' be two completely different data sets, by repeatedly applying the definition of (ϵ, δ)-DP, for any set $\mathcal{S} \subset \mathcal{H}$

$$\mathbb{P}(\mathcal{A}(Z) \in \mathcal{S}) \leq e^{n\epsilon}\mathbb{P}(\mathcal{A}(Z) \in \mathcal{S}) + \sum_{i=1}^{n} e^{(i-1)\epsilon}\delta \leq e^{n\epsilon}\mathbb{P}(\mathcal{A}(Z') \in \mathcal{S}) + ne^{(n-1)\epsilon}\delta.$$

When we shift the inequality around, we get

$$\mathbb{P}(\mathcal{A}(Z') \in \mathcal{S}) \leq e^{-n\epsilon}\mathbb{P}(\mathcal{A}(Z') \in \mathcal{S}) - e^{-\epsilon}n\delta.$$

Consider the same example in Section 2.3.3 where we hope to learn a threshold on $[0,1]$. Assuming there exists an algorithm $\mathcal{A}$ that is universally AERM and $(\epsilon(n), \delta(n))$-DP for $\epsilon(n) < \infty$ and $\delta(n) \leq 0.4ne^{-\epsilon n}$.

Everything up to (2.4) remains exactly the same. Now, apply the above implication of (ϵ, δ)-DP, we can replace (2.4) for each $i = 2, ..., K$, by

$$\mathbb{P}(\mathcal{A}(Z_1) \in [h_i - \eta/3, h_i + \eta/3]) \geq \exp(-\epsilon n)\mathbb{P}(\mathcal{A}(Z_i) \in [h_i - \eta/3, h_i + \eta/3]) - n\delta(n).$$

Then (2.5) becomes

$$\mathbb{P}(\mathcal{A}(Z_1) \notin [h_1 - \eta/3, h_1 + \eta/3]) \geq K\exp(-\epsilon n)0.9 - Ke^{-\epsilon}n\delta(n) \geq 0.9 \geq 0.5,$$

where the last inequality follows by $K > \exp(\epsilon n)$ and $\delta(n) \leq 0.4ne^{-\epsilon n}$. This yields the same contradiction to always AERM of $\mathcal{A}$ on Z_1, which requires $\mathbb{P}(\mathcal{A}(Z_1) \notin [h_1 - \eta/3, h_1 + \eta/3]) < 0.1$. Therefore, such AERM does not exist. By the contrapositive of Theorem 2.32, the problem is not approximately privately learnable for $\tilde{\delta}(n) \leq \frac{\exp(-\epsilon(n^2)n^2)}{n}$. □

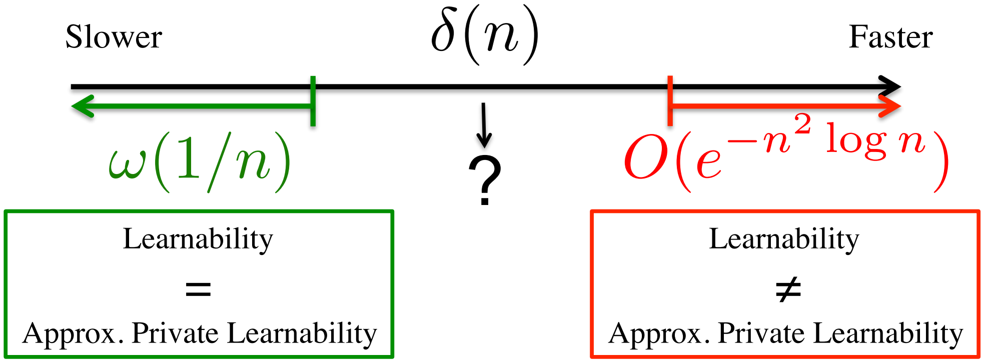

Slower
$\delta(n)$
Faster
$\omega(1/n)$
?
$O(e^{-n^2 \log n})$
Learnability = Approx. Private Learnability
Learnability ≠ Approx. Private Learnability

Figure 2.4: Illustration of Proposition 2.33 and the open problem.

The bound can be further improved to $\exp(-\epsilon(n)n)/n$ if we directly work with universal consistency on various distributions rather than through always AERM on specific data points. Even that is likely to be suboptimal as there might be more challenging problems and less favorable packings to consider.

The point of this exposition, however, is to illustrate that (ϵ, δ)-DP alone does not close the gap between learnability and private learnability. Additional relaxation on the specified rate of decay on δ does. We now know that the phase transition occurs when $\delta(n)$ is somewhere between $\Omega(\exp(-n^2 \log n))$ and $O(1/n)$; but there is still a substantial gap between the upper and lower bounds. [3].

2.7 Conclusion and future work

In this chapter, we revisited the question *"What can we learn privately?"* and considered a broader class of statistical machine learning problems than those studied previously. Specifically, we characterized the learnability under privacy constraint by showing any privately learnable problems can be learned by a private algorithm that asymptotically minimizes the empirical risk for any data, and the problem is not privately learnable otherwise. This allows us to construct a conceptual procedure that privately learns any privately learnable problem. We also propose a relaxed notion of private learnability called private $\mathfrak{D}$-learnability, which requires the existence of an algorithm that is consistent for any the distribution within a class of distributions $\mathfrak{D}$. We characterized private $\mathfrak{D}$-learnability too with a weaker notion of AERM. For problems that can be formulated as penalized empirical risk minimization, we provide a sampling algorithm with a set of meaningful sufficient conditions on the geometry of the hypothesis space and demonstrate that it covers a large class of problems. In addition, we further extended the characterization to learnability under

[3] After the paper was accepted for publication, we became aware that the phase transition occurs sharply at $O(1/n)$. The result follows from a sharp lower bound of sample complexity in learning threshold functions in Bun [45, Theorem 4.5.2], which improves over a previously published result that requires $O(n^{-1-\alpha})$ for any $\alpha > 0$ in Bun et al. [44]. The consequence is that the general learning setting is hard for (ϵ, δ)-DP too unless δ becomes meaninglessly large for privacy purposes.

(ϵ, δ)-differential privacy and provided a preliminary analysis which establishes the existence of a phase transition from all learnable problems being approximately private learnable to some learnable problems being not approximately private learnable at some non-trivial rate of decay on $\delta(n)$.

Future work includes understanding the conditions under which privacy and AERM are contradictory (recall that we only have one example on learning thresholding functions due to Chaudhuri and Hsu [54]), characterizing the rate of convergence, searching for practical algorithms that generically learns all privately learnable problems, and better understanding the gap between learnability and approximate private learnability.

2.8 Proofs of technical results

In this section, we provide detailed proofs to the technical results that in the main text.

2.8.1 Privacy in subsampling

Proof of Lemma 2.4. Let $\mathcal{A}$ be the consistent ϵ-DP algorithm. Consider $\mathcal{A}'$ that apply $\mathcal{A}$ to a random subsample of $\lfloor\sqrt{n}\rfloor$ data points. By Lemma 2.34 with $\gamma = \frac{\lfloor\sqrt{n}\rfloor}{n} \leq \frac{1}{\sqrt{n}}$, we get the privacy claim. For the consistency claim, note that the given sample is an iid sample of size $\sqrt{n}$ from the original distribution. □

Lemma 2.34 (Subsampling theorem)**.** *If Algorithm $\mathcal{A}$ is ϵ-DP for $Z \in \mathcal{Z}^n$ for any $n = 1, 2, 3, \ldots$, then the algorithm $\mathcal{A}'$ that output the result of $\mathcal{A}$ to a random subsample of size γn data points preserves $2\gamma(e^{\epsilon} - e^{-\epsilon})$-DP.*

Proof of Lemma 2.34 (Subsampling theorem). This is a corollary of Lemma 4.4 in Beimel et al. [21]. To be self-contained, we reproduce the proof here in our notation.

Recall that $\mathcal{A}'$ is the algorithm that first randomly subsample γn data points then apply $\mathcal{A}$. Let Z and Z' be any neighboring databases and assume they differ on the ith data point. Let $\mathcal{S} \subset [n]$ be the indices of the random subset of the entries that are selected, and $\mathcal{R} \subset [n]\backslash\{i\}$ be a index size of size $\gamma n - 1$. We apply the law of total expectation twice and argue that for any adjacent Z, Z', any event $E \subset \mathcal{H}$,

$$\frac{\mathbb{P}_{h\sim\mathcal{A}'(Z)}(h \in E)}{\mathbb{P}_{h\sim\mathcal{A}'(Z')}(h \in E)} = \frac{\gamma\mathbb{P}_{h\sim\mathcal{A}(Z_\mathcal{S})}(h \in E|i \in \mathcal{S}) + (1-\gamma)\mathbb{P}_{h\sim\mathcal{A}(Z_\mathcal{S})}(h \in E|i \notin \mathcal{S})}{\gamma\mathbb{P}_{h\sim\mathcal{A}(Z'_\mathcal{S})}(h \in E|i \in \mathcal{S}) + (1-\gamma)\mathbb{P}_{h\sim\mathcal{A}(Z'_\mathcal{S})}(h \in E|i \notin \mathcal{S})}$$

$$= \frac{\sum_{\mathcal{R}\in[n]\backslash\{i\}} \mathbb{P}(\mathcal{R}) \left[\gamma\mathbb{P}_{h\sim\mathcal{A}(Z_\mathcal{S})}\left(h \in E|\mathcal{S} = \mathcal{R}\cup\{i\}\right) + (1-\gamma)\mathbb{P}_{h\sim\mathcal{A}(Z_\mathcal{S})}\left(h \in E|\mathcal{S} = \mathcal{R}\cup\{j\}, j \neq i\right)\right]}{\sum_{\mathcal{R}\in[n]\backslash\{i\}} \mathbb{P}(\mathcal{R}) \left[\gamma\mathbb{P}_{h\sim\mathcal{A}(Z'_\mathcal{S})}\left(h \in E|\mathcal{S} = \mathcal{R}\cup\{i\}\right) + (1-\gamma)\mathbb{P}_{h\sim\mathcal{A}(Z'_\mathcal{S})}\left(h \in E|\mathcal{S} = \mathcal{R}\cup\{j\}, j \neq i\right)\right]}$$

By the given condition that $\mathcal{A}$ is ϵ-DP, we can replace $\mathcal{R}\cup\{i\}$ with $\mathcal{R}\cup\{j\}$ for an arbitrary j

with bounded changes in the probability and the above likelihood ratio can be upper bounded by

$$\frac{(\gamma e^{\epsilon}+1-\gamma)\mathbb{E}_{\mathcal{R}\in[n]\setminus\{i\},j\neq i}\mathbb{P}_{h\sim\mathcal{A}(Z_{\mathcal{S}})}(h\in E|\mathcal{S}=\mathcal{R}\cup\{j\})}{(\gamma e^{-\epsilon}+1-\gamma)\mathbb{E}_{\mathcal{R}\in[n]\setminus\{i\},j\neq i}\mathbb{P}_{h\sim\mathcal{A}(Z_{\mathcal{S}})}(h\in E|\mathcal{S}=\mathcal{R}\cup\{j\})} = \frac{\gamma e^{\epsilon}+1-\gamma}{\gamma e^{-\epsilon}+1-\gamma} = \frac{1+\gamma(e^{\epsilon}-1)}{1+\gamma(e^{-\epsilon}-1)}.$$

By definition, the privacy loss of the algorithm $\mathcal{A}'$ is therefore

$$\epsilon' \leq \log\left(1+\gamma[e^{\epsilon}-1]\right) - \log\left(1+\gamma\left[e^{-\epsilon}-1\right]\right).$$

Note that $\epsilon > 0$ implies that $-1 \leq e^{-\epsilon}-1 < 0$ and $0 < e^{\epsilon}-1 < \infty$. The result follows by applying the property of the natural logarithm:

$$\begin{aligned} \log(1+x) &\leq \frac{x}{2}\frac{2+x}{1+x} \leq x && \text{for } 0 \leq x < \infty \\ \log(1+x) &\geq \frac{x}{2}\frac{2+x}{1+x} \geq \frac{x}{1+x} && \text{for } -1 \leq x \leq 0 \end{aligned}$$

to upper bound the expression. □

2.8.2 Characterization of private learnability

Privacy implies stability Lemma 2.8 says that an ϵ-differentially private algorithm is $(e^{\epsilon}-1)$-stable (and also 2ϵ-stable if $\epsilon < 1$).

Proof of Lemma 2.8. Construct Z' by replacing an arbitrary data point in Z with z' and let the probability density/mass defined by $\mathcal{A}(Z)$ and $\mathcal{A}(Z')$ be $p(h)$ and $p'(h)$ respectively, then we can bound the stability as follows

$$\begin{aligned} &\left|\mathbb{E}_{h\sim\mathcal{A}(Z)}\ell(h,z) - \mathbb{E}_{h\sim\mathcal{A}(Z')}\ell(h,z)\right| \\ =&\left|\int_h \ell(h,z)p(h)dh - \int_h \ell(h,z)p'(h)dh\right| = \left|\int_h \ell(h,z)(p(h)-p'(h))dh\right| \\ \leq&\sup_{h,z}|\ell(h,z)|\int_{p(h)\geq p'(h)} p(h)-p'(h)dh \leq 1\cdot\int_{p(h)\geq p'(h)} p'(h)(\frac{p(h)}{p'(h)}-1)dh \\ \leq&(e^{\epsilon}-1)\int_{p(h)\geq p'(h)} p'(h)dh \leq (e^{\epsilon}-1). \end{aligned}$$

For $\epsilon < 1$ we have $\exp(\epsilon)-1 < 2\epsilon$.

□

Stability + AERM ⇒ consistency

Theorem 2.35 (Randomized version of Shalev-Shwartz et al. 191, Theorem 8)**.**
If any algorithm is $\xi_1(n)$-stable and $\xi_2(n)$-AERM then it is consistent with rate $\xi(n) = \xi_1(n) + \xi_2(n)$.

Proof. We will show the following the two steps as in Shalev-Shwartz et al. [191]

1. Uniform RO stability ⇒ On average stability ⇔ On average generalization
2. AERM + On average generalization ⇒ consistency

The definition of these quantities is self-explanatory.

To show that "stability implies generalization", we have

$$
\begin{aligned}
&\left|\mathbb{E}_{Z\sim\mathcal{D}^n}\left(\mathbb{E}_{h\sim\mathcal{A}(Z)}R(h) - \mathbb{E}_{h\sim\mathcal{A}(Z)}\hat{R}(h,Z)\right)\right| \\
=&\left|\mathbb{E}_{Z\sim\mathcal{D}^n}\left(\mathbb{E}_{z\sim\mathcal{D}}\mathbb{E}_{h\sim\mathcal{A}(Z)}\ell(h,z) - \frac{1}{n}\mathbb{E}_{h\sim\mathcal{A}(Z)}\sum_{i=1}^{n}\ell(h,z_i)\right)\right| \\
=&\left|\mathbb{E}_{Z\sim\mathcal{D}^n,\{z_1',\dots,z_n'\}\sim\mathcal{D}^n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{h\sim\mathcal{A}(Z)}\ell(h,z_i') - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{h\sim\mathcal{A}(Z^{(i)})}\ell(h,z_i')\right)\right| \\
\le&\sup_{Z,Z^{(i)}\in\mathcal{Z}^n,d(Z,Z^{(i)})=1,z'\in\mathcal{Z}}\left|\mathbb{E}_{h\sim\mathcal{A}(Z)}\ell(h,z') - \mathbb{E}_{h\sim\mathcal{A}(Z^{(i)})}\ell(h,z')\right| \le \xi_1(n)\,,
\end{aligned}
$$

where $Z^{(i)}$ is obtained by replacing the ith entry of Z with z_i'. Next, we show that "generalization and AERM implies consistency". Let $h^* \in \arg\inf_{h\in\mathcal{H}} R(h)$. By definition, we have $\mathbb{E}_{Z\sim\mathcal{D}^n}\hat{R}(h^*,Z) = R^*$. It follows that

$$
\begin{aligned}
&\mathbb{E}_{Z\sim\mathcal{D}^n}[\mathbb{E}_{h\in\mathcal{A}(Z)}R(h) - R^*] = \mathbb{E}_{Z\sim\mathcal{D}^n}[\mathbb{E}_{h\in\mathcal{A}(Z)}R(h) - \hat{R}(h^*,Z)] \\
=&\ \mathbb{E}_{Z\sim\mathcal{D}^n}[\mathbb{E}_{h\in\mathcal{A}(Z)}R(h) - \mathbb{E}_{h\in\mathcal{A}(Z)}\hat{R}(h,Z)] + \mathbb{E}_{Z\sim\mathcal{D}^n}[\mathbb{E}_{h\in\mathcal{A}(Z)}\hat{R}(h,Z) - \hat{R}(h^*,Z)] \\
\le&\ \mathbb{E}_{Z\sim\mathcal{D}^n}[\mathbb{E}_{h\in\mathcal{A}(Z)}R(h) - \mathbb{E}_{h\in\mathcal{A}(Z)}\hat{R}(h,Z)] + \mathbb{E}_{Z\sim\mathcal{D}^n}[\mathbb{E}_{h\in\mathcal{A}(Z)}\hat{R}(h,Z) - \hat{R}^*(Z)] \\
\le&\ \xi_1(n) + \xi_2(n)\,.
\end{aligned}
$$

□

Privacy + AERM ⇒ consistency

Proof of Corrollary 2.9. It follows by combining Lemma 2.8 and Theorem 2.35. □

Necessity

Proof of Lemma 2.10. We construct an algorithm $\mathcal{A}'$ by subsampling the data points using a random subset of $\sqrt{n}$ and then running $\mathcal{A}$. The privacy claim follows from Lemma 2.34 directly.

To prove the “always AERM” claim, we adapt the proof of Lemma 24 in Shalev-Shwartz et al. [191]. For any fixed data set $Z \in \mathcal{Z}^n$,

$$\begin{aligned}\hat{R}(\mathcal{A}'(Z), Z) - \hat{R}^*(Z) &= \mathbb{E}_{Z' \subset Z, |Z'| = \lfloor \sqrt{n} \rfloor} \left[\hat{R}(\mathcal{A}(Z'), Z) - \hat{R}^*(Z) \right] \\ &= \mathbb{E}_{Z' \sim \text{Unif}(Z)^{\lfloor \sqrt{n} \rfloor}} \left[\hat{R}(\mathcal{A}(Z'), Z) - \hat{R}^*(Z) \middle| \text{ no duplicates } \right] \\ &\leq \frac{\mathbb{E}_{Z' \sim \text{Unif}(Z)^{\lfloor \sqrt{n} \rfloor}} \left[\hat{R}(\mathcal{A}(Z'), Z) - \hat{R}^*(Z) \right]}{\mathbb{P}(\text{no duplicates})},\end{aligned}$$

where Unif(Z) is the uniform distribution defined on the n points in Z. We need to condition on the event that there are no duplicates for the second equality to hold because Z' is a subsample taken without replacements. The last inequality is by the law of total expectation and the non-negativity of the conditional expectation. But $\mathbb{P}(\text{no duplicates}) = \prod_{i=0}^{\lfloor \sqrt{n} \rfloor - 1}(1 - i/n) \geq 1 - \sum_{i=0}^{\lfloor \sqrt{n} \rfloor - 1} i/n \geq 1/2$. By universal consistency, $\mathcal{A}$ is consistent on the discrete uniform distribution defined on Z, so

$$\hat{R}(\mathcal{A}'(Z), Z) - \hat{R}^*(Z) \leq 2\mathbb{E}_{Z' \sim \text{Unif}(Z)^{\lfloor n \rfloor}} \left[\hat{R}(\mathcal{A}(Z'), Z) - \hat{R}^*(Z) \right] \leq 2\xi(\sqrt{n}).$$

It is obvious that $\mathcal{A}'$ is consistent with rate $\sqrt{n}$ as it applies $\mathcal{A}$ on a random sample of size $\sqrt{n}$. By Lemma 2.4, $\mathcal{A}'$ is $2n^{-1/2}(e^{\epsilon} - e^{-\epsilon})$ differentially private. By Corollary 2.9, the new algorithm $\mathcal{A}'$ is universally consistent. □

2.8.3 Proofs for Section 2.3.3

Proof of Proposition 2.11. If $\mathcal{A}(Z)$ is a continuous distribution, we can pick $h \in \mathcal{H}$ at any point where $\mathcal{A}(Z)$ has finite density and set $\mathcal{A}'(Z)|z \in Z$ to be h with probability $1/n$ and the same as $\mathcal{A}(Z)$ with probability $1 - 1/n$. This breaks privacy because conditioned on two databases with z or without z, $\mathcal{A}$, the probability ratio of outputting h is ∞.

If $\mathcal{A}(Z)$ is a discrete distribution or a mixed distribution, it must have the same support of the point mass for all Z. Otherwise it violates DP because we need $\frac{\mathbb{P}_{h \in \mathcal{A}(Z)}(h)}{\mathbb{P}_{h \in \mathcal{A}(Z')}} \leq \exp(n\epsilon)$ for any $Z, Z' \in \mathcal{Z}^n$. Specifically, let the discrete set of point mass be $\tilde{\mathcal{H}}$ if $\mathcal{H} \backslash \tilde{\mathcal{H}} \neq \emptyset$, then we can use the same technique as in the continuous case by adding a small probability $1/n$ on $\mathcal{H} \backslash \tilde{\mathcal{H}}$ when $z \in Z$.

If $\tilde{\mathcal{H}} = \mathcal{H}$, then $\mathcal{H}$ is a discrete set, if $|\mathcal{H}| < n$, then by boundedness and Hoeffding, ERM is a deterministic algorithm that learns any learnable problem. On the other hand, if $|\mathcal{H}| > n$, then by pigeon hole principle, there always exists a hypothesis h that has probability smaller than $1/n$ in $\mathcal{A}(Z)$ for any $Z \in \mathcal{Z}^n$ and we can construct $\mathcal{A}'$ by outputting a sample of $\mathcal{A}(Z)$ if z is not observed and outputting a sample $\mathcal{A}(Z)|\mathcal{A}(Z) \neq h$ whenever z is observed.

The consistency of $\mathcal{A}'$ follows easily as its risk is at most $1/n$ larger than that of $\mathcal{A}$. □

2.8.4 Proofs for characterization of private $\mathfrak{D}$-learnability

Proof of Lemma 2.14. Let $\mathcal{A}'$ be the algorithm that applies $\mathcal{A}$ to a random subsample of size $\lfloor\sqrt{n}\rfloor$. If we can show that, for any $\mathcal{D} \in \mathfrak{D}$,

(a) the empirical risk of $\mathcal{A}'$ converges to the the optimal population risk R^* in expectation;

(b) the empirical risk of the ERM learning rule also converges to R^* in expectation,

then by triangle inequality, the empirical risk of $\mathcal{A}'$ must also converge to the empirical risk of ERM, i.e., $\mathcal{A}'$ is $\mathfrak{D}$-universal AERM.

We will start with (a). For any distribution $\mathcal{D} \in \mathfrak{D}$, we have

$$\begin{aligned}
&\mathbb{E}_{Z\sim\mathcal{D}^n}\hat{R}(\mathcal{A}'(Z), Z) = \mathbb{E}_{Z\sim\mathcal{D}^n}\left[\mathbb{E}_{Z'\subset Z, |Z'|=\lfloor\sqrt{n}\rfloor}\hat{R}(\mathcal{A}(Z'), Z)\right]\\
=&\mathbb{E}_{Z'\sim\mathcal{D}^{\lfloor\sqrt{n}\rfloor}}\left[\frac{\lfloor\sqrt{n}\rfloor}{n}\hat{R}(\mathcal{A}(Z'), Z') + \mathbb{E}_{Z''\sim\mathcal{D}^{n-\lfloor\sqrt{n}\rfloor}}\left(\frac{n-\lfloor\sqrt{n}\rfloor}{n}\hat{R}(\mathcal{A}(Z'), Z'')\right)\right]\\
=&\mathbb{E}_{Z'\sim\mathcal{D}^{\lfloor\sqrt{n}\rfloor}}\left[\frac{\lfloor\sqrt{n}\rfloor}{n}\hat{R}(\mathcal{A}(Z'), Z') + \frac{n-\lfloor\sqrt{n}\rfloor}{n}R(\mathcal{A}(Z'))\right] \leq \frac{1}{\sqrt{n}} + R^* + \xi(\sqrt{n}). \qquad (2.14)
\end{aligned}$$

The last inequality uses the boundedness of the loss function to get $\hat{R}(\mathcal{A}(Z'), Z') \leq 1$ and the $\mathfrak{D}$-consistency of $\mathcal{A}$ to bound the excess risk of $\mathbb{E}_{Z'}R(\mathcal{A}(Z'))$.

To show (b), we need to exploit the assumption that the problem is (non-privately) learnable. By Shalev-Shwartz et al. [191, Theorem 7], the problem being learnable implies that there exists a universally consistent algorithm $\mathcal{B}$ (not restricted to $\mathfrak{D}$), that is universally AERM with rate $3\xi'(n^{\frac{1}{4}}) + \frac{8}{\sqrt{n}}$ and stable with rate $\frac{2}{\sqrt{n}}$. Moreover, by Shalev-Shwartz et al. [191, Theorem 8], $\mathcal{B}$'s stability and AERM implies that $\mathcal{B}$ is also generalizing, with rate $6\xi'(n^{\frac{1}{4}}) + \frac{18}{\sqrt{n}}$. Here the term "generalizing" means that the empirical risk is close to the population risk. Therefore, we can establish (b) via the following chain of approximations

$$\mathbb{E}_{Z\sim\mathcal{D}^n}\hat{R}^*(Z) \underset{\substack{\uparrow\\ \text{AERM of } \mathcal{B}}}{\approx} \mathbb{E}_{Z\sim\mathcal{D}^n}\hat{R}(\mathcal{B}(Z), Z) \overset{\substack{\text{Generalization of } \mathcal{B}\\ \downarrow}}{\approx} R(\mathcal{B}(Z)) \underset{\substack{\uparrow\\ \text{Consistency of } \mathcal{B}}}{\approx} R^*.$$

More precisely,

$$\begin{aligned}
&\left|\mathbb{E}_{Z\sim\mathcal{D}^n}\hat{R}^*(Z) - R^*\right|\\
\leq&\left|\mathbb{E}_{Z\sim\mathcal{D}^n}\hat{R}^*(Z) - \mathbb{E}_{Z\sim\mathcal{D}^n}\hat{R}\right| + \left|\mathbb{E}_{Z\sim\mathcal{D}^n}\hat{R} - R(\mathcal{B}(Z), Z)\right| + |R(\mathcal{B}(Z), Z) - R^*|\\
\leq&[3\xi'(n^{\frac{1}{4}}) + \frac{8}{\sqrt{n}}] + [6\xi'(n^{\frac{1}{4}}) + \frac{18}{\sqrt{n}}] + [3\xi'(n^{\frac{1}{4}}) + \frac{10}{\sqrt{n}}] = 12\xi(n^{\frac{1}{4}}) + \frac{36}{\sqrt{n}}. \qquad (2.15)
\end{aligned}$$

Combine (2.14) and (2.15), we obtain the AERM of $\mathcal{A}'$ with rate $12\xi'(n^{1/4}) + \frac{37}{\sqrt{n}} + \xi(\sqrt{n})$ as required. The privacy of $\mathcal{A}'$ follows from Lemma 2.34. □

2.8.5 Proof for Theorem 2.19

We first present the proof for Theorem 2.19. Recall that the roadmap of the proof is summarized in Figure 2.3.

For readability, we denote $\epsilon(n)$ by simply ϵ.

Recall that the objective function is $F(h, Z) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, z_i) + g_n(h)$ and the corresponding utility function $q(h, Z) = -F(h, Z)$. By the boundedness assumption, it is easy to show that if we replace one data point in any Z with something else, then sensitivity

$$\Delta q = \sup_{h \in \mathcal{H}, d(Z,Z')=1} |q(Z, h) - q(Z', h)| \leq \frac{2}{n}. \tag{2.16}$$

Then by McSherry and Talwar [156, Theorem 6], Algorithm 1 that outputs $h \in \mathcal{H}$ with $\mathbb{P}(h) \propto \exp(\frac{\epsilon}{2\Delta q} q(h, Z))$ naturally ensures ϵ-differential privacy.

Denote shorthand $F^* := \inf_{f \in \mathcal{H}} F(Z, h)$ and $q^* := -F^*$, we can state an analog of the utility theorem of the exponential mechanism in [156].

Lemma 2.36 (Utility). *Assuming $\epsilon < \log n$ (otherwise the privacy protection is meaningless anyway), if assumption A1, A2 hold for distribution $\mathcal{D}$, then*

$$\mathbb{E}_{Z\sim\mathcal{D}^n}\mathbb{E}_{h\sim\mathcal{A}(Z)} q(Z, h) \geq -\mathbb{E}_{Z\sim\mathcal{D}^n} F^* - \frac{9[(\rho + 2)\log n + \log K]}{n\epsilon}. \tag{2.17}$$

Proof. By the boundedness of ℓ and g

$$q(Z, h) = -\frac{1}{n}\sum_i \ell(h, z_i) - g_n(h) \geq -(1 + \zeta(n)).$$

By Lemma 7 in McSherry and Talwar [156] (translated to our case),

$$\mathbb{P}_{h\sim\mathcal{A}(Z)}\left[q(Z, h) < -F^* - 2t\right] \leq \frac{\mu(\mathcal{H})}{\mu(\mathcal{S}_t)} e^{-\frac{\epsilon}{2\Delta q}t}, \tag{2.18}$$

Apply (2.16), take expectation over the data distribution on both sides, and applying assumption A2, we get

$$\mathbb{E}_{Z\sim\mathcal{D}^n}\mathbb{P}_{h\sim\mathcal{A}(Z)}\left[q(Z, h) < -F^* - 2t\right] \leq Kt^{-\rho}e^{-\frac{\epsilon nt}{4}} = e^{-\frac{\epsilon nt}{4} + \log K - \rho \log t} := e^{-\gamma}. \tag{2.19}$$

Take $t = \frac{4\left[(\rho+2)\log n + \log(K)\right]}{\epsilon n}$, by the assumption that $\epsilon < \log n$, we get $\log(nt) > 0$. Substitute t into the expression of γ we obtain

$$\gamma = \frac{\epsilon n}{4}t - \log K + \rho \log t = 2\log n + \rho \log(nt) \geq 2\log n,$$

and therefore

$$\mathbb{E}_{Z\sim\mathcal{D}^n}\mathbb{P}_{h\sim\mathcal{A}(Z)}\left[q(Z, h) < -F^* - 2t\right] \leq n^{-2}.$$

Denote $\mathbb{P}_{h\sim\mathcal{A}(Z)}[q(Z,h) < -F^* - 2t] =: p$, we can then bound the expectation from below as follows:

$$\begin{aligned}\mathbb{E}_{Z\sim\mathcal{D}^n}\mathbb{E}_{h\sim\mathcal{A}(Z)}q(Z,h) \geq &\mathbb{E}_{Z\sim\mathcal{D}^n}(-F^* - 2t)(1-p) + \min_{h\in\mathcal{H},Z\in\mathcal{Z}^n} q(Z,h)\mathbb{E}_{Z\sim\mathcal{D}^n}p \\ \geq &\mathbb{E}_{Z\sim\mathcal{D}^n}(-F^* - 2t) + (-1-\zeta(n))\,n^{-2} \\ \geq & - \mathbb{E}_{Z\sim\mathcal{D}^n}F^* - \frac{8\big[(\rho+2)\log n + \log(K)\big]}{\epsilon n} - (1+\zeta(n))\,n^{-2} \\ \geq & - \mathbb{E}_{Z\sim\mathcal{D}^n}F^* - \frac{9\big[(\rho+2)\log n + \log(K)\big]}{\epsilon n}.\end{aligned}$$

□

Now we can say something about the learning problem. In particular, the AERM follows directly from the utility result and stability follows from the definition of differential privacy.

Lemma 2.37 (Universal AERM)**.** *Assume A1 and A2, and $\epsilon \leq \log n$ (so Lemma 2.36 holds), then*

$$\mathbb{E}_{Z\sim\mathcal{D}^n}\Big[\mathbb{E}_{h\sim\mathcal{A}(Z)}\hat{R}(h,Z) - \hat{R}^*(Z)\Big] \leq \frac{9[(\rho+2)\log n + \log(1/K)]}{n\epsilon} + \zeta(n).$$

Proof. This is a simple consequence of boundedness and Lemma 2.36.

$$\begin{aligned}&\mathbb{E}_{Z\sim\mathcal{D}^n}\Big[\mathbb{E}_{h\sim\mathcal{A}(Z)}\hat{R}(h,Z) - \hat{R}^*(Z)\Big] \\ =&\mathbb{E}_{Z\sim\mathcal{D}^n}\mathbb{E}_{h\sim\mathcal{A}(Z)}\frac{1}{n}\sum_i \ell(h,z_i) - \mathbb{E}_{Z\sim\mathcal{D}^n}\inf_h \frac{1}{n}\sum_i \ell(h,z_i) \\ \leq&\mathbb{E}_{Z\sim\mathcal{D}^n}\mathbb{E}_{h\sim\mathcal{A}(Z)}\left[\frac{1}{n}\sum_i \ell(h,z_i) + g_n(h)\right] - \mathbb{E}_{h\sim\mathcal{A}(Z)}g_n(h) \\ &- \mathbb{E}_{Z\sim\mathcal{D}^n}\inf_h\left[\frac{1}{n}\sum_i \ell(h,z_i) + g_n(h)\right] + \sup_h(g_n(h)) \\ =&\mathbb{E}_{Z\sim\mathcal{D}^n}(-F^* - \mathbb{E}_{h\sim\mathcal{A}(Z)}q(Z,h)) + \sup_h g_n(h) - \mathbb{E}_{h\sim\mathcal{A}(Z)}g_n(h) \\ \leq&\frac{9[(\rho+2)\log n + \log(1/K)]}{n\epsilon} + 2\zeta(n).\end{aligned}$$

The last step applies Lemma 2.36 and $\sup_h |g_n(h)| \leq \zeta(n)$ as in Assumption A2 by using the fact that $\sup_h g_n(h) - \mathbb{E}g_n(h) \leq 2\sup_h |g_n(h)|$ for any distribution of h the expectation is taken over. □

The above theorem shows that Algorithm 1 is asymptotic ERM. By Theorem 2.8, the fact that this algorithm is ϵ-differential private implies that it is 2ϵ-stable. Now the proof follows by applying Theorem 2.35 which says that stability and AERM of an algorithm certify its consistency. Noting that this holds for any distribution $\mathcal{D}$ completes our proof for learnability in Theorem 2.19.

2.8.6 Proofs of other technical results

High confidence private learning.

Proof of Theorem 2.25. The algorithm $\mathcal{A}$ privately learns the problem with rate $\xi(n)$ implies that

$$\mathbb{E}_{Z\in\mathcal{D}^n}\mathbb{E}_{h\sim\mathcal{A}(Z)}R(h) - R^* \leq \xi(n).$$

Let $h \sim \mathcal{A}(Z)$ and $Z \sim \mathcal{D}^n$, by Markov's inequality, with probability at least $1 - 1/e$,

$$R(h) - R^* \leq e\xi(n).$$

If we split the data randomly into $a+1$ parts of size $n/(a+1)$ and run $\mathcal{A}$ on the first a partitions, then we get $h_j \sim \mathcal{A}(Z_j)$. Then with probability at lest $1-(1/e)^a$, at least one of them has risk

$$\min_{j\in[a]} R(h_j) - R^* \leq e\xi(\frac{n}{a+1}). \tag{2.20}$$

Since the $(a+1)$th partition are iid data, and ℓ is bounded, we can apply Hoeffding's inequality and union bound, so that with probability $1-\delta_1$ for all $j = 1, ..., a+1$

$$\hat{R}(h_j, Z_{a+1}) - R(h_j) \leq \sqrt{\frac{\log(2a/\delta_1)}{2n}}. \tag{2.21}$$

This means that if exponential mechanism picked the one with the best validation risk it will be almost as good as the one with the best risk. Assume h_1 is the one that achieves the best validation risk.

Now it remains to bound the probability that exponential mechanism pick an $h \in \{h_1, ..., h_a\}$ that is much worse than h_1.

Recall that the utility function is the negative validation risk which depends only on the last partition I_{a+1}.

$$q(X, h) = \frac{1}{n/(a+1)} \sum_{i\in I_{a+1}} \ell_i(z_i, h).$$

This is in fact a random function of the data because we are picking the the validation set I_{a+1} randomly from the data. Suppose we arbitrarily replace one data point j from the dataset, the distribution of the output of function $q(Z, h)$ is a mixture of the two cases: $j \in I_{a+1}$ and $j \notin I_{a+1}$. Since in the first case, $q(Z, h) = q(Z', h)$ for all h, sensitivity for this case is 0. In the second case, by the boundedness assumption, the sensitivity is at most $2(a+1)/n$. For the exponential mechanism guarantee ϵ differential privacy, it suffices to take the sensitivity parameter to be $2(a+1)/n$.

By the utility theorem of the exponential mechanism,

$$\mathbb{P}\left[\hat{R}(h) > \hat{R}(h_1) + \frac{8(\eta\log n + \log a)}{\epsilon n/(a+1)}\right] \leq n^{-\eta}. \tag{2.22}$$

Combine (2.20)(2.21) and(2.22) we get

$$\mathbb{P}\left[R(h) - R^* > e\xi(\frac{n}{a+1}) + \sqrt{\frac{\log(2a/\delta_1)}{2n}} + \frac{8(\eta\log n + \log a)}{\epsilon n/(a+1)}\right] \le n^{-\eta} + \delta_1 + e^{-a}.$$

Now by appropriately choosing $\eta = \log(3/\delta)/\log n$, $a = \log(3/\delta)$, $\delta_1 = \delta/3$, we get

$$\mathbb{P}\Bigg[R(h) - R^* > e\xi(\frac{n}{\log(3/\delta)+1}) + \sqrt{\frac{\log(2\log(3/\delta)) + \log(3/\delta)}{2n}} \\ + \frac{8(\log(3/\delta) + \log\log(3/\delta))}{\epsilon n/(\log(3/\delta)+1)}\Bigg] \le \delta$$

combine the terms and take $\epsilon = \frac{\log(3/\delta)+1}{\sqrt{n}}$, we get the bound of the excess risk in the theorem.

To get the privacy claim, note that we are applying $\mathcal{A}$ on disjoint partitions of the data so the privacy parameter does not aggregate. Take the worst over all partitions, we get the overall privacy loss $\max\left\{\epsilon\left(\frac{n}{\log(3/n)+1}\right), \frac{\log(3/\delta)+1}{\sqrt{n}}\right\}$ as stated in the theorem. □

The Lipschitz example.

Proof of Example 2.21. Let $h^* \in \mathrm{argmin}_{h\in\mathcal{H}} F(Z,h)$, the Lipschitz condition dictates that for any h,

$$|F(h) - F(h^*)| \le L\|h - h^*\|_p.$$

Choose a small enough $t < t_0$ such that h is in the small neighborhood of h^*, and we can construct a function $\tilde{F}$ that within the sublevel set $\mathcal{S}_t$, such that the above inequality (when we replace F with $\tilde{F}$) is equality, then for any $h \in \mathcal{S}_{t_0}$, $\tilde{F}(h) \ge F(Z,h)$. Verify that the sublevel set of $\tilde{F}(h)$, denoted by $\tilde{\mathcal{S}_t}$ always contains $\mathcal{S}_t$. In addition, we can compute the measure $\mu(\tilde{\mathcal{S}_t})$ explicitly, since the function is a cone and

$$L\|h - h^*\|_p = |\tilde{F}(h) - \tilde{F}(h^*)| = \tilde{F}(h) - \tilde{F}(h^*) \le t,$$

therefore

$$\tilde{\mathcal{S}_t} = \{h \mid L\|h - h^*\|_p \le t\}.$$

Since $\mathcal{H}$ is β_p-regular, $\mu(B \cap \mathcal{H}) \ge \beta_p\mu(B)$ for any ℓ_p ball $B \subset \mathbb{R}^d$, the measure of the sublevel set can be lower bounded by β_p times the volume of the ℓ_p ball with radius t/L and since $\tilde{\mathcal{S}_t} \subseteq \mathcal{S}_t$, we have

$$\mu(\mathcal{S}_t) \ge \mu(\tilde{\mathcal{S}_t}) \ge \beta_p\mu\left(B(t/L)\right) = \beta_p\left(t/L\right)^d$$

as required. □

2.9 Alternative proof of Corollary 2.9 via Dwork et al. [87, Theorem 7]

In this section, we describe how the results in Dwork et al. [87] can be used to obtain the forward direction of our characterization without going through a stability argument. We first restate the result here in our notation:

Lemma 2.38 (Theorem 7 in Dwork et al. 87)**.** *Let $\mathcal{B}$ be an ϵ-DP algorithm such that given a dataset Z, $\mathcal{B}$ outputs a function from $\mathcal{Z}$ to $[0,1]$. For any distribution $\mathcal{D}$ over $\mathcal{Z}$ and random variable $Z \sim \mathcal{D}^n$, we let $\phi \sim \mathcal{B}(Z)$. Then for any $\beta > 0$, $\tau > 0$ and $n \geq 12\log(4/\beta)/\tau^2$, setting $\epsilon < \tau/2$ ensures*

$$\mathbb{P}_{\phi\sim\mathcal{B}(Z),Z\sim\mathcal{D}^n}\left[\left|\mathbb{E}_{z\sim\mathcal{D}}\phi(z) - \frac{1}{n}\sum_{z\in Z}\phi(z)\right| \geq \tau\right] \leq \beta.$$

This lemma was originally stated to prove the claim that privately generated mechanisms for answering statistical queries always generalize.

For statistical learning problems, we can simply take the statistical query ϕ to be the loss function $\ell(h,\cdot)$ parameterized by $h \in \mathcal{H}$. If an algorithm $\mathcal{A}$ that samples from a distribution on $\mathcal{H}$ upon observing data Z is ϵ-DP, then $\mathcal{B}: Z \to \ell(\mathcal{A}(Z),\cdot)$ is also ϵ-DP. The result therefore reduces to that the empirical risk and population risk are close with high probability. Due to the boundedness assumption, we can translate the high probability result to the expectation form, which verifies the definition of “generalization”.

However, “generalization” alone still does not imply “consistency”, as we also need

$$\mathbb{E}_{\phi\sim\mathcal{B}(Z)}\frac{1}{n}\sum_{z\in Z}\phi(z) \to R^* = \min_{\phi\in\Phi}\mathbb{E}_{z\sim\mathcal{D}}\phi(z)$$

as Z gets large, which does not hold for all DP-output ϕ. But when $\phi = \ell(h,\cdot)$, it can be obtained if we assume $\mathcal{A}$ is AERM. This is shown via the following inequality

$$\mathbb{E}_{Z\in\mathcal{D}^n}\mathbb{E}_{\phi\sim\mathcal{B}(Z)}\frac{1}{n}\sum_{z\in Z}\phi(z) \to \mathbb{E}_{Z\in\mathcal{D}^n}\min_{\phi\in\Phi}\frac{1}{n}\sum_{z\in Z}\phi(z) \leq \mathbb{E}_{Z\in\mathcal{D}^n}\frac{1}{n}\sum_{z\in Z}\phi^*(z) = \mathbb{E}\phi^*(z) = R^*,$$

where $\phi^* = \ell(h^*,\cdot)$ and h^* is an optimal hypothesis function. This wraps up the proof of consistency.

The above proof of “consistency” via Lemma 2.38 and “AERM”, however, leads to a looser bound comparing to our result (Corollary 2.9) when the additional assumption on n and τ (equivalently ϵ) is active, i.e., when $\frac{\epsilon(n)}{\log(1/\epsilon(n))} < O\left(\frac{1}{\sqrt{n}}\right)$. In this case it only implies a $\xi(n) + \frac{\log n}{\sqrt{n}}$ bound due to that ϵ-DP implies ϵ'-DP for any $\epsilon' > \epsilon$. Our proof of Corollary 2.9 is considerably simpler and more general in that it does not require any assumption on the number of data points n.

This can easily lead to worse overall error bound for very simple learning problems with sufficiently fast rate. For example, in the problem of learning the mean of $X \in [0,1]$, let the loss function be $|x-h|^{10}$. Consider the $\epsilon(n)$-DP algorithm that outputs ERM $+$ Laplace$(\frac{2}{\epsilon(n)n})$ where $\epsilon(n)$ is chosen to be $n^{-9/10}$. This algorithm is AERM with rate $\xi(n) = \frac{10!2!}{(\epsilon(n)n)^{10}} = O(n^{-1})$. By Corollary 2.9 we get an overall rate of $O(n^{-9/10})$ while through Lemma 2.38 and the argument that follows, we only get $\tilde{O}(n^{-1/2})$.

Chapter 3

Privacy for free: Posterior sampling and stochastic gradient Monte Carlo

In this chapter, we consider the problem of Bayesian learning on sensitive datasets and present two simple but somewhat surprising results that connect Bayesian learning to "differential privacy", a cryptographic approach to protect individual-level privacy while permitting database-level utility. Specifically, we show that under standard assumptions, getting one single sample from a posterior distribution is differentially private "for free". We will see that estimator is statistically consistent, near optimal and computationally tractable whenever the Bayesian model of interest is consistent, optimal and tractable. Similarly but separately, we show that a recent line of works that use stochastic gradient for Hybrid Monte Carlo (HMC) sampling also preserve differentially privacy with minor or no modifications of the algorithmic procedure at all, these observations lead to an "anytime" algorithm for Bayesian learning under privacy constraint. We demonstrate that it performs much better than the state-of-the-art differential private methods on synthetic and real datasets.

3.1 Introduction

Bayesian models have proven to be one of the most successful classes of tools in machine learning. It stands out as a principled yet conceptually simple pipeline for combining expert knowledge and statistical evidence, modeling with complicated dependency structures and harnessing uncertainty by making probabilistic inferences [98, 99]. In the past few decades, the Bayesian approach has been intensively used in modeling speeches [173], text documents [34], images/videos [90], social networks [6], brain activity [171], and is often considered gold standard in many of these application domains. Learning a Bayesisan model typically involves sampling from a posterior distribution, therefore the learning process is inherently randomized.

Differential privacy (DP) is a cryptography-inspired notion of privacy [77, 84]. It is designed to provide a very strong form of protection of individual user's private information and at the same time allow data analyses to be conducted with proper utility. Any algorithm that preserves differential privacy must be appropriately randomized too. For instance, one can differential-privately release the average salary of Californian males by adding a Laplace noise proportional to the sensitivity of this figure upon small perturbation of the data sample.

In this chapter, we connect the two seemingly unrelated concepts by showing that under standard assumptions, the intrinsic randomization in the Bayesian learning can be exploited to obtain a degree of differential privacy. In particular, we show that:

- Any algorithm that produces a single sample from the exact (or approximate) posterior distribution of a Bayesian model with bounded log-likelihood is ϵ (or (ϵ, δ))-differentially private[1]. By the classic results in asymptotic statistics [137, 232], we show that this posterior sample is a consistent estimator whenever the Bayesian model is consistent; and near optimal whenever the asymptotic normality and efficiency of the maximum likelihood estimate holds.

- The popular large-scale sampler Stochastic Gradient Langevin Dynamics [250] and extensions, e.g. Ahn et al. [4], Chen et al. [56], Ding et al. [65] obey (ϵ, δ)-differentially private with no algorithmic changes when the stepsize is chosen to be small. This gives us a procedure that can potentially output many (correlated) samples from an approximate posterior distribution.

These simple yet interesting findings make it possible for differential privacy to be explicitly considered when designing Bayesian models, and for Bayesian posterior sampling to be used as a valid DP mechanism. We demonstrate empirically that these methods work as well as or better than the state-of-the-art differential private empirical risk minimization (ERM) solvers using objective perturbation [55, 128].

The results presented in this chapter are closely related to a number of previous work, e.g., Bassily et al. [17], Dimitrakakis et al. [64], McSherry and Talwar [156], Mir [158]. Proper comparisons with them would require the knowledge of our results, thus we will defer detailed comparisons to Section 3.6 near the end of the paper.

3.2 Notations and preliminary

Throughout the paper, we assume data point $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\theta} \in \Theta$ is the model. This can be the finite dimensional parameter of a single exponential family model or a collection of these in a graphical model, or a function in a Hilbert space or other infinite dimensional objects if the model is nonparametric. $\pi(\boldsymbol{\theta})$ denotes a prior belief of the model parameters and $p(\boldsymbol{x}|\boldsymbol{\theta})$ and $\ell(\boldsymbol{x}|\boldsymbol{\theta})$ are the likelihood and log-likelihood of observing data point x given model parameter $\boldsymbol{\theta}$. If we

[1] Similar observations were made in Mir [158] and Dimitrakakis et al. [64] under slightly different regimes and assumptions, and we will review them among other related work in Section 3.6.

observe $X = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$, the posterior distribution

$$\pi(\boldsymbol{\theta}|X) = \frac{\pi(\boldsymbol{\theta}) \prod_{i=1}^N p(\boldsymbol{x}_i|\boldsymbol{\theta})}{\int \prod_{i=1}^N p(\boldsymbol{x}_i|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\pi}$$

denotes the updated belief conditioned on the observed data. Learning Bayesian models correspond to finding the mean or mode of the posterior distribution, but often, the entire distribution is treated as the output, which provides much richer information than just a point estimator. In particular, we get error bars of the estimators for free (credibility intervals).

Ignoring the philosophical disputes of Bayesian methods for the moment, practical challenges of Bayesian learning are often computational. As the models get more complicated, often there is not a closed-form expression for the posterior. Instead, we often rely on Markov Chain Monte Carlo methods, e.g., Metropolis-Hastings algorithm [110] to generate samples. This is often prohibitively expensive when the data is large. One recent approach to scale up Bayesian learning is to combine stochastic gradient estimation as in Robbins and Monro [177] and Monte Carlo methods that simulates stochastic differential equations, e.g. Neal [163]. These include Stochastic Gradient Langevin dynamics (SGLD) [250], Stochastic Gradient Fisher scoring (SGFS) [4], Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) [56] as well as more recent Stochastic Gradient Nosé-Hoover Thermostat (SGNHT) [65]. We will describe them with more details and show that these series of tools provide differential privacy as a byproduct of using stochastic gradient and requiring the solution to not collapse to a point estimate.

3.2.1 Differential privacy

We now restate the definition of differential privacy and approximate differential privacy that we saw in Chapter 2 and write two lemmas that we need to use in this chapter. Let the space of data be $\mathcal{X}$ and data points $X, Y \in \mathcal{X}^n$. Define $d(X, Y)$ to be the edit distance or Hamming distance between data set X and Y, for instance, if X and Y are the same except one data point, then $d(X, Y) = 1$.

Definition 3.1. *(Differential Privacy) We call a randomized algorithm $\mathcal{A}$ (ϵ, δ)-differentially private with domain $\mathcal{X}^n$ if for all measurable set $S \subset$ Range$(\mathcal{A})$ and for all $X, Y \in \mathcal{X}^n$ such that $d(X, Y) \leq 1$, we have*

$$\mathbb{P}(\mathcal{A}(X) \in S) \leq \exp(\epsilon)\mathbb{P}(\mathcal{A}(Y) \in S) + \delta.$$

If $\delta = 0$, then $\mathcal{A}$ is the called ϵ-differential private.

This definition naturally prevents linkage attacks and the identification of individual data from adversaries having arbitrary side information and infinite computational power. The promise of differential privacy has been interpreted in statistical testing, Bayesian inference and information theory for which we refer readers to Chapter 1 of [80].

There are several interesting properties of differential privacy that we will exploit here. Firstly, the definition is closed under post-processing.

Lemma 3.2 (Post-processing immunity)**.** *If $\mathcal{A}$ is an (ϵ,δ)-DP algorithm, $\mathcal{B}\circ\mathcal{A}$ is also (ϵ,δ)-DP algorithm for any $\mathcal{B}$.*

This is natural because otherwise the whole point of differential privacy will be forfeited. Also, the definition automatically allows for cases when the sensitive data are accessed more than once.

Lemma 3.3 (Composition rule)**.** *If algorithm $\mathcal{A}_1$ is (ϵ_1,δ_1)-DP, and $\mathcal{A}_2$ is (ϵ_2,δ_2)-DP then $(\mathcal{A}_1\otimes\mathcal{A}_2)$ is $(\epsilon_1+\epsilon_2,\delta_1+\delta_2)$-DP.*

We will describe more advanced properties of DP as we need in Section 3.4.

3.3 Posterior sampling and differential privacy

In this section, we make a simple observation that under boundedness condition of a log-likelihood, getting one single sample from the posterior distribution (denoted by "OPS mechanism" from here onwards) preserves a degree of differential privacy for free. Then we will cite classic results in statistics and show that this sample is a consistent estimator in a Frequentist sense and near-optimal in many cases.

3.3.1 Implicitly Preserving Differential Privacy

To begin with, we show that sampling from the posterior distribution is intrinsically differentially private.

Theorem 3.4. *If $\sup_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\theta}\in\Theta}|\log p(\boldsymbol{x}|\boldsymbol{\theta})|\leq B$, releasing one sample from the posterior distribution $p(\boldsymbol{\theta}|X^n)$ with any prior preserves $4B$-differential privacy. Alternatively, if $\mathcal{X}$ is a bounded domain (e.g., $\|x\|_*\leq R\ \forall\boldsymbol{x}\in\mathcal{X}$) and $\log p(\boldsymbol{x}|\boldsymbol{\theta})$ is an L-Lipschitz function in $\|\cdot\|_*$ for any $\boldsymbol{\theta}\in\Theta$, then releasing one sample from the posterior distribution preserves $4LR$-differential privacy.*

Proof. The posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x}_1,...,\boldsymbol{x}_n)=\frac{\prod_{i=1}^n p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}}\prod_{i=1}^n p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$. For any $\boldsymbol{x}_1,...,\boldsymbol{x}_n,x'_k$, The ratio can be factorized into

$$\frac{p(\boldsymbol{\theta}|\boldsymbol{x}_1,...,x'_k,...,\boldsymbol{x}_n)}{p(\boldsymbol{\theta}|\boldsymbol{x}_1,...,\boldsymbol{x}_k,...,\boldsymbol{x}_n)}=\underbrace{\frac{p(\boldsymbol{x}'_k|\boldsymbol{\theta})\prod_{i=1:n,i\neq k}p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\prod_{i=1}^n p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})}}_{\text{Factor 1}}\times\underbrace{\frac{\int_{\boldsymbol{\theta}}\prod_{i=1}^n p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}}p(\boldsymbol{x}'_k|\boldsymbol{\theta})\prod_{i=1:n,i\neq k}p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}}_{\text{Factor 2}}.$$

Algorithm 2 One-Posterior Sample (OPS) estimator

```
input  Data $X$, log-likelihood function $\ell(\cdot|\cdot)$ satisfying $\sup_{\boldsymbol{x},\boldsymbol{\theta}} \|\ell(\boldsymbol{x}|\boldsymbol{\theta})\| \leq B$ a prior $\pi(\cdot)$. Privacy loss $\epsilon$.
    1. Set $\rho = \min\{1, \frac{\epsilon}{4B}\}$.
    2. Re-define log-likelihood function and the prior $\ell'(\cdot|\cdot) := \rho\ell(\cdot|\cdot)$ and $\pi'(\cdot) := (\pi(\cdot))^{\rho}$.
output  $\hat{\boldsymbol{\theta}} \sim P(\boldsymbol{\theta}|X) \propto \exp\left(\sum_{i=1}^{N} \ell'(\boldsymbol{\theta}|\boldsymbol{x}_i)\right) \pi'(\boldsymbol{\theta})$.
```

It follows that

$$\begin{aligned}
\text{Factor 1} &= \frac{p(\boldsymbol{x}'_k|\boldsymbol{\theta})}{p(\boldsymbol{x}_k|\boldsymbol{\theta})} = e^{\log p(\boldsymbol{x}'_k|\boldsymbol{\theta}) - \log p(\boldsymbol{x}_k|\boldsymbol{\theta})} \leq e^{2B}, \\
\text{Factor 2} &= \frac{\int_{\boldsymbol{\theta}} \prod_{i\neq k} p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{x}_k)d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}} p(\boldsymbol{x}'_k|\boldsymbol{\theta}) \prod_{i\neq k} p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{\int_{\boldsymbol{\theta}} \prod_{i\neq k} p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{x}'_k|\boldsymbol{\theta})\frac{p(\boldsymbol{x}_k)}{p(\boldsymbol{x}'_k)}d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}} p(\boldsymbol{x}'_k|\boldsymbol{\theta}) \prod_{i\neq k} p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \\
&= \frac{\int_{\boldsymbol{\theta}} \prod_{i\neq k} p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{x}'_k|\boldsymbol{\theta})e^{\log p(\boldsymbol{x}_k|\boldsymbol{\theta}) - \log p(\boldsymbol{x}'_k|\boldsymbol{\theta})}d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}} p(\boldsymbol{x}'_k|\boldsymbol{\theta}) \prod_{i\neq k} p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \\
&\leq e^{2B} \frac{m(\boldsymbol{x}_1, ..., \boldsymbol{x}'_k, ..., \boldsymbol{x}_n)}{m(\boldsymbol{x}_1, ..., \boldsymbol{x}'_k, ..., \boldsymbol{x}_n)} = e^{2B}.
\end{aligned}$$

where we use $m(X)$ to denote the marginal distribution. As a result, the whole thing is bounded by e^{4B}.

Alternatively, we can use the Lipschitz constant and boundedness to get $\log p(\boldsymbol{x}'_k|\boldsymbol{\theta}) - \log p(\boldsymbol{x}_k|\boldsymbol{\theta}) \leq L\|x'_k - \boldsymbol{x}_k\|_* \leq 2LR$. □

Readers familiar with differential privacy must have noticed that this is actually an instance of the exponential mechanism [156], a general procedure that preserves privacy while making outputs with higher utility exponentially more likely. If one sets the utility function to be the log-likelihood and the privacy parameter being $4B$, then we get exactly the one-posterior sample mechanism. This exponential mechanism point of view provides an an simple extension which allows us to specify ϵ by simply scaling the log-likelihood (see Algorithm 2). We will overload the notation OPS to also represent this mechanism where we can specify ϵ. The nice thing about this algorithm is that there is almost zero implementation effort to extend all posterior sampling-based Bayesian learning models to have differentially privacy of any specified ϵ.

Assumption on the boundedness. The boundedness on the loss-function (log-likelihood here) is a standard assumption in many DP works [17, 55, 128, 205]. Lipschitz constant L is usually small for continuous distributions (at least when the parameter space Θ is bounded). This is a bound on $\log p(\boldsymbol{x}|\boldsymbol{\theta}))$ so as long as $p(\boldsymbol{x}|\boldsymbol{\theta})$ does not increase or decrease super exponentially fast at any point, L will be a small constant. R can also be made small by a simple preprocessing step that scales down all data points. In the aforementioned papers that assume L, it is typical that they also assume $R = 1$ for convenience. So we will do the same. In practice, we can algorithmically remove large data points from the data by some predefined threshold or using the "Propose-Test-Release" framework in [79] or perform weighted training where we can assign lower weight to data points with large magnitude. Note that this is a desirable step for the robustness to outliers too. Exponential families (in Hilbert space) are an example, see e.g. Bialek et al. [32], Hofmann et al. [115], Wainwright and Jordan [241].

3.3.2 Consistency and Near-Optimality

Now we move on to study the consistency of the OPS estimator. In great generality, we will show that the one-posterior sample estimator is consistent whenever the Bayesian model is posterior consistent. Since the consistency in Bayesian methods can have different meanings, we briefly describe two of them according to the nomenclature in Orbanz [168].

Definition 3.5 (Posterior consistency in the Bayesian Sense)**.** *For a prior π, we say the model is posterior consistent in the Bayesian sense, if $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$, $\boldsymbol{x}_1, ..., \boldsymbol{x}_n \sim p_{\boldsymbol{\theta}}$, and the posterior*

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \overset{\textit{weakly}}{\longrightarrow} \delta_{\boldsymbol{\theta}} \textit{ a.s. } \pi.$$

$\delta_{\boldsymbol{\theta}}$ is the Dirac-delta function at $\boldsymbol{\theta}$.

In great generality, Doob's well-known theorem guarantees posterior consistency in the Bayesian sense for a model with any prior under no conditions except identifiability and measurability. A concise statement of Doob's result can be found in Van der Vaart [232, Theorem 10.10]).

An arguably more reasonable definition is given below. It applies to the case when the statistician who chooses the prior π does not know about the true parameter.

Definition 3.6 (Posterior consistency in the Frequentist Sense)**.** *For a prior π, we say the model is posterior consistent in the Frequentist sense, if for every $\boldsymbol{\theta}_0 \in \Theta$, $\boldsymbol{x}_1, ..., \boldsymbol{x}_n \sim p_{\boldsymbol{\theta}}$, the posterior*

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \overset{\textit{weakly}}{\longrightarrow} \delta_{\boldsymbol{\theta}_0} \textit{ a.s. } p_{\boldsymbol{\theta}_0}.$$

This type of consistency is much harder to satisfy especially when Θ is an infinite dimensional space, in which case the consistency often depends on the specific priors to use. A promising series of results on the consistency for Bayesian nonparametric models can be found in Ghosal [100]).

Regardless which definition one favors, the key notion of consistency is that the posterior distribution to concentrates around the true underlying $\boldsymbol{\theta}$ that generates the data.

Proposition 3.7. *The one-posterior sample estimator is consistent* if and only if *the Bayesian model is posterior consistent (in either Definition 3.5 or 3.6).*

Proof. The equivalence follows from the standard equivalence of convergence weakly and convergence in probability when a random variable converges weakly to a point mass. □

How about the rate of convergence? In the low dimensional setting when $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$ and $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is suitably differentiable and the prior is supported at the neighborhood of the true parameter, then by the Bernstein-von Mises theorem [137], the posterior mean is an asymptotically efficient estimator and the posterior distribution converges in L_1-distance to a normal distribution with covariance being the inverse Fisher Information.

Proposition 3.8. *Under the regularity conditions where Bernstein-von Mises theorem holds, the One-Posterior sample $\hat{\boldsymbol{\theta}} \sim \pi(\boldsymbol{\theta}|\boldsymbol{x}_1, .., \boldsymbol{x}_n)$ obeys*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{\textit{weakly}}{\longrightarrow} \mathcal{N}(0, 2\mathbb{I}^{-1}),$$

i.e., the One-Posterior sample estimator has an asymptotic relative efficiency of 2.

Proof. Let the One-Posterior sample $\hat{\boldsymbol{\theta}} \sim \pi(\boldsymbol{\theta}|\boldsymbol{x}_1, .., \boldsymbol{x}_n)$. By Bernstein-von Mises theorem $\sqrt{n}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) \overset{\text{weakly}}{\to} \mathcal{N}(0, \mathbb{I}^{-1})$. By the asymptotic normality and efficiency of the posterior mean estimator $\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{\text{weakly}}{\to} \mathcal{N}(0, \mathbb{I}^{-1})$. The proof is complete by taking the sum of the two asymptotically independent Gaussian vectors ($\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}$ are asymptotically independent). □

The above proposition suggests that in many interesting classes of parametric Bayesian models, the One-Posterior Sample estimator is asymptotically near optimal. Similar statements can also be obtained for some classes of semi-parametric and nonparametric Bayesian models [100], which we leave as future work.

The drawback of the above two propositions is that it is only stated for the version of the OPS when $\epsilon = 4B$. Using results in De Blasi and Walker [61] and Kleijn et al. [130] for misspecified, we can prove consistency, asymptotic normality for any ϵ and parameterize the asymptotic relative efficiency of the OPS estimator as a function of ϵ. The key idea is that when scaling the log-likelihood and sample from a different distribution, we are essentially fitting a model that may not include the data-generating true distribution. De Blasi and Walker [61] shows that under mild conditions, when the model is misspecified, the posterior distribution will converge to a point mass $\boldsymbol{\theta}^*$ that minimizes the KL-divergence between between the true distribution and the corresponding distribution in the misspecified model. $\boldsymbol{\theta}^*$ is essentially MLE and in our case, since we only scaled the distribution, the MLE will remain exactly the same. De Blasi and Walker [61]'s result is quite general and covers both parametric and nonparametric Bayesian models and whenever their assumptions hold, the OPS estimator is consistent. Using a similar argument and the modified Bernstein-Von-Mises theorem in Kleijn et al. [130], we can prove asymptotic normality and near optimality for the subset of problems where regularities of MLE hold.

Proposition 3.9. *Under the same assumption as Proposition 3.8, if we set a different ϵ by rescaling the log-likelihood by a factor of $\frac{\epsilon}{4B}$, then the the One-Posterior sample estimator obeys*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{\textit{weakly}}{\longrightarrow} \mathcal{N}\left(0, (1 + \frac{4B}{\epsilon})\mathbb{I}^{-1}\right),$$

in other word, the estimator has an ARE of $(1 + \frac{4B}{\epsilon})$.

Proof. By scaling the log-likelihood, we are essentially changing the correct model $p_{\boldsymbol{\theta}}$ to a misspecified model $(p_{\boldsymbol{\theta}})^{\frac{\epsilon}{4B}}$. Let the true log-likelihood be ℓ and the misspecified log-likelihood be $\tilde{\ell} = \frac{\epsilon}{4B}\ell$, in addition, define

$$V(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}} \nabla \tilde{\ell}(\boldsymbol{\theta}) \nabla \tilde{\ell}(\boldsymbol{\theta})^T = \frac{\epsilon^2}{16B^2} \mathbb{E}_{\boldsymbol{\theta}} \nabla \ell(\boldsymbol{\theta}) \nabla \ell(\boldsymbol{\theta})^T = \frac{\epsilon^2}{16B^2} \mathbb{I}(\boldsymbol{\theta})$$

$$J(\boldsymbol{\theta}) := -\mathbb{E}_{\boldsymbol{\theta}} \nabla^2 \tilde{\ell}(\boldsymbol{\theta}) = -\frac{\epsilon}{4B} \mathbb{E}_{\boldsymbol{\theta}} \nabla^2 \ell(\boldsymbol{\theta}) = -\frac{\epsilon}{4B} \mathbb{I}(\boldsymbol{\theta}).$$

The last equality holds under the standard regularity conditions. By the sandwich formula, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ under the misspecified model is asymptotically normal:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \overset{\text{weakly}}{\to} \mathcal{N}(0, J^{-1} V J(-1)) = \mathcal{N}(0, \mathbb{I}^{-1})$$

where $\boldsymbol{\theta}^*$ defines the closest (in terms of KL-divergence) model in the misspecified class of distributions to the true distribution that generates the data. Since the difference is only in scaling, the minimum KL-divergence is obtained at $\boldsymbol{\theta}^* = \boldsymbol{\theta}$. Now under the same regularity conditions, we can invoke the modified Bernstein-Von-Mises theorem for misspecified models [130, Lemma 2.2], which says that the posterior distribution $p(\theta|X^n)$ (of the misspecified model) converges in distribution to $\mathcal{N}(\hat{\boldsymbol{\theta}}, (nJ)^{-1})$. In our case, $(nJ)^{-1} = \frac{4B}{n\epsilon}\mathbb{I}^{-1}$. The proof is concluded by noting that the posterior sample is an independent draw. □

We make a few interesting remarks about the result.

1. Proposition 3.9 suggests that for models with bounded log-likelihood, OPS is only a factor of $(1 + 4B/\epsilon)$ away from being optimal. This is in sharp contrast to most previous statistical analysis of DP methods that are only tight up to a numerical constant (and often a logarithmic term). In ℓ_2-norm, the convergence rate is $O(\frac{\sqrt{1+4B/\epsilon}\|I^{-1}\|_F}{\sqrt{n}})$. The bound depends on the dimension through the Frobenius norm which is usually $O(\sqrt{d})$. The bound can be further sharpened using assumptions on the intrinsic rank, incoherence conditions or the rate of decays in eigenvalues of the Fisher information. In ℓ_∞-norm, the convergence rate is $\frac{\sqrt{1+4B/\epsilon}\|I^{-1}\|_2}{\sqrt{n}}$, which does not depend on the dimension of the problem.

2. Another implication is on statistical inference. Proposition 3.9 essentially generalizes that classic results in hypothesis testing and confidence intervals, e.g., Wald test, generalized likelihood ratio test, can be directly adopted for the private learning problems, with an appropriate calibration using ϵ. We can control the type I error in an asymptotically exact fashion. In addition, the trade-off with ϵ and the test power is also explicitly described, so in cases where the power of the tests are well-studied [142], the same handle can be used to analyze the most-powerful-test under privacy constraints.

3. Lastly, Kleijn et al. [130]'s result is much more general. It is easy to extend the guarantee for OPS to handle private Bayesian learning in a fully agnostic setting and in non-iid cases. We will leave the formalization of these claims as future directions.

3.3.3 (Efficient) sampling from approximate posterior

The privacy guarantee in Theorem 3.4 requires sampling from the exact posterior. In practice, however, exact samplers are rare. As Bayesian models get more and more complicated, often the only viable option is to use Markov Chain Monte Carlo (MCMC) samplers which are almost never exact. There are exceptions, e.g., Propp and Wilson [172] but they only apply to problems with very special structures. A natural question to ask is whether we can still say something meaningful about privacy when the posterior sampling is approximate. It turns out that we can, and the level of approximation in privacy is the same as the level of approximation in the sampling distribution.

Proposition 3.10. *If $\mathcal{A}$ that sampling from distribution P_X preserves ϵ-differential privacy, then any approximate sampling procedures $\mathcal{A}'$ that produces a sample from P'_X such that $\|P_X - P'_X\|_{L_1} \leq \delta$ for any X preserves $(\epsilon, (1 + e^\epsilon)\delta)$-differential privacy.*

Proof. For any $S \in \text{Range}(\mathcal{A}')$, and $d(X, X') \leq 1$

$$\begin{aligned}\mathbb{P}\left(\mathcal{A}'(X) \in S\right) &= \int_S dP'_X \leq \int_S dP_X + \delta \\ &\leq e^\epsilon \int_S dP_{X'} + \delta \leq e^\epsilon \int_S dP_{X'} \\ &\leq e^\epsilon \int_S dP'_{X'} + (1 + e^\epsilon)\delta \\ &= e^\epsilon \mathbb{P}\left(\mathcal{A}'(X') \in S\right) + (1 + e^\epsilon)\delta,\end{aligned}$$

This is $(\epsilon, (1 + e^\epsilon)\delta)$-DP by definition. □

We are using L_1 distance of the distribution because it is a commonly accepted metric to measure the convergence rate MCMC [180], and Proposition 3.10 leaves a clean interface for computational analysis in determining the number of iterations needed to attain a specific level of privacy protection.

A note on computational efficiency. The (unsurprising) bad news is that even approximate sampling from the posterior is NP-Hard in general, see, e.g. Sontag and Roy [206, Theorem 8]. There are however interesting results on when we can (approximately) sample efficiently. Approximation is easy for sampling LDA when $\alpha > 1$ while NP-Hard when $\alpha < 1$. A more general result in Applegate and Kannan [11] suggests that we can get a sample with arbitrarily close approximation in polynomial time for a class of near log-concave distributions. The log-concavity of the distributions would imply convexity in the log-likelihood, thus, this essentially confirms the computational efficiency of all convex empirical risk minimization problems under differential privacy constraint (see Bassily et al. [17]).

The nice thing is that since we do not modify the form of the sampling algorithm at all, the OPS algorithm is going to be a computationally tractable DP method whenever the Bayesian learning model of interest is proven to be computationally tractable.

This observation provides an interesting insight into the problem of computational lower bound of differential private machine learning. Unlike what is conjectured in Dwork et al. [87], our observation seems to suggest that the computational barrier is not specific to differential privacy, but rather the barrier of learning in general. The argument seems to hold at least for some class of problems, where the posterior sample achieves the optimal statistical rate and is at least $4B$-DP.

3.3.4 Discussions and comparisons

OPS has a number of advantages over the state-of-the-art differentially private ERM method: objective perturbation [55, 128] (OBJPERT from here onwards). OPS works with arbitrary bounded loss functions and priors while OBJPERT needs a number of restrictive assumptions including twice differentiable loss functions, strongly convexity parameter to be greater than a threshold and so on. These restrictions rule out many commonly used loss functions, e.g., ℓ_1-loss, hinge loss, Huber function just to name a few.

Also, OBJPERT 's privacy guarantee holds only for the exact optimal solution, which is often hard to get in practice. In contrast, OPS works when the sample is drawn from an approximate posterior distribution. From a practical point of view, since OPS stems from the intrinsic privacy protection of Bayesian learning, it requires very little implementation effort to deploy it for practical applications. It also requires the problem to be strong convexity with a minimum strong convexity parameter. When the condition is not satisfied, OBJPERT will need to add additional quadratic regularization to make it so, which may bias the problem unnecessarily.

3.4 Stochastic Gradient MCMC and (ϵ, δ)-Differential privacy

Given a fixed privacy budget, we see that the single posterior sample produces an optimal point estimate, but what if we want multiple samples? Can we use the privacy budget in a different way that produces many approximate posterior samples?

In this section we will provide an answer to it by looking at a class of Stochastic Gradient MCMC techniques developed over the past few years. We will show that they are also differentially private for free if the parameters are chosen appropriately.

The idea is to simply privately release an estimate of the gradient (as in Bassily et al. [17], Song et al. [205]) and leverage upon the following two celebrated lemmas in differential privacy in the same way as Bassily et al. [17] does in deriving the near-optimal (ϵ, δ)-differentially private SGD.

The first lemma is the advanced composition which allows us to trade off a small amount of δ to get a much better bound for the privacy loss due to composition.

Lemma 3.11 (Advanced composition, c.f.,Theorem 3.20 in [80])**.** *For all $\epsilon, \delta, \delta' \geq 0$, the class of (ϵ, δ)-DP mechanisms satisfy $(\epsilon', k\delta + \delta')$-DP under k-fold adaptive composition for:*

$$\epsilon' = \sqrt{2k \log(1/\delta')}\epsilon + k\epsilon(e^{\epsilon} - 1).$$

Remark 3.12. *When $\epsilon = \frac{c}{\sqrt{2k\log(1/\delta')}} < 1$ for some constant $c < \sqrt{\log(1/\delta')}$, we can simplify the above expression into $\epsilon' \leq 2c$. To see this, apply the inequality $e^{\epsilon} - 1 \leq 2\epsilon$ (easily shown via Taylor's theorem and the assumption that $\epsilon \leq 1$).*

In addition, we will also make use of the following lemma due to Beimel et al. [21].

Lemma 3.13 (Privacy for subsampled data. Lemma 4.4 in Beimel et al. [21].)**.** *Over a domain of data sets $\mathcal{X}^N$, if an algorithm $\mathcal{A}$ is (ϵ, δ) differentially private (with $\epsilon < 1$), then for any data set $X \in \mathcal{X}^N$, running $\mathcal{A}$ on a uniform random γN-entries of X ensures $(2\gamma\epsilon, \delta)$-DP.*

To make sense of the above lemma, notice that we are subsampling uniform randomly and the probability of any single data point being sampled is only γ. Thus, if we arbitrarily perturb one of the data points, its impact is evenly spread across all data points thanks to random sampling.

Let $f : \mathcal{X}^n \to \mathbb{R}^d$ be an arbitrary d-dimensional function. Define the ℓ_2 sensitivity of f to be

$$\Delta_2 f = \sup_{Y : d(X,Y) \leq 1} \|f(X) - f(Y)\|_2.$$

Suppose we want to output $f(X)$ differential privately, "Gaussian Mechanism" output $\hat{f}(X) = f(X) + \mathcal{N}(0, \sigma^2 I_d)$ for some appropriate σ.

Theorem 3.14 (Gaussian Mechanism, c.f. Dwork and Roth [80])**.** *Let $\epsilon \in (0, 1)$ be arbitrary. "Gaussian Mechanism" with $\sigma \geq \Delta_2 f \sqrt{2\log(1.25/\delta)}/\epsilon$ is (ϵ, δ)-differentially private.*

This will be the main workhorse that we use here.

3.4.1 Stochastic Gradient Langevin Dynamics

SGLD iteratively update the parameters to by running a perturbed version of the minibatch stochastic gradient descent on the negative log-posterior objective function

$$-\sum_{i=1}^{N} \log p(\boldsymbol{x}_i|\boldsymbol{\theta}) - \log \pi(\boldsymbol{\theta}) =: \sum_{i=1}^{N} \ell(\boldsymbol{x}_i; \boldsymbol{\theta}) + r(\boldsymbol{\theta})$$

where $\ell(\boldsymbol{x}_i; \boldsymbol{\theta})$ and $r(\boldsymbol{\theta})$ are loss-function and regularizer under the empirical risk minimization.

If one were to run stochastic gradient descent or any other optimization tools on this, one would eventually a deterministic maximum a posteriori estimator. SGLD avoids this by adding noise in every iteration. At iteration t SGLD first samples uniform randomly τ data points $\{\boldsymbol{x}_{t_1}, ..., \boldsymbol{x}_{t_2}\}$ and then updates the parameter using

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \left(\nabla r(\boldsymbol{\theta}) + \frac{N}{\tau} \sum_{i=1}^{\tau} \nabla \ell(\boldsymbol{x}_{ti}|\boldsymbol{\theta}) \right) + \boldsymbol{z}_t, \tag{3.1}$$

where $\boldsymbol{z}_t \sim \mathcal{N}(0, \eta_t)$ and τ is the mini-batch size.

For the ordinary stochastic gradient descent to converge in expectation, the stepsize η_t can be chosen as anything that $\sum_{i=1}^{\infty} \eta_t = \infty$ and $\sum_{i=1}^{\infty} \eta_t^2 < \infty$ [177]. Typically, one can chooses stepsize $\eta_t = a(b+t)^{-\gamma}$ with $\gamma \in (0.5, 1]$. In fact, it is shown that for general convex functions and μ-strongly convex functions $\frac{1}{\sqrt{t}}$ and $\frac{1}{\mu t}$ can be used to obtain the minimax optimal $O(1/\sqrt{t})$ and $O(1/t)$ rate of convergence. These results substantiate the first phase of SGLD: a convergent algorithm to the optimal solution. Once it gets closer, however, it transforms into a posterior sampler. According to Welling and Teh [250] and later formally proven in Sato and Nakagawa [187], if we choose $\eta_t \to 0$, the random iterates $\boldsymbol{\theta}_t$ of SGLD converges in distribution to the $p(\boldsymbol{\theta}|X)$. The idea is that as the stepsize gets smaller, the stochastic error from the true gradient due to the random sampling of the minibatch converges to 0 faster than the injected Gaussian noise.

In addition, if we use some fixed stepsize lower bound, such that $\eta_t = \max\{1/(t+1), \eta_0\}$ (to alleviate the slow mixing problem of SGLD), the results correspond to a discretization approximation of a stochastic differential equation (Fokker-Planck equation), which obeys the following theorem due to Sato and Nakagawa [187] (simplified and translated to our notation).

Theorem 3.15 (Weak convergence [187])**.** *Assume $f(\boldsymbol{\theta}|X)$ is differentiable, $\nabla f(\boldsymbol{\theta}|X)$ is gradient Lipschitz and bounded* [2]*. Then*

$$\left|\mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|X)}[h(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}\sim SGLD}[h(\boldsymbol{\theta}(t))]\right| = O(\eta_t),$$

for any continuous and polynomial growth function h.

This theorem implies that one can approximate the posterior mean (and other estimators) using SGLD. Finite sample properties of SGLD is also studied in [238].

Now we will show that with a minor modification to just the "burn-in" phase of SGLD, we will be able to make it differentially private (see Algorithm 3).

Theorem 3.16 (Differentially private Minibatch SGLD)**.** *Assume initial $\boldsymbol{\theta}_1$ is chosen independent of the data, also assume $\ell(\boldsymbol{x}|\boldsymbol{\theta})$ is L-smooth in $\|\cdot\|_2$ for any $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\theta} \in \Theta$. In addition,*

Algorithm 3 Differentially Private Stochastic Gradient Langevin Dynamics (DP-SGLD)

```
Require: Data $X$ of size $N$, Size of minibatch $\tau$, number of data passes $T$, privacy parameter
  $\epsilon, \delta$, Lipschitz constant $L$ and initial $\boldsymbol{\theta}_1$. Set $t = 1$.
  for $t = 1 : \lfloor NT/\tau \rfloor$ do
    1. Random sample a minibatch $S \subset [N]$ of size $\tau$.
    2. Sample each coordinate of $\boldsymbol{z}_t$ iid from $\mathcal{N}\left(0, \frac{128NTL^2}{\tau\epsilon^2}\log\left(\frac{2.5NT}{\tau\delta}\right)\log(2/\delta)\eta_t^2 \vee \eta_t\right)$.
    3. Update $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta_t\left(\nabla r(\boldsymbol{\theta}) + \frac{N}{\tau}\sum_{i\in S}\nabla\ell(\boldsymbol{x}_i|\boldsymbol{\theta})\right) + \boldsymbol{z}_t$,
    4. Return $\boldsymbol{\theta}_{t+1}$ as a posterior sample (after a pre-defined burn-in period).
    5. Increment $t \leftarrow t + 1$.
  end for
```

[2] We use boundedness to make the presentation simpler. Boundedness trivially implies the linear growth condition in Sato and Nakagawa [187, Assumption 2].

let $\epsilon, \delta, \tau, T$ be chosen such that $T \geq \frac{\epsilon^2 N}{32\tau \log(2/\delta)}$. Then Algorithm 3 preserves (ϵ, δ)-differential privacy.

Proof. In every iteration, the only data access is $\sum_{i \in S} \nabla \ell(\boldsymbol{x}_i|\boldsymbol{\theta})$ and by the L-Lipschitz condition, the sensitivity of $\sum_{i \in S} \nabla \ell(\boldsymbol{x}_i|\boldsymbol{\theta})$ is at most $2L$. Get the essential noise that is added to $\sum_{i \in S} \nabla \ell(\boldsymbol{x}_i|\boldsymbol{\theta})$ by removing the $\frac{N^2 \eta_t^2}{\tau^2}$ factor from the variance σ^2 in the algorithm, and Gaussian mechanism, ensures the privacy loss to be smaller than $\frac{\epsilon\sqrt{N}}{\sqrt{32\tau T \log(2/\delta)}}$ with probability $> 1 - \frac{\tau\delta}{2NT}$.

Using the same technique in Bassily et al. [17], we can further exploit the fact that the subset S that we use to compute the stochastic gradient is chosen uniformly randomly. By Lemma 3.13, the privacy loss for this iteration is in fact

$$\frac{\epsilon\sqrt{N}}{\sqrt{32\tau T \log(2/\delta)}} \cdot \frac{2\tau}{N} = \frac{\epsilon/2}{\sqrt{2(NT/\tau)\log(2/\delta)}}.$$

Verify that we can indeed do that as $\frac{\epsilon\sqrt{N}}{\sqrt{32\tau T \log(2/\delta)}} < 1$ from the assumption on T. Note that to get T data passes with minibatches of size τ, we need to go through at most $\lfloor \frac{NT}{\tau} \rfloor \leq \frac{NT}{\tau}$ iterations. Apply the advanced composition theorem (Remark 3.12), we get an upper bound of the total privacy loss ϵ and failure probability $\delta = \frac{\delta}{2} + \frac{\tau\delta}{2NT} \cdot \frac{NT}{\tau}$ accordingly.

The proof is complete by noting that choosing a larger noise level when η_t is bigger can only reduces the privacy loss under the same failure probability. □

α-Phase transition. For any $\alpha \in (0, 1)$, if we choose $\eta_t = \frac{\alpha\epsilon^2}{128L^2 \log(2.5NT/(\tau\delta)) \log(2/\delta) t}$, then whenever $t > \alpha NT/\tau$, then we are essentially running SGLD for the last $(1 - \alpha)NT/\tau$ iterations, and we can collect approximate posterior samples from there.

Small constant η_0. Instead of making η_t to converge to 0 as t increases, we may alternatively use constant η_0 after t is larger than a threshold. This is a suggested heuristic in Welling and Teh [250] and is inline with the analysis in Sato and Nakagawa [187] and Vollmer et al. [238].

Algorithm 4 Hybrid Posterior Sampling Algorithm

```
Require: Data $X$ of size $N$, log-likelihood function $\ell(\cdot|\theta)$ with Lipschitz constant $L$ in the first
    argument, assume $\sup_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{x}\|$, a prior $\pi$. Privacy requirement $\epsilon$.
    1. Run OPS estimator: Algorithm 2 with $\epsilon/2$. Collect sample point $\theta_0$
    2. Run DP-SGLD (Algorithm 3) or other Stochastic Gradient Monte Carlo algorithms and
    collect samples.
output : Return all samples.
```

Choice of T and τ By Bassily et al. [17], it takes at least N data passes to converge in expectation to a point near the minimizer, so taking $T = 2N$ is a good choice. The variance of both random components in our stochastic gradient is smaller when we use larger τ. Smaller variances would improve the convergence of the stochastic gradient methods and make the SGLD a better approximation to the full Langevin Dynamics. The trade-off is that when τ is too large, we will use up the allowable T datapasses with just $O(T)$ iterations and the number of posterior samples we collect from the algorithm will be small.

Overcoming the large-noise in the "Burn-in" phase When the stepsize η_t is not small enough initially, we need to inject significantly more noise than what SGLD would have to ensure privacy. We can overcome this problem by initializing the SGLD sampler with a valid output of the OPS estimator, modified according to the exponential mechanism so that the privacy loss is calibrated to $\epsilon/2$. As the initial point is already in the high probability region of the posterior distribution, we no longer need to "Burn-in" the Monte Carlo sampler so we can simply choose a sufficiently small constant stepsize so that it remains a valid SGLD. This algorithm is summarized in Algorithm 4.

Comparing to OPS The privacy claim of DP-SGLD is very different from OPS . It does not require sampling to be nearly correct to ensure differential privacy. In fact, DP-SGLD privately releases the entire sequence of parameter updates, thus ensures differential privacy even if the internal state of the algorithm gets hacked. However, the quality of the samples is usually worse than OPS due to the random-walk like behavior. The interesting fact, however, is that if we run SGLD indefinitely without worrying about the stronger notion of internal privacy, it leads to a valid posterior sample. We can potentially modify the posterior distribution to sample from into the "scaled" version so as to balancing the two ways of getting privacy.

3.4.2 Hamiltonian Dynamics, Fisher Scoring and Nose-Hoover Thermostat

One of the practical drawback of SGLD is its random walk-like behavior which slows down the mixing significantly. In this section, we describe three extensions of SGLD that attempts to resolve the issue by either using auxiliary variables to counter the noise in the stochastic gradient[56, 65], or to exploit second order information so as to use Newton-like updates with large stepsize [4].

We note that in all these methods, stochastic gradients are the only form of data access, therefore similar results like what we described for SGLD follow nicely. We briefly describe each method and how to choose their parameters for differential privacy.

Stochastic Gradient Hamiltonian Monte Carlo. According to Neal [163], Langevin Dynamics is a special limiting case of Hamiltonian Dynamics, where one can simply ignore the “momentum” auxiliary variable. In its more general form, Hamiltonian Monte Carlo (HMC) is able to generate proposals from distant states and hence enabling more rapid exploration of the state space. Chen et al. [56] extends the full “leap-frog” method for HMC in Neal [163] to work with stochastic gradient and add a “friction” term in the dynamics to “de-bias” the noise in the stochastic gradient.

$$\begin{cases} \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + h_t \boldsymbol{r}_{t-1} \\ \boldsymbol{p}_t = \boldsymbol{p}_{t-1} - h_t \widehat{\nabla} - \eta_t A \boldsymbol{p}_{t-1} + \mathcal{N}(0, 2(A - \widehat{B})h_t). \end{cases} \tag{3.2}$$

where $\widehat{B}$ is a guessed covariance of the stochastic gradient (the authors recommend restricting $\hat{B}$ to a single number or a diagonal matrix) and A can be arbitrarily chosen as long as $A \succ \widehat{B}$. If the stochastic gradient $\widehat{\nabla} \sim \mathcal{N}(\nabla, B)$ for some B and $\widehat{B} = B$, then this dynamics is simulating a dynamic system that yields the correct distribution. Note that even if the normal assumption holds and we somehow set $\widehat{B} = B$, we still requires h_t to go to 0 to sample from the actual posterior distribution, and as h_t converges to 0 the additional noise we artificially inject dominates and we get privacy for free. All we need to do is to set A, $\widehat{B}$ and h_t so that $2(A - \widehat{B})/h_t \succ \frac{128NTL^2}{\tau\epsilon^2} \log\left(\frac{2.5NT}{\tau\delta}\right) \log(2/\delta) I_n$. Note that as $h_t \to 0$ this quickly becomes true.

Stochastic Gradient Nosé-Hoover Thermostat As we discussed, the key issue about SGHMC is still in choosing $\widehat{B}$. Unless $\widehat{B}$ is chosen exactly as the covariance of true stochastic gradient, it does not sample from the correct distribution even as $h_t \to 0$ unless we trivially set $\hat{B} = 0$. The Stochastic Gradient Nosé-Hoover Thermostat (SGNHT) overcomes the issue by introducing an additional auxiliary variable ξ, which serves as a thermostat to absorb the unknown noise in the stochastic gradient. The update equations of SGNHT are given below

$$\begin{cases} \boldsymbol{p}_t = \boldsymbol{p}_{t-1} - \xi_{t-1} \boldsymbol{p}_{t-1} h_t - \widehat{\nabla} h_t + \mathcal{N}(0, 2Ah_t); \\ \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + h_t \boldsymbol{p}_{t-1}; \\ \xi_t = \xi_{t-1} + (\frac{1}{n} \boldsymbol{p}_t^T \boldsymbol{p}_t - 1)h_t. \end{cases} \tag{3.3}$$

Similar to the case in SGHMC, appropriately selected discretization parameter h_t and the friction term A will imply differential privacy.

Chen et al. [56], Ding et al. [65] both described a reformulation that can be interpret as SGD with momentum. This is by setting parameters $\eta = h^2, a = hA, \hat{b} = h\widehat{B}$ for SGHMC:

$$\begin{cases} \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{v}_{t-1} \\ \boldsymbol{v}_t = \boldsymbol{v}_{t-1} - \eta_t \widehat{\nabla} - a\boldsymbol{v} + \mathcal{N}(0, 2(a - \widehat{b})\eta_t I); \end{cases} \tag{3.4}$$

and $\boldsymbol{v} = \boldsymbol{p}h, \eta_t = h_t^2, \alpha = \xi h$ and $a = Ah$ for SGNHT:

$$\begin{cases} \boldsymbol{v}_t = \boldsymbol{v}_{t-1} - \alpha_{t-1} \boldsymbol{v}_{t-1} - \widehat{\nabla}(\boldsymbol{\theta}_{t-s})\eta_t + \mathcal{N}(0, 2a\eta_t I); \\ \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{u}_{t-1}; \\ \alpha_t = \alpha_{t-1} + (\frac{1}{n} \boldsymbol{v}_t^T \boldsymbol{v}_t - \eta_t). \end{cases} \tag{3.5}$$

where $1 - a$ is the momentum parameter and η is the learning rate in the SGD with momentum. Again note that to obtain privacy, we need $\frac{2a}{\eta_t} \geq \frac{128NTL^2}{\tau\epsilon^2} \log(\frac{2NT}{\tau\delta}) \log(1/\delta)$.

Note that as η_t gets smaller, we have the flexibility of choosing a and η_t within a reasonable range.

Stochastic Gradient Fisher Scoring Another extension of SGLD is Stochastic Gradient Fisher Scoring (SGFS), where Ahn et al. [4] proposes to adaptively interpolate between a preconditioned SGLD (see preconditioning [101]) and a Markov Chain that samples from a normal approximation of the posterior distribution. For parametric problem where Bernstein-von Mises theorem holds, this may be a good idea. The heuristic used in the SGFS is that the covariance matrix of $\boldsymbol{\theta}|X$, which is also the inverse Fisher information I_N^{-1} is estimated on the fly. The key features of SGFS is that one can use the stepsize to trade off speed and accuracy, when the stepsize is large, it mixes rapidly to the normal approximation, as the stepsize gets smaller the stationary distribution converges to the true posterior. Further details of SGFS and ideas to privatize it is described in Section 3.8.

3.4.3 Discussions and caveats.

So far, we have proposed a differentially private Bayesian learning algorithm that is memory efficient, statistically near optimal for a large class of problems, and we can release many intermediate iterates to construct error bars. Given that differential privacy is usually very restrictive, some of these results may appear too good to be true. This is a reasonable suspicion due to the following caveats.

Small η helps both privacy and accuracy. It is true that as η goes to 0, the stationary distribution that these method samples from gets closer to the target distribution. On the other hand, since the variance of the noise we need to add for privacy scales in $O(\eta^2)$ and that for posterior sampling scales like $O(\eta)$, privacy and accuracy benefits from the same underlying principle. The caveat is that we also have a budget on how many samples can we collect. Also the smaller the stepsize η is, the slower it mixes, as a result, the samples we collect from the monte carlo sampler is going to be more correlated to each other.

Adaptivity of SGNHT. While SGNHT is able to adaptively adjust the temperature so that the samples that it produces remain "unbiased" in some sense as $\eta \to 0$. The reality is that if the level of noise is too large, either we adjust the stepsize to be too small to search the space at all, or the underlying stochastic differential equation becomes unstable and quickly diverges. As a result, the adaptivity of SGNHT breaks down if the privacy parameter gets to small.

Computationally efficiency. For a large problem, it is usually the case that we would like to train with only one pass of data or very small number of data passes. However, due to the condition in Lemma 3.13, our result does not apply to one pass of data unless τ is chosen to be as

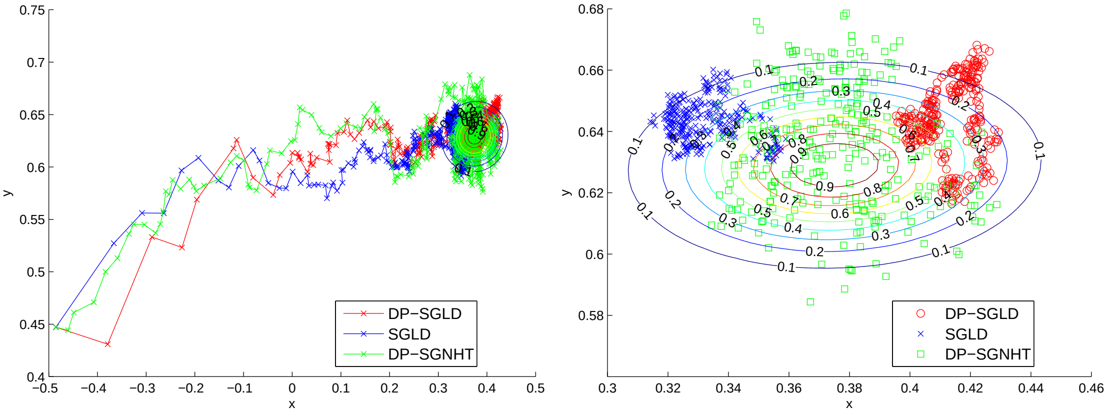


Figure 3.1: Illustration of stochastic gradient langevin dynamics and its private counterpart at $\epsilon = 10$.

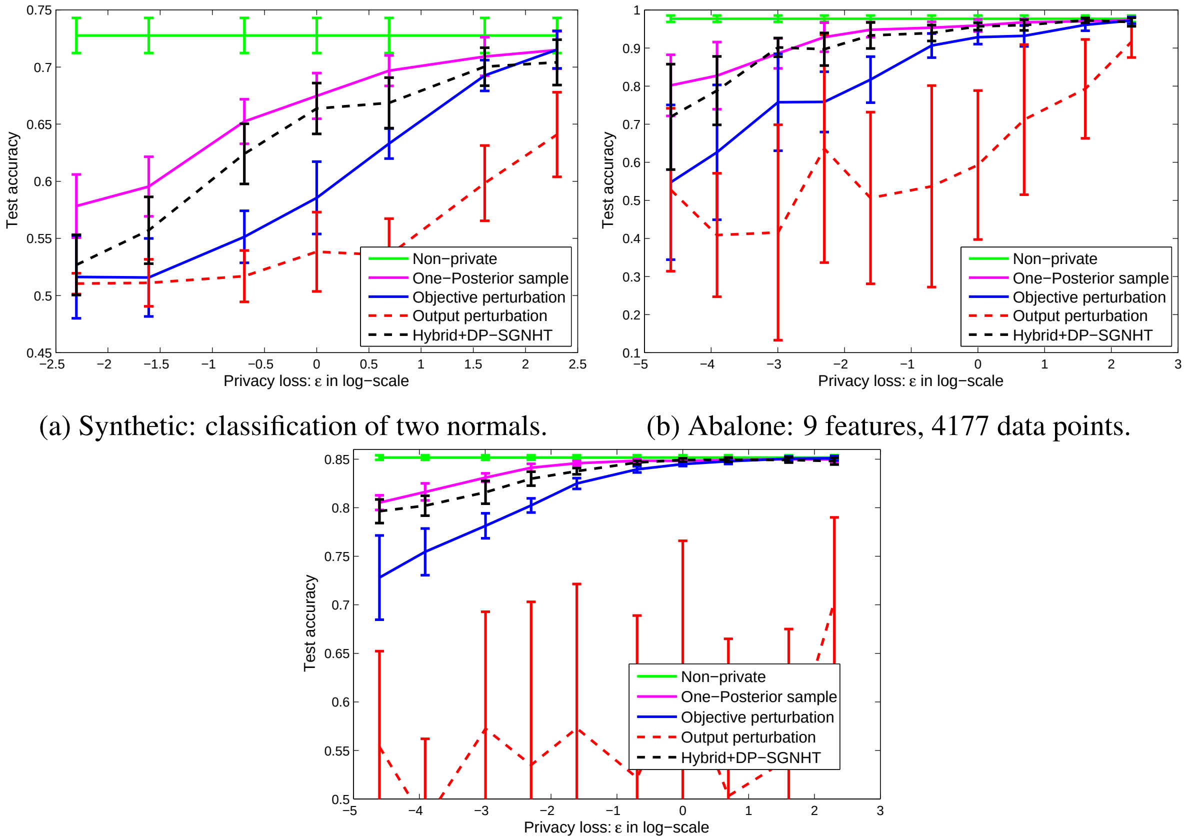
large as N. While we can still choose T to be sufficiently large and stop early, but we amount of noise that we add in each iteration will remain the same.

The Curse of Numerical constant. The analysis of algorithms often involves larger numerical constants and polylogarithmic terms in the bound. In learning algorithms these are often fine because there are more direct ways to evaluate and compare methods' performances. In differential privacy however, constants do matter. This is because we need to use these bounds (including constants) to decide how much noise or perturbation we need to inject to ensure a certain degree of privacy. These guarantees are often very conservative, but it is intractable to empirically evaluate the actual ϵ of differential privacy due to its "worst" case definition. Our stochastic gradient based differentially private sampler suffers from exactly that. For moderate data size, the product of the constant and logarithmic terms can be as large as a few thousands. That is the reason why it does not perform as well as other methods despite the theoretically being optimal in scaling (the optimality result is due to SGD [17]).

3.5 Experiments

Figure 8.5 is a plain illustration of how these stochastic gradient samplers work using a randomly generated linear regression model (note the its posterior distribution will be normal, as the contour illustrates). On the left, it shows how these methods converge like stochastic gradient descent to the basin of convergence. Then it becomes a posterior sampler. The figure on the right shows that the stochastic gradient thermostat is able to produce more accurate/unbiased result and the impact of differential privacy at the level of $\epsilon = 10$ becomes negligible.

To evaluate how our proposed methods work in practice, we selected two binary classification datasets: Abalone and Adult, from the first page of UCI Machine Learning Repository and performed privacy constrained logistic regression on them. Specifically, we compared two



(a) Synthetic: classification of two normals.

(b) Abalone: 9 features, 4177 data points.

(c) Adult: 109 features, 32561 data points.

Figure 3.2: Comparison of Differential Private methods.

of our proposed methods, OPS mechanism and hybrid algorithm against the state-of-the-art empirical risk minimization algorithm OBJPERT [55, 128] under varying level of differential privacy protection. The results are shown in Figure 3.2. As we can see from the figure, in both problems, OPS significantly improves the classification accuracy over OBJPERT . The hybrid algorithm also works reasonably well, given that it collected N samples after initializing it from the output of a run of OPS with privacy parameter $\epsilon/2$. For fairness, we used the (ϵ, δ)-DP version of the objective perturbation [128] and similarly we used Gaussian mechanism (rather than Laplace mechanism) for output perturbation. All optimization based methods are solved using BFGS algorithm to high numerical accuracy. OPS is implemented using SGNHT and we ran it long enough so that we are confident that it is a valid posterior sample. Minibatch size and number of data passes in the hybrid DP-SGNHT are chosen to be both $\sqrt{N}$.

We note that the plain DP-SGLD and DP-SGNHT without an initialization using OPS does not work nearly as well. In our experiments, it often performs equally or slightly worse than the output perturbation. This is due to the few caveats (especially "the curse of numerical constant") we described earlier.

3.6 Related work

We briefly discuss related work here. For the first part, we become aware recently that Mir [158] and Dimitrakakis et al. [64] independently developed the idea of using posterior sampling for differential privacy. Mir [158, Chapter 5] used a probabilistic bound of the log-likelihood to get $(\epsilon, \delta) - DP$ but focused mostly on conjugate priors where the posterior distribution is in closed-form. Dimitrakakis et al. [64] used Lipschitz assumption and bounded data points (implies our boundedness assumption) to obtain a generalized notion of differential privacy. Our results are different in that we also studied the statistical and computational properties. Bassily et al. [17] used exponential mechanism for empirical risk minimization and the procedure is exactly the same as OPS . Our difference is to connect it to Bayesian learning and to provide results on limiting distribution, statistical efficiency and approximate sampling. We are not aware of a similar asymptotic distribution with the exception of Smith [201], where a different algorithm (the subsample-and-aggregate scheme) is proven to give an estimator that is asymptotically normal and efficient (therefore, stronger than our result) under a different set of assumptions. Specifically, Smith [201]'s method requires boundedness of the parameter space while ours method can work with potentially unbounded space so long as the log-likelihood is bounded.

Related to the general topic, Kasiviswanathan and Smith [123] explicitly modeled the "semantics" of differential privacy from a Bayesian point of view, Xiao and Xiong [253] developed a set of tools for performing Bayesian inference under differential privacy, e.g., conditional probability and credibility intervals. Williams and McSherry [252] studied a related but completely different problem that uses posterior inference as a meta-post-processing procedure, which aims at "denoising" the privately obfuscated data when the private mechanism is known. Integrating Williams and McSherry [252] with our procedure might lead to some further performance boost, but investigating its effect is beyond the scope of the current paper.

For the second part, the idea to privately release stochastic gradient has been well-studied. Bassily et al. [17], Song et al. [205] explicitly used it for differentially private stochastic gradient descent. And Rajkumar and Agarwal [175] used it for private multi-party training. Our Theorem 3.16 is a simple modification of Theorem 2.1 in Bassily et al. [17]. Bassily et al. [17] also showed that the differential private SGD using Gaussian mechanism with $\tau = 1$ matches the lower bound up to constant and logarithmic, so we are confident that not many algorithms can do significantly better than Algorithm 3. Our contribution is to point out the interesting algorithmic structures of SGLD and extensions that preserves differential privacy. The method in Song et al. [205] requires disjoint minibatches in every data pass, and it requires adding significantly more noise in settings when Lemma 3.13 applies. Song et al. [205] are however applicable when we are doing only a small number of data passes and for these cases, it gets a much better constant. Rajkumar and Agarwal [175]'s setting is completely different as it injects a fixed amount of noise to the gradient corresponds to each data point exactly once. In this way, it replicates objective perturbation [55] (assuming the method actually finds the optimal solution).

Objective perturbation is originally proposed in Chaudhuri et al. [55] and the (ϵ, δ) version that we refer to first appears in Kifer et al. [128]. Comparing to our two mechanisms that attempts to sample from the posterior, their privacy guarantee requires the solution to be exact while ours does not. In comparison, OPS estimator is differentially private allows the distribution it samples from to be approximate, DP-SGLD on the other hand releases all intermediate results and every single iteration is public.

3.7 Conclusion and future work

In this chapter, we described two simple but conceptually interesting examples that Bayesian learning can be inherently differentially private. Specifically, we show that getting one sample from the posterior is a special case of exponential mechanism and this sample as an estimator is near-optimal for parametric learning. On the other hand, we illustrate that the algorithmic procedures of stochastic gradient Langevin Dynamics (and variants) that attempts to sample from the posterior also guarantee differential privacy as a byproduct. Preliminary experiments suggests that the One-Posterior-Sample mechanism works very well in practice and it substantially outperforms earlier privacy mechanism in logistic regression. While suffering from a large constant, our second method is also theoretically and practically meaningful in that it provides privacy protection in intermediate steps.

To carry the research forward, we think it is important to identify other cases when the existing randomness can be exploited for privacy. Randomized algorithms such as hashing and sketching, dropout and other randomization used in neural networks might be another thing to look at. More on the application end, we hope to explore the one-posterior sample approach in differentially private movie recommendation. Ultimately, the goal is to make differential privacy more practical to the extent that it can truly solve the real-life privacy problems that motivated its very advent.

3.8 Privatizing other stochastic gradient based sampling techniques

3.8.1 Fisher Scoring and Stochastic Gradient Fisher Scoring

Fisher scoring is simply the Newton's method for solving maximum likelihood estimation problem. The score function $S(\theta)$ is the gradient of the $\log$-likelihood. So intuitively, if we solve the equation $S(\theta) = 0$, we can obtain the maximum likelihood estimate. Often this equation is highly nonlinear, so we consider the an iterative update for the linearized score function (or a quadratic approximation of the likelihood) by Taylor expand it at the current point θ_0

$$S(\theta) \approx S(\theta_0) + I(\theta_0)(\theta - \theta_0)$$

where $I(\theta_0) = -\sum_{i=1}^{n} \nabla\nabla^T \ell(Z_i; \theta)$ is the observed Fisher information evaluated at θ_0.

By the fact that $S(\theta^*) = 0$, and plug in the above equation, we get $\theta^* = \theta_0 + I^{-1}(\theta_0)S(\theta_0)$ Note that this is a fix point iteration and it gives us an iterative update rule to search for θ^* via

$$\theta_{k+1} = \theta_k + I^{-1}(\theta_k)S(\theta_k).$$

Recall that S is the gradient of the score function and I^{-1} is the covariance of the score function and (under mild regularity conditions) the Hessian of the log-likelihood. As a result, this is often the same as Newton iterations.

An intuitive idea to avoid passing the entire dataset in every iteration is to simply replacing the gradient (the score function) with stochastic gradient and somehow estimate the Fisher information. Stochastic Gradient Fisher Scoring can be thought of as a Quasi-Newton method.

3.8.2 Privacy extension

By invoking a more advanced version of the Gaussian Mechanism, we will show that similar privacy guarantee can be obtained for a modified version of SGFS (described in Algorithm 5) while preserving its asymptotic properties. Specifically, under the assumption that I_N is given, when η_t is big, it also samples from a normal approximation (with larger variance), when η_t is small, the private algorithm becomes exactly the same as SGFS. Moreover, for a sequence of samples from the posterior, the online estimate in the Fisher Information converges an $O(1/N)$ approximation of true Fisher Information as in Ahn et al. [4, Theorem 1].

The privacy result relies on a more specific smoothness assumption. Assume that for any parameter $\theta \in \mathbb{R}^d$, and $X \in \mathcal{X}^N$ the ellipsoid $E = FB^d$ defined by transforming the unit ball B^d using a linear map F contains the symmetric polytope spanned by $\{\pm\nabla\ell(x_1, \theta), ..., \pm\nabla\ell(x_N, \theta)\}$. From a differential private point of view, this implies that $\nabla_\theta \ell(x, \theta)$'s sensitivity is different towards different direction. Then the non-spherical gaussian mechanism states

Algorithm 5 Differentially Private Stochastic Gradient Fisher Scoring (DP-SGFS)

```
Require: Data $X$ of size $N$, Size of minibatch $\tau$, number of data passes $T$, stepsize $\eta_t$
    for $t = 1, ..., \lfloor NT/\tau \rfloor$, a public Lipschitz matrix $F$, and initial $\theta_1$. Set $t = 1$, $\sigma^2 = \frac{32T\log(2.5NT/\tau\delta)\log(2/\delta)}{N\tau\epsilon^2}$
    for $t = 1 : \lfloor NT/\tau \rfloor$ do
        1. Random sample a minibatch $S \subset [N]$ of size $\tau$, compute $\bar{g} = \frac{1}{\tau}\sum_{i\in S}\nabla\ell(x_i|\theta)$.
        2. Sample $Z_t \sim \mathcal{N}(0, \sigma^2 \vee \frac{1}{N^2\eta_t} I_d)$, $W_{ij} \sim \mathcal{N}(0, 49\|F\|^4\sigma^2)$.
        3. Compute private stochastic gradient and sample covariance matrix
           $$\tilde{g} = \bar{g} + FZ_t, \quad \text{and} \quad V = \mathcal{P}_{S_+^d}\left\{\frac{1}{\tau-1}\sum_{i\in S}\left\{\nabla\ell_i(\theta_t) - \bar{g}\right\}\left\{\nabla\ell_i(\theta_t) - \bar{g}\right\}^T + W\right\}.$$
        4. Update the guessed Fisher Information estimate $\hat{I}_t = (1-\kappa_t)\hat{I}_{t-1} + \kappa_t V$.
        5. Update and return $\theta_{t+1} \leftarrow \theta_t + 2\left(\frac{(\tau+N)N}{\tau}\hat{I}_t + \frac{4FF^T}{\eta_t}\right)^{-1}(\nabla r(\theta_t) + N\tilde{g})$.
        6. Increment $t \leftarrow t + 1$.
    end for
```

Lemma 3.17 (Non-Spherical Gaussian Mechanism)**.** *Output* $\sum_{i=1}^N \nabla\ell(x_i, \theta) + Fw$ *where* $w \sim \mathcal{N}(0, \frac{(1+\sqrt{1\log(1/\delta)})^2}{\epsilon^2} I_d)$ *obeys* (ϵ, δ)*-DP.*

Theorem 3.18. *Let* F *be that* $\ell(x;\theta') \leq \ell_\theta + \nabla\ell(x;\theta)^T(\theta' - \theta) + \frac{1}{2}\|F(\theta' - \theta)\|^2$ *for any* $x \in \mathcal{X}, \theta \in \Theta$*. Moreover, let* $\epsilon, \delta, \tau, T$ *be chosen such that* $T \geq \frac{\epsilon^2 N}{32\tau\log(2/\delta)}$*. Then Algorithm 5 guarantees* $(2\epsilon, 2\delta)$*-differential privacy.*

Proof. First of all, $\|F\|_2$ is an upper bound for any $\nabla\ell(x|\theta)$, so by applying Lemma 3.19 on the every set of subsamples in each iteration, by Gaussian mechanism (Lemma 3.14) and the invariance to post-processing, we know that V is a private release. Then the proof follows by the same line of argument (subsampling and advanced composition) as in Theorem 3.16 for $\tilde{g}$ and V respectively, then the result follows by applying the simple composition theorem. □

Lemma 3.19 (Sensitivity of the sample covariance operator)**.** *Let* $\|x\| \leq L$ *for any* $x \in \mathcal{X}$*,* $n > 4$*, then*

$$\sup_{k,x_1,...,x_n,x_k'} \|\widehat{\mathrm{Cov}}(x_1, ..., x_k, ..., x_n) - \widehat{\mathrm{Cov}}(x_1, ..., x_k', ..., x_n)\|_F \leq \frac{7L^2}{n-1}.$$

Proof. We prove by taking the difference of two adjacent covariance matrices and bound the residual.

$$\begin{aligned}\widehat{\mathrm{Cov}}(X') =& \widehat{\mathrm{Cov}}(X) + \frac{1}{n-1}(xx^T - x'[x']^T) + \frac{1}{n(n-1)}(xx^T + x'[x']^T - x[x']^T - x'x^T) \\ &- \frac{1}{n-1}\mu(x-x')^T - \frac{1}{n-1}(x-x')\mu^T.\end{aligned}$$

Now assume $n > 4$ and take the upper bound of every term, we get $\Delta_2\left(\mathrm{Cov}(X)\right) \leq \frac{7L^2}{n-1}$. □

Chapter 4

On-Average Kullback-Leibler-privacy and its properties

In this chapter, we define On-Average KL-Privacy and present its properties and connections to differential privacy, generalization and information-theoretic quantities including max-information and mutual information. The new definition significantly weakens differential privacy, while preserving its minimalistic design features such as composition over small group and multiple queries as well as closeness to post-processing. Moreover, we show that On-Average KL-Privacy is *equivalent* to generalization for a large class of commonly-used tools in statistics and machine learning that samples from Gibbs distributions—a class of distributions that arises naturally from the maximum entropy principle. In addition, a byproduct of our analysis yields a lower bound for generalization error in terms of mutual information which reveals an interesting interplay with known upper bounds that use the same quantity.

4.1 Introduction

All previous chapters tries to understand how differential privacy work in the learning setting and come up with practical algorithms. In many cases, we can easily get a DP learning algorithm that is consistent, but when the loss function is discontinuous or the domain becomes unbounded, DP could easily become intractable. In addition, even if DP is possible, in practice it often requires adding noise with a very large magnitude, hence resulting in unsatisfactory utility, cf., [94, 228, 254].

Henceforth, a growing literature focuses on weakening the notion of differential privacy to make it applicable and for a more favorable privacy-utility trade-off. Popular attempts include (ϵ, δ)-approximate differential privacy [83], personalized differential privacy [88, 147], random differential privacy [108] and so on. They each have pros and cons and are useful in their specific contexts. There is a related literature addressing the folklore observation that "differential privacy implies generalization" [18, 85, 87, 109, 211, 246].

The implication of generalization is a minimal property that we feel any notion of privacy should have. This brings us to the natural question:

- *Is there a weak notion of privacy that is equivalent to generalization?*

In this chapter, we provide a partial answer to this question. Specifically, we define On-Average Kullback-Leibler(KL)-Privacy and show that it characterizes On-Average Generalization[1] for algorithms that draw sample from an important class of maximum entropy/Gibbs distributions, i.e., distributions with probability/density proportional to

$$\exp(-\mathcal{L}(\text{Output}, \text{Data}))\pi(\text{Output})$$

for a loss function $\mathcal{L}$ and (possibly improper) prior distribution π. Note that this is exactly the same class of sampling algorithms that we described in the previous chapter.

We argue that this is a fundamental class of algorithms that covers a big portion of tools in modern data analysis including Bayesian inference, empirical risk minimization in statistical learning as well as the private releases of database queries through Laplace and Gaussian noise adding. From here onwards, we will refer this class of distributions “MaxEnt distributions” and the algorithm that output a sample from a MaxEnt distribution “posterior sampling”.

Related work: This work is closely related to the various notions of algorithmic stability in learning theory [38, 125, 162, 191]. In fact, we can treat differential privacy as a very strong notion of stability. Thus On-average KL-privacy may well be called On-average KL-stability. Stability implies generalization in many different settings but they are often only sufficient conditions. Exceptions include [162, 191] who show that notions of stability are also necessary for the consistency of empirical risk minimization and distribution-free learnability of any algorithms. Our specific stability definition, its equivalence to generalization and its properties as a privacy measure has not been studied before. KL-Privacy first appears in [14] and is shown to imply generalization in [18]. On-Average KL-privacy further weakens KL-privacy. A high-level connection can be made to leave-one-out cross validation which is often used as a (slightly biased) empirical measure of generalization, e.g., see [161].

4.2 Symbols and Notation

We will use the standard statistical learning terminology where $z \in \mathcal{Z}$ is a data point, $h \in \mathcal{H}$ is a hypothesis and $\ell : \mathcal{Z} \times \mathcal{H} \to \mathbb{R}$ is the loss function. One can think of the negative loss function as a measure of utility of h on data point z. Lastly, $\mathcal{A}$ is a possibly randomized algorithm that maps a data set $Z \in \mathcal{Z}^n$ to some hypothesis $h \in \mathcal{H}$. For example, if $\mathcal{A}$ is the empirical risk minimization (ERM), then $\mathcal{A}$ chooses $h^* = \mathrm{argmin}_{h\in\mathcal{H}} \sum_{i=1}^n \ell(z_i, h)$.

Just to point out that many data analysis tasks can be casted in this form, e.g., in linear regression, $z_i = (x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, h is the coefficient vector and ℓ is just $\|y_i - x_i^T h\|^2$; in k-means clustering,

[1] We will formally define these quantities.

$z \in \mathbb{R}^d$ is just the feature vector, $h = \{h_1, ..., h_k\}$ is the collection of k-cluster centers and $\ell(z, h) = \min_j \|z - h_j\|^2$. Simple calculations of statistical quantities can often be represented in this form too, e.g., calculating the mean is equivalent to linear regression with identity design, and calculating the median is the same as ERM with loss function $|z - h|$.

We also consider cases when the loss function is defined over the whole data set $Z \in \mathcal{Z}$, in this case the loss function is also evaluated on the whole data set by the structured loss $\mathcal{L} : h \times \mathcal{Z} \to \mathbb{R}$. We do not require Z to be drawn from some product distribution, but rather any distribution D. Generally speaking, Z could be a string of text, a news article, a sequence of transactions of a credit card user, or rather just the entire data set of n iid samples. We will revisit this generalization with more concrete examples later. However we would like to point out that this is equivalent to the above case when we only have one (much more complicated) data point and the algorithm $\mathcal{A}$ is applied to only one sample.

4.3 Main Results

We first describe differential privacy and then it will become very intuitive where KL-privacy and On-Average KL-privacy come from. Roughly speaking, differential privacy requires that for any datasets Z and Z' that differs by only one data point, the algorithm $\mathcal{A}(Z)$ and $\mathcal{A}(Z')$ samples output h from two distributions that are very similar to each other. Recall that we use the "Hamming distance" to measure the distance between two data sets

$$d(Z, Z') := \#\{i = 1, ..., n : z_i \neq z'_i\}\,. \tag{4.1}$$

Assume the range of $\mathcal{A}$ is the whole space $\mathcal{H}$, and also assume $\mathcal{A}(Z)$ defines a density on $\mathcal{H}$ with respect to a base measure on $\mathcal{H}^2$, then ϵ-Differential Privacy (see Definition 2.2 in Chapter 2) requires

$$\sup_{Z,Z':d(Z,Z')\leq 1} \sup_{h\in\mathcal{H}} \log \frac{p_{\mathcal{A}(Z)}(h)}{p_{\mathcal{A}(Z')}(h)} \leq \epsilon.$$

Replacing the second supremum with an expectation over $h \sim \mathcal{A}(Z)$ we get the maximum KL-divergence over the output from two adjacent datasets. This is KL-Privacy as defined in Barber and Duchi [14], and by replacing both supremums with expectations we get what we call On-Average KL-Privacy. For $Z \in \mathcal{Z}^n$ and $z \in \mathcal{Z}$, denote $[Z_{-1}, z] \in \mathcal{Z}^n$ the data set obtained from replacing the first entry of Z by z. Also recall that the KL-divergence between two distributions F and G is $D_{\mathrm{KL}}(F\|G) = \mathbb{E}_F \frac{dF}{dG}$.

Definition 1 (On-Average KL-Privacy). *We say $\mathcal{A}$ obeys ϵ-On-Average KL-privacy for some distribution $\mathcal{D}$ if*

$$\mathbb{E}_{Z\sim\mathcal{D}^n, z\sim\mathcal{D}} \mathrm{D}_{\mathrm{KL}}(\mathcal{A}(Z)\|\mathcal{A}([Z_{-1}, z])) \leq \epsilon.$$

[2] These assumptions are only for presentation simplicity. The notion of On-Average KL-privacy can naturally handle mixture of densities and point masses.

Note that by the property of KL-divergence, the On-Average KL-Privacy is always nonnegative and is 0 if and only if the two distributions are the same almost everywhere. In the above case, it happens when $z = z'$.

Unlike differential privacy that provides a uniform privacy guarantee for any users in $\mathcal{Z}$, on-average KL-Privacy is a distribution-specific quantity that measures the amount of average privacy loss of an average data point $z \sim \mathcal{D}$ suffer from running data analysis $\mathcal{A}$ on an data set Z drawn iid from the same distribution $\mathcal{D}$.

We argue that this kind of average privacy protection is practically useful because it is able to adapt to benign distributions and is much less sensitive to outliers. After all, when differential privacy fails to provide a meaningful ϵ due to peculiar data sets that exist in $\mathcal{Z}^n$ but rarely appear in practice, we would still be interested to gauge how a randomized algorithm protects a typical user's privacy.

Now we define what we mean by *generalization*. Let the empirical risk $\hat{R}(h, Z) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, z_i)$ and the actual risk be $R(h) = \mathbb{E}_{z\sim\mathcal{D}}\ell(h, z)$.

Definition 2 (On-Average Generalization)**.** *We say an algorithm $\mathcal{A}$ has on-average generalization error ϵ if $\left|\mathbb{E}R(\mathcal{A}(Z)) - \mathbb{E}\hat{R}(\mathcal{A}(Z), Z)\right| \leq \epsilon$.*

This is slightly weaker than the standard notion of generalization in machine learning which requires $\mathbb{E}|R(\mathcal{A}(Z)) - \hat{R}(\mathcal{A}(Z), Z)| \leq \epsilon$. Nevertheless, on-average generalization is sufficient for the purpose of proving consistency for methods that approximately minimizes the empirical risk.

4.3.1 The equivalence to generalization

It turns out that when $\mathcal{A}$ assumes a special form, that is, sampling from a Gibbs distribution, we can completely characterize generalization of $\mathcal{A}$ using On-Average KL-Privacy. This class of algorithms include the most general mechanism for differential privacy — exponential mechanism [156], which casts many other noise adding procedures as special cases. We will discuss a more compelling reason why restricting our attention to this class is not limiting in Section 4.3.3.

Theorem 3 (On-Average KL-Privacy ⇔ Generalization)**.** *Let the loss function $\ell(z, h) = -\log p(z|h)$ for some model p parameterized by h, and let*

$$\mathcal{A}(Z) : h \sim p(h|Z) \propto \exp\left(-\sum_{i=1}^{n} \ell(z, h) - r(h)\right).$$

If in addition $\mathcal{A}(Z)$ obeys that for every Z, the distribution $p(h|Z)$ is well-defined (in that the normalization constant is finite), then $\mathcal{A}$ satisfy ϵ-On-Average KL-Privacy if and only if $\mathcal{A}$ *has on-average generalization error ϵ.*

The proof, given in the Section 9.7, uses a ghost sample trick and the fact that the expected normalization constants of the sampling distribution over Z and Z' are the same.

Remark 4.1 (Structural Loss). *Take $n = 1$, and loss function be $\mathcal{L}$. Then for an algorithm $\mathcal{A}$ that samples with probability proportional to $\exp(-\mathcal{L}(h, Z) - r(h))$: ϵ-On-Average KL-Privacy is equivalent to ϵ-generalization of the structural loss.*

Remark 4.2 (Dispersion parameter γ). *The case when $\mathcal{A} \propto \exp(-\gamma[\mathcal{L}(h, Z) - r(h)])$ for a constant γ can be handled by redefining $\mathcal{L}' = \gamma\mathcal{L}$. In that case, ϵ_γ-On-Average KL-Privacy with respect to $\mathcal{L}'$ implies ϵ_γ/γ generalization with respect to $\mathcal{L}$. For this reason, larger γ may not imply strictly better generalization.*

Remark 4.3 (Comparing to differential Privacy). *Note that here we do not require ℓ to be uniformly bounded, but if we do, i.e. $\sup_{z\in\mathcal{Z},h\in\mathcal{H}} |\ell(z, h)| \leq B$, then the same algorithm $\mathcal{A}$ above obeys $4B\gamma$-Differential Privacy [156, 245] and it implies $O(B\gamma)$-generalization. This, however, could be much larger than the actual generalization error (see our examples in Section 9.5).*

4.3.2 Preservation of other properties of DP

We now show that despite being much weaker than DP, On-Average KL-privacy does inherent some of the major properties of differential privacy (under mild additional assumptions in some cases).

Lemma 4 (Closeness to Post-processing). *Let f be any (possibly randomized) measurable function from $\mathcal{H}$ to another domain $\mathcal{H}'$, then for any Z, Z'*

$$D_{\mathrm{KL}}(f(\mathcal{A}(Z))\|f(\mathcal{A}(Z'))) \leq D_{\mathrm{KL}}(\mathcal{A}(Z)\|\mathcal{A}(Z')).$$

Proof. This directly follows from the data processing inequality for the Rényi divergence in Van Erven and Harremoës [233, Theorem 1]. □

Lemma 5 (Small group privacy). *Let $k \leq n$. Assume $\mathcal{A}$ is posterior sampling as in Theorem 3. Then for any $k = 1, ..., n$, we have*

$$\mathbb{E}_{[Z,z_{1:k}]\sim\mathcal{D}^{n+k}} D_{\mathrm{KL}}\left(\mathcal{A}(Z)\|\mathcal{A}([Z_{-1:k}, z_{1:k}])\right) = k\mathbb{E}_{[Z,z]\sim\mathcal{D}^{n+1}} D_{\mathrm{KL}}\left(\mathcal{A}(Z)\|\mathcal{A}([Z_{-1}, z])\right).$$

Lemma 6 (Composition Theorem).

1. *[Nonadaptive composition] Let $\mathcal{A}$ be ϵ_1-(On-Average) KL-Private and $\mathcal{B}$ be ϵ_2-(On-Average) KL-Privacy, then the non-adaptive composition $(\mathcal{A}, \mathcal{B})$ is $(\epsilon_1 + \epsilon_2)$-(On-Average) KL-Privacy.*
2. *[Adaptive composition] Let $\mathcal{A}$ be ϵ_1-On-Average KL Private and $\mathcal{B}(\cdot, h)$ be ϵ_2-KL-Privacy for every $h \in \Omega_{\mathcal{A}}$ where the support of random function $\mathcal{A}$ is $\Omega_{\mathcal{A}}$. Then $(\mathcal{A}, \mathcal{B})$ jointly is $(\epsilon_1 + \epsilon_2)$-On-Average KL-Privacy.*

We prove Lemma 5 and 6 in the Section 9.7. Note that On-Average KL privacy does not compose adaptively, but KL-privacy does [3]. However, if we generalize the notion of On-Average KL-privacy and parameterize $\mathcal{B}$'s on-average KL privacy loss ϵ by h, data distribution and target distribution, then an adaptively chosen $\mathcal{B}$ will lead to an on-average KL privacy of exactly

$$\epsilon_1(\mathcal{D}^n, \mathcal{D}) + \mathbb{E}_h[\epsilon_2(h, \mathcal{D}^n|h, \mathcal{D})].$$

[3] We thank Thomas Steinke and Jon Ullman for pointing out an error in a previous version.

How larger this quantity is relative to $\epsilon_1(\mathcal{D}^n, \mathcal{D}) + \epsilon_2(\mathcal{D}^n, \mathcal{D})$ depends on the extent $\mathcal{B}$ changes as a function of h.

4.3.3 Posterior Sampling as Max-Entropy solutions

In this section, we give a few theoretical justifications why restricting to posterior sampling is not limiting the applicability of Theorem 3 much. First of all, Laplace, Gaussian and Exponential Mechanism in the Differential Privacy literature are special cases of this class. Secondly, among all distributions to sample from, the Gibbs distribution is the variational solution that simultaneously maximizes the conditional entropy and utility. To be more precise on the claim, we first define conditional entropy.

Definition 7 (Conditional Entropy)**.** *Conditional entropy*

$$H(\mathcal{A}(Z)|Z) = -\mathbb{E}_Z \mathbb{E}_{h\sim\mathcal{A}(Z)} \log p(h|Z)$$

where $\mathcal{A}(Z) \sim p(h|Z)$.

Theorem 8. *Let $Z \sim \mathcal{D}^n$ for any distribution $\mathcal{D}$. A variational solution to the following convex optimization problem*

$$\min_{\mathcal{A}} \quad -\frac{1}{\gamma}\mathbb{E}_{Z\sim\mathcal{D}^n} H(\mathcal{A}(Z)|Z) + \mathbb{E}_{Z\sim\mathcal{D}^n}\mathbb{E}_{h\sim\mathcal{A}(Z)} \sum_{i=1}^{n} \ell_i(h, z_i) \tag{4.2}$$

is $\mathcal{A}$ that outputs h with distribution $p(h|Z) \propto \exp\left(-\gamma \sum_{i=1}^n \ell_i(h, z_i)\right)$.

Proof. This is an instance of Theorem 3 in Mir [159] (first appeared in Tishby et al. [225]) by taking the distortion function to be the empirical risk. Note that this is a simple convex optimization over the functions and the proof involves substituting the solution into the optimality condition with a specific Lagrange multiplier chosen to appropriately adjust the normalization constant. □

The above theorem is distribution-free, and in fact works for every instance of Z separately. Condition on each Z, the variational optimization finds the distribution with maximum information entropy among all distributions that satisfies a set of utility constraints. This corresponds to the well-known principle of maximum entropy (MaxEnt) [122]. Many philosophical justifications of this principle has been proposed, but we would like to focus on the statistical perspective and treat it as a form of regularization that penalizes the complexity of the chosen distribution (akin to Akaike Information Criterion [7]), hence avoiding overfitting to the data. For more information, we refer the readers to references on MaxEnt's connections to thermal dynamics [122], to Bayesian inference and convex duality [8] as well as its modern use in modeling natural languages [28].

Note that the above characterization also allows for any form of prior $\pi(h)$ to be assumed. Denote prior entropy $\tilde{H}(h) = -\mathbb{E}_{h\sim\pi(h)} \log(\pi(h))$, and define information gain $\tilde{I}(h; Z) = \tilde{H}(h) - H(h|Z)$. The variational solution of $p(h|Z)$ that minimizes $\tilde{I}(h; Z) + \gamma\mathbb{E}_Z\mathbb{E}_{h|Z}\mathcal{L}(Z, h)$

is proportional to $\exp(-\mathcal{L}(Z,h))\pi(h)$. This provides an alternative way of seeing the class of algorithms $\mathcal{A}$ that we consider. $\tilde{I}(h;Z)$ can be thought of as a information-theoretic quantification of privacy loss, as described in [159, 257]. As a result, we can think of the class of $\mathcal{A}$ that samples from MaxEnt distributions as the most private algorithm among all algorithms that achieves a given utility constraint.

4.4 Max-Information and Mutual Information

Recently, Dwork et al. [85] defined approximate max-information and used it as a tool to prove generalization (with high probability). Russo and Zou [183] showed that the weaker mutual information implies on-average generalization under a distribution assumption of the entire space of $\{\mathcal{L}(h,Z)|h \in \mathcal{H}\}$ induced by distribution of Z. In this section, we compare On-Average KL-Privacy with these two notions. Note that we will use Z and Z' to denote two completely different datasets rather than adjacent ones as we had in differential privacy.

Definition 9 (Max-Information, Definition 11 in [85]). *We say $\mathcal{A}$ has an β-approximate max-information of k if for every distribution $\mathcal{D}$,*

$$I^{\beta}_{\infty}(Z;\mathcal{A}(Z)) = \max_{(H,Z)\subset\mathcal{H}\times\mathcal{Z}^n:\mathbb{P}(h\in H,Z\in\tilde{Z})>\beta} \log\frac{\mathbb{P}(h\in H, Z\in\tilde{Z})-\beta}{\mathbb{P}(h\in H)p(Z\in\tilde{Z})} \leq k.$$

This is alternatively denoted by $I^{\beta}_{\infty}(\mathcal{A},n) \leq k$. We say $\mathcal{A}$ has a pure max-information of k if $\beta = 0$.

It is shown that differential privacy and short description length imply bounds on max-information [85], hence generalization. Here we show that the pure max-information implies a very strong On-Average-KL-Privacy for any distribution $\mathcal{D}$ when we take $\mathcal{A}$ to be a posterior sampling mechanism.

Lemma 10 (Relationship to max-information). *If $\mathcal{A}$ is a posterior sampling mechanism as described in Theorem 3, then $I_{\infty}(\mathcal{A},n) \leq k$ implies that $\mathcal{A}$ obeys k/n-On-Average-KL-Privacy for any data generating distribution $\mathcal{D}$.*

An immediate corollary of the above connection is that we can now significantly simplify the proof for "max-information ⇒ generalization" for posterior sampling algorithms.

Corollary 11. *Let $\mathcal{A}$ be a posterior sampling algorithm. $I_{\infty}(\mathcal{A},n) \leq k$ implies that $\mathcal{A}$ generalizes with rate k/n.*

We now compare to mutual information and draw connections to [183].

Definition 12 (Mutual Information). *The mutual information*

$$I(\mathcal{A}(Z);Z) = \mathbb{E}_Z\mathbb{E}_{h\sim\mathcal{A}(Z)}\log\frac{p(h,Z)}{p(h)p(Z)}$$

where $\mathcal{A}(Z) \sim p(h|Z)$, $p(Z) = \mathcal{D}^n$ and $p(h) = \int p(h|Z)p(Z)dZ$.

Lemma 13 (Relationship to Mutual Information). *For any randomized algorithm $\mathcal{A}$, let $\mathcal{A}(Z)$ be an RV, and Z, Z' be two datasets of size n. We have*

$$I(\mathcal{A}(Z); Z) = D_{\mathrm{KL}}(\mathcal{A}(Z) \| \mathcal{A}(Z')) + \mathbb{E}_Z \mathbb{E}_{h \sim \mathcal{A}(Z)} \left[\mathbb{E}_{Z'} \log p(\mathcal{A}(Z')) - \log \mathbb{E}_{Z'} p(\mathcal{A}(Z')) \right],$$

which by Jensen's inequality implies $I(\mathcal{A}(Z), Z) \leq D_{\mathrm{KL}}(\mathcal{A}(Z) \| \mathcal{A}(Z'))$.

A natural observation is that for MaxEnt $\mathcal{A}$ defined with $\mathcal{L}$, mutual information lower bounds its generalization error. On the other hand, Proposition 1 in Russo and Zou [183] states that under the assumption that $\mathcal{L}(h, Z)$ is σ^2-subgaussian for every h, then the on-average generalization error is always smaller than $\sigma\sqrt{2I(\mathcal{A}(Z); \mathcal{L}(\cdot, Z))}$. Similar results hold for sub-exponential $\mathcal{L}(h, Z)$ [183, Proposition 3].

Note that in their bounds, $I(\mathcal{A}(Z); \mathcal{L}(\cdot, Z))$ is the mutual information between the choice of hypothesis h and the loss function for which we are defining generalization on. By data processing inequality, we have $I(\mathcal{A}(Z); \mathcal{L}(\cdot, Z)) \leq I(\mathcal{A}(Z); Z)$. Further, when $\mathcal{A}$ is posterior distribution, it only depends on Z through $\mathcal{L}(\cdot, Z)$, namely $\mathcal{L}(\cdot, Z)$ is a sufficient statistic for $\mathcal{A}$. As a result $\mathcal{A} \perp Z | \mathcal{L}(\cdot, Z)$. Therefore, we know $I(\mathcal{A}(Z); \mathcal{L}(\cdot, Z)) = I(\mathcal{A}(Z); Z)$. Combine this observation with Lemma 13 and Theorem 3, we get the following characterization of generalization through mutual information.

Corollary 14 (Mutual information and generalization). *Let $\mathcal{A}$ be an algorithm that samples $\propto \exp(-\gamma \mathcal{L}(h, Z))$, and $\mathcal{L}(h, Z) - R(h)$ is σ^2-subgaussian for any $h \in \mathcal{H}$, then*

$$\frac{1}{\gamma} I(\mathcal{A}(Z); Z) \leq \left| \mathbb{E}_Z \mathbb{E}_{h \sim \mathcal{A}(Z)} [\mathcal{L}(h, Z) - R(h)] \right| \leq \sigma \sqrt{2I(\mathcal{A}(Z); Z)}.$$

If $\mathcal{L}(h, Z) - R(h)$ is σ^2-subexponential with parameter (σ, b) instead, then we have a weaker upper bound $bI(\mathcal{A}(Z); Z) + \sigma^2/(2b)$.

The corollary implies that for each γ we have an intriguing bound that says $I(\mathcal{A}; Z) \leq 2\gamma^2\sigma^2$ for any distribution of Z, $\mathcal{H}$ and $\mathcal{L}$ such that $\mathcal{L}(\cdot, Z)$ is σ^2-subgaussian. One interesting case is when $\gamma = 1/\sigma$. This gives

$$\sigma I(\mathcal{A}(Z); Z) \leq \left| \mathbb{E}_Z \mathbb{E}_{h \sim \mathcal{A}(Z)} [\mathcal{L}(h, Z) - R(h)] \right| \leq \sigma \sqrt{2I(\mathcal{A}(Z); Z)}.$$

The lower bound is therefore sharp up to a multiplicative factor of $\sqrt{I(\mathcal{A}(Z); Z)}$.

4.5 Connections to Other Attempts to Weaken DP

We compare and contrast the On-Average KL-Privacy with other notions of privacy that are designed to weaken the original DP. The (certainly incomplete) list includes (ϵ, δ)-approximate differential privacy (Approx-DP) [83], random differential privacy (Rand-DP) [108], Personalized Differential Privacy (Personal-DP) [88, 147] and Total-Variation-Privacy (TV-Privacy) [14, 18]. Table 4.1 summarizes and compares of these definitions.

Privacy definition	Z	z	Distance (pseudo)metric
Pure DP	$\sup_{Z\in\mathcal{Z}^n}$	$\sup_{z\in\mathcal{Z}}$	$D_\infty(P\|Q)$
Approx-DP	$\sup_{Z\in\mathcal{Z}^n}$	$\sup_{z\in\mathcal{Z}}$	$D_\infty^\delta(P\|Q)$
Personal-DP	$\sup_{Z\in\mathcal{Z}^n}$	for each z	$D_\infty(P\|Q)$ or $D_\infty^\delta(P\|Q)$
KL-Privacy	$\sup_{Z\in\mathcal{Z}^n}$	$\sup_{z\in\mathcal{Z}}$	$D_{\mathrm{KL}}(P\|Q)$
TV-Privacy	$\sup_{Z\in\mathcal{Z}^n}$	$\sup_{z\in\mathcal{Z}}$	$\|P-Q\|_{TV}$
Rand-Privacy	$1-\delta_1$ any $\mathcal{D}^n$	$1-\delta_1$ any $\mathcal{D}$	$D_\infty^{\delta_2}(P\|Q)$
On-Avg KL-Privacy	$\mathbb{E}_{Z\sim\mathcal{D}^n}$ for each $\mathcal{D}$	$\mathbb{E}_{Z\sim\mathcal{D}}$ for each $\mathcal{D}$	$D_{\mathrm{KL}}(P\|Q)$

Table 4.1: Summary of different privacy definitions.

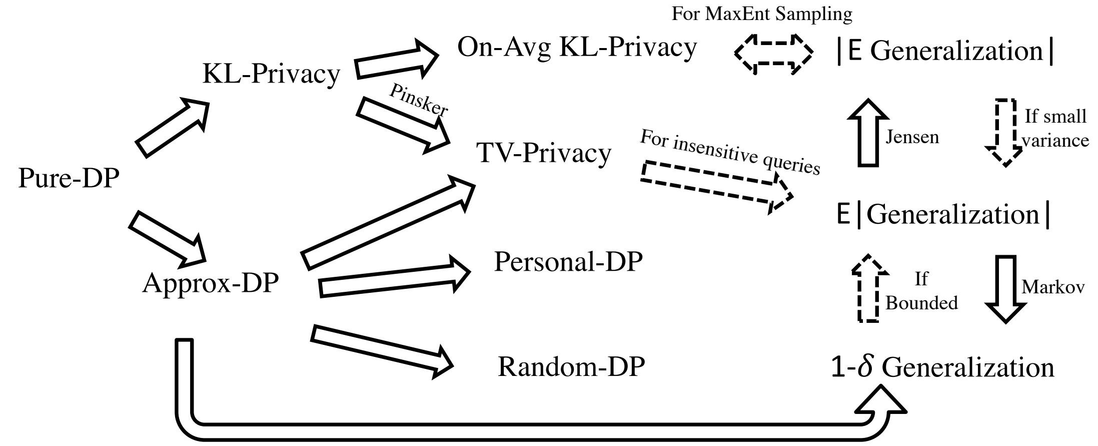


Figure 4.1: Relationship of different privacy definitions and generalization.

A key difference of On-Average KL-Privacy from almost all other previous definitions of privacy, is that the probability is defined only over the random coins of private algorithms. For this reason, even if we convert our bound into the high probability form, the meaning of the small probability δ would be very different from that in Approx-DP. The only exception in the list is Rand-DP, which assumes, like we do, the $n+1$ data points in adjacent data sets Z and Z' are draw iid from a distribution. Ours is weaker than Rand-DP in that ours is a distribution-specific quantity.

Among these notions of privacy, Pure-DP and Approx-DP have been shown to imply generalization with high probability [18, 87]; and TV-privacy was more shown to imply generalization (in expectation) for a restricted class of queries (loss functions) [18]. The relationship between our proposal and these known results are clearly illustrated in Fig. 4.1. To the best of our knowledge, our result is the first of its kind that crisply characterizes generalization.

Lastly, we would like to point out that while each of these definitions retains some properties of differential privacy, they might not possess all of them simultaneously and satisfactorily. For example, (ϵ,δ)-approx-DP does not have a satisfactory group privacy guarantee as δ grows exponentially with the group size.

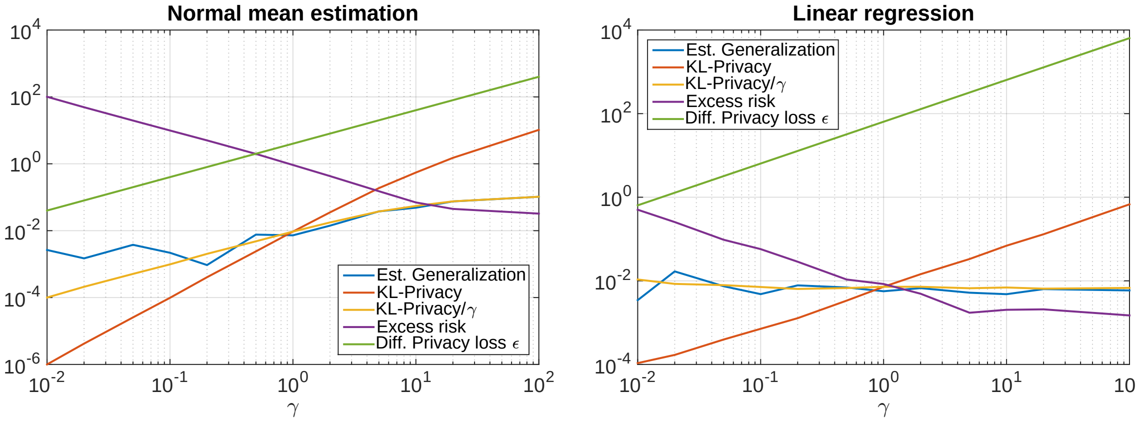


Figure 4.2: Comparison of On-Avg KL-Privacy and Differential Privacy on two examples.

4.6 Experiments

In this section, we validate our theoretical results through numerical simulation. Specifically, we use two simple examples to compare the ϵ of differential privacy, ϵ of on-average KL-privacy, the generalization error, as well as the utility, measured in terms of the excess population risk.

The first example is the private release of mean, we consider Z to be the mean of 100 samples from standard normal distribution truncated between $[-2, 2]$. Hypothesis space $\mathcal{H} = \mathbb{R}$, loss function $\mathcal{L}(Z, h) = |Z - h|$. $\mathcal{A}$ samples with probability proportional to $\exp(-\gamma|Z - h|)$. Note that this is the simple Laplace mechanism for differential privacy and the global sensitivity is 4, as a result this algorithm is 4γ-differentially private.

The second example we consider is a simple linear regression in 1D. We generate the data from a simple univariate linear regression model $y = xh +$ noise, where x and the noise are both sampled iid from a uniform distribution defined on $[-1, 1]$. The true h is chosen to be 1. Moreover, we use the standard square loss $\ell(z_i, h) = (y_i - x_i h)^2$. Clearly, the data domain $\mathcal{Z} = \mathcal{Y} \times \mathcal{X} = [-1, 1] \times [-2, 2]$ and if we constrain $\mathcal{H}$ to be within a bounded set $[-2, 2]$, $\sup_{x,y,\beta}(y - x\beta)^2 \leq 16$ and the posterior sampling with parameter γ obeys 64γ-DP.

Fig. 4.2 plots the results over an exponential grid of parameter γ. In these two examples, we calculate on-Average KL-Privacy using known formula of the KL-divergence of Laplace and Gaussian distributions. Then we stochastically estimate the expectation over data. We estimate the generalization error in the direct formula by evaluating on fresh samples. As we can see, appropriately scaled On-Average KL-Privacy characterizes the generalization error precisely as the theory predicts. On the other hand, if we just compare the privacy losses, the average ϵ from a random dataset given by On-Avg KL-Privacy is smaller than that for the worst case in DP by orders of magnitudes.

4.7 Conclusion

In this chapter, we presented On-Average KL-privacy as a new notion of privacy (or stability) on average. We showed that this new definition preserves properties of differential privacy including closedness to post-processing, small group privacy and composition over multiple data access. Moreover, we showed that On-Average KL-privacy/stability characterizes a weak form of generalization for a large class of sampling distributions that simultaneously maximize entropy and utility. This equivalence and connections to certain information-theoretic quantities allowed us to provide the first lower bound of generalization using mutual information. Lastly, we conduct numerical simulations which confirm our theory and demonstrate the substantially more favorable privacy-utility trade-off.

4.8 Proofs of technical results

Proof of Theorem 3. We prove this result using a ghost sample trick.

$$
\begin{aligned}
&\mathbb{E}_{z\sim\mathcal{D},Z\sim\mathcal{D}^n}\mathbb{E}_{h\sim\mathcal{A}(Z)}\left[\ell(z,h)-\frac{1}{n}\sum_{i=1}^{n}\ell(z_i,h)\right]\\
=&\mathbb{E}_{Z'\sim\mathcal{D}^n,Z\sim\mathcal{D}^n}\mathbb{E}_{h\sim\mathcal{A}(Z)}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(z_i',h)-\frac{1}{n}\sum_{i=1}^{n}\ell(z_i,h)\right]\\
=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{z_i'\sim\mathcal{D},Z\sim\mathcal{D}^n}\mathbb{E}_{h\sim\mathcal{A}(Z)}\left[\ell(z_i',h)-\ell(z_i,h)\right]\\
=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{z_i'\sim\mathcal{D},Z\sim\mathcal{D}^n}\mathbb{E}_{h\sim\mathcal{A}(Z)}\left[\ell(z_i',h)+\sum_{j\neq i}\ell(z_j,h)+r(h)-\ell(z_i,h)-\sum_{j\neq i}\ell(z_j,h)-r(h)\right]\\
=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{z_i'\sim\mathcal{D},Z\sim\mathcal{D}^n}\mathbb{E}_{\mathcal{A}(Z)}\left[-\log p_{\mathcal{A}([Z_{-i},z_i'])}(h)+\log p_{\mathcal{A}(Z)(h)}(h)+\log K_i-\log K_i'\right]\\
=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{z_i'\sim\mathcal{D},Z\sim\mathcal{D}^n}\mathbb{E}_{\mathcal{A}(Z)}\left[\log p_{\mathcal{A}(Z)}(h)-\log p_{\mathcal{A}([Z_{-i},z_i'])}(h)\right]\\
=&\mathbb{E}_{z\sim\mathcal{D},Z\sim\mathcal{D}^n}\mathbb{E}_{h\sim\mathcal{A}(Z)}\left[\log p_{\mathcal{A}(Z)}(h)-\log p_{\mathcal{A}([Z_{-1},z])}(h)\right]\ .
\end{aligned}
$$

The K_i and K_i' are partition functions of $p_{\mathcal{A}(Z)}(h)$ and $p_{\mathcal{A}([Z_{-i},z_i'])}(h)$ respectively. Since $z_i \sim z_i'$, we know $\mathbb{E}K_i - \mathbb{E}K_i' = 0$. The proof is complete by noting that the On-Average KL-privacy is always non-negative and so is the difference of the actual risk and expected empirical risk (therefore we can take absolute value without changing the equivalence).

□

Proof of Lemma 5. Let $k = 2$, we have

$$\begin{aligned}
&\mathbb{E}_{[Z,z_1',z_2']\sim\mathcal{D}^{n+2}}\mathbb{E}_{h\sim\mathcal{A}(Z)}\log\frac{p_{\mathcal{A}(Z)}(h)}{p_{\mathcal{A}([Z_{-1:2},z_1',z_2'])}(h)}\\
=&\mathbb{E}_{[Z,z_1']\sim\mathcal{D}^{n+1}}\mathbb{E}_{h\sim\mathcal{A}(Z)}\left[\log\frac{p_{\mathcal{A}(Z)}(h)}{p_{\mathcal{A}([Z_{-1},z_1'])}(h)}\right]\\
&+\mathbb{E}_{[Z,z_1',z_2']\sim\mathcal{D}^{n+2}}\mathbb{E}_{h\sim\mathcal{A}(Z)}\left[\log\frac{p_{\mathcal{A}([Z_{-1},z_1'])}(h)}{p_{\mathcal{A}([Z_{-1:2},z_1',z_2'])}(h)}\right]\\
\leq&\epsilon+\mathbb{E}_{[Z,z_1',z_2']\sim\mathcal{D}^{n+2}}\mathbb{E}_{h\sim\mathcal{A}(Z)}\left[\log\frac{p_{\mathcal{A}([Z_{-1},z_1'])}(h)}{p_{\mathcal{A}([Z_{-1:2},z_1',z_2'])}(h)}\right].
\end{aligned}$$

The technical issue is that the second term does not have the correct distribution to take expectation over. By the property of $\mathcal{A}$ being a posterior sampling algorithm, we can rewrite the second term of the above equation into

$$\mathbb{E}_{Z\sim\mathcal{D}^n,z_1',z_2'\sim\mathcal{D}}\mathbb{E}_{h\sim\mathcal{A}(Z)}\left[\log p(z_2,h)-\log p(z_2',h)-\log K+\log K'\right]$$

where K and K' are normalization constants of $\exp(\log p(z_1',h)+\sum_{i=2}^n \log p(z_i,h))$ and $\exp(\log p(z_1',h)+\log p(z_2',h)+\sum_{i=3}^n \log p(z_i,h))$ respectively. The expected log-partition functions are the same so we can replace them with normalization constants of $\exp(\sum_{i=1}^n \log p(z_i,h))$ and $\exp(\log p(z_2',h)+\sum_{i\neq 2}\log p(z_i,h))$. By adding and subtracting the missing log-likelihood functions on $z_1, z_3, ..., z_n$, we get

$$\begin{aligned}
&\mathbb{E}_{Z\sim\mathcal{D}^n,z_1',z_2'\sim\mathcal{D}}\mathbb{E}_{h\sim\mathcal{A}(Z)}\left[\log\frac{p_{\mathcal{A}([Z_{-1},z_1'])}(h)}{p_{\mathcal{A}([Z_{-1:2},z_2'])}(h)}\right]\\
=&\mathbb{E}_{Z\sim\mathcal{D}^n,z_1',z_2'\sim\mathcal{D}}\mathbb{E}_{h\sim\mathcal{A}(Z)}\left[\log\frac{p_{\mathcal{A}(Z)}(h)}{p_{\mathcal{A}([Z_{-2},z_2'])}(h)}\right]\leq\epsilon
\end{aligned}$$

This completes the proof for $k = 2$. Apply the same argument recursively by different decompositions of , we get the results for $k = 3, ..., n$.

The second statement follows by the same argument with all "$\leq$" changed into "$=$". □

Proof of Lemma 6.

$$\begin{aligned}
&\mathbb{E}_{h_1\sim\mathcal{A}(Z)}\mathbb{E}_{h_2\sim\mathcal{B}(Z,h_1)}\log\left[\frac{p_{\mathcal{B}(Z,h_1)}(h_2)}{p_{\mathcal{B}(Z',h_1)}(h_2)}\frac{p_{\mathcal{A}(Z)}(h_1)}{p_{\mathcal{A}(Z')}(h_1)}\right]\\
=&\mathbb{E}_{h_1\sim\mathcal{A}(Z)}\mathbb{E}_{h_2\sim\mathcal{B}(Z,h_1)}\log\left[\frac{p_{\mathcal{B}(Z,h_1)}(h_2)}{p_{\mathcal{B}(Z',h_1)}(h_2)}\right]+\mathbb{E}_{h_1\sim\mathcal{A}(Z)}\mathbb{E}_{h_2\sim\mathcal{B}(Z,h_1)}\log\left[\frac{p_{\mathcal{A}(Z)}(h_1)}{p_{\mathcal{A}(Z')}(h_1)}\right]
\end{aligned}$$

Take $\mathbb{E}$ over Z and z such that $Z' = [Z_{-1}, z]$ on both sides, we get On-Average-KL privacy of $(\mathcal{A},\mathcal{B})$ on the left, and on the right, if $\mathcal{B}$ is not adaptive to h_1, we get $\epsilon_2 + \epsilon_1$.

If $\mathcal{B}$ is adaptive to h_1, then we can swap the order of the integral in the joint distribution of Z and h_1, and get

$$\mathbb{E}_{h_1}\left[\mathbb{E}_{Z|h_1}\left[\mathbb{E}_{h_2\sim\mathcal{B}(Z,h_1)}\log\left[\frac{p_{\mathcal{B}(Z,h_1)}(h_2)}{p_{\mathcal{B}(Z',h_1)}(h_2)}\right]\big|h_1\right]\right]+\epsilon_1.$$

If $\mathcal{B}$ is KL-private, then it works for any pairs of adjacent data set, including those drawn from the conditional distribution $Z|h_1$ for any h_1, then $\mathbb{E}_{h_2\sim\mathcal{B}(Z,h_1)} \log\left[\frac{p_{\mathcal{B}(Z,h_1)}(h_2)}{p_{\mathcal{B}(Z',h_1)}(h_2)}\right] \le \epsilon_2$. This implies that we can upper bound the RHS by $\epsilon_2+\epsilon_1$.

□

Proof of Lemma 10. By Lemma 12 in Dwork et al. [85], $I_\infty(\mathcal{A},n) = \sup_{Z,Z'\in\mathcal{Z}^n} D_\infty(\mathcal{A}(Z)||\mathcal{A}(Z'))$.

$$\begin{aligned}
D_\infty(\mathcal{A}(Z)||\mathcal{A}(Z')) &= \sup_h \log\frac{p_{\mathcal{A}(Z)}(h)}{p_{\mathcal{A}(Z')}(h)}\\
&= \sup_h \sum_{i=1}^n \log\frac{p_{\mathcal{A}([Z_{-1:(i-1)},Z'_{1:(i-1)}])}(h)}{p_{\mathcal{A}([Z_{-1:(i)},Z'_{1:(i)}])}(h)}\\
&\ge \mathbb{E}_{h\sim\mathcal{A}(Z)}\sum_{i=1}^n \log\frac{p_{\mathcal{A}([Z_{-1:(i-1)},Z'_{1:(i-1)}])}(h)}{p_{\mathcal{A}([Z_{-1:(i)},Z'_{1:(i)}])}(h)}\\
&\ge \sum_{i=1}^n \mathbb{E}_{h\sim\mathcal{A}(Z)} \log\frac{p_{\mathcal{A}([Z_{-1:(i-1)},Z'_{1:(i-1)}])}(h)}{p_{\mathcal{A}([Z_{-1:(i)},Z'_{1:(i)}])}(h)}\\
&= \sum_{i=1}^n \mathbb{E}_{h\sim\mathcal{A}(Z)}(\log p(z_i;h) - \log p(z_i';h) - \log K_i + \log K_i')
\end{aligned}$$

where K_i and K_i' are normalization constants for distribution $p_{\mathcal{A}([Z_{-1:(i-1)},Z'_{1:(i-1)}])}(h)$ and $p_{\mathcal{A}([Z_{-1:(i)},Z'_{1:(i)}])}(h)$ respectively.

Take expectation over Z and Z' on both sides, by symmetry, the expected normalization constants are equal no matter which size n subset of $[Z,Z']$ this posterior distribution h is defined over. Define $Z^{(i)} := [z_1,...,z_{i-1},z_i',z_{i+1},...,z_n]$. Let K be the normalization constant of $\mathcal{A}(Z)$ and $K^{(i)}$ be the normalization constant of $\mathcal{A}(Z^{(i)})$. We get

$$\begin{aligned}
&\mathbb{E}_{Z,Z'\sim\mathcal{D}^n} D_\infty(\mathcal{A}(Z)||\mathcal{A}(Z')) \ge \mathbb{E}_{Z,Z'\sim\mathcal{D}^n}\sum_{i=1}^n \mathbb{E}_{h\sim\mathcal{A}(Z)}(\log p(z_i;h) - \log p(z_i';h))\\
=&\mathbb{E}_{Z,Z'\sim\mathcal{D}^n}\sum_{i=1}^n \mathbb{E}_{h\sim\mathcal{A}(Z)}\left[\sum_{j=1}^n \log p_h(z_j) - \sum_{j\neq i}\log p_h(z_j) - \log p_h(z_i') - \log K + \log K^{(i)}\right]\\
=&\mathbb{E}_{Z,Z'\sim\mathcal{D}^n}\sum_{i=1}^n \mathbb{E}_{h\sim\mathcal{A}(Z)}\left[\log\frac{p_{\mathcal{A}(Z)}(h)}{p_{\mathcal{A}(Z^{(i)})}(h)}\right]\\
=&\sum_{i=1}^n \mathbb{E}_{Z\sim\mathcal{D}^n,z_i'\sim\mathcal{D}}\mathbb{E}_{h\sim\mathcal{A}(Z)}\left[\log\frac{p_{\mathcal{A}(Z)}(h)}{p_{\mathcal{A}(Z^{(i)})}(h)}\right]
\end{aligned}$$

Note that

$$\text{RHS} = n\mathbb{E}_{Z\sim\mathcal{D}^n,z\sim\mathcal{D}}\text{KL}(\mathcal{A}(Z)\|\mathcal{A}([Z_{-1},z])),$$

and

$$\text{LHS} = \mathbb{E}_{Z,Z'\sim\mathcal{D}^n} D_\infty(\mathcal{A}(Z)\|\mathcal{A}(Z')) \le \sup_{Z,Z'\in\mathcal{Z}^n} D_\infty(\mathcal{A}(Z)\|\mathcal{A}(Z')) = I_\infty(\mathcal{A},n).$$

Collecting the three systems of inequalities above, we get that $\mathcal{A}$ is k/n-On-Average-KL-Privacy as claimed. □

Proof of Lemma 13. Denote $p(\mathcal{A}(Z)) = p(h|Z)$. $p(h, Z) = p(h|Z)p(Z)$. The marginal distribution of h is therefore $p(h) = \int_Z p(h, Z)dZ = \mathbb{E}_Z p(\mathcal{A}(Z))$. By definition,

$$\begin{aligned}
&I(\mathcal{A}(Z); Z) = \mathbb{E}_Z \mathbb{E}_{h|Z} \log \frac{p(h|Z)p(Z)}{p(h)p(Z)}\\
=&\mathbb{E}_Z \mathbb{E}_{h|Z} \log p(h|Z) - \mathbb{E}_Z \mathbb{E}_{h|Z} \log \mathbb{E}_{Z'} p(h|Z')\\
=&\mathbb{E}_Z \mathbb{E}_{h|Z} \log p(h|Z) - \mathbb{E}_{Z,Z'} \mathbb{E}_{h|Z} \log p(h|Z')\\
&+ \mathbb{E}_{Z,Z'} \mathbb{E}_{h|Z} \log p(h|Z') - \mathbb{E}_Z \mathbb{E}_{h|Z} \log \mathbb{E}_{Z'} p(h|Z')\\
=&\mathbb{E}_{Z,Z'} \mathbb{E}_{h|Z} \log \frac{p(h|Z)}{p(h|Z')} + \mathbb{E}_{Z,Z'} \mathbb{E}_{h|Z} \log p(h|Z') - \mathbb{E}_Z \mathbb{E}_{h|Z} \log \mathbb{E}_{Z'} p(h|Z')\\
=&D_{\mathrm{KL}}(\mathcal{A}(Z), \mathcal{A}(Z')) + \mathbb{E}_Z \mathbb{E}_{h|Z} \left[\mathbb{E}_{Z'} \log p(\mathcal{A}(Z')) - \log \mathbb{E}_{Z'} p(\mathcal{A}(Z'))\right]\\
\leq&D_{\mathrm{KL}}(\mathcal{A}(Z), \mathcal{A}(Z')).
\end{aligned}$$

The last line follows from Jensen's inequality. □

Chapter 5

Per-instance differential privacy, leverage scores and the statistical efficiency

In this chapter, we continues the pursuit on finding a more reasonable notation of privacy. Despite the usefulness of On-Average KL-privacy in some cases and its connection to learning-theoretic quantities, it is hard to justify that this is a strong enough notion of privacy to use, as the protection is for an average person from the population and says very little about individuals and could be especially bad for minorities or people with unique characteristics, as they will be the most vulnerable persons in the first place. Just imagine, if the agency wanting to collect my data tells me that most people's privacy will be protected fine, but "your mileage may vary", it will be unlikely for me to take the agency seriously.

Recall that differential privacy is too strong for two reasons. First, it requires that the probability ratio bound any potential output of the algorithm even if the output happens with negligible probability. This is already somewhat satisfactorily addressed by the widely accepted weaker notion: (ϵ, δ)-DP. Secondly, it requires the bound to hold for any two adjacent data sets, even including those that will never occur in practice. This often means that the privacy loss achieved by differential private algorithm might be overly conservative.

We address this issue by a new and more fine-grained notion of differential privacy — per instance differential privacy (pDP), which captures the privacy of a specific individual with respect to a fixed data set. We show that this is a strict generalization of the standard DP and inherits all its desirable properties, e.g., composition, invariance to side information and closedness to postprocessing, except that they all hold for every instance separately. When the data is drawn from a distribution, we show that per-instance DP implies generalization. Moreover, we provide explicit calculations of the per-instance DP for the output perturbation on a class of smooth learning problems. The result reveals an interesting and intuitive fact that an individual has stronger privacy if he/she has small "leverage score" with respect to the data set and if he/she can be predicted more accurately using the leave-one-out data set. Using the developed techniques, we provide a novel analysis of the One-Posterior-Sample (OPS) estimator and show that when the data set is well-conditioned it provides (ϵ, δ)-pDP for any target individuals and matches the exact lower bound up to a $1 + \tilde{O}(n^{-1}\epsilon^{-2})$ multiplicative factor. We also propose AdaOPS which uses adaptive regularization to achieve the same results with (ϵ, δ)-DP. Simulation shows several orders-of-magnitude more favorable privacy and utility trade-off when we consider the privacy of only the users in the data set.

5.1 Introduction

While modern statistics and machine learning had seen amazing success, their applications to sensitive domains involving personal data remain challenging due to privacy issues. Differential privacy [84] is a mathematical notion that allows strong provable protection of individuals from being identified by an arbitrarily powerful adversary, and has been increasingly popular within the machine learning community as a solution to the aforementioned problem [1, 55, 147, 155]. The strong privacy protection however comes with a steep price to pay. Differential privacy almost always lead to substantial and often unacceptable drop in utility, e.g., in contingency tables [93] and in genome-wide association studies [255]. This motivated a large body of research to focus on making differential privacy more practical [43, 79, 81, 96, 166, 197, 245] by exploiting local structures and/or revising the privacy definition.

Majority of these approaches adopt the "privacy-centric" model, which involves theoretically proving that an algorithm is differentially private for any data (within a data domain), then carefully analyzing the utility of the algorithm under additional assumptions on the data. For instance, in statistical estimation it is often assumed that the data is drawn i.i.d. from a family of distributions. In nonparametric statistics and statistical learning, the data are often assumed to having specific deterministic/structural conditions, e.g., smoothness, incoherence, eigenvalue conditions, low-rank, sparsity and so on. While these assumptions are strong and sometimes unrealistic, they are often necessary for a model to work correctly, even without privacy constraints. Take high-dimensional statistics for example, "sparsity" is almost never true, but if the true model is dense and unstructured, it is simply impossible to recover the true model anyways in the "small n large d" regime. That is why Friedman et al. [97] argued that one should "bet on sparsity" regardless and hope that it is a reasonable approximation of the reality. This is known as *adaptivity* in that an algorithm can perform provably better when some additional conditions are true.

The effect of these assumptions on privacy is unclear, mostly because there are no tools available to analyze such *adaptivity* in privacy. Since differential privacy is a worst-case quantity — a property of the randomized algorithm only (independent to the data) — it is unlikely that the obtained privacy loss ϵ could accurately quantify the privacy protection on a given data set at hand. It is always an upper bound, but the bound could be too conservative to be of any use in practice (e.g., when $\epsilon = 100$).

To make matter worse, the extent to which DP is conservative is highly problem-dependent. In cases like, releasing counting queries, the ϵ clearly measures the correct information leakage, since the sensitivity of such queries do not change with respect to the two adjacent data sets; however, in the context of machine learning and statistical estimation (as we will show later), the

ϵ of DP can be orders of magnitude larger than the actual limit of information leakage that the randomized algorithm guarantees. That is why in practice, it is challenging even for experts of differential privacy to provide a consistent recommendation on standard questions such as:

"What is the value of ϵ I should set in my application?"

In this chapter, we take a new "algorithm-centric" approach of analyzing privacy. Instead of designing algorithms that take the privacy loss ϵ as an input, we consider a fixed randomized algorithm and then analyze its privacy protection for every pair of adjacent data sets separately.

Our contribution is three-fold.

1. First, we develop per-instance differential privacy as a strict generalization of the standard pure and approximate DP. It provides a more fine-grained description of the privacy protection for each target individual and a fixed data set. We show that it inherits many desirable properties of differential privacy and can easily recover differential privacy for a given class of data and target users.
2. Secondly, we quantify the per-instance sensitivity in a class of smooth learning problems including linear and kernel machines. The result allows us to explicitly calculate per-instance DP of multivariate Gaussian mechanism. For an appropriately chosen noise covariance, the per-instance DP is proportional to the norm of the pseudo-residual in norm specified by the Hessian matrix. In particular, in linear regression, the per-instance sensitivity for a data point is proportional to its square root statistical leverage score and its leave-one-out prediction error.
3. Lastly, we analyze the procedure of releasing one sample from the posterior distribution (the OPS estimator) for ridge regression as an output perturbation procedure with a data dependent choice of covariance matrix. We show using the pDP technique that, when conditioning on a data set drawn from the linear regression model or having a well-conditioned design matrix, OPS achieves (ϵ, δ)-pDP for while matching the Cramer-Rao lower bound up to a $1 + \tilde{O}(n^{-1}\epsilon^{-2})$ multiplicative factor. OPS unfortunately cannot achieve DP with a constant ϵ while remaining asymptotically efficient. We fixed that by a new algorithm called ADAOPS , which provides (ϵ, δ)-DP and $1 + \tilde{O}(n^{-1}\epsilon^{-2})$-statistical efficiency at the same time.

5.1.1 Symbols and notations

Throughout the paper, we will use the standard notation in statistical learning. Data point $z \in \mathcal{Z}$. In supervised learning setting, $z = (x, y) \in \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$. We use $\theta \in \Theta$ to denote either the predictive function $\mathcal{X} \to \mathcal{Y}$ or the parameter vector that specifies such a function. $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ to denote the loss function or in a statistical model, ℓ represents the negative log-likelihood $-\log p_\theta(z)$. For example, in linear regression, $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} \subset \mathbb{R}$, $\Theta \subset \mathbb{R}^d$ and $\ell(\theta, (x, y)) = (y - x^T\theta)^2$. We use $\mathcal{A} : \mathcal{Z}^n \to P_\Theta$ to denote a randomized algorithm that outputs a draw from a distribution defined on a model space. Capital Z denotes a data set, ϵ and $\epsilon(Z, z)$ will be used to denote privacy loss.

5.2 Per-instance differential privacy

In this section, we define notion of per-instance differential privacy, and derive its properties. We begin by restating and parsing the standard definition of differential privacy.

Definition 5.1 (Differential privacy [84])**.** *We say a randomized algorithm $\mathcal{A}$ satisfies (ϵ, δ)-DP if for* all *data set Z and data set Z' that can be constructed by adding or removing one row z from Z,*

$$\mathbb{P}_{\theta\sim\mathcal{A}(Z)}(\theta \in \mathcal{S}) \leq e^{\epsilon}\mathbb{P}_{\theta\sim\mathcal{A}(Z')}(\theta \in \mathcal{S}) + \delta, \ \ \forall \textit{ measurable set } \mathcal{S}.$$

When $\delta = 0$, this is also known as pure differential privacy, and it is much stronger because for each data set Z, the protection holds *uniformly* over all privacy target z. When $\delta > 0$, then the protection becomes much weaker, in that the protection is stated for each privacy target separately.

It is helpful to understand what differential privacy is protecting against — a powerful adversary that knows everything in the entire universe, except one bit of information: whether a target z is in the data set or not in the data set. The optimal strategy for such an adversary is to conduct a likelihood ratio test (or posterior inference) on this bit, and differential privacy uses randomization to limit the probability of success of such test [249].

In the above, we described the original "In-or-Out" version of DP definition (see, e.g., [80, Definition 2.3, 2.4]). This is slightly different from the "Replace-One" version of the DP definition that we used in Chapter **??** and 4 that preserves the cardinality of the data set and makes it more convenient in those settings. The "replace-one" differential privacy protects against a slightly stronger adversary who know data set Z except one row and can limit the possibility of the unknown row to either z or z'. Again, this is only 1-bit of information that the adversary tries to infer and the optimal strategy for the adversary is to conduct a likelihood ratio test. In this chapter, we choose to work with the "In-or-Out" version of the differential privacy, although everything we derived can also be stated for the alternative version of differential privacy.

Note that the adversary always knows Z and has a clearly defined target z, and it is natural to evaluate the winnings and losses of the "player", the data curator by conditioning on the same data set and privacy target. This gives rise to the following generalization of DP.

Definition 5.2 (Per-instance Differential Privacy)**.** *For a fixed data set Z and a fixed data point z. We say a randomized algorithm $\mathcal{A}$ satisfy (ϵ, δ)-per-instance-DP for (Z, z) if for all measurable set $S \subset \Theta$, it holds that*

$$P_{\theta\sim\mathcal{A}(Z)}(\theta \in S) \leq e^{\epsilon}P_{\theta\sim\mathcal{A}([Z,z])}(\theta \in S) + \delta,$$
$$P_{\theta\sim\mathcal{A}([Z,z])}(\theta \in S) \leq e^{\epsilon}P_{\theta\sim\mathcal{A}(Z)}(\theta \in S) + \delta.$$

This definition is different from DP primarily because DP is the property of the $\mathcal{A}$ only and pDP is the property of both $\mathcal{A}, Z$ and z. If we take supremum over all $Z \in \mathcal{Z}^n$ and $z \in \mathcal{Z}$, then it recovers the standard differential privacy.

Similarly, we can define per-instance sensitivity for (Z, z).

Definition 5.3 (per-instance sensitivity)**.** *Let $\mathcal{H} = \mathbb{R}^d$, for a fixed Z and z. The per-instance $\|\cdot\|_*$ sensitivity of a function f : Data $\rightarrow \mathbb{R}^d$ is defined as $\|f(Z) - f([Z, z])\|_*$, where $\|\cdot\|_*$ could be ℓ_p norm or $\|\cdot\|_A = \sqrt{(\cdot)^T A(\cdot)}$ defined by a positive definite matrix A.*

This definition also generalizes quantities in the classic DP literature. If we condition on Z but maximize over all z, we get local-sensitivity [166]. If we maximize over all Z and z we get global sensitivity [80, Definition 3.1]. These two are often infinite in real-life problems, but for fixed data Z and target z to be protected, we could still get meaningful per-instance sensitivity.

Immediately, the per-instance sensitivity implies pDP for a noise adding procedure.

Lemma 5.4 (Multivariate Gaussian mechanism)**.** *Let $\hat{\theta}$ be a deterministic map from a data set to a point in Θ, e.g., a deterministic learning algorithm, and let the A-norm per-instance sensitivity $\Delta_A(Z, z)$ be $\|\hat{\theta}([Z, z]) - \hat{\theta}(Z)\|_A$. Then adding noise with covariance matrix A^{-1}/γ obeys $(\epsilon(Z, z), \delta)$-pDP for any $\delta > 0$ with*

$$\epsilon(Z, z) = \gamma \Delta_A(Z, z)\sqrt{\log(1.25/\delta)}.$$

The proof, which is standard and we omit, simply verifies the definition of (ϵ, δ)-pDP by calculating a tail bound of the privacy loss random variable and invokes Lemma 5.25.

5.2.1 Properties of pDP

We now describe properties of per-instance DP, which mostly mirror those of DP.

Fact 5.5 (Strong protection against identification)**.** *Let $\mathcal{A}$ obeys (ϵ, δ)-pDP for (Z, z), then for any measurable set $\mathcal{S} \subset \Theta$ where $\min\{\mathbb{P}_{\theta\sim\mathcal{A}(Z)}(\theta \in \mathcal{S}), \mathbb{P}_{\theta\sim\mathcal{A}(Z)}(\theta \in \mathcal{S})\} \geq \delta/\epsilon$ then given any side information* $\textsf{aux}$

$$-2\epsilon \leq \log \frac{\mathbb{P}_{\theta\sim\mathcal{A}(Z)}(\theta \in \mathcal{S}|\textsf{aux})}{\mathbb{P}_{\theta\sim\mathcal{A}([Z,z])}(\theta \in \mathcal{S}|\textsf{aux})} \leq 2\epsilon.$$

Proof. Note that after fixing Z, θ is a fresh sample from $\mathcal{A}(Z)$, as a result, $\theta \perp\!\!\!\perp \textsf{aux}|Z$. The claimed fact then directly follows from the definition. □

Note that the log-odds ratio measures how likely one is able to tell one distribution from another based on side information and an event $\mathcal{S}$ of the released result θ. When the log-odds ratio is close to 0, the outcome θ is equally likely to be drawn from either distributions.

Fact 5.6 (Convenient properties directly inherited from DP)**.** *For each (Z, z) separately we have:*

1. *Simple composition: Let $\mathcal{A}$ and $\mathcal{B}$ be two randomized algorithms, satisfying (ϵ_1, δ_1)-pDP, (ϵ_2, δ_2)-pDP, then $(\mathcal{A}, \mathcal{B})$ jointly is $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-pDP.*

2. *Advanced composition: Let $\mathcal{A}_1, ..., \mathcal{A}_k$ be a sequence of randomized algorithms, where $\mathcal{A}_i$ could depend on the realization of $\mathcal{A}_1(Z), ..., \mathcal{A}_i(Z)$, each with (ϵ, δ)-pDP, then jointly $\mathcal{A}_{1:k}$ obeys $O(\sqrt{k \log(1/\delta)}\epsilon), O(k\delta)$-pDP.*

3. *Closedness to post-processing: If $\mathcal{A}$ satisfies (ϵ_1, δ_1)-pDP, for any function f, $f(\mathcal{A}(\cdot))$ also obeys (ϵ_1, δ_1)-pDP.*

4. *Group privacy: If $\mathcal{A}$ obeys (ϵ,δ)-pDP with ϵ,δ parameterized by* $(\text{Data},\text{Target})$*, then*

$$P_{\theta\sim\mathcal{A}(Z)}(\theta\in S)\leq e^{\epsilon(Z,z_1)+\epsilon([Z,z_1],z_2)+...+\epsilon([Z,z_{1:k-1}],z_k)}P_{\theta\sim\mathcal{A}([Z,z_{1:k}])}(\theta\in S)+\tilde{\delta}.$$
$$P_{\theta\sim\mathcal{A}([Z,z_{1:k}])}(\theta\in S)\leq e^{\epsilon(Z,z_1)+\epsilon([Z,z_1],z_2)+...+\epsilon([Z,z_{1:k-1}],z_k)}P_{\theta\sim\mathcal{A}(Z)}(\theta\in S)+\tilde{\delta}.$$

for $\tilde{\delta}=\sum_{i=1:k}\left[\delta([Z,z_{1:i-1}],z_i)\prod_{j=1:i-1}e^{\epsilon([Z,z_{1:j-1}],z_j)}\right]$.

Proof. These properties all directly follow from the proof of these properties for differential privacy (see e.g., [80]), as the uniformity over data sets is never used in the proof. The only property that gets slightly different for the new definition is group privacy, since the size of the data set changes as the size of the privacy target (now a fixed group of people) gets larger. The claim follows from a simple calculation that repeatedly apply the definition of pDP for a different (Z,z). □

5.2.2 Moments of pDP, generalization and domain adaptation

One useful notion to consider in practice is to understand exactly how much privacy is provided for those who participated in the data sets. This is practically relevant, because if a cautious individual decides to not submit his/her data, he/she would necessarily do it by rejecting a data-usage agreement and therefore the data collector is not legally obligated to protect this person and in fact does not have access to his/her data in the first place. After all, the only type of identification risk that could happen to this person is that the adversary can be quite certain that he/she is not in the data set. For instance, in a study of graduate student income, a group of 200 students are polled and their average income is revealed with some small noise added to it. While an adversary can be almost certain based on the outcome that Bill Gates did not participate in the study, but that is hardly a any privacy risk to him. One advantage of pDP is that it offers a very natural way to analyze and also empirically estimate any statistics of the pDP losses over a data set or over a distribution of data points corresponding to a fixed randomized algorithm $\mathcal{A}$.
Definition 5.7 (Moment pDP for a distribution). *Let (Z,z) be drawn from some distribution (not necessarily a product distribution) $\mathcal{P}$, it induces a distribution of $\epsilon(Z,z)$. Then we say that the distribution obeys kth moment per-instance DP with parameter vector* $(\mathbb{E}\epsilon,\mathbb{E}[\epsilon^2],...,\mathbb{E}[\epsilon^k],\delta)$.

For example, one can treat the problem of estimating privacy loss for a fixed data set Z by choosing $\mathcal{P}$ to be a discrete uniform distribution supported on $\{(Z_{-i},z_i)\}_{i=1}^n$ with probability $1/n$ for each i. Taking $k=2$ allows us to calculate mean and variance of the privacy loss over the data set.

Similarly, if the data set is drawn iid from some unknown distribution $\mathcal{D}$ — a central assumption in statistical learning theory — then we can take $\mathcal{P}=\mathcal{D}^{n-1}\times\mathcal{D}$. This allows us to use the moment of pDP losses to capture on average, how well data points drawn from $\mathcal{D}$ are protected. It turns out that this also controls generalization error, and more generally cross-domain generalization.

Definition 5.8 (On-average generalization)**.** *Under the standard notations of statistical learning, the on average generalization error of an algorithm $\mathcal{A}$ is defined as*

$$Gen(\mathcal{A}, \mathcal{D}, n) = \left| \mathbb{E}_{Z\sim\mathcal{D}^n, z\sim\mathcal{D}} \mathbb{E}_{\theta\sim\mathcal{A}(Z)} \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, z_i) - \ell(\theta, z) \right|$$

Proposition 5.9 (Moment pDP implies generalization)**.** *Assume bounded loss function* $0 \leq \ell(\theta, z) \leq 1$*. Then the on-average generalization is smaller than*

$$\mathbb{E}_{Z\sim\mathcal{D}^n}(\mathbb{E}_{z\sim\mathcal{D}}[e^{\epsilon(Z,z)}|Z])^2 - 1 + \mathbb{E}_{Z\sim\mathcal{D}^n, z\sim\mathcal{D}}\delta(Z, z) + (\mathbb{E}_{Z\sim\mathcal{D}^n}\mathbb{E}_{z\sim\mathcal{D}}[e^{\epsilon(Z,z)}|Z]\mathbb{E}_{z\sim\mathcal{D}}[\delta(Z, z)|Z].$$

Note that this can also be used to capture the privacy and generalization of transfer learning (also known as domain adaptation) with a fixed data set or a fixed distribution. Let the training distribution be $\mathcal{D}$ and target distribution be $\mathcal{D}'$,

Take $\mathcal{P} = \mathcal{D}^n \otimes \mathcal{D}'$ or $\mathcal{P} = \delta_Z \otimes \mathcal{D}'$. In practice, this allows us to upper bound the generalization to the Asian demographics group, when the training data is drawn from a distribution that is dominated by white males (e.g., the current DNA sequencing data set). We formalize this idea as follows.

Definition 5.10 (Cross-domain generalization)**.** *Assume* $0 \leq \ell(\theta, z) \leq 1$*. The on-average cross-domain generalization with base distribution* $\mathcal{D}$ *to target distribution* $\mathcal{D}'$ *is defined as:*

$$Gen(\mathcal{A}, \mathcal{D}, \mathcal{D}', n) \leq \left| \mathbb{E}_{Z\sim\mathcal{D}^n, z\sim\mathcal{D}'} \mathbb{E}_{\theta\sim\mathcal{A}(Z)} \left[\frac{1}{n}\sum_{i=1}^{n} \rho_i \ell(\theta, z_i) - \ell(\theta, z) \right] \right|.$$

where $\rho_i = \mathcal{D}'(z_i)/\mathcal{D}(z_i)$ *is the inverse propensity (or importance weight) to account for the differences in the two domains.*

Proposition 5.11. *The cross-domain on-average generalization can be bounded as follows:*

$$Gen(\mathcal{A}, \mathcal{D}, \mathcal{D}', n) = \mathbb{E}_{Z\sim\mathcal{D}^{n-1}, \{z'\}\sim\mathcal{D}, z''\sim\mathcal{D}'}[(e^{\epsilon(Z,z')+\epsilon(Z,z'')} - 1) + \delta(Z, z') + \epsilon(Z, z')\delta(Z, z'')]$$

The expressions in Proposition 5.9 and 5.11 are a little complex, we will simplify it to make it more readable.

Corollary 5.12. *Let* $\sup_{Z,z}\delta(Z, z) \leq \delta$*, and* $\mathbb{E}_{\mathcal{D}}[e^{2\epsilon(Z,z)}] \leq 1$ *and for simplicity, we write* $\mathbb{E}_{Z\sim\mathcal{D}^n, z\sim\mathcal{D}}\epsilon(Z, z) = \mathbb{E}_{\mathcal{D}} f$ *and* $\mathbb{E}_{Z\sim\mathcal{D}^n, z\sim\mathcal{D}'}\epsilon(Z, z) = \mathbb{E}_{\mathcal{D}'} f$*. Then the cross domain on-average generalization is smaller than*

$$\frac{1}{2}[\mathbb{E}_{\mathcal{D}}e^{2\epsilon} + \mathbb{E}_{\mathcal{D}'}e^{2\epsilon}] - 1 + 2\delta = \frac{1}{2}\left[\sum_{i=1}^{\infty}\frac{2^i}{i!}\mathbb{E}_{\mathcal{D}}\epsilon^i + \mathbb{E}_{\mathcal{D}'}\epsilon^i\right] + 2\delta.$$

5.2.3 Related notions

We now compare the proposed privacy definition with existing ones in the literature. Most attempts to weaken differential privacy aims at more careful accounting of privacy loss by treating the $\epsilon = \log[p(\theta)/p'(\theta)]$ as a random variable. This produces nice connection of (ϵ,δ)-DP to concentration inequalities and in particular, it produces advanced composition of privacy loss via Martingale concentration. More recently, the idea is extended to defining weaker notions of privacy such as concentrated-DP [43, 81] and Rényi-DP [160] that allows for more fine-grained understanding of Gaussian mechanisms. Our work is complementary to this line of work, because we consider adaptivity of ϵ to a fixed pair of data-set and privacy target, and in some cases, ϵ being a random variable jointly parameterized by Z, z and θ. We summarize the differences of these definitions in the following table. It is clear from the table that if we ignore the differences in the probability metric used, per-instance DP is arguably the most general, and adaptive, since it depends on specific (Z, z) pairs.

	Data set	private target	probability metric	parametrized by
Pure-DP[84]	$\sup_Z$	$\sup_z$	$D_\infty(P\|Q)$	$\mathcal{A}$ only
Approx-DP[83]	$\sup_Z$	$\sup_z$	$D^\delta_\infty(P\|Q)$	$\mathcal{A}$ only
(z/m)-CDP[43, 81]	$\sup_Z$	$\sup_z$	$D_{\text{subG}}(P\|Q)$	$\mathcal{A}$ only
Rényi-DP[160]	$\sup_Z$	$\sup_z$	$D_\alpha(P\|Q)$	$\mathcal{A}$ only
Personal-DP[88, 147]	$\sup_Z$	fixed z	$D^\delta_\infty(P\|Q)$	$\mathcal{A}$ and z
TV-privacy[14]	$\sup_Z$	$\sup_z$	$\|P-Q\|_{TV}$	$\mathcal{A}$ only
KL-privacy[14]	$\sup_Z$	$\sup_z$	$D_{KL}(P\|Q)$	$\mathcal{A}$ only
On-Avg KL-privacy[247]	$\mathbb{E}_{Z\sim\mathcal{D}^n}$	$\mathbb{E}_{z\sim\mathcal{D}}$	$D_{KL}(P\|Q)$	$\mathcal{A}$ and $\mathcal{D}$
Per-instance DP	fixed Z	fixed z	$D^\delta_\infty(P\|Q)$	$\mathcal{A}$, Z and z

Table 5.1: Comparing variances of differential privacy.

The closest existing definition to ours is perhaps the personalized-DP, first seen in Ebadi et al. [88], Liu et al. [147]. It also tries to capture a personalized level of privacy. The difference is that personalized-DP requires the sensitivity of the private target to hold globally for all data sets.

On-Avg KL-privacy is also an adaptive quantity that measures the average privacy loss when the individuals in the data set and private target are drawn from the same distribution $\mathcal{D}$. pDP on the other hand measures the approximate worst-case privacy for a fixed (Z, z) pair that is not necessarily random. Not surprisingly, on-Avg KL-privacy and expected per-instance DP are intricately related to each other, as the following remark suggests.

Remark 5.13. *Let $\mathcal{A}(Z) = f(Z) + \mathcal{N}(0, \sigma^2 I)$, namely, Gaussian noise adding. Then On-avg KL-privacy is*

$$\mathbb{E}_{Z\sim\mathcal{D}^n, z\sim\mathcal{D}} D_{KL}(p(\theta|Z)\|p_{\theta|Z'}) = \mathbb{E}_{Z,Z'\sim\mathcal{D}^n}\|f(Z) - f([Z_{-1}, z])\|^2/\sigma^2.$$

Second moment of per-instance DP is

$$\mathbb{E}_{Z\sim\mathcal{D}^n, z\sim\mathcal{D}}\left[\left(\frac{\Delta_2 f(Z)\sqrt{1.25/\delta}}{\sigma}\right)^2\right] = \mathbb{E}_{Z\sim\mathcal{D}^n, z\sim\mathcal{D}}\|f(Z) - f([Z, z])\|^2 \log(1.25/\delta)/\sigma^2.$$

The two notions of privacy are almost equivalent. They differ only by a logarithmic term and by a minor change in the way perturbation is defined. In general, for Gaussian mechanism, ϵ-KL-privacy implies $(\sqrt{\epsilon\log(1.25/\delta)}, \delta)$-DP for any δ.

5.3 Per-instance sensitivity in smooth learning problems

In this section, we present our main results and give concrete examples in which per-instance sensitivity (hence per-instance privacy) can be analytically calculated. Specifically, we consider following regularized empirical risk minimization form:

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \sum_{i} \ell(\theta, z_i) + r(\theta). \tag{5.1}$$

or in the non-convex case, finding a local minimum. $\ell(\theta, z)$ and $r(\theta)$ are the loss functions and regularization terms. We make the following assumptions:

A.1. ℓ and r are differentiable in argument θ.

A.2. The partial derivatives are absolute continuous, i.e., they are twice differentiable almost everywhere and the second order partial derivatives are Lebesgue integrable.

Our results under these assumptions will cover learning problems such as linear and kernel machines as well as some neural network formulations (e.g., multilayer perceptron and convolutional net with sigmoid/tanh activation), but not non-smooth problems like lasso, ℓ_1-SVM or neural networks ReLU activation. We also note that these conditions are implied by standard assumptions of strong smoothness (gradient Lipschitz) and do not require the function to be twice differentiable everywhere. For instance, the results will cover the case when either ℓ or r is a Huber function, which is not twice differentiable.

Technically, these assumptions allow us to take Taylor expansion and have an integral form of the remainder, which allows us to prove the following stability bound.

Lemma 5.14. *Assume ℓ and r satisfy Assumption A.1 and A.2. Let $\hat{\theta}$ be a stationary point of $\sum_i \ell(\theta, z_i) + r(\theta)$, $\hat{\theta}'$ be a stationary point $\sum_i \ell(\theta, z_i) + \ell(\theta, z) + r(\theta)$ and in addition, let $\eta_t = t\hat{\theta} + (1-t)\hat{\theta}'$ denotes the interpolation of $\hat{\theta}$ and $\hat{\theta}'$. Then the following identity holds:*

$$\begin{aligned}\hat{\theta} - \hat{\theta}' &= \left[\int_0^1 \left(\sum_i \nabla^2 \ell(\eta_t, z_i) + \nabla^2 \ell(\eta_t, z) + \nabla^2 r(\eta_t)\right) dt\right]^{-1} \nabla \ell(\hat{\theta}, z) \\ &= -\left[\int_0^1 \left(\sum_i \nabla^2 \ell(\eta_t, z_i) + \nabla^2 r(\eta_t)\right) dt\right]^{-1} \nabla \ell(\hat{\theta}', z).\end{aligned}$$

The proof uses first order stationarity condition of the optimal solution and apply Taylor's theorem on the gradient. The lemma is very interpretable. It says that the perturbation of adding or removing a data point can be viewed as a one-step quasi-newton update to the parameter. Also note that $\nabla\ell(\hat{\theta}', z)$ is the "score function" in parametric statistical models, and is called a "pseudo-residual" in gradient boosting [see e.g., 97, Chapter 10].

The result implies that the per-instance sensitivity in $\|\cdot\|_A$ for some p.d. matrix A can be stated in terms of certain norm of the "score function" specified by a quadratic form $H^{-1}AH^{-1}$, and therefore by Lemma 5.4, the output perturbation algorithm:

$$\tilde{\theta} \sim \mathcal{N}(\hat{\theta}(X), A^{-1}/\gamma), \tag{5.2}$$

obeys (ϵ, δ)-pDP for any $\delta > 0$ and

$$\epsilon(Z, z) = \sqrt{\nabla\ell(\hat{\theta}', z)^T H^{-1}AH^{-1}\nabla\ell(\hat{\theta}', z)\log(1.25/\delta)}. \tag{5.3}$$

This is interesting because for most loss functions the "score function" is often proportional to the prediction error of the fitted model $\hat{\theta}'$ on data point z and this result suggests that the more accurately a model predicts a data point, the more private this data point is. This connection is made more explicit when we specialize to linear regression and the per-instance sensitivity

$$\Delta_A(Z, z) = |y - x^T\hat{\theta}|\sqrt{x^T([X']^TX')^{-1}A([X']^TX')^{-1}x} = |y - x^T\hat{\theta}'|\sqrt{x^T(X^TX)^{-1}A(X^TX)^{-1}x}. \tag{5.4}$$

is clearly proportional to prediction error. In addition, when we choose $A \approx X^TX$, the second term becomes either $\mu := x^T([X,x]^T[X,x])^{-1}x$ or $\mu' := x^T(X^TX)^{-1}x$, which are "in-sample" and "out-of-sample" *statistical leverage scores* of x. Leverage score measures the importance/uniqueness of a data point relative to the rest of the data set and it is used extensively in regression analysis [53] (for outlier detection and experiment design), compressed sensing (for adaptive sampling) [208] and numerical linear algebra (for fast matrix computation)[73]. For the best of our knowledge, this is the first time leverage scores are shown to be connected to differential privacy.

5.4 Case study: The adaptivity of OPS in ridge regression

So far we have described output perturbation algorithms with a fixed noise adding procedure. However in practice it is not known ahead of time how to choose A. Assume all x are normalized to $\|x\| = 1$, denote $\mu_2(x) := x^T(X^TX)^{-2}x$, $\mu_1(x) := x^T(X^TX)^{-1}x$. We discuss the pros and cons of the three natural choices.

- $A \approx \lambda_{\min}I$: This corresponds to the standard ℓ_2-sensitivity and it adds an isotropic noise and provides a uniform guarantee for all data-target pairs where X^TX has smallest eigenvalue $\lambda_{\min}$, because $\sup_x \sqrt{\mu_2(x)} \leq 1/\lambda_{\min}$, but it adds more noise than necessary for those with much smaller $\mu_2(x)$.
- $A \approx (X^TX)^2$: We call this the "democratic" choice conditioned on the data set, as it homogenizes the "leverage" part of the per-instance sensitivity of points to $\|x\| = 1$ so any x gets about the same level of privacy. It however is not robust to if our data-independent choice of A is in fact far away from the actual $(X^TX)^2$.

- $A \approx X^T X$: We call this the "Fisher"-choice, because the covariance matrix will be proportional to the inverse Fisher information, which is the natural estimation error of $\hat{\theta}$ under the linear regression assumption. The advantage of this choice is that conducting statistical inference, e.g., t-test and ANOVA for linear regression coefficients would be trivial.

In fact, for linear and ridge regression, the third choice is closely related to the one-posterior-sampling (OPS) mechanism proposed in [64, 245] with an important difference being that in OPS, A is not fixed, but rather depends on the data. As a result, Lemma 5.4 does not work. In fact, if the data-target can be arbitrary and $r = 0$, the data-independent choice of A could imply an unbounded ϵ (consider an arbitrarily near singular X and x in its null space).

Indeed, existing analysis of OPS requires additional assumption. [245] assumes that the loss function is bounded (by modifying it or constraining the domain Θ) so that the exponential mechanism [156] would apply. It was later pointed out in [96] that OPS is asymptotically inefficient in that it has an asymptotic relative efficiency (ARE) inversely proportional to ϵ, while simple sufficient statistics perturbation can achieve asymptotic efficiency comparable to [201]. This is far from satisfactory.

Based on insight from pDP, we propose a direct analysis of OPS which reveals that if (ϵ, δ)-DP is all we need, then OPS is also asymptotically efficient under the same data assumption in [96]. In addition, it effectively converges to the "Fisher"-choice of noise adding in the same asymptotic regime and offers dimension and condition number independent expected pDP loss.

The first result calculates the pDP loss of OPS.

Theorem 5.15 (The adaptivity of OPS in Linear/Ridge Regression)**.** *Consider the algorithm that samples from*

$$p(\theta|X, \mathbf{y}) \propto e^{-\frac{\gamma}{2}\left(\|\mathbf{y}-X\theta\|^2+\lambda\|\theta\|^2\right)}.$$

Let $\hat{\theta}$ and $\hat{\theta}'$ be the ridge regression estimate with data set $X \times \mathbf{y}$ and $[X, x] \times [\mathbf{y}, y]$ and defined the out of sample leverage score $\mu := x^T(X^TX + \lambda I)^{-1}x = x^T H^{-1} x$ and in-sample leverage score $\mu' := x^T[(X')^T X' + \lambda I]^{-1}x = x^T(H')^{-1}x$. Then for every $\delta > 0$, privacy target (x, y), the algorithm is (ϵ, δ)-pDP with

$$\epsilon(Z, z) \leq \frac{1}{2}\left|-\log(1+\mu) + \frac{\gamma\mu}{(1+\mu)}(y - x^T\hat{\theta})^2\right| + \frac{\mu}{2}\log(2/\delta) + \sqrt{\gamma\mu\log(2/\delta)}|y - x^T\hat{\theta}| \tag{5.5}$$

$$= \frac{1}{2}\left|-\log(1-\mu') - \frac{\gamma\mu'}{1-\mu'}(y - x^T\hat{\theta}')^2\right| + \frac{\mu'}{2}\log(2/\delta) + \sqrt{\gamma\mu'\log(2/\delta)}|y - x^T\hat{\theta}'|. \tag{5.6}$$

The two equivalent upper bounds are both useful. (5.5) is ideal for calculating pDP when x is not in the data set and (5.6) is perfect for the case when x is in the data set.

Remark 5.16. *The bound* (5.5) *can be simplified to*

$$\frac{\mu}{2}(1 + \log(2/\delta)) + \frac{1}{2}\gamma\min(\mu, 1)|y - x^T\hat{\theta}'|^2 + \sqrt{\gamma\mu\log(2/\delta)}|y - x^T\hat{\theta}'|.$$

If $\mu = o(\log(2/\delta))^{1}$ and we choose γ such that $\sqrt{\gamma\mu' \log(2/\delta)}|y - x^T\hat{\theta}'| \leq 1$, then the bound can be simplified to

$$\epsilon(Z, z) \leq 2\sqrt{\gamma\mu \log(2/\delta)}|y - x^T\hat{\theta}| + o(1).$$

This matches the order of Gaussian mechanism with a fixed (data-independent) covariance matrix.

The results in [96] are stated for general exponential family models under a set of assumptions that translate into the following for linear regression:

(a) data $x_1, ..., x_n$ is drawn i.i.d. from $\mathcal{D}$ supported on $\mathcal{X}$ where $\mathcal{X} \subset \mathcal{B}_{\|\cdot\|_2}(1)$.

(b) population covariance matrix $\frac{m}{d}I \preceq \mathbb{E}_{\mathcal{D}}xx^T \preceq \frac{M}{d}I$ for constant m and M,

(c) $y_i \sim \mathcal{N}(x_i^T\theta_0, \sigma^2)$ for some θ_0.

To simplify the presentation, we also assume n scales with respect to d such that

(d) with high probability, $XX^T \succ \frac{mn}{2d}I$.

The last assumption measures how quickly the empirical covariance matrix $\frac{1}{n}XX^T$ concentrates to $\mathbb{E}_{x\sim\mathcal{D}}xx^T$. It has been shown that if X is an appropriately scaled subgaussian random matrix, this happens with probability $1 - n^{-10}$ whenever $n > \max(10d, 10d^{-2/3}\log n)$.

Proposition 5.17. *The sequence of OPS algorithm with parameter γ_n, λ_n obeys the following properties.*

1. ***pDP and DP in agnostic setting.*** *Assume $\|x\| \leq 1$ for every $x \in \mathcal{X}$. The algorithm obeys (ϵ_n, δ)-pDP, for each data set $(X, \mathbf{y})$ and target (x, y),*

$$\epsilon_n = \sqrt{\frac{\gamma_n \log(2/\delta)}{\lambda_n + \lambda_{\min}}}|y - x^T\hat{\theta}| + \frac{\gamma_n|y - x^T\hat{\theta}|^2}{2\max\{\lambda_n + \lambda_{\min}, 1\}} + \frac{\gamma_n(1 + \log(2/\delta))}{2(\lambda_n + \lambda_{\min})}. \tag{5.7}$$

If we further assume $|y| < 1$, then $\sup_{(X,\mathbf{y}),(x,y)} |y - x^T\hat{\theta}| = 1 + n^{1/2}\lambda_n^{-1/2}$ and the algorithm obeys (ϵ_n, δ)-DP with

$$\epsilon_n = \sqrt{\frac{2(n + \lambda_n)\gamma_n \log(2/\delta)}{\lambda_n^2}} + \frac{2(n + \lambda_n)\gamma_n}{\lambda_n \max\{1, \lambda_n\}} + \frac{\gamma_n(1 + \log(2/\delta))}{2\lambda_n}. \tag{5.8}$$

2. ***pDP under model assumption.*** *Assume conditions (a)(b)(c)(d) above are true, and also $\gamma_n = \omega(1)$, $\lambda_n = o(\sqrt{n})$. Then with high probability over the joint distribution of $(X, \mathbf{y})$, the algorithm with $\gamma_n \leq \frac{4n\log(2/\delta)}{\max\{d, (1+\log(2/\delta))^2\}}$ obeys (ϵ_n, δ)-pDP with*

$$\epsilon_n = \begin{cases} O\left(\sqrt{\frac{(1+\|\theta_0\|)^2 d\gamma_n}{mn}} \log(\frac{2}{\delta})\right) & \textit{for all (x,y) satisfying } \|x\| = O(1) \textit{ and } y = O(1). \\ O\left(\sqrt{\frac{\sigma^2 d\gamma_n}{mn} \log(\frac{2}{\delta}) \log(\frac{2}{\delta'})}\right) & \textit{for any } x \in \mathcal{X} \textit{ with probability } 1 - \delta' \textit{ over } y \sim \mathcal{N}(\theta_0^T x, \sigma^2). \end{cases}$$

[1] This is not an unrealistic assumption because μ and μ' are $o(1)$ as long as x is bounded and the minimum eigenvalue of $X^TX + \lambda I$ is $\omega(1)$. This is required for (agnostic) linear regression to be consistent and is implied by the condition that the the population covariance matrix $\frac{1}{n}\mathbb{E}X^TX$ is full rank.

Moreover, with probability $1 - n\delta'$ over the conditional distribution $\mathbf{y}|X$, the privacy loss of $(x_1, y_1), ..., (x_n, y_n)$ obeys

$$\frac{1}{n}\sum_{i=1}^{n} \epsilon_n\big((X,\mathbf{y}),(x_i,y_i)\big)^2 = O\Big(\frac{\sigma^2 d\gamma_n}{n}\log(2/\delta)\log(2/\delta')\Big),$$

which does not depend on m — the smallest eigenvalue of dX^TX/n.

3. ***Statistical efficiency.*** *for every realization of data set X such that $n > d$ and let the smallest eigenvalue of X^TX be $\lambda_{\min}$, then*

$$\mathbb{E}_{\mathbf{y}\sim\mathcal{N}(X\theta_0,\sigma^2 I_n)}\left[\|\tilde{\theta}-\theta_0\|^2\middle|X\right] = \sigma^2\mathrm{tr}[(X^TX+\lambda_n I)^{-1}](1+\gamma_n^{-1})+\lambda_n^2\|(X^TX+\lambda_n I)^{-1}\theta_0\|^2$$

If $\lambda_{\min} = \Omega(d/n)$ (this is true with high probability under assumption (b)(d)) Then we get

$$\mathbb{E}_{\mathbf{y}\sim\mathcal{N}(X\theta_0,\sigma^2 I_n)}\left[\|\tilde{\theta}-\theta_0\|^2\middle|X\right] = \sigma^2\mathrm{tr}[(X^TX+\lambda_n I)^{-1}](1+\gamma_n^{-1})+O(\frac{\lambda_n^2 d^2\|\theta_0\|^2}{n^2})$$

In other word, the estimator is asymptotically efficient, for all $\lambda_n = o(n^{1/2})$ and $\gamma_n = \omega(1)$.

We now discuss a few aspects of the above results.

pDP vs DP in agnostic setting. Firstly, it highlights the key advantage of pDP over DP. DP is not able to take advantage of desirable structures in the data set, while pDP provides a principled framework to handle them.

In particular, let us compare the pDP and DP in the agnostic setting, for the OPS that uses the same randomization. DP measures something that is completely data independent, and corresponds specifically to a contrivedly constructed data set $(X, \mathbf{y})$ such that $\mathbf{y}$ is an eigenvector of XX^T corresponding to a specific eigenvalue of magnitude $\sqrt{\lambda_n}$, this makes $\|\hat{\theta}\|_2$ as large as $\sqrt{n}/\sqrt{\lambda_n}$. Moreover, a target data point is chosen so that x match the direction of $\hat{\theta}$. While this is a legitimate construction in theory, but it does not directly correspond to the specific data set that a statistician just spent two years collecting, and it is unreasonable that he/she will have to calibrate the amount of noise to inject to provide more reasonable protection to a pathological case that has nothing to do with the reality.

pDP on the other hand, makes it possible for the statistician to condition on the data set. If the statistician finds out that $\|\hat{\theta}\|_2 = O(1)$, then the pDP loss is as small as $\sqrt{\gamma_n \log(2/\delta)/\lambda_n}$ for *everyone* in the population. With $\gamma_n = n^{\alpha/2}$ and $\lambda_n = n^{1/2-\alpha/2}$ for any $\alpha > 0$, the algorithm remains to be statistically efficient with an ARE of $(1 + n^{-\alpha})$ yet can provide a strong privacy guarantee of $\epsilon_n = n^{-1/4+\alpha/2}$. If in addition, the statistician realized that the data set is *well-conditioned*, that is, the maximum and minimum eigenvalue of X^TX are on the same order of n/d, then we can further improve the bound by replacing λ_n with $\lambda_{\min} + \lambda_n$. The statistician can happily get away with the same privacy guarantee ($\epsilon_n = n^{-1/4}$) while not having to add too much noise or even regularize at all (setting $\gamma_n = n^{1/2}$ and $\lambda_n = 0$). Note that the condition number is a desirable property that governs how reliably one can hope to estimate the linear regression coefficients using the given data set.

We would like to emphasize that the pDP guarantee in the wo cases we discussed above applies to everyone in the population $\{(x,y)|\|x\| \leq 1, |y| \leq 1\}$, therefore such (ϵ,δ)-pDP guarantee is as powerful as (ϵ,δ)-DP after the data set is collected.

pDP of all vs average pDP on the data set. Secondly, unlike DP which always provides a crude upper bound for everyone, pDP is able to reflect the differences in the protection of different target person. Under the model assumption, the average privacy loss of people in the data set, is scale-invariant and interestingly, also independent to the condition number (smallest eigenvalue). It is a factor of $(1+|\theta_0|)^2/m$ times smaller than the pDP guarantee for everyone in the population. This is significant for finite sample performance since $(1+\|\theta_0\|)/m$ (although they do not change with n), can be quite large.

pDP under covariate shift. Lastly, if we consider a setting in between the above two, where the target x can be drawn from any distribution defined on $\mathcal{X}$ that could be arbitrarily different from the training data distribution, then the scale-invariant property remains (the factor of $(1+\|\theta_0\|)$ is dropped). This is relevant in causal learning when the $\mathbb{E}(y|x)$ is specified by some physical principles that's invariant to the distribution of x. In this case, the moments of the pDP would imply a much stronger notion of cross-domain generalization than what we show in Proposition 5.11 since it does not depend on target covariate distribution of interest.

Improved DP guarantee for OPS. The proposition also improves the existing analysis for the OPS algorithm as a byproduct. The first statement shows that OPS preserves a meaningful (almost constant) differential privacy when $\gamma_n = 1$ and $\lambda_n = \sqrt{n/d}$ without requiring a constant boundedness in the domain Θ or clipping the loss function like in Wang et al. [245]. As a matter of fact, the ridge regression solution $\hat{\theta}$ could be in a ball of radius $\Theta(n^{1/4})$, and even if we impose the smallest domain bound that covers $\hat{\theta}$, by exponential mechanism, the algorithm only obeys a pure $O(n^{1/2})$-DP, in contrast to the $(O(\log(1/\delta)),\delta)$-DP that we showed in the proposition above.

Despite the improvement, the DP guarantee is still a little unsatisfactory. If we require (ϵ,δ)-DP with constant ϵ, then the OPS algorithm with $\lambda_n = \sqrt{n}$ is not asymptotically efficient (although it does achieve the optimal $O(1/n)$ rate).

Meanwhile, there are algorithms that attain asymptotic efficiency either by subsample-and-aggregate [201] or by simply adding noise to the sufficient statistics [82, 96].

So the question becomes: can we modify OPS such that it become asymptotically efficient with (ϵ_n,δ)-differentially private with $\epsilon_n = o(1)$?

We address this issue in the next section.

5.5 Statistical efficient linear regression with differential privacy using AdaOPS

In this section, we resolve the dilemma described earlier using the idea of Dwork and Lei [79]. The new algorithm, which we call AdaOPS , adaptively and differentially privately chooses the tuning parameter λ_n and γ_n according to properties of the data set and privacy requirement. A pseudocode of AdaOPS is given in Algorithm 6. We acknowledge that the same idea of adaptively adding regularization term is not new and had been used by Blocki et al. [35], Kifer et al. [128], Sheffet [197] for analyzing other related differentially private algorithms. Our contribution here is only to assemble the ideas together into a working algorithm.

Algorithm 6 AdaOPS : One-Posterior Sample estimator with adaptive regularization

```
input  Data $X, \mathbf{y}$. Privacy target: $\epsilon$, $\delta$. And parameter $\kappa$ satisfying $\kappa \leq \frac{n\epsilon}{4d(1+\log(4/\delta))}$
  1. Calculate the minimum eigenvalue $\lambda_{\min}(X^TX)$.
  2. Private release $\tilde{\lambda}_{\min} = \lambda_{\min} + \frac{\sqrt{\log(4/\delta)}}{\epsilon/2}Z$, where $Z \sim \mathcal{N}(0,1)$.
  3. Get one sample
         $\tilde{\theta} \sim \mathbb{P}(\theta|X,\mathbf{y}) \propto e^{-\frac{\gamma_n}{2}\left(\|\mathbf{y}-X\theta\|^2+\lambda_n\|\theta\|^2\right)}$
     with parameter
         $\lambda_n = \min\left\{0, \frac{n}{d\kappa} - \tilde{\lambda}_{\min} + \frac{\log(4/\delta)}{\epsilon/2}\right\}$
         $\gamma_n = \min\left\{\frac{n\epsilon^2}{16\kappa^2 d^2\log(4/\delta)}, \frac{n\epsilon}{8\kappa^2d^2}\right\}$
output  $\tilde{\theta}$
```

The κ parameter is the largest acceptable condition number in the data set. Often it can be determined independent to the data. It is used in the algorithm to rule out the pathological case. We now analyze the properties of AdaOPS .

Proposition 5.18. *1. Assume data domain is $\|x\|_2 \leq 1$ and $|y| \leq 1$. The* AdaOPS *estimator preserves (ϵ,δ)-DP.*

2. If assumption (a)(b)(c) is true and in addition for the specific realization of X,

$$\lambda_{\min}(X^TX) > \frac{n}{\kappa d} + \frac{\sqrt{\log(10n)\log(4/\delta)}}{\epsilon/2}$$

(which is true with high probability if $m > 2/\kappa$ and X is an appropriately scaled subgaussian random matrix), then, we have

$$\mathbb{E}[\|\tilde{\theta}-\theta_0\|^2|X] = [1+\gamma_n]\,\sigma^2\mathrm{tr}[(X^TX)^{-1}] + O(n^{-10})\|\theta_0\|^2.$$

In other word, since $\gamma_n \leq \min\{\frac{\kappa^2d^2\log(4/\delta)}{n\epsilon^2}, \frac{\kappa^2d^2}{n\epsilon}\}$, the AdaOPS *estimator achieves asymptotic efficiency whenever ϵ obeys that $\min\{n\epsilon^2, s\epsilon\} = o(\kappa^2d^2\log(4/\delta)/n)$.*

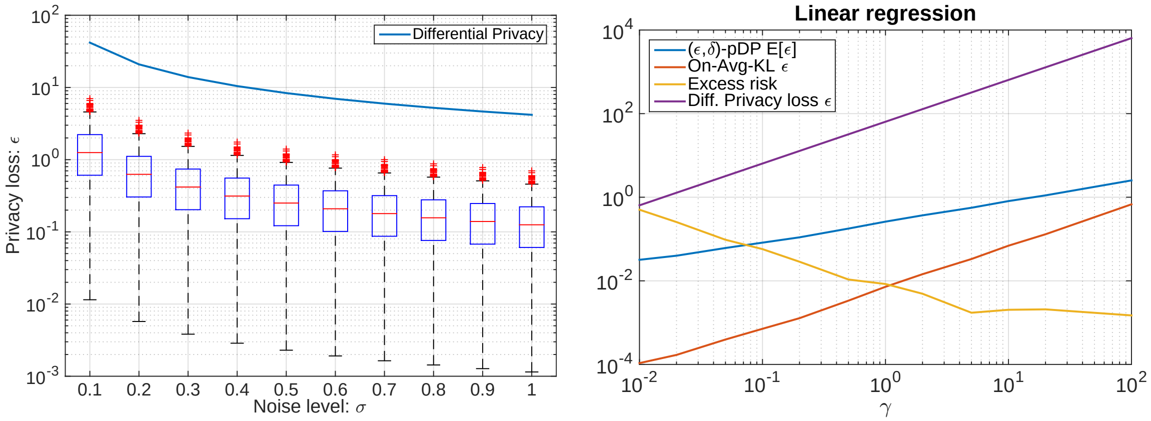


Figure 5.1: **Left:** (ϵ, δ)-DP and distribution of $(\epsilon(z, Z), \delta)$-pDP data points in linear regression with isotropic Gaussian noise adding. **Right:** Comparing the pDP privacy loss to the ϵ-DP obtained through exponential mechanism [245] using the same posterior sampling algorithm. In both experiment $\delta = 1e - 6$.

This proposition reveals that AdaOPS improves over previous results in the literature [96, 201] in several ways. First of all, we only need $n\epsilon^2 = o(1)$ to achieve asymptotic efficiency. In contrast, [96] does not provide non-asymptotic results with explicit dependence and [201]'s bound for the subsample-and-aggregate method requires $n^{-1/5}\epsilon^{-6/5} = o(1)$ to achieve asymptotic efficiency.

Secondly, both [96] and [201] suffer from additional polynomial dependence on the dimension in the supposedly lower order term, and that affects finite sample performance. While we haven't yet solved the problem, we managed to slightly improve the dimension dependence. In particular, our bound on the additive difference from exactly matching the Cramer-Rao lower bound of $(\sigma^2 \mathrm{tr}((X^T X)^{-1}))$ translates into $\sigma^2 \mathrm{tr}[(X^T X)^{-1}] + d^3/(n^2\epsilon^2)$. In contrast, there is a clear dependence of $d^5/n^2\epsilon^2$ in the sufficient statistics perturbation approach[2] and even higher polynomial dependence in the subsample-and-aggregate approach.

We conclude the section with two simulate experiments (shown in the two panes of Figure 5.1). In the first experiment, we consider the algorithm of adding isotropic Gaussian noise to linear regression coefficients, and then compare the worst-case DP and the distribution of per-instance DP for points in the data set (illustrated as box plots). In the second experiment, we compare different notions of privacy to utility (measured as excess risk) of the fixed algorithm that samples from a scaled posterior distribution. In both cases, the average per-instance differential privacy over the data sets is several orders of magnitude smaller than the worst-case differential privacy.

[2] This comes from the iid noise with standard deviation proportional to $\sqrt{d^2 \log(2/\delta)}/\epsilon$ added to $X^T X$ to obtain (ϵ, δ)-DP, which results in a random noise matrix E_1 with average eigenvalue not smaller than $d^{1.5}$. By Lemma 5.20, we see that the difference of the OLS solution on the noisy sufficient statistics $\tilde{\theta}$ and that on the true sufficient statistics $\hat{\theta}$ is $(X^T X + E_1)^{-1}(E_2 - E_1)$, which if we assume $\|\hat{\theta}\|_2$ and a well-conditioned X, gives rise to an MSE on the order of $\sigma^2 \mathrm{tr}[(X^T X)^{-1}] + O(d^5/n^2\epsilon^2)$.

5.6 Conclusion

In this chapter, we proposed to use per-instance differential privacy (pDP) for quantifying the fine-grained privacy loss of a fixed individual against randomized data analysis conducted on a fixed data set. We analyzed its properties and showed that pDP is proportional to well-studied quantities, e.g., leverage scores, residual and pseudo-residual in statistics and statistical learning theory. This formalizes the intuitive idea that the more one can "blend into the crowd" like a chameleon, the more privacy one gets; and that the better a model fits the data, the easier it is to learn the model differentially privately. Moreover, the new notion allows us to conduct statistical learning and inference and take advantage of desirable structures of the data sets to gain orders-of-magnitude more favorable privacy guarantee than the worst case. This makes it highly practical in applications.

Specifically, we conducted a detailed case-study on linear regression to illustrate how pDP can be used. The pDP analysis allows us to identify and account for key properties of the data set, like the well-conditionedness of the feature matrix and the magnitude of the fitted coefficient vector, thereby provides strong uniform differential privacy coverage to everyone in the population whenever such structures exist. As a byproduct, the analysis also leads to an improved differential privacy guarantee for the OPS algorithm [64, 245] and also a new algorithm called ADAOPS that adaptively chooses the regularization parameters and improves the guarantee further. In particular, ADAOPS achieves asymptotic statistical efficiency and differential privacy at the same time with stronger parameters than known before.

The introduction of pDP also raises many open questions for future research. First of all, how do we tell individuals what their ϵs and δs of pDP are? This is tricky because the pDP loss itself is a function of the data, thus needs to be privatized against possible malicious dummy users. Secondly, the problem gets substantially more interesting when we start to consider the economics of private data collection. For instance, what happens if what we tell the individuals would affect their decision on whether they will participate in the data set? In fact, it is unclear how to provide an estimation of pDP in the first place if we are not sure what would the data be at the end of the day. Thirdly, from the data collector's point of view, the data is going to be "easier" and the model will have a better "goodness-of-fit" on the collected data, but that will be falsely so to some extent, due to the bias incurred during data collection according to pDP. How do we correct for such bias and estimate the real performance of a model on the population of interest? Addressing these problems thoroughly would require the joint effort of the community and we hope the exposition in this chapter will encourage researchers to play with pDP in both theory and practical applications.

5.7 Proofs of technical results

Proof of Proposition 5.9. We first show that implies on-average stability and then on-average stability implies on-average generalization.

Let $Z' = [Z, z']$, $Z'' = [Z, z'']$ and fix z. We first prove stability. Let $S = \theta | p(\theta) \geq p'(\theta)$

$$\begin{aligned}
&\left|\mathbb{E}_{\theta\sim\mathcal{A}(Z')}\ell(\theta, z) - \mathbb{E}_{\theta\sim\mathcal{A}(Z'')}\ell(\theta, z)\right| \\
=&\sup_{\theta,z} \ell(\theta, z)[P_{Z'}(\theta \in S) - P_{Z''}(\theta \in S)] \\
\leq& e^{\epsilon(Z,z')}P_Z(\theta \in S) + \delta((Z, z')) - P_{Z''}(\theta \in S) \\
\leq&(e^{\epsilon(Z,z')+\epsilon(Z,z'')} - 1)P_{Z''}(\theta \in S) + \delta(Z, z') + \epsilon(Z, z')\delta(Z, z'') \\
\leq&(e^{\epsilon(Z,z')+\epsilon(Z,z'')} - 1) + \delta(Z, z') + \epsilon(Z, z')\delta(Z, z'')
\end{aligned}$$

Note that the bound is independent to z.

Now we will show stability implies generalization using a "ghost sample" trick in which we resample $Z' \sim \mathcal{D}^n$ and construct $Z^{(i)}$ by replacing the ith data point from the ith data point of Z'.

$$\begin{aligned}
&\left|\mathbb{E}_{Z\sim\mathcal{D}^n}\left(\mathbb{E}_{\theta\sim\mathcal{A}(Z)}\mathbb{E}_{z\sim\mathcal{D}}\ell(\theta, z) - \mathbb{E}_{\theta\sim\mathcal{A}(Z)}\frac{1}{n}\sum_{i=1}^{n}\ell(\theta, z_i)\right)\right| \\
=&\left|\mathbb{E}_{Z\sim\mathcal{D}^n,\{z'_1,\ldots,z'_n\}\sim\mathcal{D}^n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\theta\sim\mathcal{A}(Z)}\ell(\theta, z'_i) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\theta\sim\mathcal{A}(Z^{(i)})}\ell(\theta, z'_i)\right)\right| \\
\leq&\mathbb{E}_{Z\sim\mathcal{D}^n,\{z'_1,\ldots,z'_n\}\sim\mathcal{D}^n}\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}_{\theta\sim\mathcal{A}(Z)}\ell(\theta, z'_i) - \mathbb{E}_{\theta\sim\mathcal{A}(Z^{(i)})}\ell(\theta, z'_i)\right| \\
\leq&\mathbb{E}_{Z\sim\mathcal{D}^{n-1},\{z',z''\}\sim\mathcal{D}^2}[(e^{\epsilon(Z,z')+\epsilon(Z,z'')} - 1) + \delta(Z, z') + \epsilon(Z, z')\delta(Z, z'')]
\end{aligned}$$

The last step simply substitutes the stability bound. Take expectation on both sides, we get a generalization upper bound of form:

$$\xi = \mathbb{E}_{Z\sim\mathcal{D}^n}(\mathbb{E}_{z\sim\mathcal{D}}[e^{\epsilon(Z,z)}|Z])^2 - 1 + \mathbb{E}_{Z\sim\mathcal{D}^n,z\sim\mathcal{D}}\delta(Z, z) + (\mathbb{E}_{Z\sim\mathcal{D}^n}\mathbb{E}_{z\sim\mathcal{D}}[e^{\epsilon(Z,z)}|Z]\mathbb{E}_{z\sim\mathcal{D}}[\delta(Z, z)|Z].$$

□

Proof of Proposition 5.11. The stability argument remains the same, because it is applied to a fixed pair of (Z, z). We will modify the ghost sample arguments with and additional change of measure.

$$
\begin{aligned}
&\left|\mathbb{E}_{Z\sim\mathcal{D}^n}\left(\mathbb{E}_{\theta\sim\mathcal{A}(Z)}\mathbb{E}_{z\sim\mathcal{D}'}\ell(\theta,z)-\mathbb{E}_{\theta\sim\mathcal{A}(Z)}\frac{1}{n}\sum_{i=1}^{n}\rho(z_i)\ell(\theta,z_i)\right)\right|\\
=&\left|\mathbb{E}_{Z\sim\mathcal{D}^n}\left(\mathbb{E}_{\theta\sim\mathcal{A}(Z)}\mathbb{E}_{z\sim\mathcal{D}}\rho(z)\ell(\theta,z)-\mathbb{E}_{\theta\sim\mathcal{A}(Z)}\frac{1}{n}\sum_{i=1}^{n}\rho(z_i)\ell(\theta,z_i)\right)\right|\\
=&\left|\mathbb{E}_{Z\sim\mathcal{D}^n,\{z'_1,\ldots,z'_n\}\sim\mathcal{D}^n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\theta\sim\mathcal{A}(Z)}\rho(z'_i)\ell(\theta,z'_i)-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\theta\sim\mathcal{A}(Z^{(i)})}\rho(z'_i)\ell(\theta,z'_i)\right)\right|\\
\leq&\mathbb{E}_{Z\sim\mathcal{D}^n,\{z'_1,\ldots,z'_n\}\sim\mathcal{D}^n}\frac{1}{n}\sum_{i=1}^{n}\rho(z'_i)\left|\mathbb{E}_{\theta\sim\mathcal{A}(Z)}\ell(\theta,z'_i)-\mathbb{E}_{\theta\sim\mathcal{A}(Z^{(i)})}\ell(\theta,z'_i)\right|\\
\leq&\mathbb{E}_{Z\sim\mathcal{D}^{n-1},\{z',z''\}\sim\mathcal{D}^2}\rho(z'')[(e^{\epsilon(Z,z')+\epsilon(Z,z'')}-1)+\delta(Z,z')+\epsilon(Z,z')\delta(Z,z'')]\\
=&\mathbb{E}_{Z\sim\mathcal{D}^{n-1},z'\sim\mathcal{D},z''\sim\mathcal{D}'}[(e^{\epsilon(Z,z')+\epsilon(Z,z'')}-1)+\delta(Z,z')+\epsilon(Z,z')\delta(Z,z'')].
\end{aligned}
$$

□

Proof of Corollary 5.12.

$$
\mathbb{E}\left[\mathbb{E}_{\mathcal{D}}[e^{\epsilon(Z,z)}|Z]\mathbb{E}_{\mathcal{D}'}[e^{\epsilon(Z,z)}|Z]\right]-1+\delta(1+\mathbb{E}[e^{\epsilon}(Z,z)])\leq\sqrt{\mathbb{E}_{\mathcal{D}}e^{2\epsilon}\mathbb{E}_{\mathcal{D}'}e^{2\epsilon}}-1+2\delta.
$$

The inequality uses Jensen's inequality $\mathbb{E}\left[\mathbb{E}[e^{\epsilon(Z,z)}|Z]^2\right]\leq\mathbb{E}e^{2\epsilon(Z,z)}$ and the monotonicity of moment generating function on non-negative random variables. The statement is obtained by Taylor's series on $\mathbb{E}e^{2\epsilon(Z,z)}$. Lastly, we use the algebraic mean to upper bound the geometric mean in the first term and then use Taylor expansion. □

Proof of Lemma 5.14. By the stationarity of $\hat{\theta}$

$$
\sum_i\nabla\ell(\hat{\theta},z_i)+\nabla r(\hat{\theta})=0
$$

Add and subtract $\ell(\hat{\theta},z)$ and apply first order Taylor's Theorem centered at $\hat{\theta}'$ on $\sum_i\nabla\ell(\hat{\theta},z_i)+\nabla\ell(\hat{\theta},z_i)+\nabla r(\hat{\theta})$, we get

$$
\sum_i\nabla\ell(\hat{\theta}',z_i)+\nabla\ell(\hat{\theta}',z_i)+\nabla r(\hat{\theta}')+R-\nabla\ell(\hat{\theta},z)=0.
$$

where if we define $\eta_t=(1-t)\hat{\theta}'+t\hat{\theta}$, the remainder term $R\in\mathbb{R}^d$ can be explicitly written as

$$
R=\left[\int_0^1\left(\sum_i\nabla^2\ell(\eta_t,z_i)+\nabla^2\ell(\eta_t,z)+\nabla^2r(\eta_t)\right)dt\right](\hat{\theta}-\hat{\theta}').
$$

By the mean value theorem for Frechet differentiable functions, we know there is a t such that we can take η_t such that the integrand is equal to the integral.

Since $\hat{\theta}'$ is a stationary point, we have

$$\sum_i \nabla \ell(\hat{\theta}', z_i) + \nabla \ell(\hat{\theta}', z) + \nabla r(\hat{\theta}') = 0$$

and thus under the assumption that $\left[\int_0^1 \left(\sum_i \nabla^2 \ell(\eta_t, z_i) + \nabla^2 \ell(\eta_t, z) + \nabla^2 r(\eta_t)\right) dt\right]$ is invertible, we have

$$\hat{\theta} - \hat{\theta}' = \left[\int_0^1 \left(\sum_i \nabla^2 \ell(\eta_t, z_i) + \nabla^2 \ell(\eta_t, z) + \nabla^2 r(\eta_t)\right) dt\right]^{-1} \nabla \ell(\hat{\theta}, z).$$

The other equality follows by symmetry. □

Proof of Theorem 5.15. Let $X' = [X; x]$, $\mathbf{y}' = [\mathbf{y}; y]$. Denote $H := X^T X + \lambda I$, $H' := (X')^T X' + \lambda I$, $g := X^T \mathbf{y}$ and $g' := (X')^T \mathbf{y}'$. Correspondingly, the posterior mean $\hat{\theta} = H^{-1} g$ and $\hat{\theta}' = [H']^{-1} g'$.

The covariance matrix of the two distributions are H/γ and H'/γ. Using the fact that the normalization constant is known for Gaussian, the log-likelihood ratio at output θ is

$$\begin{aligned}&\log \frac{|H^{-1}|^{-1/2} e^{-\frac{\gamma}{2}\|\theta - \hat{\theta}\|_H^2}}{|[H']^{-1}|^{-1/2} e^{-\frac{\gamma}{2}\|\theta - \hat{\theta}'\|_{H'}^2}} \\ =&\underbrace{\log \sqrt{\frac{|H|}{|H'|}}}_{(\#)} + \frac{\gamma}{2} \underbrace{\left[\|\theta - \hat{\theta}'\|_{H'}^2 - \|\theta - \hat{\theta}\|_H^2\right]}_{(*)}.\end{aligned}$$

Note that $H' = H + xx^T$. By Lemma 5.22,

$$\frac{|H|}{|H'|} = \frac{|H|}{|H|(1+\mu)} = \frac{|H'|(1-\mu')}{|H'|},$$

so

$$(\#) = \log \sqrt{(1+\mu)^{-1}} = \log \sqrt{1-\mu'}.$$

The second term in the above equation can be expanded into

$$\begin{aligned}(*) =&\theta^T [H' - H]\theta + (\hat{\theta}')^T H' \hat{\theta}' - \hat{\theta}^T H \hat{\theta} - 2(\hat{\theta}')^T H' \theta + 2\hat{\theta}^T H \theta \\ =&(x^T\theta)^2 + \underbrace{(\mathbf{y}')^T X' [H']^{-1} X'^T \mathbf{y}' - \mathbf{y}^T X (H)^{-1} X^T \mathbf{y}}_{(**)} - 2y(x^T\theta)\end{aligned} \tag{5.9}$$

$$(**) = \left[(\mathbf{y}')^T X' [(X')^T X' + \lambda I]^{-1} X'^T \mathbf{y}' - \mathbf{y}^T X (X^T X + \lambda I)^{-1} X^T \mathbf{y}\right] = \left[(\mathbf{y}')^T \Pi' \mathbf{y}' - \mathbf{y}^T \Pi \mathbf{y}\right],$$

where we denote the“hat” matrices $\Pi := X(X^TX + \lambda I)^{-1}X^T$ and $\Pi' = X'[(X')^TX' + \lambda I]^{-1}(X')^T$. Also define $v := X(X^TX + \lambda I)^{-1}x$. By Sherman-Morrison-Woodbury formula, we can write

$$\begin{aligned}\Pi' &= \begin{bmatrix} X \\ x^T \end{bmatrix} \left[H^{-1} - H^{-1}x(1+\mu)^{-1}x^TH^{-1}\right] \begin{bmatrix} X^T & x \end{bmatrix} \\ &= \begin{bmatrix} \Pi - (1+\mu)^{-1}vv^T, & v - \mu(1+\mu)^{-1}v \\ v^T - v^T(1+\mu)^{-1}\mu, & \mu - \mu^2(1+\mu)^{-1} \end{bmatrix}\end{aligned}$$

Note that $v^Ty = x^T\hat{\theta}$ and $1 - \mu(1+\mu)^{-1} = (1+\mu)^{-1}$, therefore

$$\begin{aligned}(**) &= -(1+\mu)^{-1}(x^T\hat{\theta})^2 + 2(1+\mu)^{-1}x^T\hat{\theta} + \mu(1+\mu)^{-1}y^2 \\ &= -(1+\mu)^{-1}(y - x^T\hat{\theta})^2 + y^2.\end{aligned}$$

Substitute into (5.9), we get

$$(*) = (y - x^T\theta)^2 - (1+\mu)^{-1}(y - x^T\hat{\theta})^2.$$

And the $\log$-probability ratio is

$$\begin{aligned}\log\frac{p(\theta|X,\mathbf{y})}{p(\theta|X',\mathbf{y}')} &= \log\sqrt{(1+\mu)^{-1}} + \frac{\gamma}{2}\left[(y - x^T\theta)^2 - (1+\mu)^{-1}(y - x^T\hat{\theta})^2\right] \\ &= \log\sqrt{(1+\mu)^{-1}} + \frac{\gamma}{2}\left[(x^T\hat{\theta} - x^T\theta)^2 + 2(y - x^T\hat{\theta})(x^T\hat{\theta} - x^T\theta) + \frac{\mu}{1+\mu}(y - x^T\hat{\theta})^2\right]\end{aligned}$$

Under the distribution of θ when the data is $(X, \mathbf{y})$, $x^T\theta - x^T\hat{\theta}$ follows a univariate normal distribution with mean 0 and variance μ/γ. By the standard tail probability of normal random variable,

$$\mathbb{P}\left(|x^T\theta - x^T\hat{\theta}| > \sqrt{\frac{\mu}{\gamma}\log(2/\delta)}\right) \leq \frac{2e^{-\log(2/\delta)}}{\log(2/\delta)} = \frac{\delta}{\log(2/\delta)} \underset{\substack{\uparrow \\ \text{When } \delta < 2/e}}{\leq} \delta.$$

we can calculate (ϵ, δ)-pDP for every $\delta > 0$. In particular, under $p(\theta|X, y)$

$$\mathbb{P}(\left|\log\frac{p(\theta|X,\mathbf{y})}{p(\theta|X',\mathbf{y}')}\right| \geq \epsilon) < \delta$$

for

$$\epsilon = \frac{1}{2}\left|-\log(1+\mu) + \frac{\mu\gamma}{(1+\mu)}(y - x^T\hat{\theta})^2\right| + \frac{\mu}{2}\log(2/\delta) + |y - x^T\hat{\theta}|\sqrt{\mu\gamma\log(2/\delta)}.$$

By Lemma 5.25 this implies (ϵ, δ)-DP.

Now, we will work out an equivalent representation of the $\log$-probability ratio that depends on $\hat{\theta}'$.

Let μ' be the in-sample leverage score of x with respect to X', namely, $\mu' := x^T[H']^{-1}x$. By Sherman-Morrison-Woodbury formula

$$H^{-1} = [H' - xx^T]^{-1} = [H']^{-1} + [H']^{-1}x(1-\mu')^{-1}x^T[H']^{-1}.. \qquad (5.10)$$

Standard matrix algebra gives us

$$\begin{aligned}\mathbf{y}^T\Pi\mathbf{y} &= (\mathbf{y}')^T X' H^{-1}(X')^T\mathbf{y}' - yx^T H^{-1}xy - 2yx^T H^{-1}X^T\mathbf{y}\\ &= (\mathbf{y}')^T X' H^{-1}(X')^T\mathbf{y}' - 2yx^T H^{-1}(X')^T\mathbf{y}' + yx^T H^{-1}xy.\end{aligned}$$

Substitute (5.10) into the above, we get

$$\begin{aligned}\mathbf{y}^T\Pi\mathbf{y} &= (\mathbf{y}')^T\Pi'\mathbf{y}' + (1-\mu')^{-1}(x^T\hat\theta')^2 - 2yx^T\hat\theta'\left[1+\mu'(1-\mu')^{-1}\right] + y^2\mu' + y^2(\mu')^2(1-\mu')^{-1}\\ &= (\mathbf{y}')^T\Pi'\mathbf{y}' + (1-\mu')^{-1}(x^T\hat\theta')^2 - 2yx^T\hat\theta'(1-\mu')^{-1} + y^2(1-\mu')^{-1} - y^2\end{aligned}$$

Therefore,

$$(**) = -(y - x^T\hat\theta')^2(1-\mu')^{-1} + y^2,$$

and

$$(*) = (y - x^T\theta)^2 - (1-\mu')^{-1}(y - x^T\hat\theta')^2.$$

The corresponding log-probability ratio

$$\begin{aligned}\log\frac{p(\theta|X,\mathbf{y})}{p(\theta|X',\mathbf{y}')} &= -\log(\sqrt{1-\mu'}) + \frac{\gamma}{2}\left[(y-x^T\theta)^2 - (1-\mu')^{-1}(y-x^T\hat\theta')^2\right]\\ &= -\log(\sqrt{1-\mu'}) + \frac{\gamma}{2}\left[(x^T\hat\theta' - x^T\theta)^2 + 2(x^T\hat\theta' - x^T\theta)(y - x^T\hat\theta') - \frac{\mu'}{1-\mu'}(y-x^T\hat\theta')^2\right]\end{aligned}$$

Under the posterior distribution of $(X', \mathbf{y}')$, the mean of $x^T\theta$ is centered at $x^T\hat\theta'$ with variance μ'/γ. We can then derive a tail bound of the privacy loss random variable and it implies an $(\epsilon,\delta) - pDP$ guarantee by Lemma 5.25. Specifically, it implies that the method is (ϵ,δ)-pDP with

$$\epsilon = \frac{1}{2}\left|-\log(1-\mu') - \frac{\gamma\mu'}{1-\mu'}(y-x^T\hat\theta')^2\right| + \frac{\mu'}{2}\log(2/\delta) + \sqrt{\gamma\mu'\log(2/\delta)}|y - x^T\hat\theta'|.$$

This complete the second statement of the proof. □

Proof of Proposition 5.17. The proof mostly involves applying Theorem 5.15 and substituting bounds over either a bounded domain assumption (typical for DP analysis), or a model assumption of how data are generated (typical for statistical analysis).

Proof of Statement 1 in the agnostic setting. For any $x \in \mathcal{X}$, and any data set X, using the choice of regularization term, we can bound $\mu = 1/\lambda_n$. Substitute that into Theorem 5.15, and use the inequality that $\log(1+x) \leq x$ we get the first expression.

Now, restricting ourselves to the bounded domain. Under the choice of λ_n, we can choose an X,$\mathbf{y}$ with a singular value equal to $\sqrt{\lambda_n}$ and the corresponding singular vector $v \in \{-1,1\}^n$ such that the following upper bounds are attained

$$\|(X^TX + \lambda_n I)^{-1}X^T\| \leq \frac{1}{2\sqrt{\lambda_n}}.$$

$$\|\hat{\theta}\| \leq \|(X^TX + \lambda_n I)^{-1}X^T\|\|y\| \leq \frac{\sqrt{n}}{2\sqrt{\lambda_n}}.$$

Now choose (x, y) such that $|x^T\hat{\theta}| = \|x\|\|\hat{\theta}\|$, we get that $\sup_{(X,\mathbf{y}),(x,y)} |y - \hat{\theta}^T x| = 1 + \frac{\sqrt{n}}{2\sqrt{\lambda_n}}$. The DP claim follows by substituting the upper bound into the pDP's expression.

Proof of Statement 2 under the model assumption. To prove the second claim, note that by Assumption (b)(d), the smallest eigenvalue of X^TX is lower bounded by d/nm. Also under the model assumption, the ridge regression estimator concentrates around θ_0.

In particular, under the model assumption, the ridge regression estimate

$$\hat{\theta} = (X^TX + \lambda_n I)^{-1}X^Ty = (X^TX + \lambda_n I)^{-1}X^TX\theta_0 + (X^TX + \lambda_n I)^{-1}X^TZ \tag{5.11}$$

With high probability over the distribution of Z

$$\|\hat{\theta} - \theta_0\|^2 = O(\frac{d\sigma^2 \log(n)}{n} + \frac{\lambda_n^2 d^2 \|\theta_0\|^2}{n^2}),$$

thus for all (x, y) satisfying $\|x\| \leq 1$ $y \leq 1$, we get

$$|y - x^T\hat{\theta}| \leq |y - x^T\theta_0| + |x^T(\hat{\theta} - \theta_0)| = O(1 + \|\theta_0\|).$$

Under the assumption that $n > 10d\log n$, $\|\theta_0\| = O(1)$ and $\sigma = O(1)$ this is effectively a constant.

For $x \in \mathcal{X}$ and $y \sim \mathcal{N}(\theta_0^T x, \sigma^2)$, using standard Gaussian tail bound, with high probability the perturbation is bounded, therefore $|y - x^T\theta_0|^2 \leq \sigma^2 \log(2/\delta')$.

Lastly, for the case of the average pDP loss over the empirical data distribution. Besides taking into the above bound on $|y - x^T\hat{\theta}|$, we further consider adding the different parts over the distributions. Since this is to deal with data points in the data set, we will instantiate the bound (5.6). Our assumption on γ_n, λ_n ensures that the dominant term is the third term, thus

$$\frac{1}{n}\sum_{i=1}^{n} \epsilon_n((X, \mathbf{t}), (x_i, y_i))^2 \leq C\gamma_n \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\hat{\theta}')^2 x_i^T(X^TX + \lambda_n I)^{-1}x_i.$$

Under the high probability event that the noise is bounded by $\sigma\sqrt{2/\delta'}$ for all data points, we can extract them out then note that

$$\frac{1}{n}\sum_{i=1}^{n} x_i^T(X^TX)^{-1}x_i = \frac{1}{n}\mathrm{tr}(\sum_{i=1}^{n} x_i x_i^T(X^TX + \lambda_n I)) \leq \frac{1}{n}\mathrm{tr}(I) = \frac{d}{n}.$$

Substitute these bounds into Theorem 5.15, and we obtain the Statement 2.

Proof of Statement 3 under the model assumption. By (5.11) and the fact that OPS can be thought of as adding an independent multivariate Gaussian noise with covariance matrix $(X^TX + \lambda_n I)^{-1}X^TX(X^TX + \lambda_n I)^{-1}/\gamma_n$, we get

$$\tilde{\theta} = (X^TX + \lambda_n I)^{-1}X^TX\theta_0 + \sqrt{1+\gamma_n}(X^TX + \lambda_n I)^{-1}X^TZ.$$

By a bias-variance decomposition, we get

$$\begin{aligned}
&\mathbb{E}(\|\tilde{\theta} - \theta_0\|_2^2|X) = \mathrm{Var}(\tilde{\theta}|X) + \|\mathbb{E}\tilde{\theta} - \theta_0\|^2\\
=&(1+\gamma_n^{-1})\sigma^2\mathrm{tr}\left[(X^TX + \lambda_n I)^{-1}X^TX(X^TX + \lambda_n I)^{-1}\right] + \left\|\left[I - (X^TX + \lambda_n I)^{-1}X^TX\right]\theta_0\right\|^2\\
=&(1+\gamma_n^{-1})\sigma^2\sum_{i=1}^d \frac{\sigma_i^2}{(\sigma_i^2 + \lambda_n)^2} + \lambda_n^2\theta_0^T(X^TX + \lambda_n I)^{-2}\theta_0\\
\leq&(1+\gamma_n^{-1})\sigma^2\mathrm{tr}(X^TX + \lambda_n I)^{-1} + \lambda_n^2 m^{-2}n^{-2}\|\theta_0\|^2
\end{aligned}$$

The proof is complete by substitute the values of γ_n and λ_n into the inequality and noting that $m = \Omega(1)$ and under the model assumption $\|\theta_0\|$ does not grow with n. Clearly, if $\gamma_n = \omega(1)$ and $\lambda_n = o(\sqrt{n})$, then the algorithm is asymptotically efficient. □

Proof of Proposition 5.18. We will first prove the claim on differential privacy and then analyze the statistical efficiency.

Proof of differential privacy. First of all, by Weyl's theorem, and the assumption that $\|xx^T\|_2 \leq 1$, we get that the global sensitivity of $\lambda_{\min}(X^TX)$ is 1. We will use $\lambda_{\min}$ as the short hand of $\lambda_{\min}(X^TX)$ in the rest of the proof. So releasing $\tilde{\lambda}_{\min}$ is $(\epsilon/2, \delta/2)$-DP using the standard Gaussian mechanism. Secondly, under the same event with probability at least $1 - \delta/2$, we have

$$\lambda_{\min} - \frac{\log(4/\delta)}{\epsilon} \leq \tilde{\lambda}_{\min} \leq \lambda_{\min} + \frac{\log(4/\delta)}{\epsilon}.$$

Therefore, by our selection rule of the regularization parameter λ_n,

$$\frac{n}{d\kappa} \leq \lambda_{\min}(X^TX + \lambda_n I) \leq \max\{\lambda_{\min}, \frac{n}{d\kappa} + \frac{\log(4/\delta)}{\epsilon}\}.$$

The lower bound implies that for any (x, y) satisfying the condition, the out of sample leverage score

$$\mu = x^T(X^TX + \lambda_n I)^{-1}x \leq \frac{\kappa d}{n}. \tag{5.12}$$

It also implies an upper bound on the prediction error:

$$|y - x^T\hat{\theta}| \leq 1 + \|\hat{\theta}\| \leq 1 + \|(X^TX + \lambda_n I)^{-1}X^T\|_2\|y\|_2 \leq \min\sqrt{2d\kappa}. \tag{5.13}$$

We will prove the final inequality above using the following lemma with $h = n/d\kappa$ and then note that $\|y\|_2 \leq \sqrt{n}$.

Lemma 5.19. *For any matrix X, and any $\lambda \geq 0$. If $\lambda_{\min}(X^TX + \lambda I) \geq h$, then*

$$\|(X^TX + \lambda I)^{-1}X^T\| \leq \sqrt{2/h}.$$

The proof is technical so we defer it to later.

Now combine (5.12)(5.13) with Theorem 5.15, we get that the OPS step which obeys $(\tilde{\epsilon}, \delta/2)$-pDP with

$$\begin{aligned}\tilde{\epsilon}((X, \mathbf{y}), (x, y)) &\leq \frac{\mu}{2}(1 + \log(4/\delta)) + \frac{1}{2}\gamma_n \min(\mu, 1)(y - x^T\hat{\theta})^2 + \sqrt{\gamma\mu\log(4/\delta)}|y - x^T\hat{\theta}| \\ &\leq \frac{\kappa d(1 + \log(4/\delta))}{2n} + \frac{\gamma_n}{2}\frac{\kappa d}{n}2\kappa d + \sqrt{\frac{\gamma_n}{2}\frac{\kappa d}{n}2\kappa d\log(4/\delta)} \\ &\leq \epsilon/8 + \epsilon/8 + \epsilon/4 \leq \epsilon/2\end{aligned}$$

Note that in the last step, we made use of the choice of γ_n and the condition that concerns ϵ and κ as stated in the algorithm. Since this upper bound holds for all data set $(X, \mathbf{y})$ and all privacy target (x, y). The OPS algorithm also satisfies $(\epsilon/2, \delta/2) - DP$.

The proof of the first claim is complete when we compose the two data access.

Proof of the statistical efficiency. Now we switch gear to analyze the estimation error bound. Let event E be the event that $\tilde{\lambda}_{\min} > \lambda_{\min} - \frac{\sqrt{10\log(n)}\sqrt{\log(4/\delta)}}{\epsilon/2}$, which happens with probability $1 - n^{-10}$. Under E, we have $\lambda_0 = 0$. By our assumption, this happens with

Applying the third claim in Proposition 5.17, we get that

$$\mathbb{E}[\|\tilde{\theta} - \theta_0\|^2|X, E] \leq (1 + \gamma_n^{-1})\sigma^2\mathrm{tr}((X^TX)^{-1}).$$

Under the small probability event E^c, we use a crude upper bound that takes the sum of the maximum square bias and maximum variance.

$$\mathbb{E}[\|\tilde{\theta} - \theta_0\|^2|X, E^c] \leq (1 + \gamma_n^{-1})\sigma^2\mathrm{tr}((X^TX)^{-1}) + \|\theta_0\|^2$$

by law of total expectation, for an event E

$$\begin{aligned}\mathbb{E}[\|\tilde{\theta} - \theta_0\|^2|X] &= \mathbb{E}[\|\tilde{\theta} - \theta_0\|^2|X, E]\mathbb{P}(E|X) + \mathbb{E}[\|\tilde{\theta} - \theta_0\|^2|X, E^c]\mathbb{P}(E^c|X) \\ &\leq (1 + \gamma_n^{-1})\sigma^2\mathrm{tr}((X^TX)^{-1}) + \mathbb{P}(E^c)\|\theta_0\|^2 = (1 + \gamma_n^{-1})\sigma^2\mathrm{tr}((X^TX)^{-1}) + O(n^{-10}).\end{aligned}$$

The proof is complete by substituting γ_n into the bound. □

Proof of Lemma 5.19. Take SVD of $X = U\Sigma V^T$, we can write

$$\|(X^TX + \lambda_n I)^{-1}X^T\|_2 = \max_{i \in [d]} \frac{\Sigma_{ii}}{\Sigma_{ii}^2 + \lambda_n}$$

We now discuss two cases. First, for those $i \in [d]$ such that $\Sigma_{ii}^2 \leq \lambda_n$. In this case, adding λ_n on both sides ensures that

$$h \leq \lambda_{\min}(X^T X + \lambda_n I) = \lambda_{\min} + \lambda_n \leq \Sigma_{ii}^2 + \lambda_n \leq 2\lambda_n.$$

and therefore if $\Sigma_{ii} > 0$

$$\frac{\Sigma_{ii}}{\Sigma_{ii}^2 + \lambda_n} = \frac{1}{\Sigma_{ii} + \lambda_n / \Sigma_{ii}} \leq \frac{1}{2\sqrt{\lambda_n}} \leq \sqrt{1/(2h)}. \tag{5.14}$$

The final inequality is also true for $\Sigma_{ii} = 0$. If on the other hand, for those $i \in [d]$ such that, $\Sigma_{ii}^2 > \lambda_n$. This time by adding Σ_{ii}^2 on both sides, we get

$$2\Sigma_{ii}^2 > \lambda_n + \Sigma_{ii}^2 \geq \lambda_n + \lambda_{\min} = \lambda_{\min}(X^T X + \lambda_n I) \geq \frac{n}{\kappa}.$$

This implies that

$$\frac{\Sigma_{ii}}{\Sigma_{ii}^2 + \lambda_n} \leq \frac{1}{\Sigma_{ii}} \leq \sqrt{2/h}. \tag{5.15}$$

Combine (5.14) and (5.15) we get

$$\|(X^T X + \lambda_n I)^{-1} X^T\|_2 \leq \sqrt{2/h}$$

□

5.8 Technical lemmas

Lemma 5.20. *Let* $\hat{\theta}' = (X^T X + E_1)^{-1}(X\mathbf{y} + E_2)$ *for any matrix* E_1, E_2.

$$\hat{\theta}' - \hat{\theta} = (X^T X + E_1)^{-1}(E_2 - E_1\hat{\theta})$$

Proof.

$$\begin{aligned}
\hat{\theta}' =& (X^T X + E_1)^{-1}(X^T\mathbf{y} + E_2) \\
=& (X^T X + E_1)^{-1}(X^T X)(X^T X)^{-1} X^T \mathbf{y} + (X^T X + E_1)^{-1} E_2 \\
=& \hat{\theta} + \left[(X^T X + E_1)^{-1}(X^T X + E_1) - (X^T X + E_1)^{-1} E_1 - I_d\right] \hat{\theta} + (X^T X + E_1)^{-1} E_2 \\
=& \hat{\theta} - (X^T X + E_1)^{-1} E_1 \hat{\theta} + (X^T X + E_1)^{-1} E_2 \\
=& \hat{\theta} + (X^T X + E_1)^{-1}(E_2 - E_1\hat{\theta})
\end{aligned}$$

□

Lemma 5.21 (Sherman-Morrison-Woodbury Formula)**.** *Let* A, U, C, V *be matrices of compatible size, assume* A, C *and* $C^{-1} + VA^{-1}U$ *are all invertible, then*

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}A^{-1}.$$

Lemma 5.22 (Determinant of Rank-1 perturbation)**.** *For invertible matrix A and vector c, d of compatible dimension*

$$\det(A + cd^T) = \det(A)(1 + d^T A^{-1} c).$$

Lemma 5.23 (Weyl's eigenvalue bound [212, Theorem 1])**.** *Let $X, Y, E \in \mathbb{R}^{m \times n}$, w.l.o.g., $m \geq n$. If $X - Y = E$, then $|\sigma_i(X) - \sigma_i(Y)| \leq \|E\|_2$ for all $i = 1, ..., n$.*

Lemma 5.24 (Gaussian tail bound)**.** *Let $X \sim \mathcal{N}(0, 1)$. Then*

$$\mathbb{P}(|X| > \epsilon) \leq \frac{2e^{-\epsilon^2/2}}{\epsilon}.$$

Lemma 5.25 (Tail bound to (ϵ, δ)-DP conversion)**.** *Let $\epsilon(\theta) = \log(\frac{p(\theta)}{p'(\theta)})$ where p and p' are densities of θ. If*

$$\mathbb{P}(|\epsilon(\theta)| > t) \leq \delta$$

then for any measurable set $\mathcal{S}$

$$\mathbb{P}_p(\theta \in \mathcal{S}) \leq e^t \mathbb{P}_{p'}(\theta \in \mathcal{S}) + \delta.$$

and

$$\mathbb{P}_{p'}(\theta \in \mathcal{S}) \leq e^t \mathbb{P}_p(\theta \in \mathcal{S}) + \delta$$

Proof. Since $\log(\frac{p(\theta)}{p'(\theta)}) = -\log(\frac{p'(\theta)}{p(\theta)})$ and the tail bound is two-sided. It suffices for us to prove just one direction. Let E be the event that $|\epsilon(\theta)| > t$.

$$\mathbb{P}_p(\theta \in \mathcal{S}) = \mathbb{P}_p(\theta \in \mathcal{S} \cup E^c) + \mathbb{P}_p(\theta \in \mathcal{S} \cup E) \leq \mathbb{P}_{p'}(\theta \in \mathcal{S} \cup E) e^t + \mathbb{P}_p(\theta \in E) \leq e^t \mathbb{P}_{p'}(\theta \in \mathcal{S}) + \delta.$$

□

Lemma 5.26 (Matrix Hoeffding inequality [151])**.** *Consider a finite sequence $X_1, ..., X_n$ of independent random and self-adjoint matrices with dimension d and $A_1, ..., A_n$ be a sequence of fixed self-adjoint matrices. In addition, let $\mathbb{E} X_i = 0$ and $X_i^2 \preceq A_i^2$ almost surely for all $i = 1, ..., n$. Then, for all $t \geq 0$*

$$\mathbb{P}\left\{\lambda_{\max}(\sum_{i=1}^n X_i) \geq t\right\} \leq d e^{-t^2/2\sigma^2}$$

where $\sigma^2 \leq \| \sum_{i=1}^n A_i^2 \|$.

Subsequent work and applications

In Part I, we presented a coherent sequence of work on differentially private machine learning with contributions in both theory and practice. In this section, we briefly mention a few connections to subsequent work in the literature.

We will mostly discuss our work on the posterior sampling algorithm (OPS). It has drawn substantial practical interest mainly because it is one of the first few cases that demonstrated the success of differentially private algorithms in real-life machine learning problems; but also because of the fact that it can be efficiently implemented using stochastic gradient based MCMC methods [4, 56, 65, 250], which is an interesting area of study on its own right. Specifically, our paper [147] describes an application of the SGLD approach to privately learn a recommendation system using the matrix factorization model; and Wang et al. [243] compared the approach with a number of other methods on the problem of differentially private subspace clustering. These examples suggest that the posterior sampling can be very attractive in practice and it often achieves a reasonable level of privacy protection while not substantially affecting the model accuracy.

After our work was published, there had been a lot of interest in refining and extending the work, and coming up with various applications. Notably, Minami et al. [157] provided the (ϵ, δ)-DP of the same method, under convex and Lipschitz assumptions, Park et al. [170] used a similar subsample-and-composition approach (and a more modern way of analyzing advanced composition), to do variational inference graphical models with latent variables; Foulds et al. [96] addresses the issue of asymptotic inefficiency of our proposed approaches and used the sufficient statistics perturbation approach on the Bayesian learning problems. It allows more samples to be drawn and is asymptotically efficient, and to some extent, the work of Foulds et al. [96] motivated our refined analysis of OPS presented in Chapter 5, which showed that in more restricted setting, OPS can be made asymptotically efficient, while preserving approximate differential privacy.

We now mention two general future directions.

Practical differential privacy. As of writing, differential privacy has been slowly shifting from a purely theoretical object into a practical technology. However, the bulk of the existing literature however focuses on advancing the theory, and very few applications have been seen in the industry beyond locally differentially private data collection. There is a huge void that calls for both practical algorithms, implementation and appropriate empirical benchmarking.

Closer connections to statistics. As we have seen in the work presented in this thesis, differential privacy is very closely related to statistics and machine learning. We find it very important to shift from a privacy-centric model to an algorithm-centric model that is more utility-aware and allows assumptions on the data set to be useful to both utility and privacy. Statistics also have a role to play in coming up with sensible post-processing approaches that “denoise” the differential private releases, and in addressing the question of which ϵ to use in practice.

Part II

Towards locally adaptive nonparametric machine learning

Motivation and overview

Nonparametric regression has a rich history in statistics, carrying well over 50 years of associated literature. It solves the following fundamental problem:

- Let $y = f(x) +$ Noise. How to estimate function f using data points $(x_1, y_1), ..., (x_n, y_n)$ in conjunction with the knowledge that f belongs to some function class $\mathcal{F}$?

Nonparametric regression is closely related to the denoising problem in the signal processing community, and has seen numerous applications in science, economics and medicine.

At its heart, nonparametric regression is motivated by the need to learn from data without making strong (parametric) model assumptions. This is very similar to what motivated statistical machine learning in that both aims at using very expressive function classes, but they differ in assumptions and objective. Machine learning opted for minimizing the regret (excess risk) relative to a weaker oracle (the best within class) so that the results hold uniformly over all data distributions, while nonparametric regression aims at making weak assumptions (e.g., smooth function classes) on data distributions but at the end of the day, its success is measured by the extent to which we can estimate the true underlying functions.

We consider a recent and successful class of nonparametric regression technique called trend filtering [129, 209, 221, 244], that was shown to have properties of *local adaptivity* in both theory and practice. The term "local adaptivity" will reappear many times in this chapter. We say a nonparametric regression technique is locally adaptive if it is able to cater to local differences in smoothness, hence allowing more effective estimation of function classes that contain functions with heterogeneous smoothness. Examples of local adaptive nonparametric regression techniques include wavelet shrinkage [70], smoothing kernels with Lepski's method [145] and locally adaptive regression splines [152]. These methods, while well-developed and often generically applicable, either do not work well in finite sample or computationally expensive. Trend filtering, on the other hand, seems to be able to sidestep these issues, at least in the limited cases where one knows how to do it — smoothing univariate functions.

This part of the thesis collects a sequence of work that expands the scope of trend filtering, so that it can be used as a generic and widely applicable tools for a variety of ML problems, e.g., those that come with multivariate features variables, spatiotemporal variations, or those with signals observed on a non-Euclidean structures, such as vertices of a graph.

Before describing the content of each chapter, we first build intuition by reviewing how univariate trend filtering works. For simplicity, assume $x_1, ..., x_n \in \mathbb{R}$ are evenly spaced, say $(x_1, ..., x_n) = (1, ..., n)$, and $y_1, ..., y_n$ are observations at location $x_1, ..., x_n$. Given an integer $k \geq 0$, the kth order univariate trend filtering estimate $\hat{\theta} = (\hat{\theta}_1, \ldots \hat{\theta}_n)$ is defined as

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \ \frac{1}{2}\|y - \theta\|_2^2 + \lambda \|D^{(k+1)}\theta\|_1,$$

where $\lambda \geq 0$ is a tuning parameter, and $D^{(k+1)}$ is the discrete difference operator of order $k+1$. When $k = 0$, the problem above employs the first difference operator,

$$D^{(1)} = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 \\ 0 & -1 & 1 & \ldots & 0 \\ \vdots & & \ddots & \ddots & \\ 0 & 0 & \ldots & -1 & 1 \end{bmatrix}.$$

Therefore $\|D^{(1)}\theta\|_1 = \sum_{i=1}^{n-1} |\theta_{i+1} - \theta_i|$, and the 0th order trend filtering estimate reduces to the 1-dimensional fused lasso estimator [220], also called 1-dimensional total variation denoising [182]. For $k \geq 1$ the operator $D^{(k+1)}$ is defined recursively by

$$D^{(k+1)} = D^{(1)} D^{(k)},$$

with $D^{(1)}$ above denoting the $(n-k-1) \times (n-k)$ version of the first difference operator above. In words, $D^{(k+1)}$ is given by taking first differences of kth differences. The interpretation is hence that univariate trend filtering penalizes the changes in the kth discrete differences of the fitted trend. The estimated components $\hat{\theta}_1, \ldots \hat{\theta}_n$ exhibit the form of a kth order piecewise polynomial function, evaluated over the input locations $x_1, \ldots x_n$.

Intuitively, the ℓ_1 norm induces sparsity in the discrete $(k+1)$th order derivatives, which corresponds to the number of pieces in the piecewise polynomials estimates. The fact that the none-zero locations are chosen automatically suggests that the algorithm is able to automatically allocate more "parameters" to locations that are intrinsically more complex than other locations. The use of ℓ_1-norm is not surprising given the insight from high-dimensional statistics and compressed sensing [46, 219], however the use of finite difference estimators, the continuous extension to splines and its computational benefits are rather intriguing and delicate.

The generalization of univariate trend filtering in the subsequent chapters often involve substituting a different regularization operator that mimics a kth order "derivatives" over the observations in each respective settings.

In Chapter 6, we develop the theory and algorithms for trend filtering on signals observed on a graph. We do so by recursively specifying a sequence of regularization operators using incidence matrices and graph Laplacian matrices. These regularizers naturally induce piecewise "polynomials" on graphs. Applications to image-denoising, graph-based semi-supervised learning as well as event detection on NYC taxi data suggest that there are substantial benefits of using trend filtering over traditional approaches.

It is however, unclear whether graph trend filtering is optimal for the class of problems it tries to solve. In Chapter 7, we partially answer this question for the important special image/video

TV-denoising problems on d-dimensional grid graphs. This turns out to be a very important open problem of nonparametric regression, where the problem in 1D is completely studied in [70]. It is worth mentioning that the corresponding regularization operator used in this case is also the ℓ_1 norm of all the discrete partial derivatives of a multivariate functions defined on $[0, n^{1/d}]^d$.

In Chapter 8, we switch gears to discuss continuous extensions of the fitted discrete vectors in 1D trend filtering. We propose a spline-like basis function called the falling factorial basis, which illustrates, from an alternative point of view, why trend filtering can be solved very efficiently even in cases with nonuniform observations. The falling factorial basis can be interpreted as an "inverse" of a regularization operator that resembles the "divided difference" approximation of the discrete derivatives.

Lastly, we note that the presented results are part of a bigger agenda to develop and complete the picture of trend filtering as a full-fledged nonparametric regression methods. In Section 8, we will briefly mention other published and ongoing work that aims at (a) understanding the minimax rate of the trend filtering problems in all dimension and all order; (b) interpolating signals on grids and on graphs to continuous domain. We will also mention a few interesting applications of trend filtering, graph trend filtering.

Chapter 6

Trend filtering on Graphs

We first consider the problem of nonparametric estimation on graphs, which can be used to describe supervised/unsupervised and semisupervised learning problems through neighborhood graph embedding of the data points in possibly high-dimensional space. Graph embeddings can be used to effectively capture smooth low-dimensional manifold structures of the data, according to [25, 239].

Specifically, we introduce a family of adaptive estimators on graphs, based on penalizing the ℓ_1 norm of discrete graph differences. This generalizes the idea of trend filtering [129, 221], used for univariate nonparametric regression, to graphs. Analogous to the univariate case, graph trend filtering exhibits a level of local adaptivity unmatched by the usual ℓ_2-based graph smoothers. It is also defined by a convex minimization problem that is readily solved (e.g., by fast ADMM or Newton algorithms). We demonstrate the merits of graph trend filtering through both examples and theory.

6.1 Introduction

The goal of this chapter is to port a successful idea in univariate nonparametric regression, trend filtering [129, 209, 221, 244], to the setting of estimation on graphs. The proposed estimator, graph trend filtering, shares three key properties of trend filtering in the univariate setting.

1. **Local adaptivity:** graph trend filtering can adapt to inhomogeneity in the level of smoothness of an observed signal across nodes. This stands in contrast to the usual ℓ_2-based methods, e.g., Laplacian regularization [203], which enforce smoothness globally with a much heavier hand, and tends to yield estimates that are either smooth or else wiggly throughout.
2. **Computational efficiency:** graph trend filtering is defined by a regularized least squares problem, in which the penalty term is nonsmooth, but convex and structured enough to permit efficient large-scale computation.

3. **Analysis regularization:** the graph trend filtering problem directly penalizes (possibly higher order) differences in the fitted signal across nodes. Therefore graph trend filtering falls into what is called the *analysis* framework for defining estimators. Alternatively, in the *synthesis* framework, we would first construct a suitable basis over the graph, and then regress the observed signal over this basis; e.g., Shuman et al. [198] survey a number of such approaches using wavelets; likewise, kernel methods regularize in terms of the eigenfunctions of the graph Laplacian [131]. An advantage of analysis regularization is that it easily yields complex extensions of the basic estimator by mixing penalties.

As a motivating example, consider a denoising problem on 402 census tracts of Allegheny County, PA, arranged into a graph with 402 vertices and 2382 edges obtained by connecting spatially adjacent tracts. To illustrate the adaptive property of graph trend filtering we generated an artificial signal with inhomogeneous smoothness across the nodes, and two sharp peaks near the center of the graph, as can be seen in the top left panel of Figure 6.1. (The signal was formed using a mixture of five Gaussians, in the underlying spatial coordinates.) We drew noisy observations around this signal, shown in the top right panel of the figure, and we fit graph trend filtering, graph Laplacian smoothing, and wavelet smoothing to these observations. Graph trend filtering is to be defined in Section 6.2 (here we used $k = 2$, quadratic order); the latter two, recall, are defined by the optimization problems

$$\min_{\theta \in \mathbb{R}^n} \|y - \theta\|_2^2 + \lambda \theta^\top L \theta \quad \text{(Laplacian smoothing)},$$
$$\min_{\theta \in \mathbb{R}^n} \frac{1}{2}\|y - W\theta\|_2^2 + \lambda \|\theta\|_1 \quad \text{(wavelet smoothing)},$$

where $y \in \mathbb{R}^n$ the vector of observations measured over the $n = 402$ nodes in the graph, $L \in \mathbb{R}^{n \times n}$ is the graph Laplacian matrix, and $W \in \mathbb{R}^{n \times n}$ is a wavelet basis built over the graph. The wavelet smoothing problem displayed above is really an oversimplified representation of the class of wavelets methods, since it only encapsulates estimators that employ an orthogonal wavelet basis W (and soft-threshold the wavelet coefficients). For the present experiment, we constructed W according to the spanning tree wavelet design of Sharpnack et al. [195]; we found this construction performed best among the graph wavelet designs we considered for the data at hand. For completeness, the results from alternative wavelet designs are given in Section 6.8.

Graph trend filtering, Laplacian smoothing, and wavelet smoothing each have their own regularization parameters λ, and these parameters are not generally on the same scale. Therefore, in our comparisons we use effective degrees of freedom (df) as a common measure for the complexities of the fitted models. The top right panel of Figure 6.1 shows the graph trend filtering estimate with 68 df. We see that it adaptively fits the sharp peaks in the center of the graph, and smooths out the surrounding regions appropriately. The graph Laplacian estimate with 68 df (bottom left), substantially oversmooths the high peaks in the center, while at 132 df (bottom middle), it begins to detect the high peaks in the center, but undersmooths neighboring regions. Wavelet smoothing performs quite poorly across all df values—it appears to be most affected by the level of noise in the observations.

As a more quantitative assessment, Figure 6.2 shows the mean squared errors between the estimates and the true underlying signal. The differences in performance here are analogous to

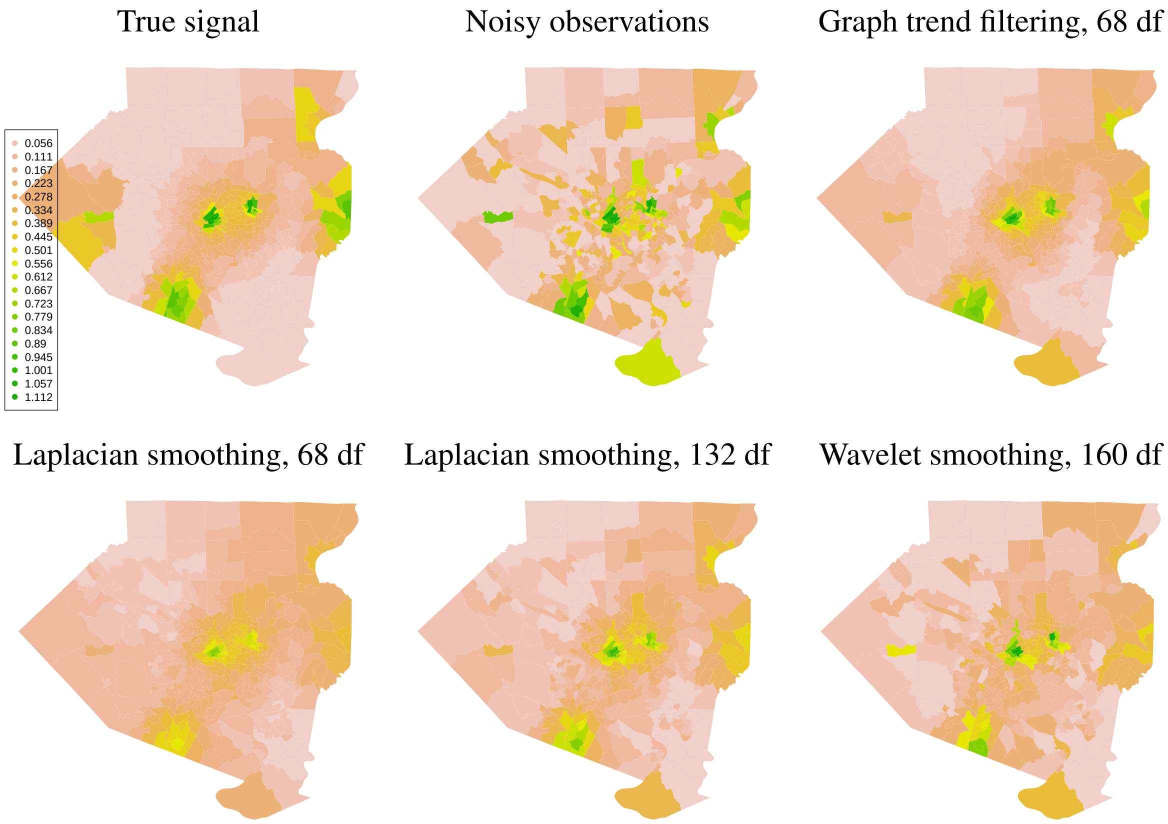


Figure 6.1: Color maps for the Allegheny County example.

the univariate case, when comparing trend filtering to smoothing splines [221]. At smaller df values, Laplacian smoothing, due to its global considerations, fails to adapt to local differences across nodes. Trend filtering performs much better at low df values, and yet it matches Laplacian smoothing when both are sufficiently complex, i.e., in the overfitting regime. This demonstrates that the local flexibility of trend filtering estimates is a key attribute.

Here is an outline for the rest of this article. Section 6.2 defines graph trend filtering and gives underlying motivation and intuition. Section 6.3 covers basic properties and extensions of the graph trend filtering estimator. Section 6.4 examines computational approaches, and Section 6.5 looks at a number of both real and simulated data examples. Section 6.6 presents asymptotic error bounds for graph trend filtering. Section 10.4 concludes with a discussion. As for notation, we write X_A to extract the rows of a matrix $X \in \mathbb{R}^{m\times n}$ that correspond to a subset $A \subseteq \{1,\dots m\}$, and X_{-A} to extract the complementary rows. We use a similar convention for vectors: x_A and x_{-A} denote the components of a vector $x \in \mathbb{R}^m$ that correspond to the set A and its complement, respectively. We write $\mathrm{row}(X)$ and $\mathrm{null}(X)$ for the row and null spaces of X, respectively, and $X^\dagger$ for the pseudoinverse of X, with $X^\dagger = (X^\top X)^\dagger X^\top$ when X is rectangular.

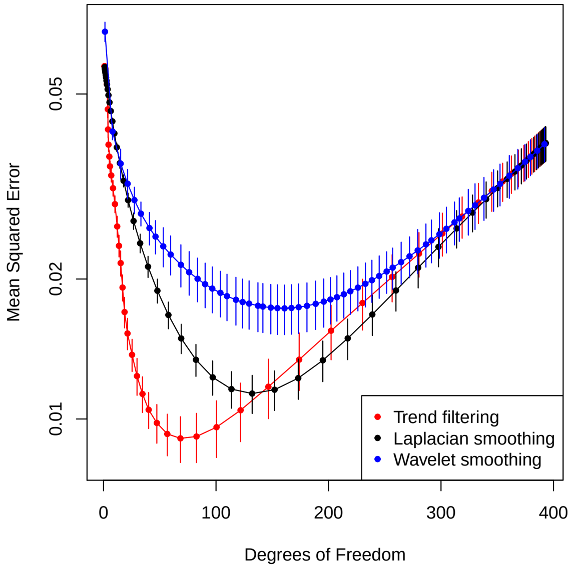


Figure 6.2: Mean squared errors for the Allegheny County example. Results were averaged over 10 simulations; the bars denote ±1 standard errors.

6.2 Trend Filtering on Graphs

In this section, we motivate and formally define graph trend filtering. As we defined in the "motivation and overview" of this part, the univariate trend filtering solves:

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^n}{\text{argmin}} \ \frac{1}{2}\|y - \theta\|_2^2 + \lambda \|D^{(k+1)}\theta\|_1, \tag{6.1}$$

where $D^{(k+1)}$ is the discrete difference operator of order $k+1$, that can be recursively constructed by

$$D^{(k+1)} = D^{(1)} D^{(k)}, \tag{6.2}$$

The estimated components $\hat{\theta}_1, \dots \hat{\theta}_n$ exhibit the form of a kth order piecewise polynomial function, evaluated over the input locations $x_1, \dots x_n$. This can be formally verified [221, 244] by examining a continuous-time analog of (8.13).

6.2.1 Higher order derivatives and trend filtering over a graph

Let $G = (V, E)$ be an graph, with vertices $V = \{1, \dots n\}$ and undirected edges $E = \{e_1, \dots e_m\}$, and suppose that we observe $y = (y_1, \dots y_n) \in \mathbb{R}^n$ over the nodes. Following the univariate definition in (8.13), we define the kth order *graph trend filtering* (GTF) estimate $\hat{\theta} = (\hat{\theta}_1, \dots \hat{\theta}_n)$ by

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^n}{\text{argmin}} \ \frac{1}{2}\|y - \theta\|_2^2 + \lambda \|\Delta^{(k+1)}\theta\|_1. \tag{6.3}$$

In broad terms, this problem (like univariate trend filtering) is a type of generalized lasso problem [222], in which the penalty matrix $\Delta^{(k+1)}$ is a suitably defined *graph difference operator*, of order $k+1$. In fact, the novelty in our proposal lies entirely within the definition of this operator.

When $k=0$, we define first order graph difference operator $\Delta^{(1)}$ in such a way it yields the graph-equivalent of a penalty on local differences:

$$\|\Delta^{(1)}\theta\|_1 = \sum_{(i,j)\in E} |\theta_i - \theta_j|.$$

so that the penalty term in (6.3) sums the absolute differences across connected nodes in G. To achieve this, we let $\Delta^{(1)} \in \{-1,0,1\}^{m\times n}$ be the oriented incidence matrix of the graph G, containing one row for each edge in the graph; specifically, if $e_\ell = (i,j)$, then $\Delta^{(1)}$ has ℓth row

$$\Delta^{(1)}_\ell = (0,\ldots \underset{\substack{\uparrow \\ i}}{-1},\ldots \underset{\substack{\uparrow \\ j}}{1},\ldots 0), \tag{6.4}$$

where the orientations of signs are arbitrary. Like trend filtering in the 1d setting, the 0th order graph trend filtering estimate coincides with the fused lasso (total variation regularized) estimate over G [114, 193, 222].

For $k \ge 1$, we use a recursion to define the higher order graph difference operators, in a manner similar to the univariate case. The recursion alternates in multiplying by the first difference operator $\Delta^{(1)}$ and its transpose (taking into account that this matrix not square):

$$\Delta^{(k+1)} = \begin{cases} (\Delta^{(1)})^\top \Delta^{(k)} = L^{\frac{k+1}{2}} & \text{for odd } k \\ \Delta^{(1)}\Delta^{(k)} = DL^{\frac{k}{2}} & \text{for even } k. \end{cases} \tag{6.5}$$

Above, we abbreviated the oriented incidence matrix $\Delta^{(1)}$ by D of G, and exploited the fact that $L = D^\top D$ is one representation for the graph Laplacian matrix. Note that $\Delta^{(k+1)} \in \mathbb{R}^{n\times n}$ for odd k, and $\Delta^{(k+1)} \in \mathbb{R}^{m\times n}$ for even k.

An important point is that our defined graph difference operators (6.4), (6.5) reduce to the univariate ones (8.9), (8.10) in the case of a chain graph (in which $V = \{1,\ldots n\}$ and $E = \{(i,i+1) : i = 1,\ldots n-1\}$), modulo boundary terms. That is, when k is even, if one removes the first $k/2$ rows and last $k/2$ rows of $\Delta^{(k+1)}$ for the chain graph, then one recovers $D^{(k+1)}$; when k is odd, if one removes the first and last $(k+1)/2$ rows of $\Delta^{(k+1)}$ for the chain graph, then one recovers $D^{(k+1)}$. Further intuition for our graph difference operators is given next.

6.2.2 Piecewise Polynomials over Graphs

We give some insight for our definition of graph difference operators (6.4), (6.5), based on the idea of piecewise polynomials over graphs. In the univariate case, sparsity of θ under the difference operator $D^{(k+1)}$ implies a specific kth order piecewise polynomial structure for the components of θ [221, 244]. Since the components of θ correspond to (real-valued) input locations

$x = (x_1, \dots x_n)$, the interpretation of a piecewise polynomial here is unambiguous. But for a graph, one might ask: does sparsity of $\Delta^{(k+1)}\theta$ mean that the components of θ are piecewise polynomial? And what does the latter even mean, as the components of θ are defined over the nodes? To address these questions, we intuitively *define* a piecewise polynomial over a graph, and show that it implies sparsity under our constructed graph difference operators.

- **Piecewise constant ($k = 0$):** we say that a signal θ is piecewise constant over a graph G if many of the differences $\theta_i - \theta_j$ are zero across edges $(i, j) \in E$ in G. Note that this is exactly the property associated with sparsity of $\Delta^{(1)}\theta$, since $\Delta^{(1)} = D$, the oriented incidence matrix of G.
- **Piecewise linear ($k = 1$):** we say that a signal θ has a piecewise linear structure over G if θ satisfies

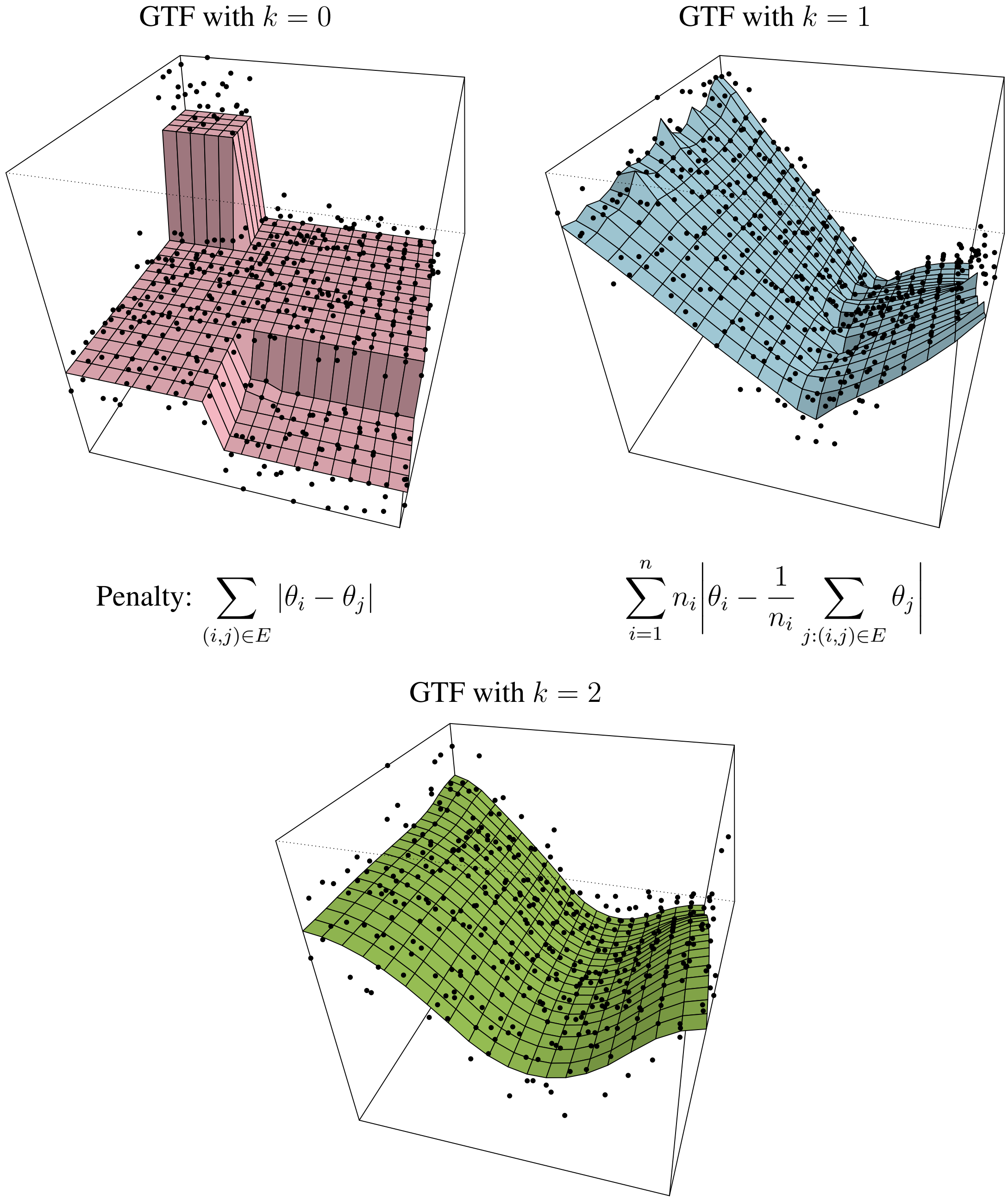
$$\theta_i - \frac{1}{n_i} \sum_{(i,j)\in E} \theta_j = 0,$$

  for many nodes $i \in V$, where n_i is the number of nodes adjacent to i. In words, we are requiring that the signal components can be linearly interpolated from its neighboring values at many nodes in the graph. This is quite a natural notion of (piecewise) linearity: requiring that θ_i be equal to the average of its neighboring values would enforce linearity at θ_i under an appropriate embedding of the points in Euclidean space. Again, this is precisely the same as requiring $\Delta^{(2)}\theta$ to be sparse, since $\Delta^{(2)} = L$, the graph Laplacian.
- **Piecewise polynomial ($k \geq 2$):** We say that θ has a piecewise quadratic structure over G if the first differences $\alpha_i - \alpha_j$ of the second differences $\alpha = \Delta^{(2)}\theta$ are mostly zero, over edges $(i, j) \in E$. Likewise, θ has a piecewise cubic structure over G if the second differences $\alpha_i - \frac{1}{n_i}\sum_{(i,j)\in E} \alpha_j$ of the second differences $\alpha = \Delta^{(2)}\theta$ are mostly zero, over nodes $i \in V$. This argument extends, alternating between leading first and second differences for even and odd k. Sparsity of $\Delta^{(k+1)}\theta$ in either case exactly corresponds to many of these differences being zero, by construction.

In Figure 6.3, we illustrate the graph trend filtering estimator on a 2d grid graph of dimension 20 × 20, i.e., a grid graph with 400 nodes and 740 edges. For each of the cases $k = 0, 1, 2$, we generated synthetic measurements over the grid, and computed a GTF estimate of the corresponding order. We chose the 2d grid setting so that the piecewise polynomial nature of GTF estimates could be visualized. Below each plot, the utilized graph trend filtering penalty is displayed in more explicit detail.

6.2.3 ℓ_1 versus ℓ_2 Regularization

It is instructive to compare the graph trend filtering estimator, as defined in (6.3), (6.4), (6.5) to Laplacian smoothing [203]. Standard Laplacian smoothing uses the same least squares loss as in (6.3), but replaces the penalty term with $\theta^\top L\theta$. A natural generalization would be to allow for a power of the Laplacian matrix L, and define kth order graph Laplacian smoothing according



GTF with $k = 0$

GTF with $k = 1$

Penalty: $\displaystyle\sum_{(i,j)\in E} |\theta_i - \theta_j|$

$$\sum_{i=1}^{n} n_i \left| \theta_i - \frac{1}{n_i} \sum_{j:(i,j)\in E} \theta_j \right|$$

GTF with $k = 2$

$$\sum_{(i,j)\in E} \left| n_i \left(\theta_i - \frac{1}{n_i} \sum_{\ell:(i,\ell)\in E} \theta_\ell \right) - n_j \left(\theta_j - \frac{1}{n_j} \sum_{\ell:(j,\ell)\in E} \theta_\ell \right) \right|$$

Figure 6.3: Graph trend filtering estimates of orders $k = 0, 1, 2$ on a 2d grid. The utilized ℓ_1 graph difference penalties are shown in elementwise detail below each plot (first, second, and third order graph differences).

to

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \; \|y - \theta\|_2^2 + \lambda \theta^\top L^{k+1} \theta. \tag{6.6}$$

The above penalty term can be written as $\|L^{(k+1)/2}\theta\|_2^2$ for odd k, and $\|DL^{k/2}\theta\|_2^2$ for even k; i.e., this penalty is exactly $\|\Delta^{(k+1)}\theta\|_2^2$ for the graph difference operator $\Delta^{(k+1)}$ defined previously.

As we can see, the critical difference between graph Laplacian smoothing (6.6) and graph trend filtering (6.3) lies in the choice of penalty norm: ℓ_2 in the former, and ℓ_1 in the latter. The effect of the ℓ_1 penalty is that the GTF program can set many (higher order) graph differences to zero exactly, and leave others at large nonzero values; i.e., the GTF estimate can simultaneously be smooth in some parts of the graph and wiggly in others. On the other hand, due to the (squared) ℓ_2 penalty, the graph Laplacian smoother cannot set any graph differences to zero exactly, and roughly speaking, must choose between making all graph differences small or large. The relevant analogy here is the comparison between the lasso and ridge regression, or univariate trend filtering and smoothing splines [221], and the suggestion is that GTF can adapt to the proper local degree of smoothness, while Laplacian smoothing cannot. This is strongly supported by the examples given throughout this chapter.

6.2.4 Related Work

Some authors from the signal processing community, e.g., Bredies et al. [40], Setzer et al. [190], have studied total generalized variation (TGV), a higher order variant of total variation regularization. Moreover, several discrete versions of these operators have been proposed. They are often similar to the construction that we have. However, the focus of these works is mostly on how well a discrete functional approximates its continuous counterpart. This is quite different from our concern, as a signal on a graph (say a social network) may not have any meaningful continuous-space embedding at all. In addition, we are not aware of any study on the statistical properties of these regularizers. In fact, our theoretical analysis in Section 6.6 may be extended to these methods too.

6.3 Properties and Extensions

We first study the structure of graph trend filtering estimates, then discuss interpretations and extensions.

6.3.1 Basic Structure and Degrees of Freedom

We describe the basic structure of graph trend filtering estimates and present an unbiased estimate for their degrees of freedom. Let the tuning parameter λ be arbitrary but fixed. By virtue of the ℓ_1 penalty in (6.3), the solution $\hat{\theta}$ satisfies $\operatorname{supp}(\Delta^{(k+1)}\hat{\theta}) = A$ for some active set A (typically A is smaller when λ is larger). Trivially, we can reexpress this as $\Delta^{(k+1)}_{-A}\hat{\theta} = 0$, or $\hat{\theta} \in \mathrm{null}(\Delta^{(k+1)}_{-A})$. Therefore, the basic structure of GTF estimates is revealed by analyzing the null space of the suboperator $\Delta^{(k+1)}_{-A}$.

Lemma 6.1. *Assume without a loss of generality that G is connected (otherwise the results apply to each connected component of G). Let D, L be the oriented incidence matrix and Laplacian matrix of G. For even k, let $A \subseteq \{1, \dots m\}$, and let G_{-A} denote the subgraph induced by removing the edges indexed by A (i.e., removing edges e_ℓ, $\ell \in A$). Let $C_1, \dots C_s$ be the connected components of G_{-A}. Then*

$$\mathrm{null}(\Delta^{(k+1)}_{-A}) = \mathrm{span}\{\mathbb{1}\} + (L^\dagger)^{\frac{k}{2}}\mathrm{span}\{\mathbb{1}_{C_1}, \dots \mathbb{1}_{C_s}\},$$

where $\mathbb{1} = (1, \dots 1) \in \mathbb{R}^n$, and $\mathbb{1}_{C_1}, \dots \mathbb{1}_{C_s} \in \mathbb{R}^n$ are the indicator vectors over connected components. For odd k, let $A \subseteq \{1, \dots n\}$. Then

$$\mathrm{null}(\Delta^{(k+1)}_{-A}) = \mathrm{span}\{\mathbb{1}\} + \{(L^\dagger)^{\frac{k+1}{2}} v : v_{-A} = 0\}.$$

The proof of Lemma 6.1 appears in the Section 9.7. The lemma is useful for a few reasons. First, as motivated above, it describes the coarse structure of GTF solutions. When $k = 0$, we can see (as $(L^\dagger)^{0/2} = I$) that $\hat{\theta}$ will indeed be piecewise constant over groups of nodes $C_1, \dots C_s$ of G. For $k = 2, 4, \dots$, this structure is smoothed by multiplying such piecewise constant levels by $(L^\dagger)^{k/2}$. Meanwhile, for $k = 1, 3 \dots$, the structure of the GTF estimate is based on assigning nonzero values to a subset A of nodes, and then smoothing through multiplication by $(L^\dagger)^{(k+1)/2}$. Both of these smoothing operations, which depend on $L^\dagger$, have interesting interpretations in terms of to the electrical network perspective for graphs. This is developed in the next subsection.

A second use of Lemma 6.1 is that it leads to a simple expression for the degrees of freedom, i.e., the effective number of parameters, of the GTF estimate $\hat{\theta}$. From results on generalized lasso problems [222, 223], we have $\mathrm{df}(\hat{\theta}) = \mathbb{E}[\mathrm{nullity}(\Delta^{(k+1)}_{-A})]$, with A denoting the support of $\Delta^{(k+1)}\hat{\theta}$, and $\mathrm{nullity}(X)$ the dimension of the null space of a matrix X. Applying Lemma 6.1 then gives the following.

Lemma 6.2. *Assume that G is connected. Let $\hat{\theta}$ denote the GTF estimate at a fixed but arbitrary value of λ. Under the normal error model $y \sim \mathcal{N}(\theta_0, \sigma^2 I)$, the GTF estimate $\hat{\theta}$ has degrees of freedom given by*

$$\mathrm{df}(\hat{\theta}) = \begin{cases} \mathbb{E}\left[\max\left\{|A|, 1\right\}\right] & \textit{odd } k, \\ \mathbb{E}\left[\textit{number of connected components of } G_{-A}\right] & \textit{even } k. \end{cases}$$

Here $A = \mathrm{supp}(\Delta^{(k+1)}\hat{\theta})$ denotes the active set of $\hat{\theta}$.

As a result of Lemma 6.2, we can form simple unbiased estimate for $\mathrm{df}(\hat{\theta})$; for k odd, this is $\max\{|A|, 1\}$, and for k even, this is the number of connected components of G_{-A}, where A is the support of $\Delta^{(k+1)}\hat{\theta}$. When reporting degrees of freedom for graph trend filtering (as in the example in the introduction), we use these unbiased estimates.

6.3.2 Electrical Network Interpretation

Lemma 6.1 reveals a mathematical structure for GTF estimates $\hat{\theta}$, which satisfy $\hat{\theta} \in \text{null}(\Delta_{-A}^{(k+1)})$ for some set A. It is interesting to interpret the results using the electrical network perspective for graphs [236]. In this perspective, we imagine replacing each edge in the graph with a resistor of value 1. If $u \in \mathbb{R}^n$ describes how much current is going in at each node in the graph, then $v = Lu$ describes the induced voltage at each node. Provided that $\mathbb{1}^\top c = 0$, which means that the total accumulation of current in the network is 0, we can solve for the current values from the voltage values: $u = L^\dagger v$.

The odd case in Lemma 6.1 asserts that

$$\text{null}(\Delta_{-A}^{(k+1)}) = \text{span}\{\mathbb{1}\} + \{(L^\dagger)^{\frac{k+1}{2}} v : v_{-A} = 0\}.$$

For $k = 1$, this says that GTF estimates are formed by assigning a sparse number of nodes in the graph a nonzero voltage v, then solving for the induced current $L^\dagger v$ (and shifting this entire current vector by a constant amount). For $k = 3$, we assign a sparse number of nodes a nonzero voltage, solve for the induced current, and then *repeat this*: we relabel the induced current as input voltages to the nodes, and compute the new induced current. This process is again iterated for $k = 5, 7, \ldots$.

The even case in Lemma 6.1 asserts that

$$\text{null}(\Delta_{-A}^{(k+1)}) = \text{span}\{\mathbb{1}\} + (L^\dagger)^{\frac{k}{2}} \text{span}\{\mathbb{1}_{C_1}, \ldots \mathbb{1}_{C_s}\}.$$

For $k = 2$, this result says that GTF estimates are given by choosing a partition $C_1, \ldots C_s$ of the nodes, and assigning a constant input voltage to each element of the partition. We then solve for the induced current (and potentially shift this by an overall constant amount). The process is iterated for $k = 4, 6, \ldots$ by relabeling the induced current as input voltage.

The comparison between the structure of estimates for $k = 2$ and $k = 3$ is informative: in a sense, the above tells us that 2nd order GTF estimates will be *smoother* than 3rd order estimates, as a sparse input voltage vector need not induce a current that is piecewise constant over nodes in the graph. For example, an input voltage vector that has only a few nodes with very large nonzero values will induce a current that is peaked around these nodes, but not piecewise constant.

6.3.3 Extensions

Several extensions of the proposed graph trend filtering model are possible. Trend filtering over a weighted graph, for example, could be performed by using a properly weighted version of the edge incidence matrix in (6.4), and carrying forward the same recursion in (6.5) for the higher order difference operators. As another example, the Gaussian regression loss in (6.3) could be changed to another suitable likelihood-derived losses in order to accommodate a different data type for y, say, logistic regression loss for binary data, or Poisson regression loss for count data.

In Section 6.5.2, we explore a modest extension of GTF, where we add a strongly convex prior term to the criterion in (6.3) to assist in performing graph-based imputation from partially observed data over the nodes. In Section 6.5.3, we investigate a modification of the proposed regularization scheme, where we add a pure ℓ_1 penalty on θ in (6.3), hence forming a sparse variant of GTF. Other potentially interesting penalty extensions include: mixing graph difference penalties of various orders, and tying together several denoising tasks with a group penalty. Extensions such as these are easily built, recall, as a result of the analysis framework used by the GTF program, wherein the estimate defined through direct regularization via an analyzing operator, the ℓ_1-based graph difference penalty $\|\Delta^{(k+1)}\theta\|_1$.

6.4 Computation

Graph trend filtering is defined by a convex optimization problem (6.3). In principle this means that, at least for small or moderately sized problems, its solutions can be reliably computed using a variety of standard algorithms. In order to handle larger scale problems, we describe two specialized algorithms that improve on generic procedures by taking advantage of the structure of $\Delta^{(k+1)}$.

6.4.1 A Fast ADMM Algorithm

We reparametrize (6.3) by introducing auxiliary variables, so that we can apply ADMM. For even k, we use a special transformation that is critical for fast computation (following Ramdas and Tibshirani [176] in univariate trend filtering); for odd k, this is not possible. The reparametrizations for even and odd k are

$$\min_{\theta,z\in\mathbb{R}^n} \frac{1}{2}\|y-\theta\|_2^2 + \lambda\|Dz\|_1 \ \text{ s.t. } \ z = L^{\frac{k}{2}}x,$$
$$\min_{\theta,z\in\mathbb{R}^n} \frac{1}{2}\|y-\theta\|_2^2 + \lambda\|z\|_1 \ \text{ s.t. } \ z = L^{\frac{k+1}{2}}x,$$

respectively. Recall that D is the oriented incidence matrix and L is the graph Laplacian. The augmented Lagrangian is

$$\frac{1}{2}\|y-\theta\|_2^2 + \lambda\|Sz\|_1 + \frac{\rho}{2}\|z - L^q\theta + u\|_2^2 - \frac{\rho}{2}\|u\|_2^2,$$

where $S = D$ or $S = I$ depending on whether k is even or odd, and likewise $q = k/2$ or $q = (k+1)/2$. ADMM then proceeds by iteratively minimizing the augmented Lagrangian over θ, minimizing over z, and performing a dual update over u. The θ and z updates are of the form, for some b,

$$\theta \leftarrow (I + \rho L^{2q})^{-1}b, \tag{6.7}$$
$$z \leftarrow \operatorname*{argmin}_{x\in\mathbb{R}^n} \frac{1}{2}\|b-x\|_2^2 + \frac{\lambda}{\rho}\|Sx\|_1, \tag{6.8}$$

The linear system in (6.7) is well-conditioned, sparse, and can be solved efficiently using the preconditioned conjugate gradient method. This involves only multiplication with Laplacian matrices. For a small enough choices of $\rho > 0$ (the augmented Lagrangian parameter), the system in (6.7) is diagonally dominant, special Laplacian/SDD solvers can be applied, which run in almost linear time [127, 133, 207].

For $S = I$, the update in (6.8) is simply given by soft-thresholding, and for $S = D$, it is given by graph TV denoising, i.e., given by solving a graph fused lasso problem. Note that this subproblem has the exact structure of the graph trend filtering problem (6.3) with $k = 0$. A direct approach for graph TV denoising is available based on parametric max-flow [51], and this algorithm is empirically much faster than its worst-case complexity [39]. In the special case that the underlying graph is a grid, a promising alternative method employs proximal stacking techniques [16].

6.4.2 A Fast Newton Method

As an alternative to ADMM, a projected Newton-type method [15, 30] can be used to solve (6.3) via its dual problem:

$$\hat{v} = \operatorname*{argmin}_{v \in \mathbb{R}^r} \|y - (\Delta^{(k+1)})^\top v\|_2^2 \ \text{ s.t. } \ \|v\|_\infty \le \lambda.$$

The solution of (6.3) is then given by $\hat{\theta} = y - (\Delta^{(k+1)})^\top \hat{v}$. (For univariate trend filtering, Kim et al. [129] adopt a similar strategy, but instead use an interior point method.) The projected Newton method performs updates using a reduced Hessian, so abbreviating $\Delta = \Delta^{(k+1)}$, each iteration boils down to

$$v \leftarrow a + (\Delta_I^\top)^\dagger b, \tag{6.9}$$

for some a, b and set of indices I. The linear system in (6.9) is always sparse, but conditioning becomes an issue as k grows (note that the same problem does not occur in (6.7) because of the addition of the identity matrix I). We have found empirically that a preconditioned conjugate gradient method works quite well for (6.9) for $k = 1$, but struggles for larger k.

6.4.3 Computation Summary

In our experience, the following algorithms work well for the various order k of graph trend filtering. We remark that orders $k = 0, 1, 2$ are of most practical interest (and solutions of polynomial order $k \ge 3$ are less likely to be sought in practice).[1]

[1] Loosely speaking, each order $k = 0, 1, 2$ provides solutions that exhibit a different class of structure: $k = 0$ gives piecewise constant solutions, $k = 1$ gives piecewise linear, and $k = 2$ gives piecewise smooth. All orders $k \ge 3$ continue to give piecewise smooth fits, with less and less transparent differences (the practical differences between piecewise quadratic versus piecewise linear fits is greater than piecewise cubic versus piecewise quadratic, etc.). Since the conditioning of the graph trend filtering operator $\Delta^{(k+1)}$ worsens as k increases, which makes computation more difficult, it makes most practical sense to simply choose $k = 2$ whenever a piecewise smooth fit is desired.

Order	Algorithm
$k = 0$	Parametric max-flow
$k = 1$	Projected Newton method
$k = 2, 4, \ldots$	ADMM with parametric max-flow
$k = 3, 5, \ldots$	ADMM with soft-thresholding

Figure 6.4 compares performances of the described algorithms on a moderately large simulated example, using a 2d grid graph. We see that when $k = 1$, the projected Newton method converges faster than ADMM (superlinear versus at best linear convergence). When $k = 2$, the story is reversed, as the projected Newton iterations quickly become stagnant, and the ADMM enjoys better convergence.

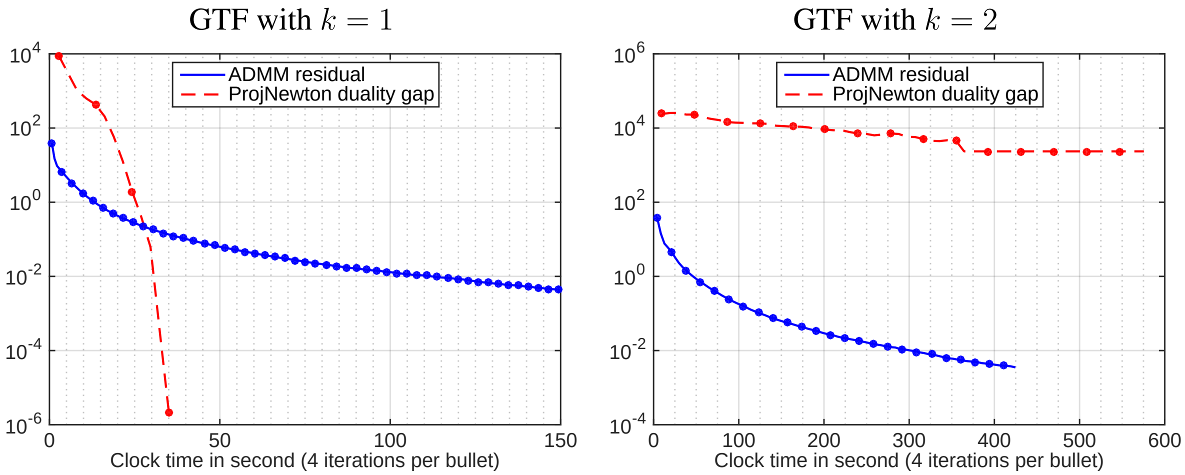


Figure 6.4: Convergence plots for projected Newton method and ADMM for solving GTF with $k = 1$ and $k = 2$. The algorithms are all run on a 2d grid graph (an 512×512 image) with 262,144 nodes and 523,264 edges. For projected Newton, we plot the duality gap across iterations; for ADMM, we plot the average of the primal and dual residuals (which also serves as a valid suboptimality bound in the ADMM framework).

6.5 Examples

In this section, we present a variety of examples of running graph trend filtering on real graphs.

6.5.1 Trend Filtering over the Facebook Graph

In the Introduction, we examined the denoising power of graph trend filtering in a spatial setting. Here we examine the behavior of graph trend filtering on a nonplanar graph: the Facebook graph from the Stanford Network Analysis Project (`http://snap.stanford.edu`). This is composed of 4039 nodes representing Facebook users, and 88,234 edges representing friendships, collected from real survey participants; the graph has one connected component, but the observed degree sequence is very mixed, ranging from 1 to 1045 (refer to McAuley and Leskovec [154] for more details).

We generated synthetic measurements over the Facebook nodes (users) based on three different ground truth models, so that we can precisely evaluate and compare the estimation accuracy of GTF, Laplacian smoothing, and wavelet smoothing. For the latter, we again used the spanning tree wavelet design of Sharpnack et al. [195], because it performed among the best of wavelets designs in all data settings considered here. Results from other wavelet designs are presented in the Section 6.8. The three ground truth models represent very different scenarios for the underlying signal x, each one favorable to different estimation methods. These are:

1. **Dense Poisson equation:** we solved the Poisson equation $Lx = b$ for x, where b is arbitrary and dense (its entries were i.i.d. normal draws).
2. **Sparse Poisson equation:** we solved the Poisson equation $Lx = b$ for x, where b is sparse and has 30 nonzero entries (again i.i.d. normal draws).
3. **Inhomogeneous random walk:** we ran a set of decaying random walks at different starter nodes in the graph, and recorded in x the total number of visits at each node. Specifically, we chose 10 nodes as starter nodes, and assigned each starter node a decay probability uniformly at random between 0 and 1 (this is the probability that the walk terminates at each step instead of travelling to a neighboring node). At each starter node, we also sent out a varying number of random walks, chosen uniformly between 0 and 1000.

In each case, the synthetic measurements were formed by adding noise to x. We note that model 1 is designed to be favorable for Laplace smoothing; model 2 is designed to be favorable for GTF; and in the inhomogeneity in model 3 is designed to be challenging for Laplacian smoothing, and favorable for the more adaptive GTF and wavelet methods.

Figure 6.5 shows the performance of the three estimation methods, over a wide range of noise levels in the synthetic measurements; performance here is measured by the best achieved mean squared error, allowing each method to be tuned optimally at each noise level. The summary: GTF estimates are (expectedly) superior when the Laplacian-based sparsity pattern is in effect (model 2), but are nonetheless highly competitive in both other settings—the dense case, in which Laplacian smoothing thrives, and the inhomogeneous random walk case, in which wavelets thrive.

6.5.2 Graph-Based Transductive Learning over UCI Data

Graph trend filtering can used for graph-based transductive learning, as motivated by the work of Talukdar and Crammer [213], Talukdar and Pereira [214], who rely on Laplacian regularization. Consider a semi-supervised learning setting, where we are given only a small number of seed labels over nodes of a graph, and the goal is to impute the labels on the remaining nodes. Write $O \subseteq \{1, \ldots n\}$ for the set of observed nodes, and assume that each observed label falls into

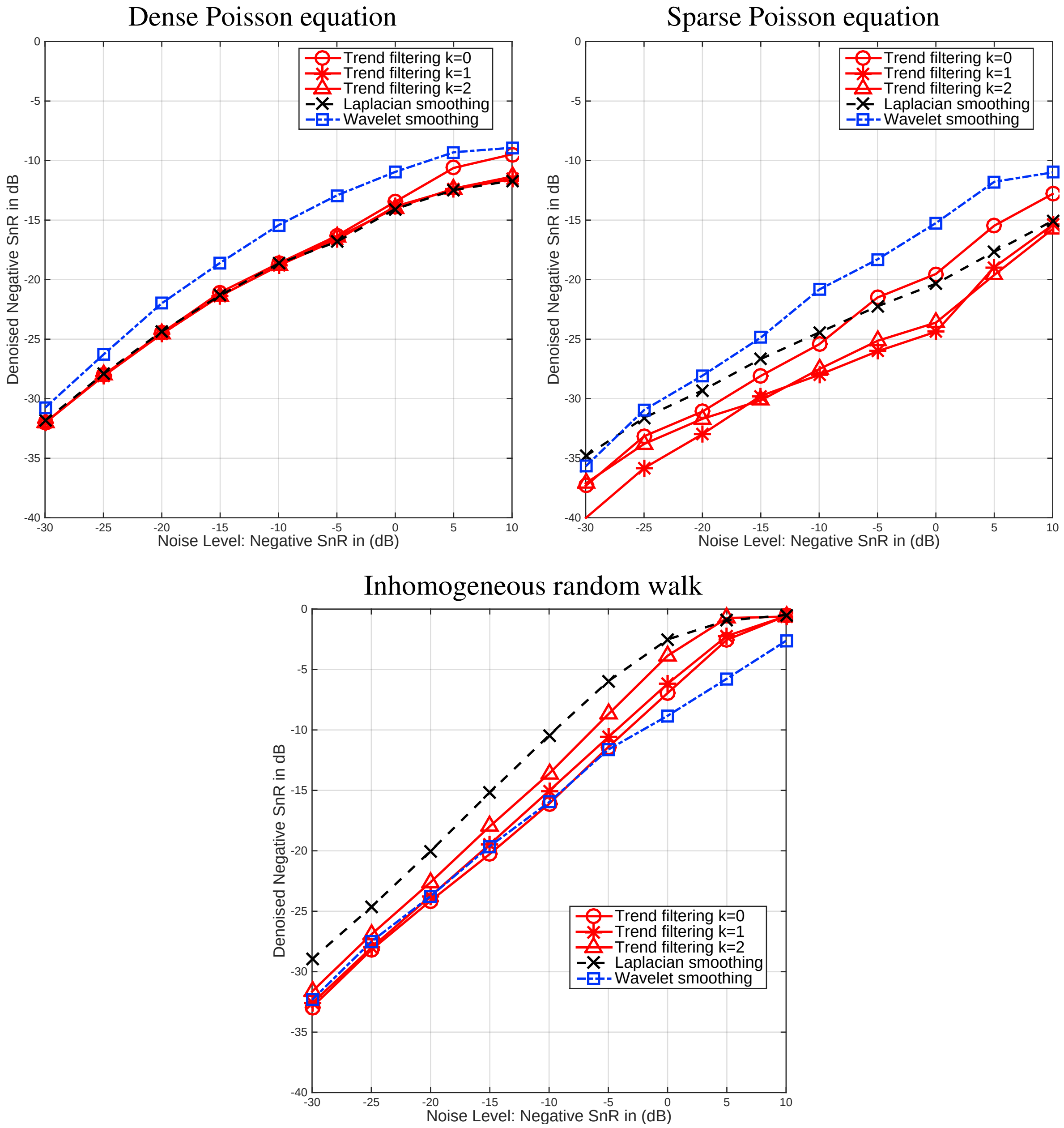


Figure 6.5: Performance of GTF and others for three generative models on the Facebook graph. The x-axis shows the negative SnR: $10\log_{10}(n\sigma^2/\|x\|_2^2)$, where $n = 4039$, x is the underlying signal, and σ^2 is the noise variance. Hence the noise level is increasing from left to right. The y-axis shows the denoised negative SnR: $10\log_{10}(\mathrm{MSE}/\|x\|_2^2)$, where MSE denotes mean squared error, so the achieved MSE is increasing from bottom to top.

$\{1, \ldots K\}$. Then we can define the modified absorption problem under graph trend filtering regularization (MAD-GTF) by

$$\hat{B} = \underset{B \in \mathbb{R}^{n \times K}}{\text{argmin}} \sum_{j=1}^{K} \sum_{i \in O} (Y_{ij} - B_{ij})^2 + \lambda \sum_{j=1}^{K} \|\Delta^{(k+1)} B_j\|_1 + \epsilon \sum_{j=1}^{K} \|R_j - B_j\|_2^2. \tag{6.10}$$

The matrix $Y \in \mathbb{R}^{n \times K}$ is an indicator matrix: each observed row $i \in O$ is described by $Y_{ij} = 1$ if class j is observed and $Y_{ij} = 0$ otherwise. The matrix $B \in \mathbb{R}^{n \times K}$ contains fitted probabilities, with B_{ij} giving the probability that node i is of class j. We write B_j for its jth column, and hence the middle term in the above criterion encourages each set of class probabilities to behave smoothly over the graph. The last term in the above criterion ties the fitted probabilities to some given prior weights $R \in \mathbb{R}^{n \times K}$. In principle ϵ could act as a second tuning parameter, but for simplicity we take ϵ to be small and fixed, with any $\epsilon > 0$ guaranteeing that the criterion in (6.10) is strictly convex, and thus has a unique solution $\hat{B}$. The entries of $\hat{B}$ need not be probabilites in any strict sense, but we can still interpret them as relative probabilities, and imputation can be performed by assigning each unobserved node $i \notin O$ a class label j such that $\hat{B}_{ij}$ is largest.

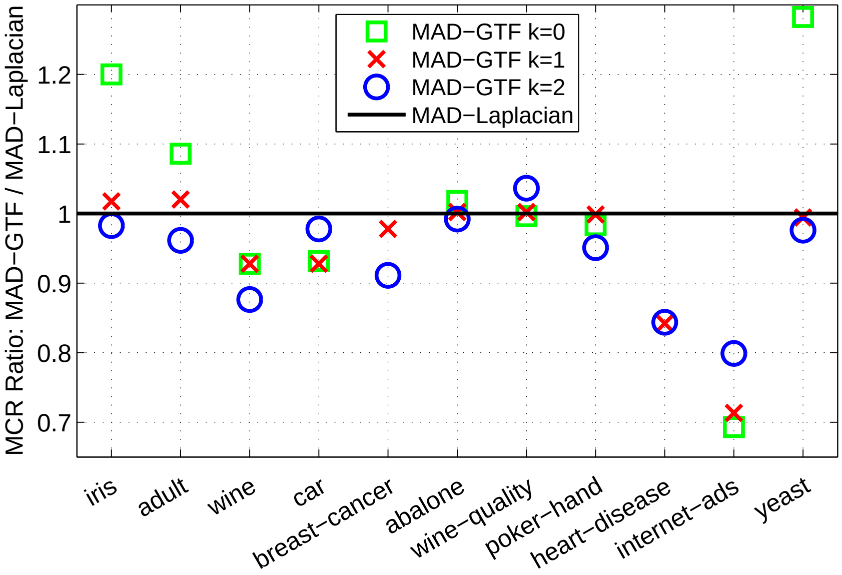


Figure 6.6: Ratio of the misclassification rate of MAD-GTF to MAD-Laplacian, for graph-based imputation, on the 11 most popular UCI classification data sets.

Our specification of MAD-GTF only deviates from the MAD proposal of Talukdar and Crammer [213] in that these authors used the Laplacian regularization term $\sum_{j=1}^{K} B_j^\top L B_j$, in place of ℓ_1-based graph difference regularizer in (6.10). If the underlying class probabilities are thought to have heterogeneous smoothness over the graph, then replacing the Laplacian regularizer with the GTF-designed one might lead to better performance. As a broad comparison of the two methods, we ran them on the 11 most popular classification data sets from the UCI Machine Learning repository (`http://archive.ics.uci.edu/ml/`).[2] For each data set, we constructed a 5-

[2] We used all data sets here, except the "forest-fires" data set, which is a regression problem. Also, we zero-filled the missing data in "internet-ads" data set and used a random one third of the data in the "poker" data set.

	iris	adult	wine	car	breast	abalone	wine-qual.	poker	heart	ads	yeast
# of classes	3	2	3	4	2	29	6	10	2	2	10
# of samples	150	32,561	178	1,728	569	4,177	1,599	3,000	303	3,279	1,484
Laplacian	0.085	0.270	0.060	0.316	0.064	0.872	0.712	0.814	0.208	0.306	0.566
GTF, $k = 0$	0.102	0.293	0.055	0.294	0.500	0.888	0.709	0.801	0.472	0.212	0.726
p-value	0.254	0.648	0.406	0.091	0.000	0.090	0.953	0.732	0.000	0.006	0.000
GTF, $k = 1$	0.087	0.275	0.055	0.293	0.063	0.874	0.713	0.813	0.175	0.218	0.563
p-value	0.443	0.413	0.025	0.012	0.498	0.699	0.920	0.801	0.134	0.054	0.636
GTF, $k = 2$	0.084	0.259	0.052	0.309	0.059	0.865	0.738	0.774	0.175	0.244	0.552
p-value	0.798	0.482	0.024	0.523	0.073	0.144	0.479	0.138	0.301	0.212	0.100

Table 6.1: Misclassification rates of MAD-Laplacian and MAD-GTF for imputation in the UCI data sets. We also compute p-values over the 10 repetitions for each data set (10 draws of nodes to serve as seed labels) via paired t-tests. Cases where MAD-GTF achieves significantly better misclassification rate, at the 0.1 level, are highlighted in green; cases where MAD-GTF achieves a significantly worse miclassification rate, at the 0.1 level, are highlighted in red.

nearest-neighbor graph based on the Euclidean distance between provided features, and randomly selected 5 seeds per class to serve as the observed class labels. Then we set $\epsilon = 0.01$, used prior weights $R_{ij} = 1/K$ for all i and j, and chose the tuning parameter λ over a wide grid of values to represent the best achievable performance by each method, on each experiment. Figure 6.6 and Table 6.1 summarize the misclassification rates from imputation using MAD-Laplacian and MAD-GTF, averaged over 10 repetitions of the randomly selected seed labels. We see that MAD-GTF with $k = 0$ (basically a graph partition akin to MRF-based graph cut, using an Ising model) does not seem to work as well as the smoother alternatives. Importantly, MAD-GTF with $k = 1$ and $k = 2$ both perform at least as well, and sometimes better, than MAD-Laplacian on each one of the UCI data sets. Recall that these data sets were selected entirely based on their popularity, and not at all on the belief that they might represent favorable scenarios for GTF (i.e., not on the prospect that they might exhibit some heterogeneity in the distribution of class labels over their respective graphs). Therefore, the fact that MAD-GTF nonetheless performs competitively in such a broad range of experiments is convincing evidence for the utility of the GTF regularizer.

6.5.3 Event Detection with NYC Taxi Trips Data

We illustrate a sparse variant of our proposed regularizers, given by adding a pure ℓ_1 penalty to the coefficients in (6.3), as in

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \; \frac{1}{2}\|y - \theta\|_2^2 + \lambda_1 \|\Delta^{(k+1)}\theta\|_1 + \lambda_2 \|\theta\|_1. \tag{6.11}$$

We call this *sparse graph trend filtering*, now with two tuning parameters $\lambda_1, \lambda_2 \geq 0$. Under the proper tuning, the sparse GTF estimate will be zero at many nodes in the graph, and will otherwise deviate smoothly from zero. This can be useful in situations where the observed signal represents a difference between two smooth processes that are mostly similar, but exhibit (perhaps significant) differences over a few regions of the graph. Here we apply it to the problem of detecting events based on abnormalities in the number of taxi trips at different locations of New York city. This data set was kindly provided by authors of Doraiswamy et al. [72], who obtained the data from NYC Taxi & Limosine Commission.[3] Specifically, we consider the graph to be the road network of Manhattan, which contains 3874 nodes (junctions) and 7070 edges (sections of roads that connect two junctions). For measurements over the nodes, we used the number of taxi pickups and dropoffs over a particular time period of interest: 12:00–2:00 pm on June 26, 2011, corresponding to the Gay Pride parade. As pickups and dropoffs do not generically occur at road junctions, we used interpolation to form counts over the graph nodes. A baseline seasonal average was calculated by considering data from the same time block 12:00–2:00 pm on the same day of the week across the nearest eight weeks. Thus the measurements y were then taken to be the difference between the counts observed during the Gay Pride parade and the seasonal averages.

Note that the nonzero node estimates from sparse GTF applied to y, after proper tuning, mark events of interest, because they convey substantial differences between the observed and expected taxi counts. According to descriptions in the news, we know that the Gay Pride parade was a giant march down at noon from 36th St. and Fifth Ave. all the way to Christopher St. in Greenwich Village, and traffic was blocked over the entire route for two hours (meaning no pickups and dropoffs could occur). We therefore hand-labeled this route as a crude “ground truth” for the event of interest, illustrated in the left panel of Figure 6.7.

In the bottom two panels of Figure 6.7, we compare sparse GTF with $k = 0$ (i.e., the sparse graph fused lasso) and a sparse variant of Laplacian smoothing, obtained by replacing the first regularization term in (6.11) by $\theta^\top L\theta$. For a qualitative visual comparison, the smoothing parameter λ_1 was chosen so that both methods have 200 degrees of freedom (without any sparsity imposed). The sparsity parameter was then set as $\lambda_2 = 0.2$. Similar to what we have seen already, GTF is able to better localize its estimates around strong inhomogenous spikes in the measurements, and is able to better capture the event of interest. The result of sparse Laplacian smoothing is far from localized around the ground truth event, and displays many nonzero node estimates throughout distant regions of the graph. If we were to decrease its flexibility (increase the smoothing parameter λ_1 in its problem formulation), then the sparse Laplacian output would display more smoothness over the graph, but the node estimates around the ground truth region would also be grossly shrunken.

6.6 Estimation Error Bounds

In this section, we assume that $y \sim \mathcal{N}(\theta_0, \sigma^2 I)$, and study asymptotic error rates for graph trend filtering. (The assumption of a normal error model could be relaxed, but is used for simplicity).

[3] These authors also considered event detection, but their topological definition of an “event” is very different from what we considered here, and hence our results not directly comparable.

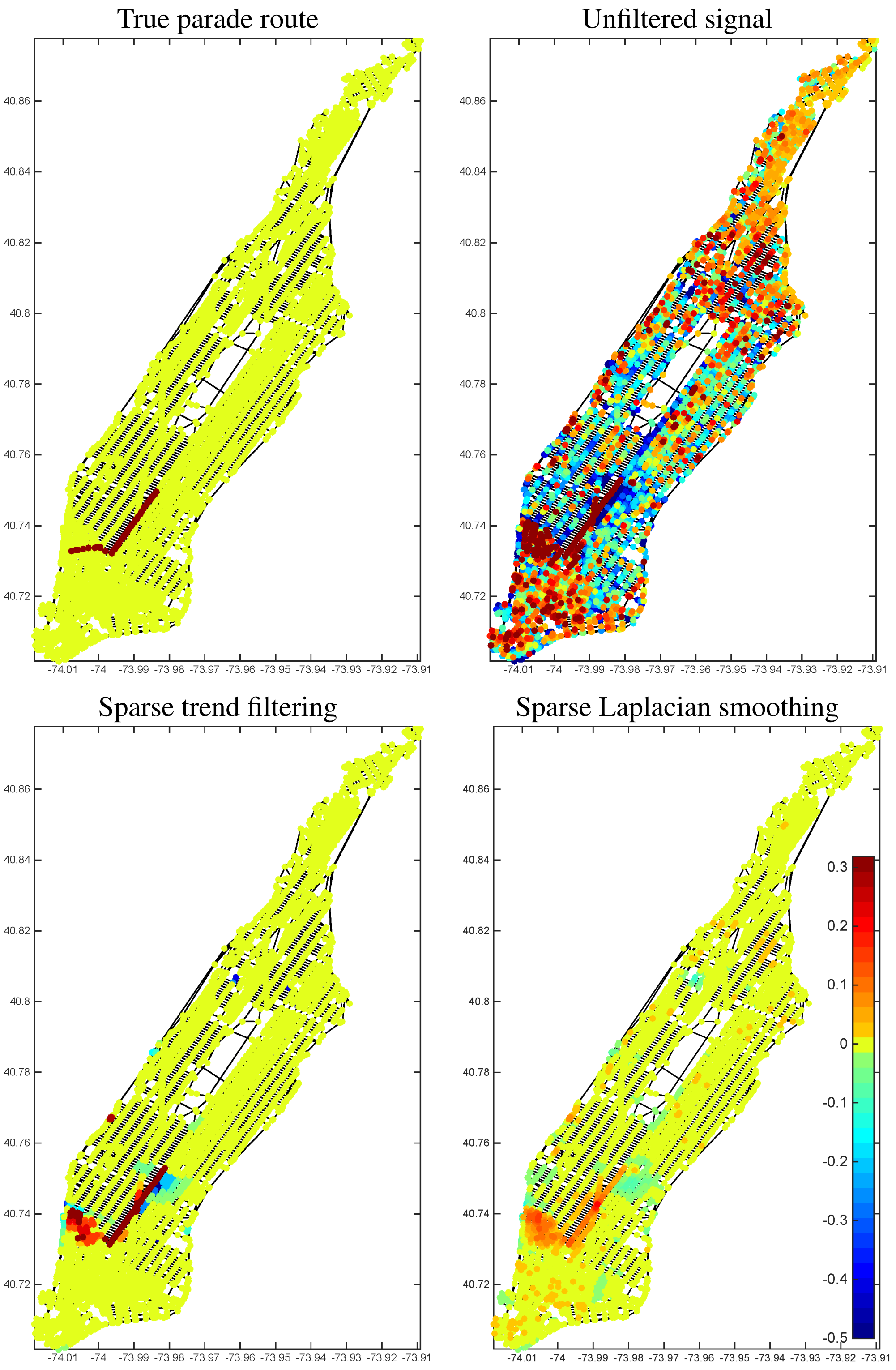


Figure 6.7: Comparison of sparse GTF and sparse Laplacian smoothing. We can see qualitatively that sparse GTF delivers better event detection with fewer false positives (zoomed-in, the sparse Laplacian plot shows a scattering of many non-zero colors).

Our analysis actually focuses more broadly on the generalized lasso problem

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \frac{1}{2}\|y - \theta\|_2^2 + \lambda \|\Delta\theta\|_1, \tag{6.12}$$

where $\Delta \in \mathbb{R}^{r \times n}$ is an arbitrary linear operator, and r denotes its number of rows. Throughout, we specialize the derived results to the graph difference operator $\Delta = \Delta^{(k+1)}$, to obtain concrete statements about GTF over particular graphs. All proofs are deferred to Section 9.7.

6.6.1 Basic Error Bounds

Using similar arguments to the basic inequality for the lasso [42], we have the following preliminary bound.

Theorem 6.3. *Let M denote the maximum ℓ_2 norm of the columns of $\Delta^\dagger$. Then for a tuning parameter value $\lambda = \Theta(M\sqrt{\log r})$, the generalized lasso estimate $\hat{\theta}$ in* (6.12) *has average squared error*

$$\frac{\|\hat{\theta} - \theta_0\|_2^2}{n} = O_{\mathbb{P}}\left(\frac{\mathrm{nullity}(\Delta)}{n} + \frac{M\sqrt{\log r}}{n} \cdot \|\Delta\theta_0\|_1\right).$$

Recall that $\mathrm{nullity}(\Delta)$ denotes the dimension of the null space of Δ. For the GTF operator $\Delta^{(k+1)}$ of any order k, note that $\mathrm{nullity}(\Delta^{(k+1)})$ is the number of connected components in the underlying graph.

When both $\|\Delta\theta_0\|_1 = O(1)$ and $\mathrm{nullity}(\Delta) = O(1)$, Theorem 6.3 says that the estimate $\hat{\theta}$ converges in average squared error at the rate $M\sqrt{\log r}/n$, in probability. This theorem is quite general, as it applies to any linear operator Δ, and one might therefore think that it cannot yield fast rates. Still, as we show next, it does imply consistency for graph trend filtering in certain cases.

Corollary 6.4. *Consider the trend filtering estimator $\hat{\theta}$ of order k, and the choice of the tuning parameter λ as in Theorem 6.3. Then:*

1. *for univariate trend filtering (i.e., essentially GTF on a chain graph),*

$$\frac{\|\hat{\theta} - \theta_0\|_2^2}{n} = O_{\mathbb{P}}\left(\sqrt{\frac{\log n}{n}} \cdot n^k \|D^{(k+1)}\theta_0\|_1\right);$$

2. *for GTF on an Erdos-Renyi random graph, with edge probability p, and expected degree $d = np \geq 1$,*

$$\frac{\|\hat{\theta} - \theta_0\|_2^2}{n} = O_{\mathbb{P}}\left(\frac{\sqrt{\log(nd)}}{nd^{\frac{k+1}{2}}} \cdot \|\Delta^{(k+1)}\theta_0\|_1\right);$$

3. *for GTF on a Ramanujan d-regular graph, and $d \geq 1$,*

$$\frac{\|\hat{\theta} - \theta_0\|_2^2}{n} = O_{\mathbb{P}}\left(\frac{\sqrt{\log(nd)}}{nd^{\frac{k+1}{2}}} \cdot \|\Delta^{(k+1)}\theta_0\|_1\right).$$

Cases 2 and 3 of Corollary 6.4 stem from the simple inequality $M \leq \|\Delta^\dagger\|_2$, the largest singular value of $\Delta^\dagger$. When $\Delta = \Delta^{(k+1)}$, the GTF operator of order $k+1$, we have

$$\|(\Delta^{(k+1)})^\dagger\|_2 \leq 1/\lambda_{\min}(L)^{(k+1)/2},$$

where $\lambda_{\min}(L)$ is the smallest nonzero eigenvalue of the Laplacian L (also known as the Fiedler eigenvalue [92]). In general, $\lambda_{\min}(L)$ can be very small, leading to loose error bounds, but for the particular graphs in question, it is well-controlled. When $\|\Delta^{(k+1)}\theta_0\|_1$ is bounded, cases 2 and 3 of the corollary show that the average squared error of GTF converges at the rate $\sqrt{\log(nd)}/(nd^{(k+1)/2})$. As k increases, this rate is stronger, but so is the assumption that $\|\Delta^{(k+1)}\theta_0\|_1$ is bounded.

Case 1 in Corollary 6.4 covers univariate trend filtering (which, recall, is basically the same as GTF over a chain graph; the only differences between the two are boundary terms in the construction of the difference operators). The result in case 1 is based on direct calculation of M, using specific facts that are known about the univariate difference operators. It is natural in the univariate setting to assume that $n^k\|D^{(k+1)}\theta_0\|_1$ is bounded (this is the scaling that would link θ_0 to the evaluations of a piecewise polynomial function f_0 over $[0,1]$, with $\mathrm{TV}(f_0^{(k)})$ bounded). Under such an assumption, the above corollary yields a convergence rate of $\sqrt{\log n/n}$ for univariate trend filtering, which is not tight. A more refined analysis shows the univariate trend filtering estimator to converge at the minimax optimal rate $n^{-(2k+2)/(2k+3)}$, proved in Tibshirani [221] by using a connection between univariate trend filtering and locally adaptive regression splines, and relying on sharp entropy-based rates for locally adaptive regression splines from Mammen and van de Geer [152]. We note that in a pure graph-centric setting, the latter strategy is not generally applicable, as the notion of a spline function does not obviously extend to the nodes of an arbitrary graph structure.

In the next subsections, we develop more advanced strategies for deriving fast GTF error rates, based on incoherence, and entropy. These can provide substantial improvements over the basic error bound established in this subsection, but are only applicable to certain graph models. Fortunately, this includes common graphs of interest, such as regular grids. To verify the sharpness of these alternative strategies, we will show that they can be used to recover optimal rates of convergence for trend filtering in the 1d setting.

6.6.2 Strong Error Bounds Based on Incoherence

A key step in the proof of Theorem 6.3 argues, roughly speaking, that

$$\epsilon^\top \Delta^\dagger \Delta x \leq \|(\Delta^\dagger)^\top \epsilon\|_\infty \|\Delta x\|_1 = O_{\mathbb{P}}(M\sqrt{\log r}\|\Delta x\|_1), \tag{6.13}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. The second bound holds by a standard result on maxima of Gaussians (recall that M is largest ℓ_2 norm of the columns of $\Delta^\dagger$). The first bound above uses Holder's inequality; note that this applies to any ϵ, Δ, i.e., it does not use any information about the distribution of ϵ, or the properties of Δ. The next lemma reveals a potential advantage that can be gained from replacing the bound (6.13), stemming from Holder's inequality, with a "linearized" bound.

Lemma 6.5. *Denote $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, and assume that*

$$\max_{x \in \mathcal{S}_\Delta(1)} \frac{\epsilon^\top x - A}{\|x\|_2} = O_{\mathbb{P}}(B), \tag{6.14}$$

where $\mathcal{S}_\Delta(1) = \{x \in \mathrm{row}(\Delta) : \|\Delta x\|_1 \leq 1\}$. With $\lambda = \Theta(A)$, the generalized lasso estimate $\hat{\theta}$ satisfies

$$\frac{\|\hat{\theta} - \theta_0\|_2^2}{n} = O_{\mathbb{P}} \left(\frac{\mathrm{nullity}(\Delta)}{n} + \frac{B^2}{n} + \frac{A}{n} \cdot \|\Delta \theta_0\|_1 \right).$$

The inequality in (6.14) is referred to as a "linearized" bound because it implies that for $x \in \mathcal{S}_\Delta(1)$,

$$\epsilon^\top x = O_{\mathbb{P}}(A + B\|x\|_2),$$

and the right-hand side is a linear function of $\|x\|_2$. Indeed, for $A = M\sqrt{2 \log r}$ and $B = 0$, this encompasses the bound in (6.13) as a special case, and the result of Lemma 6.5 reduces to that of Theorem 6.3. But the result in Lemma 6.5 can be much stronger, if A, B can be adjusted so that A is smaller than $M\sqrt{2 \log r}$, and B is also small. Such an arrangement is possible for certain operators Δ; e.g., it is possible under an incoherence-type assumption on Δ.

Theorem 6.6. *Let $q = \mathrm{rank}(\Delta)$, and let $\xi_1 \leq \ldots \leq \xi_q$ denote the singular values of Δ, in increasing order. Also let $u_1, \ldots u_q$ be the corresponding left singular vectors. Assume that these vectors are incoherent:*

$$\|u_i\|_\infty \leq \mu/\sqrt{n}, \quad i = 1, \ldots q,$$

for some constant $\mu \geq 1$. For $i_0 \in \{1, \ldots q\}$, let

$$\lambda = \Theta \left(\mu \sqrt{\frac{\log r}{n} \sum_{i=i_0+1}^{q} \frac{1}{\xi_i^2}} \right).$$

Then the generalized lasso estimate $\hat{\theta}$ has average squared error

$$\frac{\|\hat{\theta} - \theta_0\|_2^2}{n} = O_{\mathbb{P}} \left(\frac{\mathrm{nullity}(\Delta)}{n} + \frac{i_0}{n} + \frac{\mu}{n} \sqrt{\frac{\log r}{n} \sum_{i=i_0+1}^{q} \frac{1}{\xi_i^2}} \cdot \|\Delta \theta_0\|_1 \right).$$

Theorem 6.6 is proved by leveraging the linearized bound (6.14), which holds under the incoherence condition on the singular vectors of Δ. Compared to the basic result in Theorem 6.3, the result in Theorem 6.6 is clearly stronger as it allows us to replace M—which can grow like the reciprocal of the minimum nonzero singular value of Δ—with something akin to the average reciprocal of larger singular values. But it does, of course, also make stronger assumptions (incoherence). It is interesting to note that the functional in the theorem, $\sum_{i=i_0+1}^{q} \xi_i^{-2}$, was also determined to play a leading role in error bounds for a graph Fourier based scan statistic in the hypothesis testing framework [196].

Applying the above theorem to the GTF estimator requires knowledge of the singular vectors of $\Delta = \Delta^{(k+1)}$, the $(k+1)$st order graph difference operator. The validity of an incoherence assumption on these singular vectors depend on the graph G in question. When k is odd, these singular vectors are eigenvectors of the Laplacian L; when k is even, they are left singular vectors of the edge incidence matrix D. Loosely speaking, these vectors will be incoherent when neighborhoods of different vertices look roughly the same. Most social networks will have this property for the bulk of their vertices (i.e., with the exception of a small number of high degree vertices). Grid graphs also have this property. First, we consider trend filtering over a 1d grid, i.e., a chain (which, recall, is essentially equivalent to univariate trend filtering).

Corollary 6.7. *Consider the GTF estimator $\hat{\theta}$ of order k, over a chain graph, i.e., a 1d grid graph. Letting*

$$\lambda = \Theta\left(n^{\frac{2k+1}{2k+3}} (\log n)^{\frac{1}{2k+3}} \|\Delta^{(k+1)}\theta_0\|_1^{-\frac{2k+1}{2k+3}} \right),$$

the estimator $\hat{\theta}$ (here, essentially, the univariate trend filtering estimator) satisfies

$$\frac{\|\hat{\theta}-\theta_0\|_2^2}{n} = O_{\mathbb{P}}\left(n^{-\frac{2k+2}{2k+3}} (\log n)^{\frac{1}{2k+3}} \cdot \left(n^k \|\Delta^{(k+1)}\theta_0\|_1 \right)^{\frac{2}{2k+3}} \right).$$

We note that the above corollary essentially recovers the optimal rate of convergence for the univariate trend filtering estimator, for all orders k. (To be precise, it studies GTF on a chain graph instead, but this is basically the same problem.) When $n^k\|\Delta^{(k+1)}\theta_0\|_1$ is assumed to be bounded, a natural assumption in the univariate setting, the corollary shows the estimator to converge at the rate $n^{-(2k+2)/(2k+3)}(\log n)^{1/(2k+3)}$. Ignoring the log factor, this matches the minimax optimal rate as established in Tibshirani [221], Wang et al. [244]. Importantly, the proof of Corollary 6.7, unlike that used in previous works, is free from any dependence on univariate spline functions; it is completely graph-theoretic, and only uses on the incoherence properties of the 1d grid graph. The strength of this approach is its wider applicability, which we demonstrate by moving up to 2d grids.

Corollary 6.8. *Consider the GTF estimator $\hat{\theta}$ of order k, over a 2d grid graph, of size $\sqrt{n}\times\sqrt{n}$. Letting*

$$\lambda = \Theta\left(n^{\frac{2k+1}{2k+5}} (\log n)^{\frac{1}{2k+5}} \|\Delta^{(k+1)}\theta_0\|_1^{-\frac{2k+1}{2k+5}} \right),$$

the estimator $\hat{\theta}$ satisfies

$$\frac{\|\hat{\theta}-\theta_0\|_2^2}{n} = O_{\mathbb{P}}\left(n^{-\frac{2k+4}{2k+5}} (\log n)^{\frac{1}{2k+5}} \cdot \left(n^{\frac{k}{2}} \|\Delta^{(k+1)}\theta_0\|_1 \right)^{\frac{4}{2k+5}} \right).$$

The 2d result in Corollary 6.8 is written in a form that mimics the 1d result in Corollary 6.7, as we claim that the analog of boundedness of $n^k\|\Delta^{(k+1)}\theta_0\|_1$ in 1d is boundedness of $n^{k/2}\|\Delta^{(k+1)}\theta_0\|_1$ in 2d.[4] Thus, under the appropriate boundedness condition, the 2d rate shows improvement over the 1d rate, which makes sense, since regularization here is being enforced in a richer manner. It is worthwhile highlighting the result for $k = 0$ in particular: this says that, when the sum of absolute discrete differences $\|\Delta^{(1)}\theta_0\|_1$ is bounded over a 2d grid, the 2d fused lasso (i.e., 2d total variation denoising) has error rate $n^{-4/5}$. This is faster than the $n^{-2/3}$ rate for the 1d fused lasso, when the sum of absolute differences $\|D^{(1)}\theta_0\|_1$ is bounded. Rates for higher dimensional grid graphs (for all k) follow from analogous arguments, but we omit the details.

[4] This is because $1/\sqrt{n}$ is the distance between adjacent 2d grid points, when viewed as a 2d lattice over $[0,1]^2$.

6.6.3 Strong Error Bounds Based on Entropy

A different "fractional" bound on the Gaussian contrast $\epsilon^\top x$, over $x \in \mathcal{S}_\Delta(1)$, provides an alternate route to deriving sharper rates. This style of bound is inspired by the seminal work of van de Geer [230].

Lemma 6.9. *Denote $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, and assume that for a constant $w < 2$,*

$$\max_{x \in \mathcal{S}_\Delta(1)} \frac{\epsilon^\top x}{\|x\|_2^{1-w/2}} = O_{\mathbb{P}}(K), \tag{6.15}$$

where recall $\mathcal{S}_\Delta(1) = \{x \in \mathrm{row}(\Delta) : \|\Delta x\|_1 \leq 1\}$. Then with

$$\lambda = \Theta\left(K^{\frac{2}{1+w/2}} \cdot \|\Delta\theta_0\|_1^{-\frac{1-w/2}{1+w/2}} \right),$$

the generalized lasso estimate $\hat{\theta}$ satisfies

$$\frac{\|\hat{\theta} - \theta_0\|_2^2}{n} = O_{\mathbb{P}}\left(\frac{\mathrm{nullity}(\Delta)}{n} + \frac{K^{\frac{2}{1+w/2}}}{n} \cdot \|\Delta\theta_0\|_1^{\frac{w}{1+w/2}} \right).$$

The main motivation for bounds of the form (6.15) is that they follow from entropy bounds on the set $\mathcal{S}_\Delta(1)$. Recall that for a set S, the covering number $N(\delta, S, \|\cdot\|)$ is the fewest number of balls of radius δ that cover S, with respect to the norm $\|\cdot\|$. The log covering or entropy number is $\log N(\delta, S, \|\cdot\|)$. In the next result, we make the connection between between entropy and fractional bounds precise; this follows closely from Lemma 3.5 of van de Geer [230].

Theorem 6.10. *Suppose that there exist a constant $w < 2$ such that for n large enough,*

$$\log N(\delta, \mathcal{S}_\Delta(1), \|\cdot\|_2) \leq E\Big(\frac{\sqrt{n}}{\delta}\Big)^w, \tag{6.16}$$

for $\delta > 0$, where E can depend on n. Then the fractional bound in (6.15) *holds with $K = \sqrt{E} n^{w/4}$, and as a result, for*

$$\lambda = \Theta\left(E^{\frac{1}{1+w/2}} n^{\frac{w/2}{1+w/2}} \|\Delta\theta_0\|_1^{-\frac{1-w/2}{1+w/2}} \right),$$

the generalized lasso estimate $\hat{\theta}$ has average squared error

$$\frac{\|\hat{\theta} - \theta_0\|_2^2}{n} = O_{\mathbb{P}}\left(\frac{\mathrm{nullity}(\Delta)}{n} + E^{\frac{1}{1+w/2}} n^{-\frac{1}{1+w/2}} \cdot \|\Delta\theta_0\|_1^{\frac{w}{1+w/2}} \right).$$

To make use of the result in Theorem 6.10, we must obtain an entropy bound as in (6.16), on the set $\mathcal{S}_\Delta(1)$. The literature on entropy numbers is rich, and there are various methods for computing entropy bounds, any of which can be used for these purposes as long as the bounds fit the form of (6.16), as required by the theorem. For bounding the entropy of a set like $\mathcal{S}_\Delta(1)$, two common techniques are to use a characterization of the spectral decay of $\Delta^\dagger$, or an analysis of the correlations between columns of $\Delta^\dagger$. For a nice review of such strategies and their applications, we refer the reader to Section 6 of van de Geer and Lederer [231] and Section 14.12 of Buhlmann and van de Geer [42]. We do not pursue either of these two strategies in the current paper. We instead consider a third, somewhat more transparent strategy, based on a covering number bound of the columns of $\Delta^\dagger$.

Lemma 6.11. *Let $g_1, \dots g_r$ denote the "atoms" associated with the operator Δ, i.e., the columns of $\Delta^\dagger$, and let $\mathcal{G} = \{\pm g_i : i = 1, \dots r\}$ denote the symmetrized set of atoms. Suppose that there exists constants ζ, C_0 with the following property: for each $j = 1, \dots 2r$, there is an arrangement of j balls having radius at most*

$$C_0 \sqrt{n} j^{-1/\zeta},$$

with respect to the norm $\|\cdot\|_2$, that covers $\mathcal{G}$. Then the entropy bound in (6.16) *is met with $w = 2\zeta/(2+\zeta)$ and $E = O(1)$. Therefore, the generalized lasso estimate $\hat{\theta}$, with*

$$\lambda = \Theta\left(n^{\frac{\zeta}{2+2\zeta}} \|\Delta\theta_0\|_1^{-\frac{1}{1+\zeta}} \right),$$

satisfies

$$\frac{\|\hat{\theta} - \theta_0\|_2^2}{n} = O_{\mathbb{P}}\left(\frac{\text{nullity}(\Delta)}{n} + n^{-\frac{2+\zeta}{2+2\zeta}} \cdot \|\Delta\theta_0\|_1^{\frac{\zeta}{1+\zeta}} \right).$$

The entropy-based results in this subsection (Lemma 6.9, Theorem 6.10, and Lemma 6.11) may appear more complex than those involving incoherence in the previous subsection (Lemma 6.5 and Theorem 6.6). Indeed, the same can be said of their proofs, which can be found in the Section 9.7. But after all this entropy machinery has all been established, it can actually be remarkably easy to use, say, Lemma 6.11 to produce sharp results. We conclude by giving an example.

Corollary 6.12. *Consider the 1d fused lasso, i.e., the GTF estimator with $k = 0$ over a chain graph. In this case, we have $\Delta = D^{(1)}$, the univariate difference operator, and the symmetrized set $\mathcal{G}$ of atoms can be covered by j balls with radius at most $\sqrt{2n/j}$, for $j = 1, \dots 2(n-1)$. Hence, with $\lambda = \Theta(n^{1/3} \|D^{(1)}\theta_0\|_1^{-1/3})$, the 1d fused lasso estimate $\hat{\theta}$ satisfies*

$$\frac{\|\hat{\theta} - \theta_0\|_2^2}{n} = O_{\mathbb{P}}\left(n^{-2/3} \cdot \|D^{(1)}\theta_0\|_1^{2/3} \right).$$

This corollary rederives the optimal convergence rate of $n^{-2/3}$ for the univariate fused lasso, assuming boundedness of $\|D^{(1)}\theta_0\|_1$, as has been already shown in Mammen and van de Geer [152], Tibshirani [221]. Like Corollary 6.7 (but unlike previous works), its proof does not rely on any special facts about 1d functions of bounded variation. It only uses a covering number bound on the columns of the operator $(D^{(1)})^+$, a strategy that, in principle, extends to many other settings (graphs). It is worth emphasizing just how simple this covering number construction is, compared to the incoherence-based arguments that lead to the same result; we invite the curious reader to compare the proofs of Corollaries 6.7 and 6.12.

6.7 Discussion

In this work, we proposed graph trend filtering as a useful alternative to Laplacian and wavelet smoothers on graphs. This is analogous to the usefulness of univariate trend filtering in nonparametric regression, as an alternative to smoothing splines and wavelets [221]. We have documented empirical evidence for the superior local adaptivity of the ℓ_1-based GTF over the ℓ_2-based graph Laplacian smoother, and the superior robustness of GTF over wavelet smoothing in high-noise scenarios. Our theoretical analysis provides a basis for a deeper understanding of the estimation properties of GTF. More precise theoretical characterizations involving entropy will be the topic of future work, as will comparisons between the error rates achieved by GTF and other common estimators, such as Laplacian smoothing. These extensions, and many others, are well within reach.

6.8 Additional Analysis from Alternative Wavelet Designs

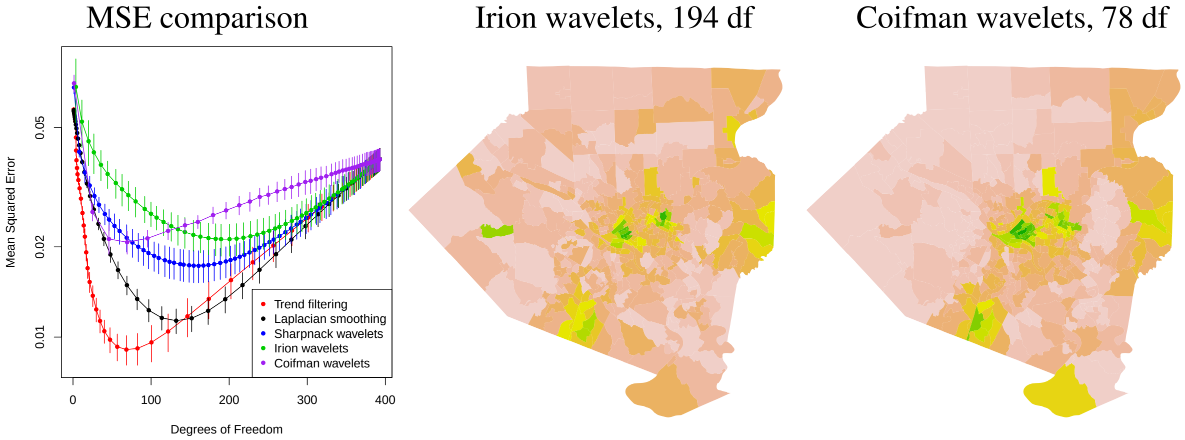
We provide detailed comparisons to a few recently proposed wavelet approaches for graph smoothing.

6.8.1 Allegheny County Example

In addition to considering the wavelet design of Sharpnack et al. [195] for the Allegheny County example, we also considered designs of Coiman and Maggioni [58]—a classic method that builds diffusion wavelets on a graph, and Irion [120]—a more recent graph wavelet construction. In contrast to Sharpnack et al. [195], which produces a single signal-independent orthogonal basis for a graph, both Coiman and Maggioni [58], Irion [120] build wavelet packets from a given graph structure. A wavelet packet is an overcomplete basis indexed by a hierarchical data structure that can be used to generate an exponential number of orthogonal bases. This construction is computationally expensive as it typically involves computing eigendecompositions of large matrices. Once the wavelet packet has been constructed, for each input signal that one observes over the graph in question, one runs a "best basis" algorithm to choose a particular orthogonal basis from the wavelet packet by optimizing a particular cost function of the eventual wavelet coefficients. This is based on a message-passing-like dynamic programming algorithm, and can be quite efficient. Lastly, the denoising procedure is defined as usual (e.g., as in Donoho and Johnstone [69]), namely, one performs the basis transformation, soft-thresholds (or hard-thresholds) the coefficients, and then reconstructs the denoised signal.

In our experiments, we used the wavelet implementations released by the authors of Coiman and Maggioni [58], Irion [120] with their default settings. In particular, the former implementation of Coiman and Maggioni [58] builds wavelets from a diffusion operator constructed from the adjacency matrix of a graph, and the cost function for the best basis is defined by the ℓ_1 norm of the wavelet coefficients. The latter implementation of Irion [120] uses a more exhaustive search, building wavelet packets through a hierarchical partitioning and eigentransform of three different Laplacian matrices and a fourth generalized Haar-Walsh transform (GHWT), then choosing the best basis from all four collections by optimizing a meta cost function of the ℓ_p norm of wavelet coefficients over $p \in \{0.1, 0.2, \ldots 2\}$. This is the "cumulative relative error" defined in equation (7.5) of Irion [120].

In the left panel of Figure 6.8, we plot the mean squared errors for these new wavelet methods over the same 10 simulations from the Allegheny County example in Figure 6.2 of Section 6.8.1. The middle and right panels of the figure show the denoised signals from the new methods fit to the data in Figure 6.1, at their optimal degrees of freedom (df) values (in terms of the achieved MSE). We can see that the spanning tree wavelet design of Sharpnack et al. [195] is the best performer among the three candidate wavelet designs. In a rough sense, the construction of Irion [120] seems to perform similarly to that of Sharpnack et al. [195], in that the MSE is best for larger df values (corresponding to more nonzero wavelet coefficients, i.e., complex fitted models), whereas the construction of Coiman and Maggioni [58] performs best for smaller df values (fewer nonzero wavelet coefficients, i.e., simpler fitted models).


MSE comparison

Irion wavelets, 194 df

Coifman wavelets, 78 df

Figure 6.8: Additional wavelet analysis of the Allegheny County example.

6.8.2 Facebook Graph Example

Again, we consider the designs of Coiman and Maggioni [58], Irion [120] for the Facebook graph example of Section 6.5.1. Due to practical reasons, we had to change some of the default settings in the implementations provided by the authors of these wavelet methods; in the wavelet implementation of Coiman and Maggioni [58], we took the power of the diffusion operator to be 1 instead of 4 (since the latter choice threw an error in the provided code); and in the wavelet implementation of Irion [120], we used another "best basis" algorithm that only searches within the basis collection from the GHWT eigendecomposition, as the original algorithm was too slow due to the larger scale considered in this example. (In most examples in Irion [120], the chosen bases come from the GHWT eigendecomposition.) We view these changes as minor, because when the same changes were applied to the methods of Coiman and Maggioni [58], Irion [120] on the smaller Allegheny County example, there are no obvious differences in the results.

Figure 6.9 shows the results for the two new wavelet methods on the Facebook graph simulation, using the same setup as in Figure 6.5. Once again, we find that the spanning tree wavelets of Sharpnack et al. [195] perform better or on par with the other two wavelet methods across essentially all scenarios.

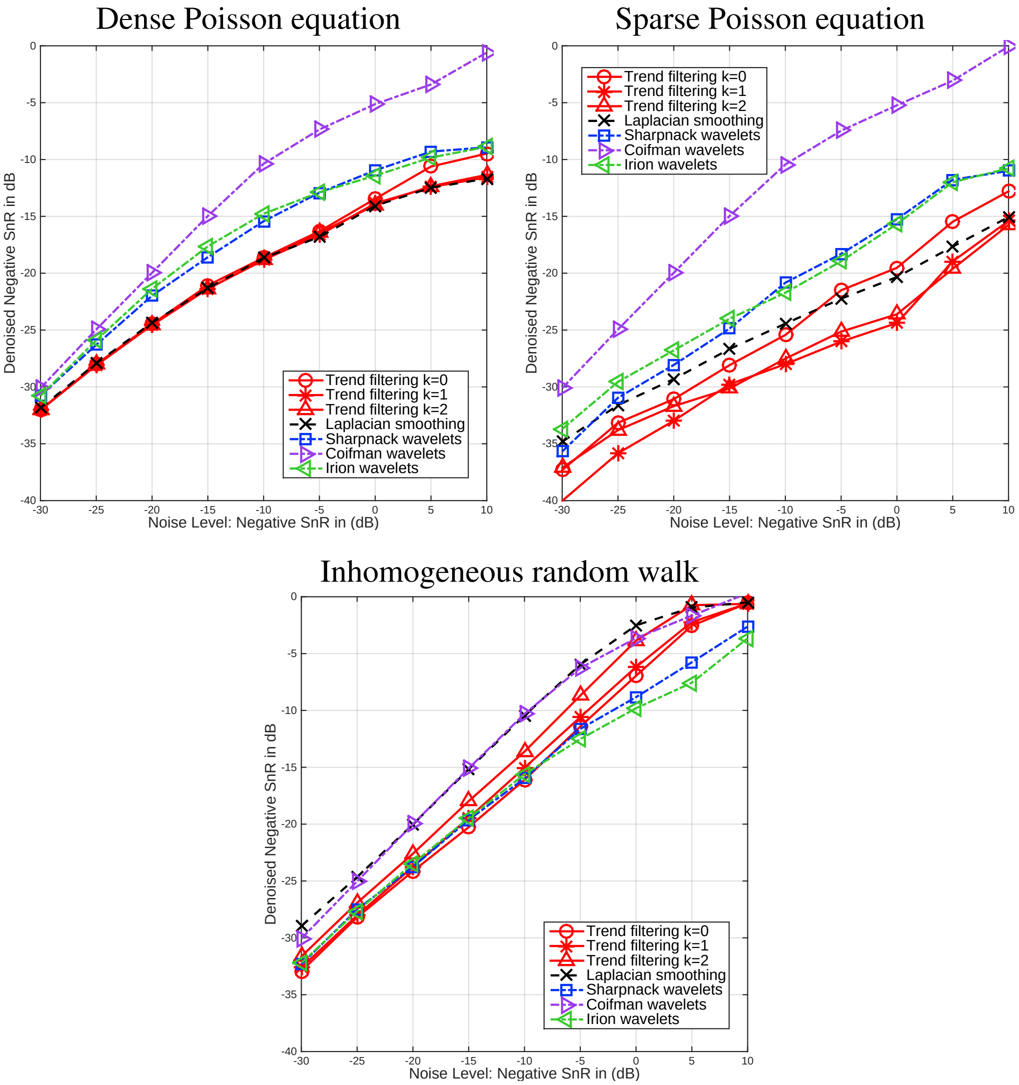


Figure 6.9: Additional wavelet analysis of the Facebook graph example.

6.9 Proofs of Theoretical Results

Here we present proofs of our theoretical results presented in Sections 6.3 and 6.6.

6.9.1 Proof of Lemma 6.1

For even k, we have $\Delta^{(k+1)} = DL^{k/2}$, so if A denotes a subset of edges, then $\Delta^{(k+1)}_{-A} = D_{-A}L^{k/2}$. Recall that for a connected graph, $\mathrm{null}(L) = \mathrm{span}\{\mathbb{1}\}$, and the same is true for any power of L. This means that we can write

$$\mathrm{null}(\Delta^{(k+1)}) = \mathrm{span}\{\mathbb{1}\} + \mathrm{span}\{\mathbb{1}\}^{\perp} \cap \{u : DL^{\frac{k}{2}}u = 0\}.$$

Note that if $\mathbb{1}^\top u = 0$, then $v = L^{\frac{k}{2}}u \iff u = (L^\dagger)^{\frac{k}{2}}u$. Moreover, if G_{-A} has connected components $C_1, \dots C_s$, then $\mathrm{null}(D_{-A}) = \mathrm{span}\{\mathbb{1}_{C_1}, \dots \mathbb{1}_{C_s}\}$. Putting these statements together proves the result for even k. For k odd, the arguments are similar.

6.9.2 Proof of Theorem 6.3

By assumption we can write

$$y = \theta_0 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I).$$

Denote $R = \mathrm{row}(\Delta)$, the row space of Δ, and $R^\perp = \mathrm{null}(\Delta)$, the null space of Δ. Also let P_R be the projection onto R, and $P_{R^\perp}$ the projection onto $R^\perp$. Consider

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \frac{1}{2}\|y - \theta\|_2^2 + \lambda\|\Delta\theta\|_1,$$
$$\tilde{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \frac{1}{2}\|P_R y - \theta\|_2^2 + \lambda\|\Delta\theta\|_1.$$

The first quantity $\hat{\theta} \in \mathbb{R}^n$ is the estimate of interest, the second one $\tilde{\theta} \in R$ is easier to analyze. Note that

$$\hat{\theta} = P_{R^\perp} y + \tilde{\theta},$$

and write $\|x\|_R = \|P_R x\|_2$, $\|x\|_{R^\perp} = \|P_{R^\perp} x\|_2$. Then

$$\|\hat{\theta} - \theta_0\|_2^2 = \|\epsilon\|_{R^\perp}^2 + \|\tilde{\theta} - \theta_0\|_R^2.$$

The first term is on the order $\dim(R^\perp) = \mathrm{nullity}(\Delta)$, and it suffices to bound the second term. Now we establish a basic inequality for $\tilde{\theta}$. By optimality of $\tilde{\theta}$, we have

$$\frac{1}{2}\|y - \tilde{\theta}\|_R^2 + \lambda\|\Delta\tilde{\theta}\|_1 \le \frac{1}{2}\|y - \theta_0\|_R^2 + \lambda\|\Delta\theta_0\|_1,$$

and after rearranging terms,

$$\|\tilde{\theta} - \theta_0\|_R^2 \le 2\epsilon^\top P_R(\tilde{\theta} - \theta_0) + 2\lambda\|\Delta\theta_0\|_1 - 2\lambda\|\Delta\tilde{\theta}\|_1. \tag{6.17}$$

This is our basic inequality. In the first term above, we use $P_R = \Delta^\dagger\Delta$, and apply Holder's inequality:

$$\epsilon^\top \Delta^\dagger \Delta(\tilde{\theta} - \theta_0) \le \|(\Delta^\dagger)^\top \epsilon\|_\infty \|\Delta(\tilde{\theta} - \theta_0)\|_1. \tag{6.18}$$

If $\lambda \geq \|(\Delta^\dagger)^\top \epsilon\|_\infty$, then from (6.17), (6.18), and the triangle inequality, we see that

$$\|\tilde{\theta} - \theta_0\|_R^2 \leq 4\lambda \|\Delta\theta_0\|_1.$$

Well, $\|(\Delta^\dagger)^\top \epsilon\|_\infty = O_{\mathbb{P}}(M\sqrt{\log r})$ by a standard result on the maximum of Gaussians (derived using the union bound, and Mills' bound on the Gaussian tail), where recall M is the maximum ℓ_2 norm of the columns of $\Delta^\dagger$. Thus with $\lambda = \Theta(M\sqrt{\log r})$, we have from the above that

$$\|\tilde{\theta} - \theta_0\|_R^2 = O_{\mathbb{P}}\big(M\sqrt{\log r}\|\Delta\theta_0\|_1\big),$$

as desired.

6.9.3 Proof of Corollary 6.4

Case 1. When $\hat{\theta}$ is the univariate trend filtering estimator of order k, we are considering a penalty matrix $\Delta = D^{(k+1)}$, the univariate difference operator of order $k+1$. Note that $D^{(k+1)} \in \mathbb{R}^{(n-k-1)\times n}$, and its null space has constant dimension $k+1$. We show in Lemma 6.13 of Section 9.7 6.9.4 that $(D^{(k+1)})^\dagger = P_R H_2^{(k)}/k!$, where $R = \mathrm{row}(D^{(k+1)})$, and $H_2^{(k)} \in \mathbb{R}^{n\times(n-k-1)}$ contains the last $n-k-1$ columns of the order k falling factorial basis matrix [244], evaluated over the input points $x_1 = 1, \ldots x_n = n$. The largest column norm of $P_R H_2^{(k)}/k!$ is on the order of $n^{k+1/2}$, which proves the result.

Cases 2 and 3. When G is the Ramanujan d-regular graph, the number of edges in the graph is $O(nd)$. The operator $\Delta = \Delta^{(k+1)}$ has number of rows $r = n$ when k is odd and $r = O(nd)$ when k is even; overall this is $O(nd)$. The dimension of the null space of Δ is constant (it is in fact 1, since the graph is connected). When G is the Erdos-Renyi random graph, the same bounds apply to the number of rows and the dimension of the null space, except that the bounds become probabilistic ones.

Now we apply the crude inequality, with e_i, $i = 1, \ldots r$ denoting the standard basis vectors,

$$M = \max_{i=1,\ldots r} \Delta^\dagger e_i \leq \max_{\|x\|_2 \leq 1} \Delta^\dagger x = \|\Delta^\dagger\|_2,$$

the right-hand side being the maximum singular value of $\Delta^\dagger$. As $\Delta = \Delta^{(k+1)}$, the graph difference operator of order $k+1$, we claim that

$$\|\Delta^\dagger\|_2 \leq 1/\lambda_{\min}(L)^{\frac{k+1}{2}}, \qquad (6.19)$$

where $\lambda_{\min}(L)$ denotes the smallest nonzero eigenvalue of the graph Laplacian L. To see this, note first that $\|\Delta^\dagger\|_2 = 1/\sigma_{\min}(\Delta)$, where the denominator is the smallest nonzero singular value of Δ. Now for odd k, we have $\Delta^{(k+1)} = L^{(k+1)/2}$, and the claim follows as

$$\sigma_{\min}(L^{\frac{k+1}{2}}) = \min_{x\in R: \|x\|_2 \leq 1} \|L^{\frac{k+1}{2}} x\|_2 \geq \big(\sigma_{\min}(L)\big)^{\frac{k+1}{2}},$$

and $\sigma_{\min}(L) = \lambda_{\min}(L)$, since L is symmetric. Above, R denotes the row space of L (the space orthogonal to the vector $\mathbb{1}$ of all 1s). For even k, we have $\Delta^{(k+1)} = DL^{k/2}$, and again

$$\sigma_{\min}(DL^{\frac{k}{2}}) = \min_{x \in R: \|x\|_2 \leq 1} \|DL^{\frac{k+1}{2}} x\|_2 \geq \sigma_{\min}(D)\big(\sigma_{\min}(L)\big)^{\frac{k}{2}},$$

where $\sigma_{\min}(D) = \sqrt{\lambda_{\min}(L)}$, since $D^\top D = L$. This verifies the claim.

Having established (6.19), it suffices to lower bound $\lambda_{\min}(L)$ for the two graphs in question. Indeed, for both graphs, we have the lower bound

$$\lambda_{\min}(L) = \Omega(d - \sqrt{d}).$$

e.g., see Lubotzky et al. [150], Marcus et al. [153] for the Ramanujan graph and Chung and Radcliffe [57], Feige and Ofek [91] for the Erdos-Renyi graph. This completes the proof.

6.9.4 Calculation of $(D^{(k+1)})^\dagger$

Lemma 6.13. *The $(k+1)$st order discrete difference operator has pseudoinverse*

$$(D^{(k+1)})^\dagger = P_R H_2^{(k)}/k!,$$

where we denote $R = \mathrm{row}(D^{(k+1)})$, and $H_2^{(k)} \in \mathbb{R}^{n \times (n-k-1)}$ the last $n-k-1$ columns of the kth order falling factorial basis matrix.

Proof. We abbreviate $D = D^{(k+1)}$, and consider the linear system

$$DD^\top x = Db \tag{6.20}$$

in x, where $b \in \mathbb{R}^n$ is arbitrary. We seek an expression for $x = (DD^\top)^{-1}D^\top = (D^\dagger)^\top b$, and this will tell us the form of $D^\dagger$. Define

$$\tilde{D} = \begin{bmatrix} C \\ D \end{bmatrix} \in \mathbb{R}^{n \times n},$$

where $C \in \mathbb{R}^{(k+1)\times n}$ is the matrix that collects the first row of each lower order difference operator, defined in Lemma 2 of Wang et al. [244]. From this same lemma, we know that

$$\tilde{D}^{-1} = H/k!,$$

where $H = H^{(k)}$ is falling factorial basis matrix of order k, evaluated over $x_1, \ldots x_n$. With this in mind, consider the expanded linear system

$$\begin{bmatrix} CC^\top & CD^\top \\ DC^\top & DD^\top \end{bmatrix} \begin{bmatrix} w \\ x \end{bmatrix} = \begin{bmatrix} a \\ Db \end{bmatrix}. \tag{6.21}$$

The second equation reads

$$DC^\top w + DD^\top x = Db,$$

and so if we can choose a in (6.21) so that at the solution we have $w = 0$, then x is the solution in (6.20). The first equation in (6.21) reads

$$CC^\top w + CD^\top x = a,$$

i.e.,

$$w = (CC^\top)^{-1}(a - CD^\top x).$$

That is, we want to choose

$$a = CD^\top x = CD^\top (DD^\top)^{-1} Db = CP_R b,$$

where P_R is the projection onto row space of D. Thus we can reexpress (6.21) as

$$\tilde{D}\tilde{D}^\top \begin{bmatrix} w \\ x \end{bmatrix} = \begin{bmatrix} CP_R b \\ Db \end{bmatrix} = \tilde{D} P_R b$$

and, using $\tilde{D}^{-1} = H/k!$,

$$\begin{bmatrix} w \\ x \end{bmatrix} = H^\top P_R b/k!.$$

Finally, writing H_2 for the last $n-k-1$ columns of H, we have $x = H_2^\top P_R b/k!$, as desired. □

Remark. The above proof did not rely on the input points $x_1, \dots x_n$; indeed, the result holds true for any sequence of inputs used to define the discrete difference matrix and falling factorial basis matrix.

6.9.5 Proof of Lemma 6.5

We follow the proof of Theorem 6.3, up until the application of Holder's inequality in (6.18). In place of this step, we use the linearized bound in (6.14), which we claim implies that

$$\epsilon^\top P_R(\tilde{\theta} - \theta_0) \le \tilde{B} \|\tilde{\theta} - \theta_0\|_R + A \|\Delta(\tilde{\theta} - \theta_0)\|_1,$$

where $\tilde{B} = O_{\mathbb{P}}(B)$. This simply follows from applying (6.14) to $x = P_R(\tilde{\theta} - \theta_0)/\|\Delta(\tilde{\theta} - \theta_0)\|_1$, which is easily seen to be an element of $\mathcal{S}_\Delta(1)$. Hence we can take take $\lambda = \Theta(A)$, and argue as in the proof of Theorem 6.3 to arrive at

$$\|\tilde{\theta} - \theta_0\|_R^2 \le \tilde{B} \|\tilde{\theta} - \theta_0\|_R + \tilde{A} \|\Delta\theta_0\|_1,$$

where $\tilde{A} = O_{\mathbb{P}}(A)$. Note that the above is a quadratic inequality of the form $ax^2 - bx - c \le 0$ with $x = \|\tilde{\theta} - \theta_0\|_R$. As $a > 0$, the larger of its two roots serves as a bound for x, i.e., $x \le (b + \sqrt{b^2 + 4ac})/(2a) \le b/a + \sqrt{c/a}$, or $x^2 \le 2b^2/a^2 + 2c/a$, which means that

$$\|\tilde{\theta} - \theta_0\|_R^2 \le 2\tilde{B}^2 + 2\tilde{A}\|\Delta\theta_0\|_1 = O_{\mathbb{P}}\big(B^2 + A\|\Delta\theta_0\|_1\big),$$

completing the proof.

6.9.6 Proof of Theorem 6.6

For an index $i_0 \in \{1, \ldots q\}$, let

$$C = \mu \sqrt{\frac{2\log 2r}{n} \sum_{i=i_0+1}^{q} \frac{1}{\xi_i^2}}.$$

We will show that

$$\max_{x \in \mathcal{S}_\Delta(1)} \frac{\epsilon^\top x - 1.001\sigma C}{\|x\|_2} = O_{\mathbb{P}}(\sqrt{i_0}).$$

Invoking Lemma 6.5 with $A = 1.001\sigma C$ and $b = \sqrt{i_0}$ would then give the result.

Henceforth we denote $[i] = \{1, \ldots i\}$. Recall that $q = \mathrm{rank}(\Delta)$. Let the singular value decomposition of Δ be

$$\Delta = U\Sigma V^\top,$$

where $U \in \mathbb{R}^{r\times q}$, $V \in \mathbb{R}^{n\times q}$ are orthogonal, and $\Sigma \in \mathbb{R}^{q\times q}$ has diagonal elements $(\Sigma)_{ii} = \xi_i > 0$ for $i \in [q]$. First, let us establish that

$$\Delta^\dagger = V\Sigma^{-1}U^\top.$$

Consider an arbitrary point $x = P_R z \in \mathcal{S}_\Delta(1)$. Denote the projection $P_{[i_0]} = V_{[i_0]}V_{[i_0]}^\top$ where $V_{[i_0]}$ contains the first i_0 right singular vectors. We can decompose

$$\epsilon^\top P_R z = \epsilon^\top P_{[i_0]} P_R z + \epsilon^\top (I - P_{[i_0]}) P_R z.$$

The first term can be bounded by

$$\epsilon^\top P_{[i_0]} P_R z \le \|P_{[i_0]}\epsilon\|_2 \|z\|_R = O_{\mathbb{P}}(\sqrt{i_0}\|z\|_R),$$

using the fact that $\|P_{[i_0]}\epsilon\|_2^2 \overset{d}{=} \sum_{i=1}^{i_0} \epsilon_i^2$. We can bound the second term by

$$\epsilon^\top (I - P_{[i_0]}) P_R z = \epsilon^\top (I - P_{[i_0]}) \Delta^\dagger \Delta z \le \|(\Delta^\dagger)^\top (I - P_{[i_0]})\epsilon\|_\infty,$$

using $P_R = \Delta^\dagger\Delta$, Holder's inequality, and the fact that $\|\Delta z\|_1 \le 1$. Define $g_j = (I - P_{[i_0]})\Delta^\dagger e_j$ for $j \in [r]$ with e_j the jth canonical basis vector. So,

$$\|g_j\|_2^2 = \|[\,0 \;\; V_{[n]\setminus[i_0]}\,] \cdot \Sigma^{-1} U^\top e_j\|_2^2 \le \frac{\mu^2}{n} \sum_{i=i_0+1}^{q} \frac{1}{\xi_i^2},$$

by rotational invariance of $\|\cdot\|_2$ and the incoherence assumption on the columns of U. By a standard result on maxima of Gaussians,

$$\|(\Delta^\dagger)^\top (I - P_{[i_0]})\epsilon\|_\infty = \max_{j\in[r]} |g_j^\top \epsilon| \le 1.001\sigma \sqrt{2\log(2r)\frac{\mu^2}{n} \sum_{i=i_0+1}^{q} \frac{1}{\xi_i^2}} = 1.001\sigma C,$$

with probability approaching 1. Putting these two terms together completes the proof, as we have shown that

$$\frac{\epsilon^\top P_R z - 1.001\sigma C}{\|z\|_R} = O_{\mathbb{P}}(\sqrt{i_0}),$$

with the probability bound on the right-hand side not depending on z.

6.9.7 Proof of Corollary 6.7

We focus on the k odd and k even cases separately.

Case for k odd. When k is odd, we have $\Delta = \Delta^{(k+1)} = L^{(k+1)/2}$, where L the graph Laplacian of a chain graph (i.e., 1d grid graph), to be perfectly explicit,

$$L = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 \\ \vdots & & \ddots & \ddots & \ddots & \\ 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & \dots & 0 & -1 & 1 \end{bmatrix}.$$

In numerical methods for differential equations, this matrix L is called the finite difference operator for the 1d Laplace equation with Neumann boundary conditions [e.g., 59, 102], and is known to have eigenvalues and eigenvectors

$$\xi_i = 4\sin^2\Big(\frac{\pi(i-1)}{2n}\Big), \quad \text{for } i = 1, \dots n,$$
$$u_{ij} = \begin{cases} \frac{1}{\sqrt{n}} & \text{if } i = 1 \\ \sqrt{\frac{2}{n}}\cos\Big(\frac{\pi(i-1)(j-1/2)}{n}\Big) & \text{otherwise} \end{cases}, \quad \text{for } i, j = 1, \dots n.$$

Therefore, the eigenvectors of L are incoherent with constant $\mu = \sqrt{2}$. This of course implies the same of $L^{(k+1)/2}$, which shares the eigenvectors of L. Meanwhile, the eigenvalues of $L^{(k+1)/2}$ are just given by raising those of L to the power of $(k+1)/2$, and for $i_0 \in \{1, \dots n\}$, we compute the partial sum of their squared reciprocals, as in

$$\frac{1}{n}\sum_{i=i_0+1}^{n} \frac{1}{\xi_i^{k+1}} = \frac{1}{n}\sum_{i=i_0+1}^{n} \frac{1}{4^{k+1}\sin^{2k+2}(\pi(i-1)/(2n))} \le \int_{(i_0-1)/n}^{(n-2)/n} \frac{1}{4^{k+1}\sin^{2k+2}(\pi x/2)} dx,$$

where we have used the fact that the right-endpoint Riemann sum, for a monotone nonincreasing function, is an underestimate of its integral. Continuing on, the above integral can be bounded by

$$\frac{1}{4^{k+1}\sin^{2k}(\pi i_0/(2n))}\int_{(i_0-1)/n}^{1} \frac{1}{\sin^2(\pi x/2)} dx = \frac{2\cot(\pi i_0/(2n))}{4^{k+1}\pi\sin^{2k}(\pi i_0/(2n))} \le \frac{1}{4^{k+1}\pi}\Big(\frac{2n}{\pi i_0}\Big)^{2k+1},$$

the last step using a Taylor expansion around 0. Hence to choose a tight a bound as possible in Theorem 6.6, we seek to balance i_0 with $\sqrt{(n/i_0)^{2k+1}\log n} \cdot \|\Delta^{(k+1)}\theta_0\|_1$. This is accomplished by choosing

$$i_0 = n^{\frac{2k+1}{2k+3}}(\log n)^{\frac{1}{2k+3}}\|\Delta^{(k+1)}\theta_0\|_1^{\frac{2}{2k+3}},$$

and applying Theorem 6.6 gives the result for k odd.

Case for k even. When k is even, we instead have $\Delta = \Delta^{(k+1)} = DL^{k/2}$, where D is the edge incidence matrix of a 1d chain, and $L = D^\top D$. It is clear that the left singular vectors of $DL^{k/2}$ are simply the left singular vectors of D, or equivalently, the eigenvectors of $DD^\top$. To be explicit,

$$DD^\top = \begin{bmatrix} 2 & -1 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 \\ \vdots & & \ddots & \ddots & \ddots & \\ 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & \dots & 0 & -1 & 2 \end{bmatrix},$$

which is called the finite difference operator associated with the 1d Laplace equation under Dirichlet boundary conditions in numerical methods [e.g., 59, 102], and is known to have eigenvectors

$$u_{ij} = \sqrt{\frac{2}{n}} \sin\Big(\frac{\pi ij}{n}\Big), \quad \text{for } i,j = 1, \dots n-1.$$

It is evident that these vectors are incoherent, with constant $\mu = \sqrt{2}$. Furthermore, the singular values of $DL^{k/2}$ are exactly the eigenvalues of L raised to the power of $(k+1)/2$, and the remainder of the proof goes through as in the k odd case.

6.9.8 Proof of Corollary 6.8

Again we treat the k odd and even cases separately.

Case for k odd. As k is odd, the GTF operator is $\Delta = \Delta^{(k+1)} = L^{(k+1)/2}$, where the L is the Laplacian matrix of a 2d grid graph. Writing $L_{\mathrm{1d}} \in \mathbb{R}^{\ell\times\ell}$ for the Laplacian matrix over a 1d grid of size $\ell = \sqrt{n}$ (and $I \in \mathbb{R}^{\ell\times\ell}$ for the identity matrix), we note that

$$L = I \otimes L_{\mathrm{1d}} + L_{\mathrm{1d}} \otimes I,$$

i.e., the 2d grid Laplacian L is the Kronecker sum of the 1d grid Laplacian L_{1d}, so its eigenvectors are given by all pairwise Kronecker products of eigenvectors of L_{1d}, of the form $u_i \otimes u_j$. Moreover, it is not hard to see that each $u_i \otimes u_j$ has unit norm (since u_i, u_j do) and $\|u_i \otimes u_j\|_\infty \le 2/\sqrt{n}$. This allows us to conclude that the eigenvectors of L obey the incoherence property with $\mu = 2$.

The eigenvalues of L are given by all pairwise sums of eigenvalues in the 1d case. Indexing by 2d grid coordinates, we may write these as

$$\xi_{j_1,j_2} = 4\sin^2\Big(\frac{\pi(j_1-1)}{2\ell}\Big) + 4\sin^2\Big(\frac{\pi(j_2-1)}{2\ell}\Big), \quad \text{for } j_1, j_2 = 1, \dots \ell.$$

Eigenvalues of $L^{(k+1)/2}$ are just given by raising the above to the power of $(k+1)/2$, and for $j_0 \in \{1, \dots \ell\}$, we let $i_0 = j_0^2$, and compute the sum

$$\frac{1}{n} \sum_{\max\{j_1 j_2\} > j_0} \frac{1}{\xi_{j_1,j_2}^{k+1}} \le \frac{2}{n} \sum_{j_1=j_0+1}^{\ell} \sum_{j_2=1}^{\ell} \frac{1}{\xi_{j_1,j_2}^{k+1}} \le \frac{2}{\ell} \sum_{j_1=j_0+1}^{\ell} \frac{1}{4^{k+1}\sin^{2k+2}(\pi(j_1-1)/(2\ell))}.$$

Just as we argued in the 1d case (for k odd), the above is bounded by

$$\frac{2}{4^{k+1}\pi}\left(\frac{2\ell}{\pi j_0}\right)^{2k+1},$$

and thus we seek to balance $i_0 = j_0^2$ with $\sqrt{(\ell/j_0)^{2k+1}\log n} \cdot \|\Delta^{(k+1)}\theta_0\|_1$. This yields

$$j_0 = \ell^{\frac{2k+1}{2k+5}}(\log n)^{\frac{1}{2k+5}}\|\Delta^{(k+1)}\theta_0\|_1^{\frac{2}{2k+5}},$$

i.e.,

$$i_0 = n^{\frac{2k+1}{2k+5}}(\log n)^{\frac{2}{2k+5}}\|\Delta^{(k+1)}\theta_0\|_1^{\frac{4}{2k+5}},$$

and applying Theorem 6.6 gives the result for k odd.

Case for k even. For k even, we have the GTF operator being $\Delta = \Delta^{(k+1)} = DL^{k/2}$, where D is the edge incidence matrix of a 2d grid, and $L = D^\top D$. It will be helpful to write

$$D = \begin{bmatrix} I \otimes D_{\mathrm{1d}} \\ D_{\mathrm{1d}} \otimes I \end{bmatrix},$$

where $D_{\mathrm{1d}} \in \mathbb{R}^{(\ell-1)\times\ell}$ is the difference operator for a 1d grid of size $\ell = \sqrt{n}$ (and $I \in \mathbb{R}^{\ell\times\ell}$ is the identity matrix). It suffices to check the incoherence of the left singular vectors of $DL^{k/2}$, since the eigenvalues of $DL^{k/2}$ are those of L raised to the power of $(k+1)/2$, and so the rest of the proof then follows precisely as in the case when k is odd. The left singular vectors of $DL^{k/2}$ are the same as the left singular vectors of D, which are the eigenvectors of $DD^\top$. Observe that

$$DD^\top = \begin{bmatrix} I \otimes D_{\mathrm{1d}}D_{\mathrm{1d}}^\top & D_{\mathrm{1d}}^\top \otimes D_{\mathrm{1d}} \\ D_{\mathrm{1d}} \otimes D_{\mathrm{1d}}^\top & D_{\mathrm{1d}}D_{\mathrm{1d}}^\top \otimes I \end{bmatrix}.$$

Let u_i, $i = 1, \ldots \ell-1$ be the eigenvectors of $D_{\mathrm{1d}}D_{\mathrm{1d}}^\top$, corresponding to eigenvalues λ_i, $i = 1, \ldots \ell-1$. Define $v_i = D_{\mathrm{1d}}^\top u_i/\sqrt{\lambda_i}$, $i = 1, \ldots \ell-1$, and $e = \mathbb{1}/\sqrt{\ell}$, where $\mathbb{1} = (1, \ldots 1) \in \mathbb{R}^\ell$ is the vector of all 1s. A straightforward calculation verifies that

$$\begin{aligned} DD^\top \begin{bmatrix} v_i \otimes u_i \\ u_i \otimes v_i \end{bmatrix} &= 2\lambda_i \begin{bmatrix} v_i \otimes u_i \\ u_i \otimes v_i \end{bmatrix}, \quad \text{for } i = 1, \ldots \ell-1, \\ DD^\top \begin{bmatrix} e \otimes u_i \\ 0 \end{bmatrix} &= \lambda_i \begin{bmatrix} e \otimes u_i \\ 0 \end{bmatrix}, \quad \text{for } i = 1, \ldots \ell-1, \\ DD^\top \begin{bmatrix} 0 \\ u_i \otimes e \end{bmatrix} &= \lambda_i \begin{bmatrix} 0 \\ u_i \otimes e \end{bmatrix}, \quad \text{for } i = 1, \ldots \ell-1. \end{aligned}$$

Hence we have derived $3(\ell-1)$ eigenvectors of $DD^\top$. Note that the vectors v_i, $i = 1, \ldots \ell-1$ are actually the eigenvectors of $L_{\mathrm{1d}} = D_{\mathrm{1d}}^\top D_{\mathrm{1d}}$ (corresponding to the $\ell-1$ nonzero eigenvalues), and from our work in the 1d case, recall, both v_i, $i = 1, \ldots \ell-1$ (studied for k odd) and u_i, $i = 1, \ldots \ell-1$ (studied for k even) are unit vectors satisfying the incoherence property with $\mu = \sqrt{2}$. This means that the above eigenvectors are all unit norm, and are also incoherent, with constant $\mu = 2$.

There are $(\ell-1)(\ell-2)$ more eigenvectors of $DD^\top$, as the rank of $DD^\top$ is $n-1=\ell^2-1$. A somewhat longer but still straightforward calculation verifies that

$$DD^\top \begin{bmatrix} v_i \otimes u_j + v_j \otimes u_i \\ \sqrt{\frac{\lambda_i}{\lambda_j}} u_i \otimes v_j + \sqrt{\frac{\lambda_j}{\lambda_i}} u_j \otimes v_i \end{bmatrix} = (\lambda_i + \lambda_j) \begin{bmatrix} v_i \otimes u_j + v_j \otimes u_i \\ \sqrt{\frac{\lambda_i}{\lambda_j}} u_i \otimes v_j + \sqrt{\frac{\lambda_j}{\lambda_i}} u_j \otimes v_i \end{bmatrix}, \text{ for } i<j,$$

$$DD^\top \begin{bmatrix} \sqrt{\frac{\lambda_j}{\lambda_i}} v_i \otimes u_j + \sqrt{\frac{\lambda_i}{\lambda_j}} v_j \otimes u_i \\ u_i \otimes v_j + u_j \otimes v_i \end{bmatrix} = (\lambda_i + \lambda_j) \begin{bmatrix} \sqrt{\frac{\lambda_j}{\lambda_i}} v_i \otimes u_j + \sqrt{\frac{\lambda_i}{\lambda_j}} v_j \otimes u_i \\ u_i \otimes v_j + u_j \otimes v_i \end{bmatrix}, \text{ for } i<j.$$

Modulo the appropriate normalization, we have derived the remaining $(\ell-1)(\ell-2)$ eigenvectors of $DD^\top$. It remains to check their incoherence, once we have normalized them (to have unit norm). As the eigenvectors in the first and second expressions above are simply (block) rearrangements of each other, it does not matter which form we study; consider, say, those in the second expression, and fix $i<j$. The entrywise absolute maximum of the eigenvector in question is at most $\sqrt{\lambda_j/\lambda_i}(4/\sqrt{n})$. Thus it suffices show that the normalization constant for this eigenvector is on the order of $\sqrt{\lambda_j/\lambda_i}$. Observe that

$$\left\| \begin{bmatrix} \sqrt{\frac{\lambda_j}{\lambda_i}} v_i \otimes u_j + \sqrt{\frac{\lambda_i}{\lambda_j}} v_j \otimes u_i \\ u_i \otimes v_j + u_j \otimes v_i \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} \sqrt{\frac{\lambda_j}{\lambda_i}} v_i \otimes u_j \\ u_i \otimes v_j \end{bmatrix} \right\|_2^2 + \left\| \begin{bmatrix} \sqrt{\frac{\lambda_i}{\lambda_j}} v_j \otimes u_i \\ u_j \otimes v_i \end{bmatrix} \right\|_2^2.$$

Here the cross-term is $(v_i^\top \otimes u_j^\top)(v_j \otimes u_i) = (v_i^\top v_j)(u_i^\top u_j) = 0$, as $v_i^\top v_j = 0$ and $u_i^\top u_j = 0$. This means that the normalization constant lies within $[\sqrt{\lambda_j/\lambda_i + 2}, \sqrt{2\lambda_j/\lambda_i + 2}]$. In particular, the lower bound shows that the incoherence property holds with $\mu = 4$. This completes the proof.

6.9.9 Proof of Lemma 6.9

As before, we follow the proof of Theorem 6.3 up until the application of Holder's inequality in (6.18), but we use the fractional bound in (6.15) instead. We claim that this implies

$$\epsilon^\top P_R(\tilde{\theta} - \theta_0) \le \tilde{K} \|\tilde{\theta} - \theta_0\|_R^{1-w/2} (\|\Delta\tilde{\theta}\|_1 + \|\Delta\theta_0\|_1)^{w/2},$$

where $\tilde{K} = O_{\mathbb{P}}(K)$. This is verified by noting that $x = P_R(\tilde{\theta} - \theta_0)/(\|\Delta\tilde{\theta}\|_1 + \|\Delta\theta_0\|_1) \in \mathcal{S}_\Delta(1)$, applying (6.15) to x, and then rearranging. Therefore, as in the proof of Theorem 6.3, we have

$$\|\tilde{\theta} - \theta_0\|_R^2 \le 2\tilde{K} \|\tilde{\theta} - \theta_0\|_R^{1-w/2} (\|\Delta\tilde{\theta}\|_1 + \|\Delta\theta_0\|_1)^{w/2} + 2\lambda(\|\Delta\theta_0\|_1 - \|\Delta\tilde{\theta}\|_1), \qquad (6.22)$$

We now set

$$\lambda = \Theta\left(K^{\frac{2}{1+w/2}} \|\Delta\theta_0\|_1^{-\frac{1-w/2}{1+w/2}} \right),$$

and in the spirit of Mammen and van de Geer [152], van de Geer [230], we proceed to argue in cases.

Case 1. Suppose that $\frac{1}{2}\|\Delta\tilde{\theta}\|_1 \geq \|\Delta\theta_0\|_1$. Then we see that (6.22) implies

$$0 \leq \|\tilde{\theta} - \theta_0\|_R^2 \leq \tilde{K}\|\tilde{\theta} - \theta_0\|_R^{1-w/2}\Big(\frac{3}{2}\Big)^{w/2}\|\Delta\tilde{\theta}\|_1^{w/2} - \lambda\|\Delta\tilde{\theta}\|_1, \tag{6.23}$$

so that

$$\lambda\|\Delta\tilde{\theta}\|_1 \leq \tilde{K}\|\tilde{\theta} - \theta_0\|_R^{1-w/2}\|\Delta\tilde{\theta}\|_1^{w/2},$$

where for simplicity have absorbed a constant factor $2(3/2)^{w/2}$ into $\tilde{K}$ (since this does not change the fact that $\tilde{K} = O_{\mathbb{P}}(K)$), and thus

$$\|\Delta\tilde{\theta}\|_1 \leq \Big(\frac{\tilde{K}}{\lambda}\Big)^{\frac{1}{1-w/2}}\|\tilde{\theta} - \theta_0\|_R.$$

Plugging this back into (6.23) gives

$$\|\tilde{\theta} - \theta_0\|_R^2 \leq \tilde{K}\|\tilde{\theta} - \theta_0\|_R^{1-w/2}\Big(\frac{\tilde{K}}{\lambda}\Big)^{\frac{w/2}{1-w/2}}\|\tilde{\theta} - \theta_0\|_R^{w/2},$$

or

$$\|\tilde{\theta} - \theta_0\|_R \leq \tilde{K}^{\frac{1}{1+w/2}}\Big(\frac{1}{\lambda}\Big)^{\frac{w/2}{1-w/2}} = O_{\mathbb{P}}\left(K^{\frac{1}{1+w/2}}\|\Delta\theta_0\|_1^{\frac{w/2}{1+w/2}}\right),$$

as desired.

Case 2. Suppose that $\frac{1}{2}\|\Delta\tilde{\theta}\|_1 \leq \|\Delta\theta_0\|_1$. Then from (6.22),

$$\|\tilde{\theta} - \theta_0\|_R^2 \leq \underbrace{2\lambda\|\Delta\theta_0\|_1}_{a} + \underbrace{2\tilde{K}\|\tilde{\theta} - \theta_0\|_R^{1-w/2}3^{w/2}\|\Delta\theta_0\|_1^{w/2}}_{b},$$

and hence either $\|\tilde{\theta} - \theta_0\|_R^2 \leq 2a$, or $\|\tilde{\theta} - \theta_0\|_R^2 \leq 2b$, and $a \leq b$. The first subcase is straightforward and leads to

$$\|\tilde{\theta} - \theta_0\|_R \leq 2\sqrt{\lambda\|\Delta\theta_0\|_1} = O_{\mathbb{P}}\left(K^{\frac{1}{1+w/2}}\|\Delta\theta_0\|_1^{\frac{w/2}{1+w/2}}\right),$$

as desired. In the second subcase, we have by assumption

$$\|\tilde{\theta} - \theta_0\|_R^2 \leq 2\tilde{K}\|\tilde{\theta} - \theta_0\|_R^{1-w/2}\|\Delta\theta_0\|_1^{w/2}, \tag{6.24}$$

$$2\lambda\|\Delta\theta_0\|_1 \leq \tilde{K}\|\tilde{\theta} - \theta_0\|_R^{1-w/2}\|\Delta\theta_0\|_1^{w/2}, \tag{6.25}$$

where again we have absorbed a constant factor $2(3^{w/2})$ into $\tilde{K}$. Working from (6.25), we derive

$$\|\Delta\theta_0\|_1 \leq \Big(\frac{\tilde{K}}{2\lambda}\Big)^{\frac{1}{1-w/2}}\|\tilde{\theta} - \theta_0\|_R,$$

and plugging this back into (6.24), we see

$$\|\tilde{\theta} - \theta_0\|_R^2 \leq 2\tilde{K}\|\tilde{\theta} - \theta_0\|_R^{1-w/2}\Big(\frac{\tilde{K}}{2\lambda}\Big)^{\frac{w/2}{1-w/2}}\|\tilde{\theta} - \theta_0\|_R^{w/2},$$

and finally

$$\|\tilde{\theta} - \theta_0\|_R \leq 2\tilde{K}^{\frac{1}{1+w/2}}\Big(\frac{1}{\lambda}\Big)^{\frac{w/2}{1-w/2}} = O_{\mathbb{P}}\left(K^{\frac{1}{1+w/2}}\|\Delta\theta_0\|_1^{\frac{w/2}{1+w/2}}\right).$$

This completes the second case, and the proof.

6.9.10 Proof of Theorem 6.10

The proof follows closely from Lemma 3.5 of van de Geer [230]. However, this author uses a different problem scaling than ours, so some care must be taken in applying the lemma. First we abbreviate $\mathcal{S} = \mathcal{S}_\Delta(1)$, and define $\tilde{\mathcal{S}} = \mathcal{S} \cdot \sqrt{n}/M$, where recall M is the maximum column norm of $\Delta^\dagger$. Now it is not hard to check that

$$\mathcal{S} = \{x \in \mathrm{row}(\Delta) : \|\Delta x\|_1 \le 1\} = \Delta^\dagger\{\alpha \in \mathrm{col}(\Delta) : \|\alpha\|_1 \le 1\},$$

so that $\max_{x\in\mathcal{S}} \|x\|_2 \le M$, and $\max_{x\in\tilde{\mathcal{S}}} \|x\|_2 \le \sqrt{n}$. This is important because Lemma 3.5 of van de Geer [230] concerns a form of "local" entropy that allows for deviations on the order of $\sqrt{n}$ in the norm $\|\cdot\|_2$, or equivalently, constant order in the scaled metric $\|\cdot\|_n = \|\cdot\|_2/\sqrt{n}$. Hence, the entropy bound in (6.16) translates into

$$\log N(\delta, \tilde{\mathcal{S}}, \|\cdot\|_2) \le E\Big(\frac{\sqrt{n}}{M}\Big)^w \Big(\frac{\sqrt{n}}{\delta}\Big)^w,$$

that is,

$$\log N(\delta, \tilde{\mathcal{S}}, \|\cdot\|_n) \le E\Big(\frac{\sqrt{n}}{M}\Big)^w \delta^{-w}.$$

Now we apply Lemma 3.5 of van de Geer [230]: in the scaled metric used by this author,

$$\max_{x\in\tilde{\mathcal{S}}} \frac{\epsilon^\top x}{\sqrt{n}\|x\|_n^{1-w/2}} = O_{\mathbb{P}}\left(\sqrt{E}\Big(\frac{\sqrt{n}}{M}\Big)^{w/2}\right),$$

that is,

$$\max_{x\in\tilde{\mathcal{S}}} \frac{\epsilon^\top x}{\|x\|_2^{1-w/2}} = O_{\mathbb{P}}\left(\sqrt{E}\big(\sqrt{n}\big)^{w/2}\Big(\frac{\sqrt{n}}{M}\Big)^{w/2}\right),$$

and finally,

$$\max_{x\in\mathcal{S}} \frac{\epsilon^\top x}{\|x\|_2^{1-w/2}} = O_{\mathbb{P}}\left(\sqrt{E}\big(\sqrt{n}\big)^{w/2}\right),$$

as desired.

6.9.11 Proof of Corollary 6.11

For each $j = 1, \ldots 2r$, if $\mathcal{G}$ is covered by j balls having radius at most $C_0\sqrt{n}j^{-1/\zeta}$, with respect to the norm $\|\cdot\|_2$, then it is covered by j balls having radius at most $C_0 j^{-1/\zeta}$, with respect to the scaled norm $\|\cdot\|_n = \|\cdot\|_2/\sqrt{n}$. By Theorem 1 of Carl [47], this implies that for each $j = 1, 2, 3, \ldots$, the convex hull $\mathrm{conv}(\mathcal{G})$ is covered by 2^j balls having radius at most $C_0' j^{-(1/2+1/\zeta)}$, with respect to $\|\cdot\|_n$, for another constant C_0'. Converting this back to an entropy bound in our original metric, and noting that $\mathrm{conv}(\mathcal{G}) = \mathcal{S}_\Delta(1)$, we have

$$\log(\delta, \mathcal{S}_\Delta(1), \|\cdot\|_2) \le C_0'' \Big(\frac{\sqrt{n}}{\delta}\Big)^{\frac{1}{1/2+1/\zeta}},$$

for a constant C_0'', as needed. This proves the lemma.

6.9.12 Proof of Corollary 6.12

According to Lemma 6.13, we know that $(D^{(1)})^\dagger = P_{\mathbb{1}}^\perp H$, where H is an $n \times (n-1)$ lower triangular matrix with $H_{ij} = 1$ if $i > j$ and 0 otherwise, and $P_{\mathbb{1}}^\perp$ is the projection map orthogonal to the all 1s vector. Thus $g_i = P_{\mathbb{1}}^\perp h_i$, $i = 1, \dots n-1$, with $h_1, \dots h_{n-1}$ denoting the columns of H. It is immediately apparent that

$$\|g_i - g_\ell\|_2 \le \|h_i - h_\ell\|_2 \le \sqrt{i - \ell},$$

for all $i > \ell$. Now, given $2j$ balls at our disposal, consider centering the first j balls at

$$g_d, g_{2d}, \dots g_{jd},$$

where $d = \lfloor n/j \rfloor$. Also let these balls have radius $\sqrt{n/j}$. By construction, then, we see that

$$\|g_1 - g_d\|_2 \le \sqrt{n/j},\ \|g_d - g_{2d}\|_2 \le \sqrt{n/j},\ \dots \|g_{jd} - g_{n-1}\|_2 \le \sqrt{n/j},$$

which means that we have covered $g_1, \dots g_{n-1}$ with j balls of radius $\sqrt{n/j}$.

We can cover $-g_1, \dots, -g_{n-1}$ with the remaining j balls analogously. Therefore, we have shown that $2j$ balls require a radius of $\sqrt{n/j}$, or in other words, j balls require a radius of $\sqrt{2n/j}$.

Chapter 7

Discrete TV-classes and Minimax Denoising on kD-Grids

In the previous section, we showed that in many commonly used graphs, we obtain faster rates using GTF then the standard Laplacian smoothing, but it is unclear whether GTF is optimal. In this chapter, we answer this question for a specific subset of the problems where the underlying graphs are regular lattice grids.

More specifically, we consider the problem of estimating a function defined over n locations on a d-dimensional grid (having all side lengths equal to $n^{1/d}$). When the function is constrained to have discrete total variation bounded by C_n, we derive the minimax optimal (squared) ℓ_2 estimation error rate, parametrized by n and C_n. Total variation denoising, also known as the fused lasso, is seen to be rate optimal. Several simpler estimators exist, such as Laplacian smoothing and Laplacian eigenmaps. A natural question is: can these simpler estimators perform just as well? We prove that these estimators, and more broadly all estimators given by linear transformations of the input data, are suboptimal over the class of functions with bounded variation. This extends fundamental findings of Donoho and Johnstone [70] on 1-dimensional total variation spaces to higher dimensions. The implication is that the computationally simpler methods cannot be used for such sophisticated denoising tasks, without sacrificing statistical accuracy. We also derive minimax rates for discrete Sobolev spaces over d-dimensional grids, which are, in some sense, smaller than the total variation function spaces. Indeed, these are small enough spaces that linear estimators can be optimal—and a few well-known ones are, such as Laplacian smoothing and Laplacian eigenmaps, as we show. Lastly, we investigate the problem of adaptivity of the total variation denoiser to these smaller Sobolev function spaces.

7.1 Introduction

Let $G = (V, E)$ be a d-dimensional grid graph, i.e., a lattice graph, with equal side lengths. Label the nodes as $V = \{1, \ldots, n\}$, and edges as $E = \{e_1, \ldots, e_m\}$. Consider data $y = (y_1, \ldots, y_n) \in$

$\mathbb{R}^n$ observed over the nodes, from a model

$$y_i \sim N(\theta_{0,i}, \sigma^2), \quad \text{i.i.d., for } i = 1, \dots, n, \tag{7.1}$$

where $\theta_0 = (\theta_{0,1}, \dots, \theta_{0,n}) \in \mathbb{R}^n$ is an unknown mean parameter to be estimated, and $\sigma^2 > 0$ is the marginal noise variance. It is assumed that θ_0 displays some kind of regularity over the grid G, e.g., $\theta_0 \in \mathcal{T}_d(C_n)$ for some $C_n > 0$, where

$$\mathcal{T}_d(C_n) = \big\{\theta : \|D\theta\|_1 \le C_n\big\}, \tag{7.2}$$

and $D \in \mathbb{R}^{m \times n}$ is the edge incidence matrix of G. This has ℓth row $D_\ell = (0, \dots, -1, \dots, 1, \dots, 0)$, with a -1 in the ith location, and 1 in the jth location, provided that the ℓth edge is $e_\ell = (i, j)$ with $i < j$. Equivalently, $L = D^T D$ is the graph Laplacian matrix of G, and thus

$$\|D\theta\|_1 = \sum_{(i,j) \in E} |\theta_i - \theta_j|, \qquad \text{and} \qquad \|D\theta\|_2^2 = \theta^T L \theta = \sum_{(i,j) \in E} (\theta_i - \theta_j)^2.$$

We will refer to the class in (7.2) as a *discrete total variation (TV) class*, and to the quantity $\|D\theta_0\|_1$ as the discrete total variation of θ_0, though for simplicity we will often drop the word "discrete".

The problem of estimating θ_0 given a total variation bound as in (7.2) is of great importance in both nonparametric statistics and signal processing, and has many applications, e.g., changepoint detection for 1d grids, and image denoising for 2d and 3d grids. There has been much methodological and computational work devoted to this problem, resulting in practically efficient estimators in dimensions 1, 2, 3, and beyond. However, theoretical performance, and in particularly optimality, is only really well-understood in the 1-dimensional setting. Results presented in this chapter seek to change that, and offers theory in d-dimensions that parallel more classical results known in the 1-dimensional case.

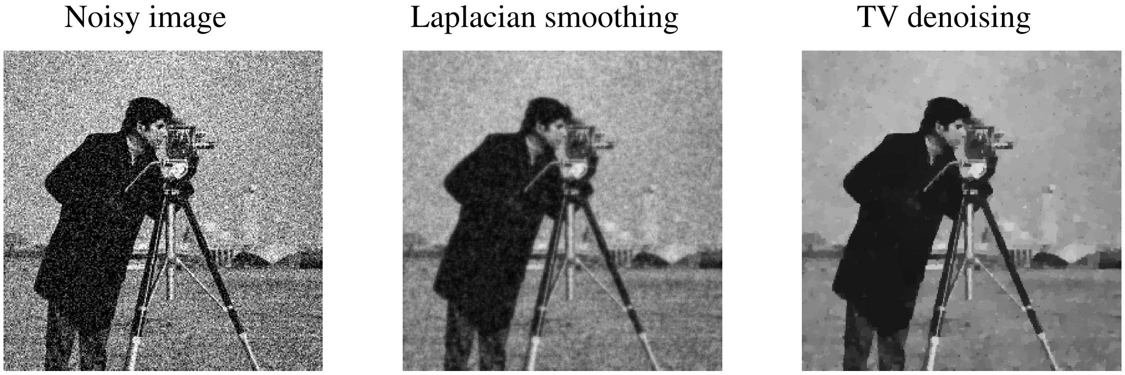
Estimators under consideration. Central role to our work is the *total variation (TV) denoising* or *fused lasso* estimator (e.g., [16, 50, 114, 182, 194, 220, 222, 237]), defined by the convex optimization problem

$$\hat{\theta}^{\mathrm{TV}} = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \; \|y - \theta\|_2^2 + \lambda \|D\theta\|_1, \tag{7.3}$$

where $\lambda \ge 0$ is a tuning parameter. Another pair of methods that we study carefully are *Laplacian smoothing* and *Laplacian eigenmaps*, which are most commonly seen in the context of clustering, dimensionality reduction, and semi-supervised learning, but are also useful tools for estimation in a regression setting like ours (e.g., [10, 22, 23, 24, 25, 26, 192, 204, 256, 258]). The Laplacian smoothing estimator is given by

$$\hat{\theta}^{\mathrm{LS}} = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \; \|y - \theta\|_2^2 + \lambda \|D\theta\|_2^2, \quad \text{i.e.,} \quad \hat{\theta}^{\mathrm{LS}} = (I + \lambda L)^{-1} y, \tag{7.4}$$

for a tuning parameter $\lambda \ge 0$, where in the second expression we have written $\hat{\theta}^{\mathrm{LS}}$ in closed-form (this is possible since it is the minimizer of a convex quadratic). For Laplacian eigenmaps,


Noisy image

Laplacian smoothing

TV denoising

Figure 7.1: *Comparison of Laplacian smoothing and TV denoising for the common "cameraman" image. TV denoising provides a more visually appealing result, and also achieves about a 35% reduction in MSE compared to Laplacian smoothing (MSE being measured to the original image). Both methods were tuned optimally.*

we must introduce the eigendecomposition of the graph Laplacian, $L = V \Sigma V^T$, where $\Sigma = \mathrm{diag}(\rho_1, \dots, \rho_n)$ with $0 = \rho_1 < \rho_2 \le \dots \le \rho_n$, and where $V = [V_1, V_2, \dots, V_n] \in \mathbb{R}^{n \times n}$ has orthonormal columns. The Laplacian eigenmaps estimator is

$$\hat{\theta}^{\mathrm{LE}} = V_{[k]} V_{[k]}^T y, \quad \text{where} \quad V_{[k]} = [V_1, V_2, \dots, V_k] \in \mathbb{R}^{n \times k}, \tag{7.5}$$

where now $k \in \{1, \dots, n\}$ acts as a tuning parameter.

Laplacian smoothing and Laplacian eigenmaps are appealing because they are (relatively) simple: they are just linear transformations of the data y. Indeed, as we are considering G to be a grid, both estimators in (7.4), (7.5) can be computed very quickly, in nearly $O(n)$ time, since the columns of V here are discrete cosine transform (DCT) basis vectors when $d = 1$, or Kronecker products thereof, when $d \ge 2$ (e.g., [59, 102, 134, 165, 242]). The TV denoising estimator in (7.3), on the other hand, cannot be expressed in closed-form, and is much more difficult to compute, especially when $d \ge 2$, though several advances have been made over the years (see the references above, and in particular [16] for an efficient operator-splitting algorithm and nice literature survey). Importantly, these computational difficulties are often worth it: TV denoising often practically outperforms ℓ_2-regularized estimators like Laplacian smoothing (and also Laplacian eigenmaps) in image denoising tasks, as it is able to better preserve sharp edges and object boundaries (this is now widely accepted, early references are, e.g., [3, 51, 66]). See Figure 7.1 for an example, using the often-studied "cameraman" image.

In the 1d setting, classical theory from nonparametric statistics draws a clear distinction between the performance of TV denoising and estimators like Laplacian smoothing and Laplacian eigenmaps. Perhaps surprisingly, this theory has not yet been fully developed in dimensions $d \ge 2$. Arguably, the comparison between TV denoising and Laplacian smoothing and Laplacian eigenmaps is even more interesting in higher dimensions, because the computational gap between the methods is even larger (the former method being much more expensive, say in 2d and 3d, than the latter two). Shortly, we review the 1d theory, and what is known in d-dimensions, for $d \ge 2$. First, we introduce notation.

Notation. For deterministic (nonrandom) sequences a_n, b_n we write $a_n = O(b_n)$ to denote that a_n/b_n is upper bounded for all n large enough, and $a_n \asymp b_n$ to denote that both $a_n = O(b_n)$ and $a_n^{-1} = O(b_n^{-1})$. Also, for random sequences A_n, B_n, we write $A_n = O_{\mathbb{P}}(B_n)$ to denote that A_n/B_n is bounded in probability. We abbreviate $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For an estimator $\hat{\theta}$ of the parameter θ_0 in (7.1), we define its mean squared error (MSE) to be

$$\mathrm{MSE}(\hat{\theta}, \theta_0) = \frac{1}{n}\|\hat{\theta} - \theta_0\|_2^2.$$

The risk of $\hat{\theta}$ is the expectation of its MSE, and for a set $\mathcal{K} \subseteq \mathbb{R}^n$, we define the minimax risk and minimax linear risk to be

$$R(\mathcal{K}) = \inf_{\hat{\theta}} \sup_{\theta_0 \in \mathcal{K}} \mathbb{E}\big[\mathrm{MSE}(\hat{\theta}, \theta_0)\big] \quad \text{and} \quad R_L(\mathcal{K}) = \inf_{\hat{\theta} \text{ linear}} \sup_{\theta_0 \in \mathcal{K}} \mathbb{E}\big[\mathrm{MSE}(\hat{\theta}, \theta_0)\big],$$

respectively, where the infimum on in the first expression is over all estimators $\hat{\theta}$, and in the second expression over all *linear estimators* $\hat{\theta}$, meaning that $\hat{\theta} = Sy$ for a matrix $S \in \mathbb{R}^{n \times n}$. We will also refer to linear estimators as *linear smoothers*. Note that both Laplacian smoothing in (7.4) and Laplacian eigenmaps in (7.5) are linear smoothers, but TV denoising in (7.3) is not. Lastly, in somewhat of an abuse of nomenclature, we will often call the parameter θ_0 in (7.1) a function, and a set of possible values for θ_0 as in (7.2) a function space; this comes from thinking of the components of θ_0 as the evaluations of an underlying function over n locations on the grid. This embedding has no formal importance, but it is convenient notationally, and matches the notation in nonparametric statistics.

Review: TV denoising in 1d. The classical nonparametric statistics literature [70, 71, 152] provides a more or less complete story for estimation under total variation constraints in 1d. See also [221] for a translation of these results to a setting more consistent (notationally) to that in the current paper. Assume that $d = 1$ and $C_n = C > 0$, a constant (not growing with n). The results in [70] imply that

$$R(\mathcal{T}_1(C)) \asymp n^{-2/3}. \tag{7.6}$$

Further, [152] showed that the TV denoiser $\hat{\theta}^{\mathrm{TV}}$ in (7.3), with $\lambda \asymp n^{1/3}$, satisfies

$$\mathrm{MSE}(\hat{\theta}^{\mathrm{TV}}, \theta_0) = O_{\mathbb{P}}(n^{-2/3}), \tag{7.7}$$

for all $\theta_0 \in \mathcal{T}_1(C)$, and is thus minimax rate optimal over $\mathcal{T}_1(C)$. (In assessing rates here and throughout, we do not distinguish between convergence in expectation versus convergence in probability.) Wavelet denoising, under various choices of wavelet bases, also achieves the minimax rate. However, many simpler estimators do not. To be more precise, it is shown in [70] that

$$R_L(\mathcal{T}_1(C)) \asymp n^{-1/2}. \tag{7.8}$$

Therefore, a substantial number of commonly used nonparametric estimators—such as running mean estimators, smoothing splines, kernel smoothing, Laplacian smoothing, and Laplacian eigenmaps, which are all linear smoothers—have a major deficiency when it comes to estimating functions of bounded variation. Roughly speaking, they will require many more samples to estimate θ_0 within the same degree of accuracy as an optimal method like TV or wavelet denoising (on the order of $\epsilon^{-1/2}$ times more samples to achieve an MSE of ϵ). Further theory and empirical examples (e.g., [67, 70, 221]) offer the following perspective: linear smoothers cannot cope with functions in $T(C)$ that have spatially inhomogeneous smoothness, i.e., that vary smoothly at some locations and vary wildly at others. Linear smoothers can only produce estimates that are smooth throughout, or wiggly throughout, but not a mix of the two. They can hence perform well over smaller, more homogeneous function classes like Sobolev or Holder classes, but not larger ones like total variation classes (or more generally, Besov and Triebel classes), and for these, one must use more sophisticated, nonlinear techniques. A motivating question: does such a gap persist in higher dimensions, between optimal nonlinear and linear estimators, and if so, how big is it?

Review: TV denoising in multiple dimensions. Recently, [248] established rates for TV denoising over various graph models, including grids, and [118] made improvements, particularly in the case of d-dimensional grids with $d \geq 2$. We can combine Propositions 4 and 6 of [118] with Theorem 3 of [248] to give the following result: if $d \geq 2$, and C_n is an arbitrary sequence (potentially unbounded with n), then the TV denoiser $\hat{\theta}^{\mathrm{TV}}$ in (7.3) satisfies, over all $\theta_0 \in \mathcal{T}_d(C_n)$,

$$\mathrm{MSE}(\hat{\theta}^{\mathrm{TV}}, \theta_0) = O_{\mathbb{P}}\bigg(\frac{C_n \log n}{n}\bigg) \text{ for } d = 2, \text{ and } \mathrm{MSE}(\hat{\theta}^{\mathrm{TV}}, \theta_0) = O_{\mathbb{P}}\bigg(\frac{C_n \sqrt{\log n}}{n}\bigg) \text{ for } d \geq 3, \tag{7.9}$$

with $\lambda \asymp \log n$ for $d = 2$, and $\lambda \asymp \sqrt{\log n}$ for $d \geq 3$. Note that, at first glance, this is a very different result from the 1d case. We expand on this next.

7.2 Summary of results

A gap in multiple dimensions. For estimation of θ_0 in (7.1) when $d \geq 2$, consider, e.g., the simplest possible linear smoother: the mean estimator, $\hat{\theta}^{\mathrm{mean}} = \bar{y}\mathbb{1}$ (where $\mathbb{1} = (1, \dots, 1) \in \mathbb{R}^n$, the vector of all 1s). Lemma 7.8, given below, implies that over $\theta_0 \in \mathcal{T}_d(C_n)$, the MSE of the mean estimator is bounded in probability by $C_n^2 \log n/n$ for $d = 2$, and C_n^2/n for $d \geq 3$. Compare this to (7.9). When $C_n = C > 0$ is a constant, i.e., when the TV of θ_0 is assumed to be bounded (which is assumed for the 1d results in (7.6), (7.7), (7.8)), this means that the TV denoiser and the mean estimator converge to θ_0 *at the same rate*, basically (ignoring log terms), the "parametric rate" of $1/n$, for estimating a finite-dimensional parameter! That TV denoising and such a trivial linear smoother perform comparably over 2d and 3d grids could not be farther from the story in 1d, where TV denoising is separated by an unbridgeable gap from *all* linear smoothers, as shown in (7.6), (7.7), (7.8).

Our results in Section 7.3 clarify this conundrum, and can be summarized by three points.

- We argue in Section 7.3.1 that there is a proper "canonical" scaling for the TV class defined in (7.2). E.g., when $d = 1$, this yields $C_n \asymp 1$, a constant, but when $d = 2$, this yields

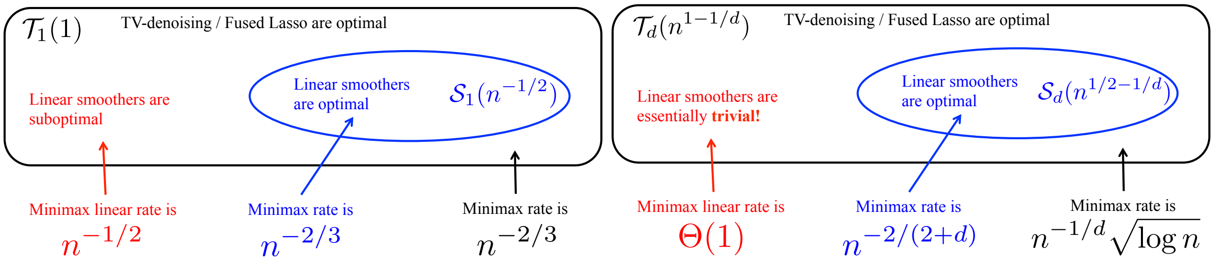

$\mathcal{T}_1(1)$
TV-denoising / Fused Lasso are optimal
Linear smoothers are suboptimal
Linear smoothers are optimal
$\mathcal{S}_1(n^{-1/2})$
Minimax linear rate is $n^{-1/2}$
Minimax rate is $n^{-2/3}$
Minimax rate is $n^{-2/3}$
$\mathcal{T}_d(n^{1-1/d})$
TV-denoising / Fused Lasso are optimal
Linear smoothers are essentially **trivial!**
Linear smoothers are optimal
$\mathcal{S}_d(n^{1/2-1/d})$
Minimax linear rate is $\Theta(1)$
Minimax rate is $n^{-2/(2+d)}$
Minimax rate is $n^{-1/d}\sqrt{\log n}$

Figure 7.2: Comparison of minimax rate and minimax linear rates in 1 dimension and d-dimension. **Left:** well-known results in 1-dim due to [70]. **Right:** summary of our results in d-dim under canonical scaling $C_n = O(n^{1-1/d})$.

$C_n \asymp \sqrt{n}$, and C_n also diverges with n for all $d \geq 3$. Sticking with $d = 2$ as an interesting example, we see that under such a scaling, the MSE rates achieved by TV denoising and the mean estimator respectively, are drastically different; ignoring log terms, these are

$$\frac{C_n}{n} \asymp \frac{1}{\sqrt{n}} \quad \text{and} \quad \frac{C_n^2}{n} \asymp 1, \tag{7.10}$$

respectively. Hence, TV denoising has an MSE rate of $1/\sqrt{n}$, in a setting where the mean estimator has a *constant* rate, i.e., a setting where it is not even known to be consistent.

- We show in Section 7.3.3 that our choice to study the mean estimator here is not somehow "unlucky" (it is not a particularly bad linear smoother, nor is the upper bound on its MSE loose): the minimax linear risk over $\mathcal{T}_d(C_n)$ is on the order C_n^2/n, for all $d \geq 2$. Thus, even the best linear smoothers have the same poor performance as the mean over $\mathcal{T}_d(C_n)$.
- We show in Section 7.3.2 that the TV estimator is (essentially) minimax optimal over $\mathcal{T}_d(C_n)$, as the minimax risk over this class scales as C_n/n (ignoring log terms).

To summarize, these results reveal a significant gap between linear smoothers and optimal estimators like TV denoising, for estimation over $\mathcal{T}_d(C_n)$ in d dimensions, with $d \geq 2$, as long as C_n scales appropriately. Roughly speaking, the TV classes encompass a challenging setting for estimation because they are very broad, containing a wide array of functions—both globally smooth functions, said to have homogeneous smoothness, and functions with vastly different levels of smoothness at different grid locations, said to have heterogeneous smoothness. Linear smoothers cannot handle heterogeneous smoothness, and only nonlinear methods can enjoy good estimation properties over the entirety of $\mathcal{T}_d(C_n)$ (see a summary in Figure 7.2). To reiterate, a telling example is $d = 2$ with the canonical scaling $C_n \asymp \sqrt{n}$, where we see that TV denoising achieves the optimal $1/\sqrt{n}$ rate (up to log factors), meanwhile, the best linear smoothers have max risk that is constant over $\mathcal{T}_2(\sqrt{n})$. See Figure 7.3 for an illustration.

Minimax rates over smaller function spaces, and adaptivity. Sections 7.4 and 7.5 are focused on different function spaces, discrete Sobolev spaces, which are ℓ_2 analogs of discrete TV spaces as we have defined them in (7.2). Under the canonical scaling of Section 7.3.1, Sobolev

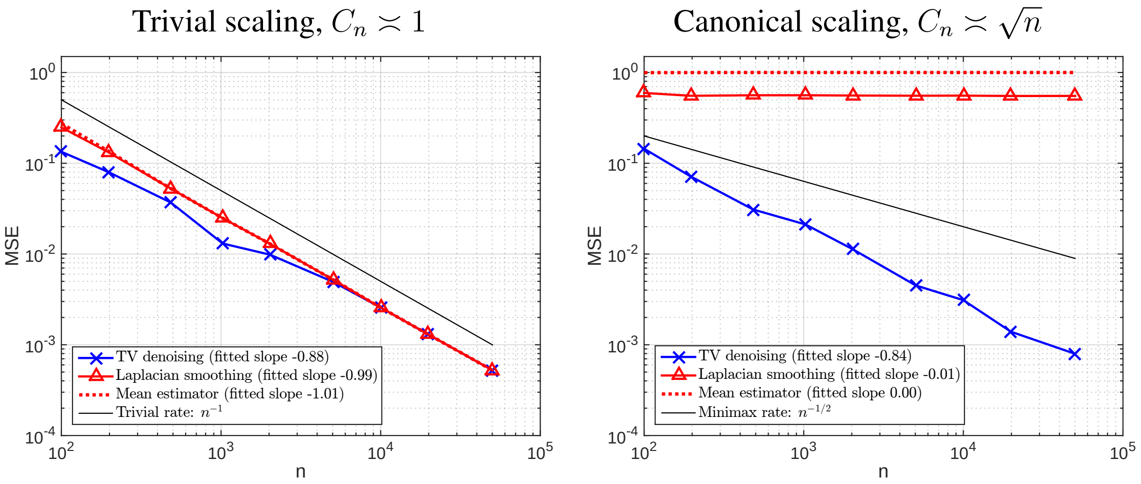


Figure 7.3: *MSE curves for estimation over a 2d grid, under two very different scalings of C_n: constant and $\sqrt{n}$. The parameter θ_0 was a "one-hot" signal, with all but one component equal to 0. For each n, the results were averaged over 5 repetitions, and Laplacian smoothing and TV denoising were tuned for optimal average MSE.*

spaces are contained in TV spaces, and the former can be roughly thought of as containing functions of more homogeneous smoothness. The story now is more optimistic for linear smoothers, and the following is a summary.

- In Section 7.4, we derive minimax rates for Sobolev spaces, and prove that linear smoothers—in particular, Laplacian smoothing and Laplacian eigenmaps—are optimal over these spaces.
- In Section 7.5, we discuss an interesting phenomenon, a phase transition of sorts, at $d = 3$ dimensions. When $d = 1$ or 2, the minimax rates for a TV space and its inscribed Sobolev space match; when $d \geq 3$, they do not, and the inscribed Sobolev space has a faster minimax rate. Aside from being an interesting statement about the TV and Sobolev function spaces in high dimensions, this raises an important question of adaptivity over the smaller Sobolev function spaces. As the minimax rates match for $d = 1$ and 2, any method optimal over TV spaces in these dimensions, such as TV denoising, is automatically optimal over the inscribed Sobolev spaces. But the question remains open for $d \geq 3$—does, e.g., TV denoising adapt to the faster minimax rate over Sobolev spaces? We present empirical evidence to suggest that this may be true, and leave a formal study to future work.

Other considerations and extensions. There are many problems related to the one that we study in this chapter. Clearly, minimax rates for the TV and Sobolev classes over general graphs, not just d-dimensional grids, are of interest. Our minimax lower bounds for TV classes actually apply to generic graphs with bounded max degree, though it is unclear whether to what extent they are sharp beyond grids; a detailed study will be left to future work. Another related topic is that of higher-order smoothness classes, i.e., classes containing functions whose *derivatives* are of (say) bounded variation. The natural extension of TV denoising here is called *trend filtering*, defined via the regularization of discrete higher-order derivatives. In the 1d setting, minimax rates, the optimality of trend filtering, and the suboptimality of linear smoothers is already well-understood [221]. Trend filtering has been defined and studied to some extent on general graphs [248], but no notions of optimality have been investigated beyond 1d. This will also be left to future work. Lastly, it is worth mentioning that there are other estimators (i.e., other than the ones we study in detail) that attain or nearly attain minimax rates over various classes we consider in this chapter. E.g., wavelet denoising is known to be optimal over TV classes in 1d [70]; and comparing recent upper bounds from [118, 164] with the lower bounds in this work, we see that wavelet denoising is also nearly minimax in 2d (ignoring log terms).

7.3 Analysis over TV classes

7.3.1 Canonical scalings for TV and Sobolev classes

We start by establishing what we call a "canonical" scaling for the radius C_n of the TV ball $\mathcal{T}_d(C_n)$ in (7.2), as well as the radius C'_n of the Sobolev ball $\mathcal{S}_d(C'_n)$, defined as

$$\mathcal{S}_d(C'_n) = \big\{\theta : \|D\theta\|_2 \le C'_n\big\}. \tag{7.11}$$

Proper scalings for C_n, C'_n will be critical for properly interpreting our new results in d dimensions, in a way that is comparable to known results for $d = 1$ (which are usually stated in terms of the 1d scalings $C_n \asymp 1$, $C'_n \asymp 1/\sqrt{n}$). To study (7.2), (7.11), it helps to introduce a third function space,

$$\mathcal{H}_d(1) = \Big\{\theta : \theta_i = f(i_1/\ell \ldots, i_d/\ell),\ i = 1, \ldots, n,\ \text{for some } f \in \mathcal{H}_d^{\text{cont}}(1)\Big\}. \tag{7.12}$$

Above, we have mapped each location i on the grid to a multi-index $(i_1, \ldots, i_d) \in \{1, \ldots, \ell\}^d$, where $\ell = n^{1/d}$, and $\mathcal{H}_d^{\text{cont}}(1)$ denotes the (usual) continuous Holder space on $[0, 1]^d$, i.e., functions that are 1-Lipschitz with respect to the ℓ_∞ norm. We seek an embedding that is analogous to the embedding of continuous Holder, Sobolev, and total variation spaces in 1d functional analysis, namely,

$$\mathcal{H}_d(1) \subseteq \mathcal{S}_d(C'_n) \subseteq \mathcal{T}_d(C_n). \tag{7.13}$$

Our first lemma provides a choice of C_n, C'_n that makes the above true. Its proof, as with all proofs in this chapter, can be found in Section 9.7.

Lemma 7.1. *For $d \ge 1$, the embedding in* (7.13) *holds with choices $C_n \asymp n^{1-1/d}$ and $C'_n \asymp n^{1/2-1/d}$. Such choices are called the* canonical scalings *for the function classes in* (7.2)*,* (7.11).

As a sanity check, both the (usual) continuous Holder and Sobolev function spaces in d dimensions are known to have minimax risks that scale as $n^{-2/(2+d)}$, in a standard nonparametric regression setup (e.g., [106]). Under the canonical scaling $C'_n \asymp n^{1/2-1/d}$, our results in Section 7.4 show that the discrete Sobolev class $\mathcal{S}_d(n^{1/2-1/d})$ also admits a minimax rate of $n^{-2/(2+d)}$.

7.3.2 Minimax rates over TV classes

The following is a lower bound for the minimax risk of the TV class $\mathcal{T}_d(C_n)$ in (7.2).

Theorem 7.2. *Assume $n \geq 2$, and denote $d_{\max} = 2d$. Then, for constants $c > 0$, $\rho_1 \in (2.34, 2.35)$,*

$$R(\mathcal{T}_d(C_n)) \geq c \cdot \begin{cases} \dfrac{\sigma C_n \sqrt{1 + \log(\sigma d_{\max} n / C_n)}}{d_{\max} n} & \text{if } C_n \in [\sigma d_{\max} \sqrt{\log n}, \sigma d_{\max} n / \sqrt{\rho_1}] \\ C_n^2 / (d_{\max}^2 n) \vee \sigma^2 / n & \text{if } C_n < \sigma d_{\max} \sqrt{\log n} \\ \sigma^2 / \rho_1 & \text{if } C_n > \sigma d_{\max} n / \sqrt{\rho_1} \end{cases}. \tag{7.14}$$

The proof uses a simplifying reduction of the TV class, via $\mathcal{T}_d(C_n) \supseteq B_1(C_n/d_{\max})$, the latter set denoting the ℓ_1 ball of radius $C_n/d_{\max}$ in $\mathbb{R}^n$. It then invokes a sharp characterization of the minimax risk in normal means problems over ℓ_p balls due to [33]. Several remarks are in order.

Remark 7.3. *The first line on the right-hand side in* (7.14) *often provides the most useful lower bound. To see this, recall that under the canonical scaling for TV classes, we have $C_n = n^{1-1/d}$. For all $d \geq 2$, this certainly implies $C_n \in [\sigma d_{\max} \sqrt{\log n}, \sigma d_{\max} n / \sqrt{\rho_1}]$, for large n.*

Remark 7.4. *Even though its construction is very simple, the lower bound on the minimax risk in* (7.14) *is sharp or nearly sharp in many interesting cases. Assume that $C_n \in [\sigma d_{\max} \sqrt{\log n}, \sigma d_{\max} n / \sqrt{\rho_1}]$. The lower bound rate is $C_n \sqrt{\log(n/C_n)}/n$. When $d = 2$, we see that this is very close to the upper bound rate of $C_n \log n / n$ achieved by the TV denoiser, as stated in* (7.9)*. These two differ by at most a $\log n$ factor (achieved when $C_n \asymp n$). When $d \geq 3$, we see that the lower bound rate is even closer to the upper bound rate of $C_n \sqrt{\log n}/n$ achieved by the TV denoiser, as in* (7.9)*. These two now differ by at most a $\sqrt{\log n}$ factor (again achieved when $C_n \asymp n$). We hence conclude that the TV denoiser is essentially minimax optimal in all dimensions $d \geq 2$.*

Remark 7.5. *When $d = 1$, and (say) $C_n \asymp 1$, the lower bound rate of $\sqrt{\log n}/n$ given by Theorem 7.2 is not sharp; we know from [70] (recall* (7.6)*) that the minimax rate over $\mathcal{T}_1(1)$ is $n^{-2/3}$. The result in the theorem (and also Theorem 7.6) in fact holds more generally, beyond grids: for an arbitrary graph G, its edge incidence matrix D, and $\mathcal{T}_d(C_n)$ as defined in* (7.2)*, the result holds for $d_{\max}$ equal to the max degree of G. It is unclear to what extent this is sharp, for different graph models.*

7.3.3 Minimax linear rates over TV classes

We now turn to a lower bound on the minimax linear risk of the TV class $\mathcal{T}_d(C_n)$ in (7.2).

Theorem 7.6. *Recall the notation $d_{\max} = 2d$. Then*

$$R_L(\mathcal{T}_d(C_n)) \geq \frac{\sigma^2 C_n^2}{C_n^2 + \sigma^2 d_{\max}^2 n} \vee \frac{\sigma^2}{n} \geq \frac{1}{2} \left(\frac{C_n^2}{d_{\max}^2 n} \wedge \sigma^2 \right) \vee \frac{\sigma^2}{n}. \tag{7.15}$$

The proof relies on an elegant meta-theorem on minimax rates from [71], which uses the concept of a "quadratically convex" set, whose minimax linear risk is the same as that of its hardest rectangular subproblem. An alternative proof can be given entirely from first principles. See Section 9.7.

Remark 7.7. *When C_n^2 grows with n, but not too fast (scales as $\sqrt{n}$, at most), the lower bound rate in* (7.15) *will be C_n^2/n. Compared to the C_n/n minimax rate from Theorem 7.2 (ignoring log terms), we see a clear gap between optimal nonlinear and linear estimators. In fact, under the canonical scaling $C_n \asymp n^{1-1/d}$, for any $d \geq 2$, this gap is seemingly huge: the lower bound for the minimax linear rate will be a constant, whereas the minimax rate from Theorem 7.2 (ignoring log terms) will be $n^{-1/d}$.*

We now show that the lower bound in Theorem 7.6 is essentially tight, and remarkably, it is certified by analyzing two trivial linear estimators: the mean estimator and the identity estimator.

Lemma 7.8. *Let M_n denote the largest column norm of $D^\dagger$. For the mean estimator $\hat{\theta}^{\mathrm{mean}} = \bar{y}\mathbb{1}$,*

$$\sup_{\theta_0 \in \mathcal{T}_d(C_n)} \mathbb{E}\big[\mathrm{MSE}(\hat{\theta}^{\mathrm{mean}}, \theta_0)\big] \leq \frac{\sigma^2 + C_n^2 M_n^2}{n},$$

From Proposition 4 in [118], we have $M_n = O(\sqrt{\log n})$ when $d = 2$ and $M_n = O(1)$ when $d \geq 3$.

The risk of the identity estimator $\hat{\theta}^{\mathrm{id}} = y$ is clearly σ^2. Combining this logic with Lemma 7.8 gives the upper bound $R_L(\mathcal{T}_d(C_n)) \leq (\sigma^2 + C_n^2 M_n^2)/n \wedge \sigma^2$. Comparing this with the lower bound described in Remark 7.7, we see that the two rates basically match, modulo the M_n^2 factor in the upper bound, which only provides an extra $\log n$ factor when $d = 2$. The takeaway message: in the sense of max risk, the best linear smoother does not perform much better than the trivial estimators.

We summarize the complete picture of the minimax rate in Figure 7.4. Additional empirical experiments, similar to those shown in Figure 7.3, are given in Section 9.7.

7.4 Analysis over Sobolev classes

Our first result here is a lower bound on the minimax risk of the Sobolev class $\mathcal{S}_d(C_n')$ in (7.11).

Theorem 7.9. *For a universal constant $c > 0$,*

$$R(\mathcal{S}_d(C_n')) \geq \frac{c}{n}\Big((n\sigma^2)^{\frac{2}{d+2}}(C_n')^{\frac{2d}{d+2}} \wedge n\sigma^2 \wedge n^{2/d}(C_n')^2\Big) + \frac{\sigma^2}{n}.$$

Elegant tools for minimax analysis from [71], which leverage the fact that the ellipsoid $\mathcal{S}_d(C_n')$ is orthosymmetric and quadratically convex (after a rotation), are used to prove the result.

The next theorem gives upper bounds, certifying that the above lower bound is tight, and showing that Laplacian eigenmaps and Laplacian smoothing, both linear smoothers, are optimal over $\mathcal{S}_d(C_n')$ for all d and for $d = 1, 2$, or 3 respectively.

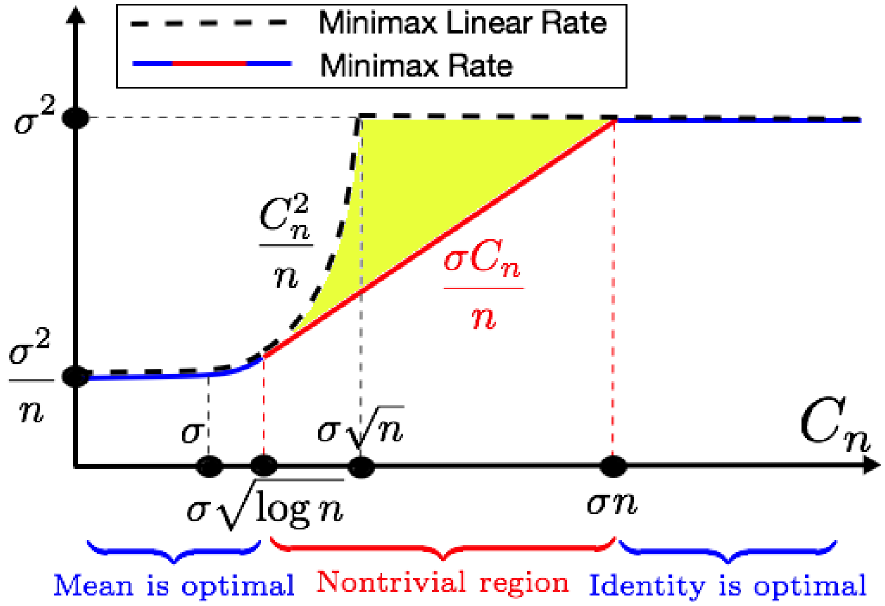


Figure 7.4: Illustration of the matching lower bounds for TV-class with every pair of C_n and σ. The yellow region indicates the statistical gain of using more computationally expensive nonlinear estimators.

Theorem 7.10. *For Laplacian eigenmaps, $\hat{\theta}^{\mathrm{LE}}$ in* (7.5)*, with $k \asymp ((n(C_n')^d)^{2/(d+2)} \vee 1) \wedge n$, we have*

$$\sup_{\theta_0 \in \mathcal{S}_d(C_n')} \mathbb{E}\big[\mathrm{MSE}(\hat{\theta}^{\mathrm{LE}}, \theta_0)\big] \leq \frac{c}{n}\Big((n\sigma^2)^{\frac{2}{d+2}}(C_n')^{\frac{2d}{d+2}} \wedge n\sigma^2 \wedge n^{2/d}(C_n')^2\Big) + \frac{c\sigma^2}{n},$$

for a universal constant $c > 0$, and n large enough. When $d = 1, 2,$ or 3, the same bound holds for Laplacian smoothing $\hat{\theta}^{\mathrm{LS}}$ in (7.5)*, with $\lambda \asymp (n/(C_n')^2)^{2/(d+2)}$ (and a possibly different constant c).*

Remark 7.11. *As shown in the proof, Laplacian smoothing is nearly minimax rate optimal over $\mathcal{S}_d(C_n')$ when $d = 4$, just incurring an extra log factor. It is unclear to us whether this method is still (nearly) optimal when $d \geq 5$; based on insights from our proof technique, we conjecture that it is not.*

7.5 A phase transition, and adaptivity

The TV and Sobolev classes in (7.2) and (7.11), respectively, display a curious relationship. We reflect on Theorems 7.2 and 7.9, using, for concreteness, the canonical scalings $C_n \asymp n^{1-1/d}$ and $C_n' \asymp n^{1/2-1/d}$, that, recall, guarantee $\mathcal{S}_d(C_n') \subseteq \mathcal{T}_d(C_n)$. (Similar statements could also be made outside of this case, subject to an appropriate relationship with $C_n/C_n' \asymp \sqrt{n}$.) When $d = 1$, both the TV and Sobolev classes have a minimax rate of $n^{-2/3}$ (this TV result is actually due to [70],

Function class	Dimension 1	Dimension 2	Dimension $d \geq 3$
TV ball $\mathcal{T}_d(n^{1-1/d})$	$n^{-2/3}$	$n^{-1/2}\sqrt{\log n}$	$n^{-1/d}\sqrt{\log n}$
Sobolev ball $\mathcal{S}_d(n^{1/2-1/d})$	$n^{-2/3}$	$n^{-1/2}$	$n^{-\frac{2}{2+d}}$

Table 7.1: *Summary of rates for canonically-scaled TV and Sobolev spaces.*

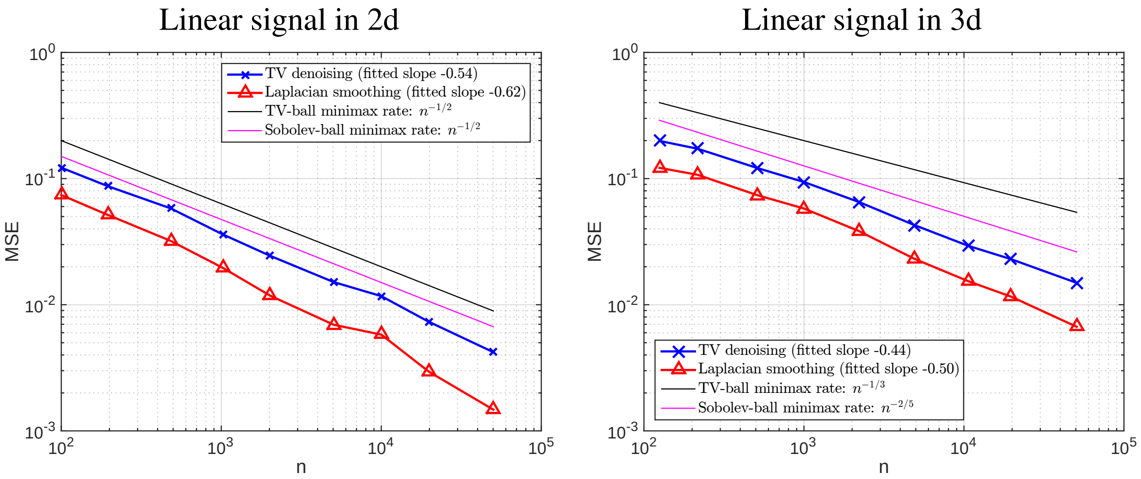


Figure 7.5: *MSE curves for estimating a "linear" signal, a very smooth signal, over 2d and 3d grids. For each n, the results were averaged over 5 repetitions, and Laplacian smoothing and TV denoising were tuned for best average MSE performance. The signal was set to satisfy $\|D\theta_0\|_2 \asymp n^{1/2-1/d}$, matching the canonical scaling.*

as stated in (7.6), not Theorem 7.2). When $d = 2$, both the TV and Sobolev classes again have the same minimax rate of $n^{-1/2}$, the caveat being that the rate for TV class has an extra $\sqrt{\log n}$ factor. But for all $d \geq 3$, the rates for the canonical TV and Sobolev classes differ, and the smaller Sobolev spaces have faster rates than their inscribing TV spaces. This may be viewed as a phase transition at $d = 3$; see Table 7.1.

We may paraphrase to say that 2d is just like 1d, in that expanding the Sobolev ball into a larger TV ball does not hurt the minimax rate, and methods like TV denoising are automatically *adaptive*, i.e., optimal over both the bigger and smaller classes. However, as soon as we enter the 3d world, it is no longer clear whether TV denoising can adapt to the smaller, inscribed Sobolev ball, whose minimax rate is faster, $n^{-2/5}$ versus $n^{-1/3}$ (ignoring log factors). Theoretically, this is an interesting open problem that we do not approach in this chapter and leave to future work.

We do, however, investigate the matter empirically: see Figure 7.5, where we run Laplacian smoothing and TV denoising on a highly smooth "linear" signal θ_0. This is constructed so that each component θ_i is proportional to $i_1 + i_2 + \ldots + i_d$ (using the multi-index notation $(i_1, \ldots, i_d)$ of (7.12) for grid location i), and the Sobolev norm is $\|D\theta_0\|_2 \asymp n^{1/2-1/d}$. Arguably, these are among the "hardest" types of functions for TV denoising to handle. The left panel, in 2d, is a case in which we know that TV denoising attains the minimax rate; the right panel, in 3d, is a case in which we do not, though empirically, TV denoising surely seems to be doing better than the slower minimax rate of $n^{-1/3}$ (ignoring log terms) that is associated with the larger TV ball.

Even if TV denoising is shown to be minimax optimal over the inscribed Sobolev balls when $d \geq 3$, note that this does not necessarily mean that we should scrap Laplacian smoothing in favor of TV denoising, in all problems. Laplacian smoothing is the unique Bayes estimator in a normal means model under a certain Markov random field prior (e.g., [192]); statistical decision theory therefore tells that it is *admissible*, i.e., no other estimator—TV denoising included—can uniformly dominate it.

7.6 Discussion

We conclude with a quote from Albert Einstein: “Everything should be made as simple as possible, but no simpler”. In characterizing the minimax rates for TV classes, defined over d-dimensional grids, we have shown that simple methods like Laplacian smoothing and Laplacian eigenmaps—or even in fact, all linear estimators—must be passed up in favor of more sophisticated, nonlinear estimators, like TV denoising, if one wants to attain the optimal max risk. Such a result was previously known when $d = 1$; our work has extended it to all dimensions $d \geq 2$. We also characterized the minimax rates over discrete Sobolev classes, revealing an interesting phase transition where the optimal rates over TV and Sobolev spaces, suitably scaled, match when $d = 1$ and 2 but diverge for $d \geq 3$. It is an open question as to whether an estimator like TV denoising can be optimal over both spaces, for all d.

7.7 Proofs

We present proofs of all results, according to the order in which they appear in the paper.

7.7.1 Proof of Lemma 7.1 (canonical scaling)

Suppose that $\theta \in \mathcal{H}_d(1)$ that is a discretization of a 1-Lipschitz function f, i.e., $\theta_i = f(i_1/\ell \ldots, i_d/\ell)$, $i = 1, \ldots, n$. We first we compute and bound its squared Sobolev norm

$$\begin{aligned} \|D\theta\|_2^2 = \sum_{(i,j)\in E} (\theta_i - \theta_j)^2 &= \sum_{(i,j)\in E} \big(f(i_1/\ell, \ldots, i_d/\ell) - f(j_1/\ell, \ldots, j_d/\ell)\big)^2 \\ &\leq \sum_{(i,j)\in E} \big\|(i_1/\ell, \ldots, i_d/\ell) - (j_1/\ell, \ldots, j_d/\ell)\big\|_\infty^2 \\ &= m/\ell^2, \end{aligned}$$

where, recall, we denote by $m = |E|$ the number of edges in the grid. In the second line we used the 1-Lipschitz property of f, and in the third we used that multi-indices corresponding to adjacent locations on the grid are exactly 1 apart, in ℓ_∞ distance. Thus we see that setting $C_n' = \sqrt{m}/\ell$

gives the desired containment $\mathcal{S}_d(C_n') \supseteq \mathcal{H}_d(1)$. It is always true that $m \asymp n$ for a d-dimensional grid (though the constant may depend on d), so that $\sqrt{m}/\ell \asymp n^{1/2-1/d}$. This completes the proof for the Sobolev class scaling.

As for TV class scaling, the result follows from the simple fact that $\|x\|_1 \le \sqrt{m}\|x\|_2$ for any $x \in \mathbb{R}^m$, so that we may take $C_n = \sqrt{m}C_n' = n^{1-1/d}$. □

7.7.2 Proof of Theorem 7.2 (minimax rates over TV classes)

Here and henceforth, we use the notation $B_p(r) = \{x : \|x\|_p \le r\}$ for the ℓ_p ball of radius r, where $p, r > 0$ (and the ambient dimension will be determined based on the context).

We begin with a very simple lemma, that embeds an ℓ_1 ball inside the TV ball $\mathcal{T}_d(C_n)$.

Lemma 7.12. *Let G be a graph with maximum degree $d_{\max}$, and let $D \in \mathbb{R}^{m\times n}$ be its incidence matrix. Then for any $r > 0$, it holds that $B_1(r/d_{\max}) \subseteq T_d(r)$.*

Proof. Write D_i for the ith column of D. The proof follows from the observation that, for any θ,

$$\|D\theta\|_1 = \Big\| \sum_{i=1}^n D_i\theta_i \Big\|_1 \le \sum_{i=1}^n \|D_i\|_1 |\theta_i| \le \Big(\max_{i=1,\dots,n} \|D_i\|_1 \Big) \|\theta\|_1 = d_{\max}\|\theta\|_1.$$

□

To prove Theorem 7.2, we will rely on a result from Birge and Massart [33], which gives a lower bound for the risk in a normal means problem, over ℓ_p balls. Another related, earlier result is that of Donoho and Johnstone [68]; however, the Birge and Massart result places no restrictions on the radius of the ball in question, whereas the Donoho and Johnstone result does. Translated into our notation, the Birge and Massart result is as follows.

Lemma 7.13 (Proposition 5 of Birge and Massart [33])**.** *Assume i.i.d. observations $y_i \sim N(\theta_{0,i}, \sigma^2)$, $i = 1, \dots, n$, and $n \ge 2$. Then the minimax risk over the ℓ_p ball $B_p(r_n)$, where $0 < p < 2$, satisfies*

$$n \cdot R(B_p(r_n)) \ge c \cdot \begin{cases} \sigma^{2-p} r_n^p \left[1 + \log\left(\dfrac{\sigma^p n}{r_n^p} \right) \right]^{1-p/2} & \textit{if } \sigma\sqrt{\log n} \le r_n \le \sigma n^{1/p}/\sqrt{\rho_p} \\ r_n^2 & \textit{if } r_n < \sigma\sqrt{\log n} \\ \sigma^2 n/\rho_p & \textit{if } r_n > \sigma n^{1/p}/\sqrt{\rho} \end{cases}.$$

Here $c > 0$ is a universal constant, and $\rho_p > 1.76$ is the unique solution of $\rho_p \log \rho_p = 2/p$.

Finally, applying Lemma 7.13 to $B_1(C_n/d_{\max})$ almost gives the lower bound as stated in Theorem 7.2. However, note that the minimax risk in question is trivially lower bounded by σ^2/n, because

$$\begin{aligned}\inf_{\hat\theta} \sup_{\theta_0 \in \mathcal{T}_d(C_n)} \frac{1}{n}\, \mathbb{E}\|\hat\theta - \theta_0\|_2^2 &\geq \inf_{\hat\theta} \sup_{\theta_0 : \theta_{0,1} = \ldots = \theta_{0,n}} \frac{1}{n} \sum_{i=1}^n \mathbb{E}(\hat\theta_i - \theta_{0,1})^2 \\ &= \inf_{\hat\theta_1} \sup_{\theta_{0,1}} \mathbb{E}(\hat\theta_1 - \theta_{0,1})^2 \\ &= \frac{\sigma^2}{n}.\end{aligned}$$

In the second to last line, the problem is to estimate a 1-dimensional mean parameter $\theta_{0,1}$, given the observations $y_i \sim N(\theta_{0,1}, \sigma^2)$, i.i.d., for $i = 1, \ldots, n$; this has a well-known minimax risk of σ^2/n. What this means for our TV problem: to derive a lower bound for the minimax rate over $\mathcal{T}_d(C_n)$, we may take the maximum of the result of applying Lemma 7.13 to $B_1(C_n/d_{\max})$ and σ^2/n. One can see that the term σ^2/n only plays a role for small C_n, i.e., it effects the case when $C_n < \sigma d_{\max}\sqrt{\log n}$, where the lower bound becomes $C_n^2/(d_{\max}^2 n) \vee \sigma^2/n$. □

7.7.3 Proof of Theorem 7.6 (minimax linear rates over TV classes)

First we recall a few definitions, from Donoho et al. [71]. Given a set $A \subseteq \mathbb{R}^k$, its *quadratically convex hull* $\mathrm{qconv}(A)$ is defined as

$$\begin{aligned}\mathrm{qconv}(A) &= \big\{(x_1, \ldots, x_k) : (x_1^2, \ldots, x_k^2) \in \mathrm{conv}(A_+^2)\big\}, \quad \text{where} \\ A_+^2 &= \big\{(a_1^2, \ldots, a_k^2) : a \in A,\ a_i \geq 0,\ i = 1, \ldots, k\big\}.\end{aligned}$$

(Here $\mathrm{conv}(B)$ denotes the convex hull of a set B.) Furthermore, the set A is called *quadratically convex* provided that $\mathrm{qconv}(A) = A$. Also, A is called *orthosymmetric* provided that $(a_1, \ldots, a_k) \in A$ implies $(\sigma_1 a_1, \ldots, \sigma_k a_k) \in A$, for any choice of signs $\sigma_1, \ldots, \sigma_k \in \{-1, 1\}$.

Now we proceed with the proof. Following from equation (7.2) of Donoho et al. [71],

$$\mathrm{qconv}\big(B_1(C_n/d_{\max})\big) = B_2(C_n/d_{\max}).$$

Theorem 11 of Donoho et al. [71] states that, for orthosymmetric, compact sets, such as $B_1(C_n/d_{\max})$, the minimax linear risk equals that of its quadratically convex hull. Moreover, Theorem 7 of Donoho et al. [71] tells us that for sets that are orthosymmetric, compact, convex, and quadratically convex, such as $B_2(C_n/d_{\max})$, the minimax linear risk is the same as the minimax linear risk over the worst rectangular subproblem. We consider $B_\infty(C_n/(d_{\max}\sqrt{n}))$, and abbreviate $r_n = C_n/(d_{\max}\sqrt{n})$. It is fruitful to study rectangles because the problem separates across dimensions, as in

$$\begin{aligned}\inf_{\hat\theta \text{ linear}} \sup_{\theta_0 \in B_\infty(r_n)} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\hat\theta_i - \theta_{0,i})^2\right] &= \frac{1}{n}\sum_{i=1}^n \left[\inf_{\hat\theta_i \text{ linear}} \sup_{|\theta_{0,i}| \leq r_n} \mathbb{E}(\hat\theta_i - \theta_{0,i})^2\right] \\ &= \inf_{\hat\theta_1 \text{ linear}} \sup_{|\theta_{0,1}| \leq r_n} \mathbb{E}(\hat\theta_1 - \theta_{0,1})^2.\end{aligned}$$

Thus it suffices to compute the minimax linear risk over the 1d class $\{\theta_{0,1} : |\theta_{0,1}| \le r_n\}$. It is easily shown (e.g., see Section 2 of Donoho et al. [71]) that this is $r_n^2\sigma^2/(r_n^2+\sigma_2^2)$, and so this is precisely the minimax linear risk for $B_2(C_n/d_{\max})$, and for $B_1(C_n/d_{\max})$.

To get the first lower bound as stated in the theorem, we simply take a maximum of $r_n^2\sigma^2/(r_n^2+\sigma_2^2)$ and σ^2/n, as the latter is the minimax risk for estimating a 1-dimensional mean parameter given n observations in a normal model with variance σ^2, recall the end of the proof of Theorem 7.2. To get the second, we use the fact that $2ab/(a+b) \ge \min\{a,b\}$. This completes the proof. □

7.7.4 Alternative proof of Theorem 7.6

Here, we reprove Theorem 7.6 using elementary arguments. We write $y = \theta_0 + \epsilon$, for $\epsilon \sim N(0,\sigma^2 I)$. Given an arbitary linear estimator, $\hat\theta = Sy$ for a matrix $S \in \mathbb{R}^{n\times n}$, observe that

$$\begin{aligned}\mathbb{E}\big[\mathrm{MSE}(\hat\theta,\theta_0)\big] = \frac{1}{n}\mathbb{E}\|\hat\theta-\theta_0\|_2^2 &= \frac{1}{n}\mathbb{E}\|S(\theta_0+\epsilon)-\theta_0\|_2^2 \\ &= \frac{1}{n}\mathbb{E}\|S\epsilon\|_2^2 + \frac{1}{n}\|(S-I)\theta_0\|_2^2 \\ &= \frac{\sigma^2}{n}\|S\|_F^2 + \frac{1}{n}\|(S-I)\theta_0\|_2^2, \end{aligned} \tag{7.16}$$

which we may view as the variance and (squared) bias terms, respectively. Now denote by e_i the ith standard basis vector, and consider

$$\begin{aligned}\frac{\sigma^2}{n}\|S\|_F^2 + \left(\sup_{\theta_0:\|D\theta_0\|_1\le C_n} \frac{1}{n}\|(S-I)\theta_0\|_2^2\right) &\ge \frac{\sigma^2}{n}\|S\|_F^2 + \frac{C_n^2}{d_{\max}^2 n}\left(\max_{i=1,\dots,n} \|(I-S)e_i\|_2^2\right) \\ &\ge \frac{\sigma^2}{n}\|S\|_F^2 + \frac{C_n^2}{d_{\max}^2 n^2}\sum_{i=1}^n \|(I-S)e_i\|_2^2 \\ &= \frac{\sigma^2}{n}\|S\|_F^2 + \frac{C_n^2}{d_{\max}^2 n^2}\|(I-S)\|_F^2 \\ &\ge \frac{\sigma^2}{n}\sum_{i=1}^n S_{ii}^2 + \frac{C_n^2}{d_{\max}^2 n^2}\sum_{i=1}^n (1-S_{ii})^2 \\ &= \frac{1}{n}\sum_{i=1}^n \left(\sigma^2 S_{ii}^2 + \frac{C_n^2}{d_{\max}^2 n}(1-S_{ii})^2\right). \end{aligned}$$

Here S_{ii}, $i=1,\dots,n$ denote the diagonal entries of S. To bound each term in the sum, we apply the simple inequality $ax^2 + b(1-x)^2 \ge ab/(a+b)$ for all x (since a short calculation shows that the quadratic in x here is minimized at $x = b/(a+b)$). We may continue on lower bounding the last displayed expression, giving

$$\frac{\sigma^2}{n}\|S\|_F^2 + \left(\sup_{\theta_0:\|D\theta_0\|_1\le C_n} \frac{1}{n}\|(S-I)\theta_0\|_2^2\right) \ge \frac{\sigma^2 C_n^2}{C_n^2 + \sigma^2 d_{\max}^2 n}.$$

Lastly, we may take the maximum of this with σ^2/n in order to derive a final lower bound, as argued in the proof of Theorem 7.6. □

7.7.5 Proof of Lemma 7.8 (mean estimator over TV classes)

For this estimator, the smoother matrix is $S = \mathbb{1}\mathbb{1}^T/n$ and so $\|S\|_F^2 = 1$. From (7.16), we have

$$\mathbb{E}\big[\mathrm{MSE}(\hat{\theta}^{\mathrm{mean}}, \theta_0)\big] = \frac{\sigma^2}{n} + \frac{1}{n}\|\theta_0 - \bar{\theta}_0\mathbb{1}\|_2^2,$$

where $\bar{\theta}_0 = (1/n)\sum_{i=1}^n \theta_{0,i}$. Now

$$\begin{aligned}
\sup_{\theta_0 : \|D\theta_0\|_1 \le C_n} \frac{1}{n}\|\theta_0 - \bar{\theta}_0\mathbb{1}\|_2^2 &= \sup_{x \in \mathrm{row}(D) : \|Dx\|_1 \le C_n} \frac{1}{n}\|x\|_2^2 \\
&= \sup_{z \in \mathrm{col}(D) : \|z\|_1 \le C_n} \frac{1}{n}\|D^\dagger z\|_2^2 \\
&\le \sup_{z : \|z\|_1 \le C_n} \frac{1}{n}\|D^\dagger z\|_2^2 \\
&= \frac{C_n^2}{n} \max_{i=1,\ldots,n} \|D_i^\dagger\|_2^2 \\
&\le \frac{C_n^2 M_n^2}{n},
\end{aligned}$$

which establishes the desired bound. □

7.7.6 Proof of Theorem 7.9 (minimax rates over Sobolev classes)

Recall that we denote by $L = V\Sigma V^T$ the eigendecomposition of the graph Laplacian $L = D^T D$, where $\Sigma = \mathrm{diag}(\rho_1, \ldots, \rho_n)$ with $0 = \rho_1 < \rho_2 \le \ldots \le \rho_n$, and where $V \in \mathbb{R}^{n \times n}$ has orthonormal columns. Also denote by $D = U\Sigma^{1/2}V^T$ the singular value decomposition of the edge incidence matrix D, where $U \in \mathbb{R}^{m \times n}$ has orthonormal columns.[1] First notice that

$$\|D\theta_0\|_2 = \|U\Sigma^{1/2}V^T\theta_0\|_2 = \|\Sigma^{1/2}V^T\theta_0\|_2.$$

This suggests that a rotation by V^T will further simplify the minimax risk over $\mathcal{S}_d(C_n')$, i.e.,

$$\begin{aligned}
\inf_{\hat{\theta}} \sup_{\theta_0 : \|\Sigma^{1/2}V^T\theta_0\|_2 \le C_n'} \frac{1}{n}\mathbb{E}\|\hat{\theta} - \theta_0\|_2^2 &= \inf_{\hat{\theta}} \sup_{\theta_0 : \|\Sigma^{1/2}V^T\theta_0\|_2 \le C_n'} \frac{1}{n}\mathbb{E}\|V^T\hat{\theta} - V^T\theta_0\|_2^2 \\
&= \inf_{\hat{\gamma}} \sup_{\gamma_0 : \|\Sigma^{1/2}\gamma_0\|_2 \le C_n'} \frac{1}{n}\mathbb{E}\|\hat{\gamma} - \gamma_0\|_2^2, \qquad (7.17)
\end{aligned}$$

where we have rotated and now consider the new parameter $\gamma_0 = V^T\theta_0$, constrained to lie in

$$\mathcal{E}_d(C_n') = \Big\{\gamma : \sum_{i=2}^n \rho_i\gamma_i^2 \le (C_n')^2\Big\}.$$

[1] When $d = 1$, we have $m = n - 1$ edges, and so it is not be possible for U to have orthonormal columns; however, we can just take its first column to be all 0s, and take the rest as the eigenbasis for $\mathbb{R}^{n-1}$, and all the arguments given here will go through.

To be clear, in the rotated setting (7.17) we observe a vector $y' = V^T y \sim N(\gamma_0, \sigma^2 I)$, and the goal is to estimate the mean parameter γ_0. Since there are no constraints along the first dimension, we can separate out the MSE in (7.17) into that incurred on the first component, and all other components. Decomposing $\gamma_0 = (\alpha_0, \beta_0) \in \mathbb{R}^{1\times(n-1)}$, with similar notation for an estimator $\hat{\gamma}$,

$$\begin{aligned} \inf_{\hat{\gamma}} \sup_{\gamma_0 \in \mathcal{E}_d(C_n')} \frac{1}{n}\mathbb{E}\|\hat{\gamma}-\gamma_0\|_2^2 &= \inf_{\hat{\alpha}} \sup_{\alpha_0} \frac{1}{n}\mathbb{E}(\hat{\alpha}-\alpha_0)^2 + \inf_{\hat{\beta}} \sup_{\beta_0 \in P_{-1}(\mathcal{E}_d(C_n'))} \frac{1}{n}\mathbb{E}\|\hat{\beta}-\beta_0\|_2^2 \\ &= \frac{\sigma^2}{n} + \inf_{\hat{\beta}} \sup_{\beta_0 \in P_{-1}(\mathcal{E}_d(C_n'))} \frac{1}{n}\mathbb{E}\|\hat{\beta}-\beta_0\|_2^2, \end{aligned} \tag{7.18}$$

where P_{-1} projects onto all coordinate axes but the 1st, i.e., $P_{-1}(x) = (0, x_2, \ldots, x_n)$, and in the second line we have used the fact that the minimax risk for estimating a 1-dimensional parameter α_0 given an observation $z \sim N(\alpha_0, \sigma^2)$ is simply σ^2.

Let us lower bound the second term in (7.18), i.e., $R(P_{-1}(\mathcal{E}_d(C_n')))$. The ellipsoid $P_{-1}(\mathcal{E}_d(C_n'))$ is orthosymmetric, compact, convex, and quadratically convex, hence Theorem 7 in Donoho et al. [71] tells us that its minimax linear risk is the minimax linear risk of its hardest rectangular subproblem. Further, Lemma 6 in Donoho et al. [71] then tells us the minimax linear risk of its hardest rectangular subproblem is, up to a constant factor, the same as the minimax (nonlinear) risk of the full problem. More precisely, Lemma 6 and Theorem 7 from Donoho et al. [71] imply

$$\frac{5}{4}R(P_{-1}(\mathcal{E}_d(C_n'))) \geq R_L(P_{-1}(\mathcal{E}_d(C_n'))) = \sup_{H \subseteq P_{-1}(\mathcal{E}_d(C_n'))} R_L(H), \tag{7.19}$$

where the supremum above is taken over all rectangular subproblems, i.e., all rectangles H contained in $P_{-1}(\mathcal{E}_d(C_n'))$.

To study rectangular subproblems, it helps to reintroduce the multi-index notation for a location i on the d-dimensional grid, writing this as $(i_1, \ldots, i_d) \in \{1, \ldots, \ell\}^d$, where $\ell = n^{1/d}$. For a parameter $2 \leq \tau \leq \ell$, we consider rectangular subsets of the form[2]

$$\begin{aligned} H(\tau) &= \big\{\beta \in \mathbb{R}^{n-1} : |\beta_i| \leq t_i(\tau),\ i = 2, \ldots, n\big\}, \quad \text{where} \\ t_i(\tau) &= \begin{cases} C_n' / (\sum_{j_1,\ldots,j_d \leq \tau} \rho_{j_1,\ldots,j_d})^{1/2} & \text{if } i_1, \ldots, i_d \leq \tau \\ 0 & \text{otherwise} \end{cases}, \quad \text{for } i = 2, \ldots, n. \end{aligned}$$

It is not hard to check that $H(\tau) \subseteq \{\beta \in \mathbb{R}^{n-1} : \sum_{i=2}^n \rho_i \beta_i^2 \leq (C_n')^2\} = P_{-1}(\mathcal{E}_d(C_n'))$. Then, from (7.19),

$$\begin{aligned} \frac{5}{4}R(P_{-1}(\mathcal{E}_d(C_n'))) \geq \sup_{\tau} R_L(H(\tau)) &= \sup_{\tau} \frac{1}{n}\sum_{i=1}^n \frac{t_i(\tau)^2\sigma^2}{t_i(\tau)^2+\sigma^2} \\ &= \sup_{\tau} \frac{1}{n}\frac{(\tau^d-1)\sigma^2(C_n')^2}{(C_n')^2 + \sum_{j_1,\ldots,j_d \leq \tau} \rho_{j_1,\ldots,j_d}}. \end{aligned}$$

[2] Here, albeit unconvential, it helps to index $\beta \in H(\tau) \subseteq \mathbb{R}^{n-1}$ according to components $i = 2, \ldots, n$, rather than $i = 1, \ldots, n-1$. This is so that we may keep the index variable i to be in correspondence with positions on the grid.

The first equality is due to the fact that the minimax risk for rectangles decouples across dimensions, and the 1d minimax linear risk is straightforward to compute for an interval, as argued in the proof Theorem 7.6; the second equality simply comes from a short calculation following the definition of $t_i(\tau)$, $i = 2, \dots, n$. Applying Lemma 7.15, on the eigenvalues of the graph Laplacian matrix L for a d-dimensional grid, we have that for a constant $c > 0$,

$$\frac{(\tau^d - 1)\sigma^2(C_n')^2}{(C_n')^2 + \sum_{j_1,\dots,j_d \le \tau} \rho_{j_1,\dots,j_d}} \ge \frac{(\tau^d - 1)\sigma^2(C_n')^2}{(C_n')^2 + c\sigma^2\tau^{d+2}/\ell^2} \ge \frac{1}{2}\frac{\sigma^2(C_n')^2}{(C_n')^2\tau^{-d} + c\sigma^2\tau^2/\ell^2}.$$

We can choose τ to maximize the expression on the right above, given by

$$\tau^* = \left(\frac{\ell^2(C_n')^2}{c\sigma^2}\right)^{\frac{1}{d+2}}.$$

When $2 \le \tau^* \le \ell$, this provides us with the lower bound on the minimax risk

$$\frac{5}{4}R(P_{-1}(\mathcal{E}_d(C_n'))) \ge R_L(H(\tau^*)) \ge \frac{1}{2n}\frac{\tau^d\sigma^2(C_n')^2}{2(c\sigma^2)^{\frac{d}{d+2}}(C_n')^{\frac{4}{d+2}}\ell^{-\frac{2d}{d+2}}} = \frac{c_1}{n}(n\sigma^2)^{\frac{2}{d+2}}(C_n')^{\frac{2d}{d+2}}, \tag{7.20}$$

for a constant $c_1 > 0$. When $\tau^* < 2$, we can use $\tau = 2$ as lower bound on the minimax risk,

$$\frac{5}{4}R(P_{-1}(\mathcal{E}_d(C_n'))) \ge R_L(H(2)) \ge \frac{1}{2n}\frac{\sigma^2\ell^2(C_n')^2}{\ell^2(C_n')^2 2^{-d} + c\sigma^2 2^2} \ge \frac{c_2}{n}\ell^2(C_n')^2, \tag{7.21}$$

for a constant $c_2 > 0$, where in the last inequality, we used the fact that $\ell^2(C_n')^2 \le c\sigma^2 2^{d+2}$ (just a constant) since we are in the case $\tau^* < 2$. Finally, when $\tau^* > \ell$, we can use $\tau = \ell$ as a lower bound on the minimax risk,

$$\frac{5}{4}R(P_{-1}(\mathcal{E}_d(C_n'))) \ge R_L(H(\ell)) \ge \frac{1}{2n}\frac{\sigma^2(C_n')^2}{\ell^{-d}(C_n')^2 + c\sigma^2} \ge c_3\sigma^2, \tag{7.22}$$

for a constant $c_3 > 0$, where in the last inequality, we used that $c\sigma^2 \le \ell^{-d}(C_n')^2$ as we are in the case $\tau^* > \ell$. Taking a minimum of the lower bounds in (7.20), (7.21), (7.22), as a way to navigate the cases, gives us a final lower bound on $R(P_{-1}(\mathcal{E}_d(C_n')))$, and completes the proof.

7.7.7 Proof of Theorem 7.10 (Laplacian eigenmaps and Laplacian smoothing over Sobolev classes)

We will prove the results for Laplacian eigenmaps and Laplacian separately.

Laplacian eigenmaps. The smoother matrix for this estimator is $S_k = V_{[k]}V_{[k]}^T$, for a tuning parameter $k = 1, \dots, n$. From (7.16),

$$\mathbb{E}\big[\mathrm{MSE}(\hat{\theta}^{\mathrm{LE}}, \theta_0)\big] = \frac{\sigma^2}{n}k + \frac{1}{n}\|(I - S_k)\theta_0\|_2^2.$$

Now we write $k = \tau^d$, and analyze the max risk of the second term,

$$\begin{aligned}
\sup_{\theta_0 : \|D\theta_0\|_2 \le C'_n} \frac{1}{n}\|(I - S_k)\theta_0\|_2^2 &= \sup_{z : \|z\|_2 \le C'_n} \frac{1}{n}\|(I - S_k)D^\dagger z\|_2^2 \\
&= \frac{(C'_n)^2}{n}\sigma^2_{\max}\big((I - S_k)D^\dagger\big) \\
&\le \frac{(C'_n)^2}{n}\frac{1}{4\sin^2(\pi\tau/(2\ell))} \\
&\le \frac{(C'_n)^2}{n}\frac{4\ell^2}{\pi^2\tau^2}.
\end{aligned}$$

Here we denote by $\sigma_{\max}(A)$ the maximum singular value of a matrix A. The last inequality above used the simple lower bound $\sin(x) \ge x/2$ for $x \in [0, \pi/2]$. The earlier inequality used that

$$(I - S_k)D^\dagger = (I - V_{[k]}V_{[k]}^T)V^T(\Sigma^\dagger)^{1/2}U^T = \big[0, \ldots, 0, V_{k+1}, \ldots, V_n\big](\Sigma^\dagger)^{1/2}U^T,$$

where we have kept the same notation for the singular value decomposition of D as in the proof of Theorem 7.9. Therefore $\sigma^2_{\max}((I - S_k)D^\dagger)$ is the reciprocal of the $(k+1)$st smallest eigenvalue ρ_{k+1} of the graph Laplacian L. For any subset A of the set of eigenvalues $\lambda(L) = \{\rho_1, \ldots, \rho_n\}$ of the Laplacian, with $|A| = k$, note that $\rho_{k+1} \ge \min\, \lambda(L) \setminus A$. This means that, for our d-dimensional grid,

$$\begin{aligned}
\rho_{k+1} &\ge \min\ \lambda(L) \setminus \{\rho_{i_1,\ldots,i_d} : i_1, \ldots, i_d \le \tau\} \\
&= 4\sin^2(\pi\tau/(2\ell)),
\end{aligned}$$

where recall $\ell = n^{1/d}$, as explained by (7.23), in the proof of Lemma 7.15.

Hence, we have established

$$\sup_{\theta_0 : \|D\theta_0\|_2 \le C'_n} \mathbb{E}\big[\mathrm{MSE}(\hat\theta^{\mathrm{LE}}, \theta_0)\big] \le \frac{\sigma^2}{n} + \frac{\sigma^2}{n}\tau^d + \frac{(C'_n)^2}{n}\frac{4\ell^2}{\pi^2\tau^2}.$$

Choosing τ to balance the two terms on the right-hand side above results in $\tau^* = (2\ell C'_n/(\pi\sigma))^{\frac{2}{d+2}}$. Plugging in this choice of τ, while utilizing the bounds $1 \le \tau \le \ell$, very similar to the arguments given at the end of the proof of Theorem 7.9, gives the result for Laplacian eigenmaps.

Laplacian smoothing. The smoother matrix for this estimator is $S_\lambda = (I + \lambda L)^{-1}$, for a tuning parameter $\lambda \ge 0$. From (7.16),

$$\mathbb{E}\big[\mathrm{MSE}(\hat\theta^{\mathrm{LS}}, \theta_0)\big] = \frac{\sigma^2}{n}\sum_{i=1}^n \frac{1}{(1 + \lambda\rho_i)^2} + \frac{1}{n}\|(I - S_\lambda)\theta_0\|_2^2.$$

When $d = 1, 2$, or 3, the first term upper is bounded by $c_1\sigma^2/n + c_2\sigma^2/\lambda^{d/2}$, for some constants $c_1, c_2 > 0$, by Lemma 7.16. As for the second term,

$$\begin{aligned}
\sup_{\theta_0 : \|D\theta_0\|_2 \le C'_n} \frac{1}{n}\|(I - S_\lambda)\theta_0\|_2^2 &= \sup_{z : \|z\|_2 \le C'_n} \|(I - S_\lambda)D^\dagger z\|_2^2 \\
&= \frac{(C'_n)^2}{n}\sigma^2_{\max}\big((I - S_\lambda)D^\dagger\big) \\
&= \frac{(C'_n)^2}{n} \max_{i=2,\ldots,n} \left(1 - \frac{1}{1 + \lambda\rho_i}\right)^2 \frac{1}{\rho_i} \\
&= \frac{(C'_n)^2}{n}\lambda \max_{i=2,\ldots,n} \frac{\lambda\rho_i}{(1 + \lambda\rho_i)^2} \\
&\le \frac{(C'_n)^2\lambda}{4n}.
\end{aligned}$$

In the third equality we have used the fact the eigenvectors of $I - S_\lambda$ are the left singular vectors of $D^\dagger$, and in the last inequailty we have used the simple upper bound $f(x) = x/(1 + x)^2 \le 1/4$ for $x \ge 0$ (this function being maximized at $x = 1$).

Therefore, from what we have shown,

$$\sup_{\theta_0 : \|D\theta_0\|_2 \le C'_n} \mathbb{E}\big[\mathrm{MSE}(\hat{\theta}^{\mathrm{LS}}, \theta_0)\big] \le \frac{c_1\sigma^2}{n} + \frac{c_2\sigma^2}{\lambda^{d/2}} + \frac{(C'_n)^2\lambda}{4n}.$$

Choosing λ to balance the two terms on the right-hand side above gives $\lambda^* = c(n/(C'_n)^2)^{2/(d+2)}$, for a constant $c > 0$. Plugging in this choice, and using upper bounds from the trivial cases $\lambda = 0$ and $\lambda = \infty$ when C'_n is very small or very large, respectively, gives the result for Laplacian smoothing. □

Remark 7.14. *When $d = 4$, Lemma 7.16 gives a slightly worse upper bound on $\sum_{i=1}^n 1/(1 + \lambda\rho_i)^2$, with an "extra" term $(nc_2/\lambda^{d/2}))\log(1 + c_3\lambda)$, for constants $c_2, c_3 > 0$. It is not hard to show, by tracing through the same arguments as given above that we can use this to establish an upper bound on the max risk of*

$$\sup_{\theta_0 \in \mathcal{S}_d(C'_n)} \mathbb{E}\big[\mathrm{MSE}(\hat{\theta}^{\mathrm{LE}}, \theta_0)\big] \le \frac{c}{n}\Big((n\sigma^2)^{\frac{2}{d+2}}(C'_n)^{\frac{2d}{d+2}} \log(n/(C'_n)^2) \wedge n\sigma^2 \wedge n^{2/d}(C'_n)^2\Big) + \frac{c\sigma^2}{n},$$

only slightly worse than the minimax optimal rate, by a log factor.

When $d \ge 5$, our analysis provides a much worse bound for the max risk of Laplacian smoothing, as the integral denoted $I(d)$ in the proof of Lemma 7.16 grows very large when $d \ge 5$. We conjecture that this not due to slack in our proof technique, but rather, to the Laplacian smoothing estimator itself, since all inequalities the proof are fairly tight.

7.8 Utility lemmas used in the proofs of Theorems 7.9 and 7.10

This section contains some calculations on the partial sums of eigenvalues of the Laplacian matrix L, for d-dimensions grids. These are useful for the proofs of both Theorem 7.9 and Theorem 7.10.

Lemma 7.15. *Let $L \in \mathbb{R}^{n\times n}$ denote the graph Laplacian matrix of a d-dimensional grid graph, and $\rho_{i_1,\ldots,i_d}$, $(i_1,\ldots,i_d) \in \{1,\ldots,\ell\}^d$ be its eigenvalues, where $\ell = n^{1/d}$. Then there exists a constant $c > 0$ (dependent on d) such that, for any $1 \le \tau \le \ell$,*

$$\sum_{(i_1,\ldots,i_d)\in\{1,\ldots,\tau\}^d} \rho_{i_1,\cdots,i_d} \le c\frac{\tau^{d+2}}{\ell^2}.$$

Proof. The eigenvalues of L can be written explicitly as

$$\rho_i = 4\sin^2\Big(\frac{\pi(i_1-1)}{2\ell}\Big) + \ldots + 4\sin^2\Big(\frac{\pi(i_d-1)}{2\ell}\Big), \quad (i_1,\ldots,i_d) \in \{1,\ldots,\ell\}^d. \tag{7.23}$$

This follows from known facts about the eigenvalues for the Laplacian matrix of a 1d grid, and the fact that the Laplacian matrix for higher-dimensional grids can be expressed in terms of a Kronecker sum of the Laplacian matrix of an appropriate 1d grid (e.g., [59, 118, 134, 165, 242, 248]). We now use the fact that $\sin(x) \le x$ for all $x \ge 0$, which gives us the upper bound

$$\begin{aligned}
\sum_{(i_1,\ldots,i_d)\in\{1,\ldots,\tau\}^d} \rho_{i_1,\cdots,i_d} &\le \frac{\pi^2}{\ell^2} \sum_{(i_1,\ldots,i_d)\in\{1,\ldots,\tau\}^d} \Big((i_1-1)^2 + \ldots + (i_d-1)^2\Big) \\
&\le \frac{\pi^2 d}{\ell^2}\tau^{d-1}\sum_{i=1}^{\tau}(i-1)^2 \\
&\le \frac{\pi^2 d}{\ell^2}\tau^{d-1}\tau^3 \\
&= \frac{\pi^2 d}{\ell^2}\tau^{d+2},
\end{aligned}$$

as desired. □

Lemma 7.16. *Let $L \in \mathbb{R}^{n\times n}$ denote the graph Laplacian matrix of a d-dimensional grid graph, and ρ_i, $i = 1,\ldots,n$ be its eigenvalues. Let $\lambda \ge 0$ be arbitrary. For $d = 1, 2,$ or 3, there are constants $c_1, c_2 > 0$ such that*

$$\sum_{i=1}^{n} \frac{1}{(1+\lambda\rho_i)^2} \le c_1 + c_2\frac{n}{\lambda^{d/2}}.$$

For $d = 4$, there are constants $c_1, c_2, c_3 > 0$ such that

$$\sum_{i=1}^{n} \frac{1}{(1+\lambda\rho_i)^2} \le c_1 + c_2\frac{n}{\lambda^{d/2}}\Big(1 + \log(1 + c_3\lambda)\Big).$$

Proof. We will use the explicit form of the eigenvalues as given in the proof of Lemma 7.15. In the expressions below, we use $c > 0$ to denote a constant whose value may change from line to line. Using the inequality $\sin x \geq x/2$ for $x \in [0, \pi/2]$,

$$\begin{aligned}
\sum_{i=1}^{n} \frac{1}{(1+\lambda\rho_i)^2} &\leq \sum_{(i_1,\ldots,i_d)\in\{1,\ldots,\ell\}^d} \frac{1}{\left(1+\lambda\frac{\pi^2}{4\ell^2}\sum_{j=1}^{d}(i_j-1)^2\right)^2} \\
&\leq 1 + \int_{[0,\ell]^d} \frac{1}{\left(1+\lambda\frac{\pi^2}{4}\sum_{j=1}^{d} x_j^2/\ell^2\right)^2}\,dx \\
&= 1 + c\int_0^{\ell\sqrt{d}} \frac{1}{\left(1+\lambda\frac{\pi^2}{4}r^2/\ell^2\right)^2} r^{d-1}\,dr \\
&= 1 + c\frac{n}{\lambda^{d/2}} \underbrace{\int_0^{\frac{\pi}{2}\sqrt{\lambda d}} \frac{u^{d-1}}{(1+u^2)^2}\,du}_{I(d)}\,.
\end{aligned}$$

In the second inequality, we used the fact that the right-endpoint Riemann sum is always an underestimate for the integral of a function that is monotone nonincreasing in each coordinate. In the third, we made a change to spherical coordinates, and suppressed all of the angular variables, as they contribute at most a constant factor. It remains to compute $I(d)$, which can be done by symbolic integration:

$$\begin{aligned}
I(1) &= \frac{\pi\sqrt{d}}{4\left(1+\frac{\pi^2}{4}\lambda d\right)} + \frac{1}{2}\tan^{-1}\left(\frac{\pi}{2}\sqrt{\lambda d}\right) \leq \frac{1}{4} + \frac{\pi}{4}, \\
I(2) &= \frac{1}{2} - \frac{1}{2\left(1+\frac{\pi^2}{4}\lambda d\right)} \leq \frac{1}{2}, \\
I(3) &= \frac{1}{2}\tan^{-1}\left(\frac{\pi}{2}\sqrt{\lambda d}\right) \leq \frac{\pi}{4}, \quad \text{and} \\
I(4) &= \frac{1}{2}\log\left(1+\frac{\pi^2}{4}\lambda d\right) + \frac{1}{2\left(1+\frac{\pi^2}{4}\lambda d\right)} - \frac{1}{2} \leq \frac{1}{2}\log\left(1+\frac{\pi^2}{4}\lambda d\right) + \frac{1}{2}.
\end{aligned}$$

This completes the proof. □

7.9 Additional experiments comparing TV denoising and Laplacian smoothing for piecewise constant functions

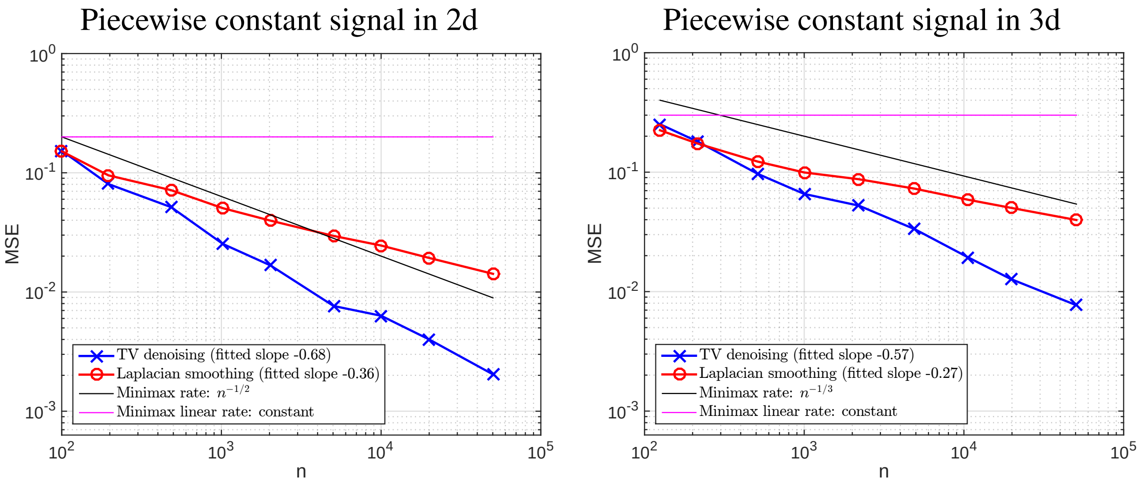


Figure 7.6: *MSE curves for estimating a "piecewise constant" signal, having a single elevated region, over 2d and 3d grids. For each n, the results were averaged over 5 repetitions, and the Laplacian smoothing and TV denoising estimators were tuned for best average MSE performance. We set θ_0 to satisfy $\|D\theta_0\|_1 \asymp n^{1-1/d}$, matching the canonical scaling. Note that all estimators achieve better performance than that dictated by their minimax rates.*

Chapter 8

Falling factorial basis and continuous extensions

In this section, we backtrack a little bit to the univariate trend filtering and discuss ways to extrapolate from the discrete fitted values of trend filtering to functions. This is needed because if you recall, nonparametric regression in its canonical form, demands the output to be a function..

This is a natural feat for any methods that starts from a variational formulation and then discretized through certain "representer theorem". However, trend filtering starts with discrete domain to begin with and it is unclear whether there is a natural set of basis functions that it actually corresponds to in the continuous domain.

Specifically, we study a novel spline-like basis, which we name the "falling factorial basis", bearing many similarities to the classic truncated power basis. The advantage of the falling factorial basis is that it enables rapid, linear-time computations in basis matrix multiplication and basis matrix inversion. The falling factorial functions are not actually splines, but are close enough to splines that they provably retain some of the favorable properties of the latter functions. We examine their application in two problems: trend filtering over arbitrary input points, and a higher-order variant of the two-sample Kolmogorov-Smirnov test.

8.1 Introduction

Splines are an old concept, and they play important roles in various subfields of mathematics and statistics; see e.g., de Boor [62], Wahba [240] for two classic references. In words, a spline of order k is a piecewise polynomial of degree k that is continuous and has continuous derivatives of orders $1, 2, \dots k-1$ at its knot points. In this chapter, we look at a new twist on an old problem: we examine a novel set of spline-like basis functions with sound computational and statistical properties. This basis, which we call the *falling factorial basis*, is particularly attractive when assessing higher order of smoothness via the total variation operator, due to the capability for sparse decompositions. A summary of our main findings is as follows.

- The falling factorial basis and its inverse both admit a linear-time transformation, i.e., much faster decompositions than the spline basis, and even faster than, e.g., the fast Fourier transform.
- For all practical purposes, the falling factorial basis shares the statistical properties of the spline basis. We derive a sharp characterization of the discrepancy between the two bases in terms of the polynomial degree and the distance between sampling points.
- We simplify and extend known convergence results on trend filtering, a nonparametric regression technique that implicitly employs the falling factorial basis.
- We also extend the Kolmogorov-Smirnov two-sample test to account for higher order differences, and utilize the falling factorial basis for rapid computations. We provide no theory but demonstrate excellent empirical results, improving on, e.g., the maximum mean discrepancy [104] and Anderson-Darling [9] tests.

In short, the falling factorial function class offers an exciting prospect for univariate function regularization.

Now let us review some basics. Recall that the set of kth order splines with knots over a fixed set of n points forms an $(n+k+1)$-dimensional subspace of functions. Here and throughout, we assume that we are given ordered input points $x_1 < x_2 < \ldots < x_n$ and a polynomial order $k \geq 0$, and we define a set of knots $T = \{t_1, \ldots t_{n-k-1}\}$ by excluding some of the input points at the left and right boundaries, in particular,

$$T = \begin{cases} \{x_{k/2+2}, \ldots x_{n-k/2}\} & \text{if } k \text{ is even,} \\ \{x_{(k+1)/2+1}, \ldots x_{n-(k+1)/2}\} & \text{if } k \text{ is odd.} \end{cases} \tag{8.1}$$

The set of kth order splines with knots in T hence forms an n-dimensional subspace of functions. The canonical parametrization for this subspace is given by the truncated power basis, $g_1, \ldots g_n$, defined as

$$\begin{gathered} g_1(x) = 1,\ g_2(x) = x,\ \ldots\ g_{k+1}(x) = x^k, \\ g_{k+1+j}(x) = (x - t_j)^k \cdot 1\{x \geq t_j\},\ \ j = 1, \ldots n-k-1. \end{gathered} \tag{8.2}$$

These functions can also be used to define the truncated power basis matrix, $G \in \mathbb{R}^{n \times n}$, by

$$G_{ij} = g_j(x_i), \quad i, j = 1, \ldots n, \tag{8.3}$$

i.e., the columns of G give the evaluations of the basis functions $g_1, \ldots g_n$ over the inputs $x_1, \ldots x_n$. As $g_1, \ldots g_n$ are linearly independent functions, G has linearly independent columns, and hence G is invertible.

As noted, our focus is a related but different set of basis functions, named the falling factorial basis functions. We define these functions, for a given order $k \geq 0$, as

$$\begin{gathered} h_j(x) = \prod_{\ell=1}^{j-1} (x - x_\ell), \quad j = 1, \ldots k+1, \\ h_{k+1+j}(x) = \prod_{\ell=1}^{k} (x - x_{j+\ell}) \cdot 1\{x \geq x_{j+k}\},\ \ j = 1, \ldots n-k-1. \end{gathered} \tag{8.4}$$

(Our convention is to take the empty product to be 1, so that $h_1(x) = 1$.) The falling factorial basis functions are piecewise polynomial, and have an analogous form to the truncated power basis functions in (8.2). Loosely speaking, they are given by replacing an rth order power function in the truncated power basis with an appropriate r-term product, e.g., replacing x^2 with $(x-x_2)(x-x_1)$, and $(x-t_j)^k$ with $(x-x_{j+k})(x-x_{j+k-1}) \cdot \ldots (x-x_{j+1})$. Similar to the above, we can define the falling factorial basis matrix, $H \in \mathbb{R}^{n\times n}$, by

$$H_{ij} = h_j(x_i), \quad i, j = 1, \ldots n, \tag{8.5}$$

and the linear independence of $h_1, \ldots h_n$ implies that H too is invertible.

Note that the first $k+1$ functions of either basis, the truncated power or falling factorial basis, span the same space (the space of kth order polynomials). But this is not true of the last $n-k-1$ functions. Direct calculation shows that, while continuous, the function h_{j+k+1} has discontinuous derivatives of all orders $1, \ldots k$ at the point x_{j+k}, for $j = 1, \ldots n-k-1$. This means that the falling factorial functions $h_{k+2}, \ldots h_n$ are not actually kth order splines, but are instead continuous kth order piecewise polynomials that are “close to” splines. Why would we ever use such a seemingly strange basis as that defined in (8.4)? To repeat what was summarized above, the falling factorial functions allow for linear-time (and closed-form) computations with the basis matrix H and its inverse. Meanwhile, the falling factorial functions are close enough to the truncated power functions that using them in several spline-based problems (i.e., using H in place of G) can be statistically legitimized. We make this statement precise in the sections that follow.

As we see it, there is really nothing about their form in (8.4) that suggests a particularly special computational structure of the falling factorial basis functions. Our interest in these functions arose from a study of trend filtering, a nonparametric regression estimator, where the inverse of H plays a natural role. The inverse of H is a kind of discrete derivative operator of order $k+1$, properly adjusted for the spacings between the input points $x_1, \ldots x_n$. It is really the special, banded structure of this derivative operator that underlies the computational efficiency surrounding the falling factorial basis; all of the computational routines proposed in this chapter leverage this structure.

Here is an outline for rest of this article. In Section 8.2, we describe a number of basic properties of the falling factorial basis functions, culminating in fast linear-time algorithms for multiplication H and H^{-1}, and tight error bounds between H and the truncated power basis matrix G. Section 8.3 discusses B-splines, which provide another highly efficient basis for spline manipulations; we explain why the falling factorial basis offers a preferred parametrization in some specific statistical applications, e.g., the ones we present in Sections 8.4 and 8.5. Section 8.4 covers trend filtering, and extends a known convergence result for trend filtering over evenly spaced input points [221] to the case of arbitrary input points. The conclusion is that trend filtering estimates converge at the minimax rate (over a large class of true functions) assuming only mild conditions on the inputs. In Section 8.5, we consider a higher order extension of the classic two-sample Kolmogorov-Smirnov test. We find this test to have better power in detecting higher order (tail) differences between distributions when compared to the usual Kolmogorov-Smirnov test; furthermore, by employing the falling factorial functions, it can computed in linear time. In Section 8.6, we end with some discussion.

8.2 Basic properties

Consider the falling factorial basis matrix $H \in \mathbb{R}^{n \times n}$, as defined in (8.5), over input points $x_1 < \ldots < x_n$. The following subsections describe a recursive decomposition for H and its inverse, which lead to fast computational methods for multiplication by H and H^{-1} (as well as H^T and $(H^T)^{-1}$). The last subsection bounds the maximum absolute difference bewteen the elements of H and G, the truncated power basis matrix (also defined over $x_1, \ldots x_n$). Lemmas 8.1, 8.2, 8.4 below were derived in Tibshirani [221] for the special case of evenly spaced inputs, $x_i = i/n$ for $i = 1, \ldots n$. We reiterate that here we consider generic input points $x_1, \ldots x_n$. In the interest of space, we defer all proofs to Section 8.7.

8.2.1 Recursive decomposition

Our first result shows that H decomposes into a product of simpler matrices. It helpful to define, for $k \geq 1$,

$$\Delta^{(k)} = \mathrm{diag}\big(x_{k+1} - x_1,\, x_{k+2} - x_2,\, \ldots\, x_n - x_{n-k}\big),$$

the $(n-k) \times (n-k)$ diagonal matrix whose diagonal elements contain the k-hop gaps between input points.

Lemma 8.1. *Let I_m denote the $m \times m$ identity matrix, and L_m the $m \times m$ lower triangular matrix of 1s. If we write $H^{(k)}$ for the falling factorial basis matrix of order k, then in this notation, we have $H^{(0)} = L_n$, and for $k \geq 1$,*

$$H^{(k)} = H^{(k-1)} \cdot \begin{bmatrix} I_k & 0 \\ 0 & \Delta^{(k)} L_{n-k} \end{bmatrix}. \tag{8.6}$$

Lemma 8.1 is really a key workhorse behind many properties of the falling factorial basis functions. E.g., it acts as a building block for results to come: immediately, the representation (8.6) suggests both an analogous inverse representation for $H^{(k)}$, and a computational strategy for matrix multiplication by $H^{(k)}$. These are discussed in the next two subsections. We remark that the result in the lemma may seem surprising, as there is not an apparent connection between the falling factorial functions in (8.4) and the recursion in (8.6), which is based on taking cumulative sums at varying offsets (the rightmost matrix in (8.6)). We were led to this result by studying the evenly spaced case; its proof for the present case is considerably longer and more technical, but the statement of the lemma is still quite simple.

8.2.2 The inverse basis

The result in Lemma 8.1 clearly also implies a result on the inverse operators, namely, that $(H^{(0)})^{-1} = L_n^{-1}$, and

$$(H^{(k)})^{-1} = \begin{bmatrix} I_k & 0 \\ 0 & L_{n-k}^{-1}(\Delta^{(k)})^{-1} \end{bmatrix} \cdot (H^{(k-1)})^{-1} \tag{8.7}$$

for all $k \geq 1$. We note that

$$L_m^{-1} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ \vdots & & & \\ 1 & 1 & \dots & 1 \end{bmatrix}^{-1} = \begin{bmatrix} e_1^T \\ D^{(1)} \end{bmatrix}, \tag{8.8}$$

with $e_1 = (1, 0, \dots 0) \in \mathbb{R}^m$ being the first standard basis vector, and $D^{(1)} \in \mathbb{R}^{(m-1)\times m}$ the first discrete difference operator

$$D^{(1)} = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix}, \tag{8.9}$$

With this in mind, the recursion in (8.7) now looks like the construction of the higher order discrete difference operators, over the input $x_1, \dots x_n$. To define these operators, we start with the first order discrete difference operator $D^{(1)} \in \mathbb{R}^{(n-1)\times n}$ as in (8.9), and define the higher order difference discrete operators according to

$$D^{(k+1)} = D^{(1)} \cdot k \cdot (\Delta^{(k)})^{-1} \cdot D^{(k)}, \tag{8.10}$$

for $k \geq 1$. As $D^{(k+1)} \in \mathbb{R}^{(n-k-1)\times n}$, leading matrix $D^{(1)}$ above denotes the $(n-k-1) \times (n-k)$ version of the first order difference operator in (8.9).

To gather intuition, we can think of $D^{(k)}$ as a type of discrete kth order derivative operator across the underlying points $x_1, \dots x_n$; i.e., given an arbitrary sequence $u = (u_1, \dots u_n) \in \mathbb{R}^n$ over the positions $x_1, \dots x_n$, respectively, we can think of $(D^{(k)}u)_i$ as the discrete kth derivative of the sequence u evaluated at the point x_i. It is not difficult to see, from its definition, that $D^{(k)}$ is a banded matrix with bandwidth $k + 1$. The middle (diagonal) term in (8.10) accounts for the fact that the underlying positions $x_1, \dots x_n$ are not necessarily evenly spaced. When the input points are evenly spaced, this term contributes only a constant factor, and the difference operators $D^{(k)}$, $k = 1, 2, 3, \dots$ take a very simple form, where each row is a shifted version of the previous, and the nonzero elements are given by the kth order binomial coefficients (with alternating signs); see Tibshirani [221].

By staring at (8.7) and (8.10), one can see that the falling factorial basis matrices and discrete difference operators are essentially inverses of each other. The story is only slightly more complicated because the difference matrices are not square.

Lemma 8.2. *If $H^{(k)}$ is the kth order falling factorial basis matrix defined over the inputs $x_1, \ldots x_n$, and $D^{(k+1)}$ is the $(k+1)$st order discrete difference operator defined over the same inputs $x_1 \ldots x_n$, then*

$$(H^{(k)})^{-1} = \begin{bmatrix} C \\ \frac{1}{k!} \cdot D^{(k+1)} \end{bmatrix}, \tag{8.11}$$

for an explicit matrix $C \in \mathbb{R}^{(k+1) \times n}$. If we let A_i denote the ith row of a matrix A, then C has first row $C_1 = e_1^T$, and subsequent rows

$$C_{i+1} = \left[\frac{1}{(i-1)!} \cdot (\Delta^{(i)})^{-1} \cdot D^{(i)} \right]_1, \quad i = 1, \ldots k.$$

Lemma 8.2 shows that the last $n-k-1$ rows of $(H^{(k)})^{-1}$ are given exactly by $D^{(k+1)}/k!$. This serves as the crucial link between the falling factorial basis functions and trend filtering, discussed in Section 8.4. The route to proving this result revealed the recursive expressions (8.6) and (8.7), and in fact these are of great computational interest in their own right, as we discuss next.

8.2.3 Fast matrix multiplication

The recursions in (8.6) and (8.7) allow us to apply $H^{(k)}$ and $(H^{(k)})^{-1}$ with specialized linear-time algorithms. Further, these algorithms are completely in-place: we do not need to form the matrices $H^{(k)}$ or $(H^{(k)})^{-1}$, and the algorithms operate entirely by manipulating the input vector (the vector to be multiplied).

Lemma 8.3. *For the kth order falling factorial basis matrix $H^{(k)} \in \mathbb{R}^{n \times n}$, over arbitrary sorted inputs $x_1, \ldots x_n$, multiplication by $H^{(k)}$ and $(H^{(k)})^{-1}$ can each be computed in $O(nk)$ in-place operations with zero memory requirements (aside from storing the input points and the vector to be multiplied), i.e., we do not need to form $H^{(k)}$ or $(H^{(k)})^{-1}$. Algorithms 7 and 8 give the details. The same is true for matrix multiplication by $(H^{(k)})^T$ and $[(H^{(k)})^T]^{-1}$; Algorithms 9 and 10, found in Section 8.7.3, give the details.*

Note that the lemma assumes presorted inputs $x_1, \ldots x_n$ (sorting requires an extra $O(n \log n)$ operations). The routines for multiplication by $H^{(k)}$ and $(H^{(k)})^{-1}$, in Algorithms 7 and 8, are really just given by inverting each term one at a time in the product representations (8.6) and (8.7). They are composed of elementary in-place operations, like cumulative sums and pairwise differences. This brings to mind a comparison to wavelets, as both the wavelet and inverse wavelets operators can be viewed as highly specialized linear-time matrix multplications.

Borrowing from the wavelet perspective, given a sampled signal $y_i = f(x_i)$, $i = 1, \ldots n$, the action $(H^{(k)})^{-1}y$ can be thought of as the forward transform under the piecewise polynomial falling factorial basis, and $H^{(k)}y$ as the backward or inverse transform under this basis. It might be interesting to consider the applicability of such transforms to signal processing tasks, but this is beyond the scope of the current paper, and we leave it to potential future work.

We do however include a computational comparison between the forward and backward falling factorial transforms, in Algorithms 8 and 7, and the well-studied Fourier and wavelet transforms.

Algorithm 7 Multiplication by $H^{(k)}$

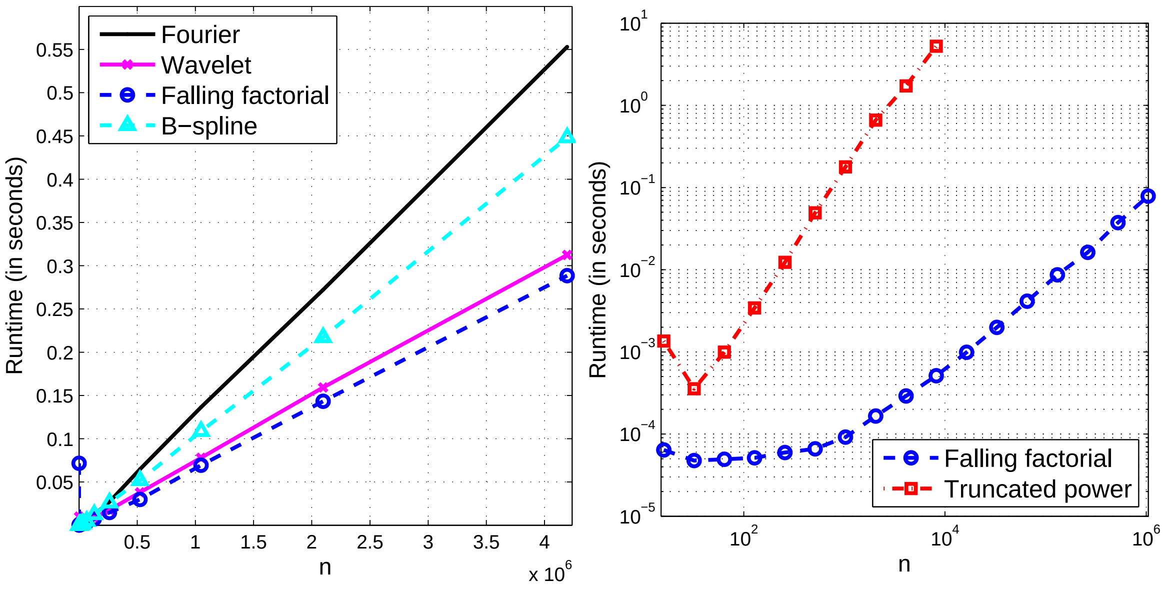
```
**Input:** Vector to be multiplied $y \in \mathbb{R}^n$, order $k \geq 0$, sorted inputs vector $x \in \mathbb{R}^n$.
**Output:** $y$ is overwritten by $H^{(k)}y$.
for $i = k$ to $0$ do
  $y_{(i+1):n} = \mathrm{cumsum}(y_{(i+1):n})$,
  where $y_{a:b}$ denotes the subvector $(y_a, y_{a+1}, ..., y_b)$ and $\mathrm{cumsum}$ is the cumulative sum operator.
  if $i \neq 0$ then
    $y_{(i+1):n} = \big(x_{(i+1):n} - x_{1:(n-i)}\big) \,.\!*\; y_{(i+1):n}$,
    where $.*$ denotes entrywise multiplication.
  end if
end for
Return $y$.
```

Algorithm 8 Multiplication by $(H^{(k)})^{-1}$

```
**Input:** Vector to be multiplied $y \in \mathbb{R}^n$, order $k \geq 0$, sorted inputs vector $x \in \mathbb{R}^n$.
**Output:** $y$ is overwritten by $(H^{(k)})^{-1}y$.
for $i = 0$ to $k$ do
  if $i \neq 0$ then
    $y_{(i+1):n} = y_{i+1:n} \,./\, \big(x_{(i+1):n} - x_{1:(n-i]}\big)$,
  end if
  $y_{(i+2):n} = \mathrm{diff}(y_{(i+1):n})$,
  where $\mathrm{diff}$ is the pairwise difference operator.
end for
Return $y$.
```

Figure 8.1a shows the runtimes of one complete cycle of falling factorial transforms (i.e., one forward and one backward transform), with $k = 3$, versus one cycle of fast Fourier transforms and one cycle of wavelet transforms (using symmlets). The comparison was run in Matlab, and we used Matlab's "fft" and "ifft" functions for the fast Fourier transforms, and the Stanford WaveLab's "FWT_PO" and "IWT_PO" functions (with symmlet filters) for the wavelet transforms [41]. These functions all call on C implementations that have been ported to Matlab using MEX-functions, and so we did the same with our falling factorial transforms to even the comparison. For each problem size n, we chose evenly spaced inputs (this is required for the Fourier and wavelet transforms, but recall, not for the falling factorial transform), and averaged the results over 10 repetitions. The figure clearly demonstrates a linear scaling for the runtimes of the falling factorial transform, which matches their theoretical $O(n)$ complexity; the wavelet and fast fourier transforms also behave as expected, with the former having $O(n)$ complexity, and the latter $O(n \log n)$. In fact, a raw comparison of times shows that our implementation of the falling factorial transforms runs slightly faster than the highly-optimized wavelet transforms from the Stanford WaveLab.

For completeness, Figure 8.1b displays a comparison between the falling factorial transforms and the corresponding transforms using the truncated power basis (also with $k = 3$). We see that



(a) Falling factorial vs. Fourier, wavelet, and B-spline transforms (linear scale)

(b) Falling factorial (H) vs. truncated power (G) transforms (log-log scale)

Figure 8.1: Comparison of runtimes for different transforms. The experiments were performed on a laptop computer.

the latter scale quadratically with n, which is again to be expected, as the truncated power basis matrix is essentially lower triangular.

8.2.4 Proximity to truncated power basis

With computational efficiency having been assured by the last lemma, our next lemma lays the footing for the statistical credibility of the falling factorial basis.

Lemma 8.4. *Let $G^{(k)}$ and $H^{(k)}$ be the kth order truncated power and falling factorial matrices, defined over inputs $0 \leq x_1 < \ldots < x_n \leq 1$. Let $\delta = \max_{i=1,\ldots n}(x_i - x_{i-1})$, where we write $x_0 = 0$. Then*

$$\max_{i,j=1,\ldots n} |G_{ij}^{(k)} - H_{ij}^{(k)}| \leq k^2\delta.$$

This tight elementwise bound between the two basis matrices will be used in Section 8.4 to prove a result on the convergence of trend filtering estimates. We will also discuss its importance in the context of a fast nonparametric two-sample test in Section 8.5. To give a preview: in many problem instances, the maximum gap δ between adjacent sorted inputs $x_1, \ldots x_n$ is of the order $\log n/n$ (for a more precise statement see Lemma 8.6), and this means that the maximum absolute discrepancy between the elements of $G^{(k)}$ and $H^{(k)}$ decays very quickly.

8.3 Why not just use B-splines?

B-splines already provide a computationally efficient parametrization for the set of kth order splines; i.e., since they produce banded basis matrices, we can already perform linear-time basis matrix multiplication and inversion with B-splines. To confirm this point empirically, we included B-splines in the timing comparison of Section 8.2.3, refer to Figure 8.1a for the results. So, why not always use B-splines in place of the falling factorial basis, which only approximately spans the space of splines?

A major reason is that the falling factorial functions (like the truncated power functions) admit a sparse representation under the total variation operator, whereas the B-spline functions do not. To be more specific, suppose that $f_1, \ldots f_m$ are kth order piecewise polynomial functions with knots at the points $0 \leq z_1 < \ldots < z_r \leq 1$, where $m = r + k + 1$. Then, for $f = \sum_{j=1}^m \alpha_j f_j$, we have

$$\mathrm{TV}(f^{(k)}) = \sum_{i=1}^r \left| \sum_{j=1}^m \Big(f_j^{(k)}(z_i) - f_j^{(k)}(z_{i-1}) \Big) \cdot \alpha_j \right|,$$

denoting $z_0 = 0$ for ease of notation. If $f_1, \ldots f_m$ are the falling factorial functions defined over the points $z_1, \ldots z_r$, then the term $f_j^{(k)}(z_i) - f_j^{(k)}(z_{i-1})$ is equal to 0 for all i, j, except when $i = j - k - 1$ and $j \geq k + 2$, in which case it equals 1. Therefore, $\mathrm{TV}(f^{(k)}) = \sum_{j=k+2}^m |\alpha_j|$, a simple sum of absolute coefficients in the falling factorial expansion. The same result holds for the truncated power basis functions. But if $f_1, \ldots f_m$ are B-splines, then this is not true; one can show that in this case $\mathrm{TV}(f^{(k)}) = \|C\alpha\|_1$, where C is a (generically) dense matrix. The fact that C is dense makes it cumbersome, both mathematically and computationally, to use the B-spline parametrization in spline problems involving total variation, such as those discussed in Sections 8.4 and 8.5.

8.4 Trend filtering for arbitrary inputs

We will now apply the results in previous sections to generalize the univariate trend filtering as we described in the "motivation and overview". We first quickly recap the setting. Suppose that we observe

$$y_i = f_0(x_i) + \epsilon_i, \quad i = 1, \ldots n, \tag{8.12}$$

for a true (unknown) regression function f_0, inputs $x_1 < \ldots < x_n \in \mathbb{R}$, and errors $\epsilon_1, \ldots \epsilon_n$. The task is to come up with estimators of f_0.

The trend filtering estimator was first proposed by Kim et al. [129], and further studied by Tibshirani [221] with a spline-like continuous extension specified for the case where $x_i = i/n$, $i = 1, \ldots n$.

In the present section, we allow $x_1, \ldots x_n$ to be arbitrary, and extend the convergence guarantees for trend filtering, utilizing the properties of the falling factorial basis derived in Section 8.2.

The new trend filtering estimate $\hat{\theta}$ of order $k \geq 0$ is defined by

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \; \frac{1}{2}\|y - \theta\|_2^2 + \lambda \cdot \frac{1}{k!}\|D^{(k+1)}\theta\|_1, \tag{8.13}$$

where $y = (y_1, \ldots y_n) \in \mathbb{R}^n$, $D^{(k+1)} \in \mathbb{R}^{(n-k-1)\times n}$ is the $(k+1)$st order discrete difference operator defined in (8.10) over the input points $x_1, \ldots x_n$, and $\lambda \geq 0$ is a tuning parameter. We can think of the components of $\hat{\theta}$ as defining an estimated function $\hat{f}$ over the input points. To give an example, in Figure 8.2, we drew noisy observations from a smooth underlying function, where the input points $x_1, \ldots x_n$ were sampled uniformly at random over $[0, 1]$, and we computed the trend filtering estimate $\hat{\theta}$ with $k = 3$ and a particular choice of λ. From the plot (where we interpolated between $(x_1, \hat{\theta}_1), \ldots (x_n, \hat{\theta}_n)$ for visualization purposes), we can see that the implicitly defined trend filtering function $\hat{f}$ displays a piecewise cubic structure, with adaptively chosen knot points. Lemma 8.2 makes this connection precise by showing that such a function $\hat{f}$ is indeed a linear combination of falling factorial functions. Letting $\theta = H^{(k)}\alpha$, where $H^{(k)} \in \mathbb{R}^{n\times n}$ is the kth order falling factorial basis matrix defined over the inputs $x_1, \ldots x_n$, the trend filtering problem in (8.13) becomes

$$\hat{\alpha} = \underset{\alpha \in \mathbb{R}^n}{\operatorname{argmin}} \; \frac{1}{2}\|y - H^{(k)}\alpha\|_2^2 + \lambda \cdot \sum_{j=k+2}^{n} |\alpha_j|, \tag{8.14}$$

equivalent to the functional minimization problem

$$\hat{f} = \underset{f \in \mathcal{H}_k}{\operatorname{argmin}} \; \frac{1}{2}\sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 + \lambda \cdot \mathrm{TV}\big(f^{(k)}\big), \tag{8.15}$$

where $\mathcal{H}_k = \mathrm{span}\{h_1, \ldots h_n\}$ is the span of the kth order falling factorial functions in (8.4), $\mathrm{TV}(\cdot)$ denotes the total variation operator, and $f^{(k)}$ denotes the kth weak derivative of f. In other words, the solutions of problems (8.13) and (8.15) are related by $\hat{\theta}_i = \hat{f}(x_i)$, $i = 1, \ldots n$. The trend filtering estimate hence verifiably exhibits the structure of a kth order piecewise polynomial function, with knots at a subset of $x_1, \ldots x_n$, and this function is not necessarily a spline, but is close to one (since it lies in the span of the falling factorial functions $h_1, \ldots h_n$).

In Figure 8.2, we also fit a smoothing spline estimate to the same example data. A striking difference: the trend filtering estimate is far more locally adaptive towards the middle of plot, where the underlying function is less smooth (the two estimates were tuned to have the same degrees of freedom, to even the comparison). This phenomenon is investigated in Tibshirani [221], where it is shown that trend filtering estimates attain the minimax convergence rate over a large class of underlying functions, a class for which it is known that smoothing splines (along with any other estimator linear in y) are suboptimal. This latter work focused on evenly spaced inputs, $x_i = i/n$, $i = 1, \ldots n$, and the next two subsections extend the trend filtering convergence theory to cover arbitrary inputs $x_1, \ldots x_n \in [0, 1]$. We first consider the input points as fixed, and then random. All proofs are deferred until Section 8.7.

8.4.1 Fixed input points

The following is our main result on trend filtering.

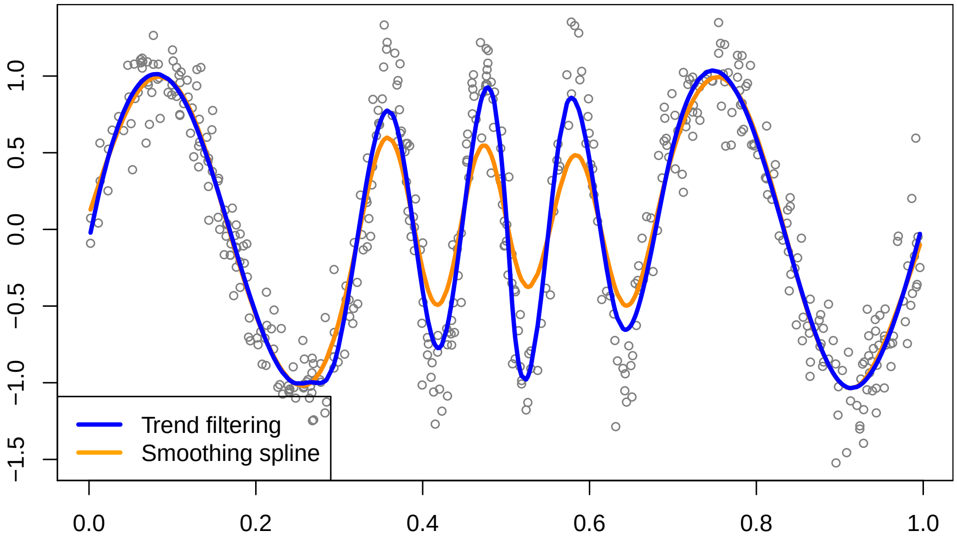


Figure 8.2: Example trend filtering and smoothing spline estimates.

Theorem 8.5. *Let $y \in \mathbb{R}^n$ be drawn from* (8.12)*, with fixed inputs $0 \leq x_1 < \ldots < x_n \leq 1$, having a maximum gap*

$$\max_{i=1,\ldots n} (x_i - x_{i-1}) = O(\log n/n), \tag{8.16}$$

and i.i.d., mean zero sub-Gaussian errors. Assume that, for an integer $k \geq 0$ and constant $C > 0$, the true function f_0 is k times weakly differentiable, with $\mathrm{TV}(f_0^{(k)}) \leq C$. Then the kth order trend filtering estimate $\hat{\theta}$ in (8.13)*, with tuning parameter value $\lambda = \Theta(n^{1/(2k+3)})$, satisfies*

$$\frac{1}{n}\sum_{i=1}^{n}\big(\hat{\theta}_i - f_0(x_i)\big)^2 = O_{\mathbb{P}}(n^{-(2k+2)/(2k+3)}). \tag{8.17}$$

Remark 1. The rate $n^{-(2k+2)/(2k+3)}$ is the minimax rate of convergence with respect to the class of k times weakly differentiable functions f such that $\mathrm{TV}(f^{(k)}) \leq C$ (see, e.g., Nussbaum [167], Tibshirani [221]). Hence Theorem 8.5 shows that trend filtering estimates converge at the minimax rate over a broad class of true functions f_0, assuming that the fixed input points are not too irregular, in that the maximum adjacent gap between points must satisfy (8.16). This condition is not stringent and is naturally satisfied by continuously distributed random inputs, as we show in the next subsection. We note that Tibshirani [221] proved the same conclusion (as in Theorem 8.5) for unevenly spaced inputs $x_1, \ldots x_n$, but placed very complicated and basically uninterpretable conditions on the inputs. Our tighter analysis of the falling factorial functions yields the simple sufficient condition (8.16).

Remark 2. The conclusion in the theorem can be strengthened, beyond the the convergence of $\hat{\theta}$ to f_0 in (8.17); under the same assumptions, the trend filtering estimate $\hat{\theta}$ also converges to $\hat{f}^{\text{spline}}$ at the same rate $n^{-(2k+2)/(2k+3)}$, where we write $\hat{f}^{\text{spline}}$ to denote the solution in (8.15) with $\mathcal{H}_k$ replaced by $\mathcal{G}_k = \mathrm{span}\{g_1, \ldots g_n\}$, the span of the truncated power basis functions in (8.2).

This asserts that the trend filtering estimate is indeed "close to" a spline, and here the bound in Lemma 8.4, between the truncated power and falling factorial basis matrices, is key. Moreover, we actually rely on the convergence of $\hat{\theta}$ to $\hat{f}^{\text{spline}}$ to establish (8.17), as the total variation regularized spline estimator $\hat{f}^{\text{spline}}$ is already known to converge to f_0 at the minimax rate [152].

8.4.2 Random input points

To analyze trend filtering for random inputs, $x_1, \ldots x_n$, we need to bound the maximum gap between adjacent points with high probability. Fortunately, this is possible for a large class of distributions, as shown in the next lemma.

Lemma 8.6. *If $0 \le x_1 < \ldots < x_n \le 1$ are sorted i.i.d. draws from an arbitrary continuous distribution supported on $[0,1]$, whose density is bounded below by $p_0 > 0$, then with probability at least $1 - 2p_0 n^{-10}$,*

$$\max_{i=1,\ldots n} (x_i - x_{i-1}) \le \frac{c_0 \log n}{p_0 n},$$

for a universal constant c_0.

The proof of this result is readily assembled from classical results on order statistics; we give a simple alternate proof in Section 8.7. Lemma 8.6 implies the next corollary.

Corollary 8.7. *Let $y \in \mathbb{R}^n$ be distributed according to the model (8.12), where the inputs $0 \le x_1 < \ldots < x_n \le 1$ are sorted i.i.d. draws from an arbitrary continuous distribution on $[0,1]$, whose density is bounded below. Assume again that the errors are i.i.d., mean zero sub-Gaussian variates, independent of the inputs, and that the true function f_0 has k weak derivatives and satisfies $\mathrm{TV}(f_0^{(k)}) \le C$. Then, for $\lambda = \Theta(n^{1/(2k+3)})$, the kth order trend filtering estimate $\hat{\theta}$ converges at the same rate as in Theorem 8.5.*

8.5 A higher order Kolmogorov-Smirnov test

The two-sample Kolmogorov-Smirnov (KS) test is a standard nonparametric hypothesis test of equality between two distributions, say $\mathbb{P}_X$ and $\mathbb{P}_Y$, from independent samples $x_1, \ldots x_m \sim \mathbb{P}_X$ and $y_1, \ldots y_n \sim \mathbb{P}_Y$. Writing $X_{(m)} = (x_1, \ldots x_m)$, $Y_{(n)} = (y_1, \ldots y_n)$, and $Z_{(m+n)} = (z_1, \ldots z_{m+n}) = X_{(m)} \cup Y_{(n)}$ for the joined samples, the KS statistic can be expressed as

$$\mathrm{KS}(X_{(m)}, Y_{(n)}) = \max_{z_j \in Z_{(m+n)}} \left| \frac{1}{m} \sum_{i=1}^{m} 1\{x_i \le z_j\} - \frac{1}{n} \sum_{i=1}^{n} 1\{y_i \le z_j\} \right|. \qquad (8.18)$$

This examines the maximum absolute difference between the empirical cumulative distribution functions from $X_{(m)}$ and $Y_{(n)}$, across all points in the joint set $Z_{(m+n)}$, and so the test rejects for large values of (8.18). A well-known alternative (variational) form for the KS statistic is

$$\mathrm{KS}(X_{(m)}, Y_{(n)}) = \max_{f : \mathrm{TV}(f) \le 1} \left| \hat{\mathbb{E}}_{X_{(m)}}[f(X)] - \hat{\mathbb{E}}_{Y_{(n)}}[f(Y)] \right|, \qquad (8.19)$$

where $\hat{\mathbb{E}}_{X_{(m)}}$ denotes the empirical expectation under $X_{(m)}$, so that $\hat{\mathbb{E}}_{X_{(m)}}[f(X)] = 1/m \sum_{i=1}^m f(x_i)$, and similarly for $\hat{\mathbb{E}}_{Y_{(n)}}$. The equivalence between (8.19) and (8.18) comes from the fact that maximum in (8.19) is achieved by taking f to be a step function, with its knot (breakpoint) at one of the joined samples $z_1, \ldots z_{m+n}$.

The KS test is perhaps one of the most widely used nonparametric tests of distributions, but it does have its shortcomings. Loosely speaking, it is known to be sensitive in detecting differences between the centers of distributions $\mathbb{P}_X$ and $\mathbb{P}_Y$, but much less sensitive in detecting differences in the tails. In this section, we generalize the KS test to "higher order" variants that are more powerful than the original KS test in detecting tail differences (when, of course, such differences are present). We first define the higher order KS test, and describe how it can be computed in linear time with the falling factorial basis. We then empirically compare these higher order versions to the original KS test, and several other commonly used nonparametric two-sample tests of distributions.

8.5.1 Definition of the higher order KS tests

For a given order $k \geq 0$, we define the kth order KS test statistic between $X_{(m)}$ and $Y_{(n)}$ as

$$\mathrm{KS}_G^{(k)}(X_{(m)}, Y_{(n)}) = \left\| (G_2^{(k)})^T \left(\frac{\mathbb{1}_{X_{(m)}}}{m} - \frac{\mathbb{1}_{Y_{(n)}}}{n} \right) \right\|_\infty . \tag{8.20}$$

Here $G^{(k)} \in \mathbb{R}^{(m+n)\times(m+n)}$ is the kth order truncated power basis matrix over the joined samples $z_1 < \ldots < z_{m+n}$, assumed sorted without a loss of generality, and $G_2^{(k)}$ is the submatrix formed by excluding its first $k+1$ columns. Also, $\mathbb{1}_{X_{(m)}} \in \mathbb{R}^{(m+n)}$ is a vector whose components indicate the locations of $x_1 < \ldots < x_m$ among $z_1 < \ldots < z_{m+n}$, and similarly for $\mathbb{1}_{Y_{(n)}}$. Finally, $\|\cdot\|_\infty$ denotes the ℓ_∞ norm, $\|u\|_\infty = \max_{i=i,\ldots r} |u_i|$ for $u \in \mathbb{R}^r$.

As per the spirit of our paper, an alternate definition for the kth order KS statistic uses the falling factorial basis,

$$\mathrm{KS}_H^{(k)}(X_{(m)}, Y_{(n)}) = \left\| (H_2^{(k)})^T \left(\frac{\mathbb{1}_{X_{(m)}}}{m} - \frac{\mathbb{1}_{Y_{(n)}}}{n} \right) \right\|_\infty , \tag{8.21}$$

where now $H^{(k)} \in \mathbb{R}^{(m+n)\times(m+n)}$ is the kth order falling factorial basis matrix over the joined samples $z_1 < \ldots < z_{m+n}$. Not surprisingly, the two definitions are very close, and Hölder's inequality shows that

$$|\mathrm{KS}_G^{(k)}(X_{(m)}, Y_{(n)}) - \mathrm{KS}_H^{(k)}(X_{(m)}, Y_{(n)})| \leq \max_{i,j=1,\ldots m+n} 2|G_{ij}^{(k)} - H_{ij}^{(k)}| \leq 2k^2\delta,$$

the last inequality due to Lemma 8.4, with δ the maximum gap between $z_1, \ldots z_{m+n}$. Recall that Lemma 8.6 shows δ to be of the order $\log(m+n)/(m+n)$ for continuous distributions $\mathbb{P}_X, \mathbb{P}_Y$ supported nontrivially on $[0,1]$, which means that with high probability, the two definitions differ by at most $2k^2 \log(m+n)/(m+n)$, in such a setup.

The advantage to using the falling factorial definition is that the test statistic in (8.21) can be computed in $O(k(m+n))$ time, without even having to form the matrix $H_2^{(k)}$ (this is assuming

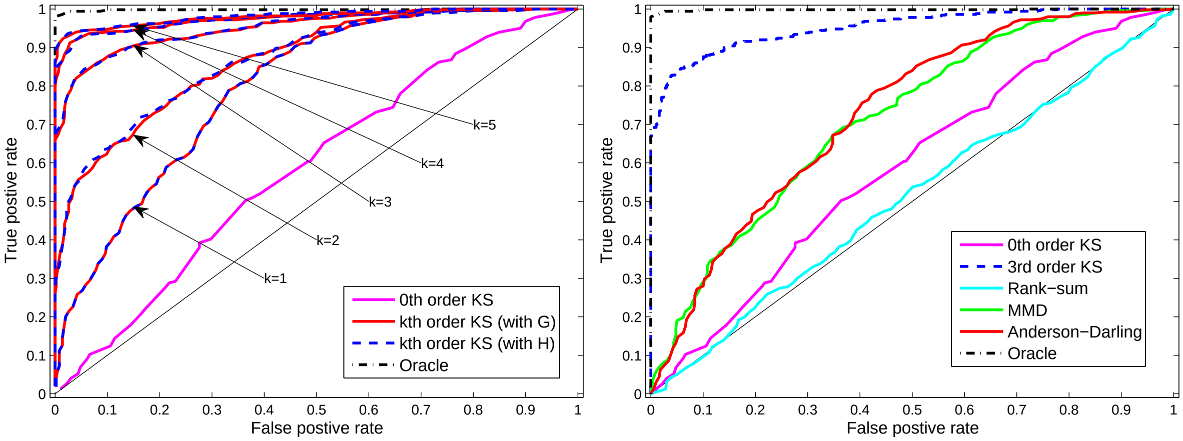


Figure 8.3: ROC curves for experiment 1, normal vs. t.

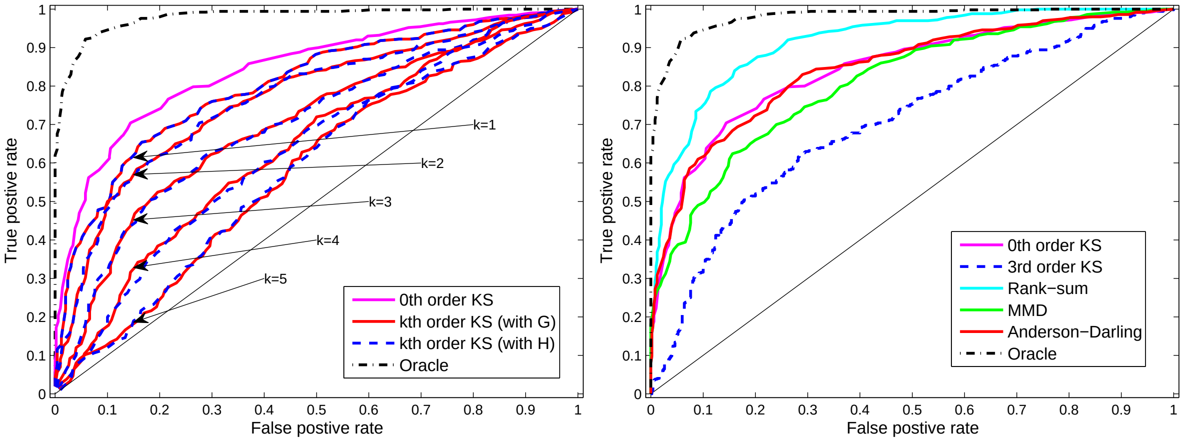


Figure 8.4: ROC curves for experiment 2, Laplace vs. Laplace.

sorted points $z_1, \dots z_{m+n}$). See Lemma 8.3, and Algorithm 9 in Section 8.7.3. By comparison, the statistic in (8.20) requires $O((m+n)^2)$ operations. In addition to the theoretical bound described above, we also find empirically that the two definitions perform quite similarly, as shown in the next subsection, and hence we advocate the use of $\mathrm{KS}_H^{(k)}$ for computational reasons.

A motivation for our proposed tests is as follows: it can be shown that (8.20), and therefore (8.21), approximately take a variational form similar to (8.19), but where the constraint is over functions whose kth (weak) derivative has total variation at most 1. See Section 8.7.

8.5.2 Numerical experiments

We examine the higher order KS tests by simulation. The setup: we fix two distributions P, Q. We draw n i.i.d. samples $X_{(n)}, Y_{(n)} \sim P$, calculate a test statistic, and repeat this $R/2$ times; we also draw n i.i.d. samples $X_{(n)} \sim P$, $Y_{(n)} \sim Q$, calculate a test statistic, and repeat $R/2$ times. We then construct an ROC curve, i.e., the true positive rate versus the false positive rate of the test, as we vary its rejection threshold. For the test itself, we consider our kth order KS test, in both its G and H forms, as well as the usual KS test, and a number of other popular two-sample tests: the Anderson-Darling test [9, 189], the Wilcoxon rank-sum test [251], and the maximum mean discrepancy (MMD) test, with RBF kernel [104].

Figures 8.3 and 8.4 show the results of two experiments in which $n = 100$ and $R = 1000$. (See Section 8.8 for more experiments.) In the first we used $P = N(0, 1)$ and $Q = t_3$ (t-distribution with 3 degrees of freedom), and in the second $P = \mathrm{Laplace}(0)$ and $Q = \mathrm{Laplace}(0.3)$ (Laplace distributions of different means). We see that our proposed kth order KS test performs favorably in the first experiment, with its power increasing with k. When $k = 3$, it handily beats all competitors in detecting the difference between the standard normal distribution and the heavier-tailed t-distribution. But there is no free lunch: in the second experiment, where the differences between P, Q are mostly near the centers of the distributions and not in the tails, we can see that increasing k only decreases the power of the kth order KS test. In short, one can view our proposal as introducing a family of tests parametrized by k, which offer a tradeoff in center versus tail sensitivity. A more thorough study will be left to future work.

8.6 Discussion

We formally proposed and analyzed the spline-like falling factorial basis functions. These basis functions admit attractive computational and statistical properties, and we demonstrated their applicability in two problems: trend filtering, and a novel higher order variant of the KS test. These examples, we feel, are just the beginning. As typical operations associated with the falling factorial basis scale merely linearly with the input size (after sorting), we feel that this basis may be particularly well-suited to a rich number of large-scale applications in the modern data era, a direction that we are excited to pursue in the future.

8.7 Proofs and technical details

8.7.1 Proof of Lemma 8.1 (recursive decomposition)

The falling factorial basis matrix, as defined in (8.4), (8.5), can be expressed as $H^{(k)} = [H_1^{(k)} \; H_2^{(k)}]$, where

$$H_1^{(k)} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & x_2 - x_1 & 0 & \cdots & 0 \\ 1 & x_3 - x_1 & (x_3 - x_2)(x_3 - x_1) & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{k+1} - x_1 & (x_{k+1} - x_2)(x_{k+1} - x_1) & \cdots & \prod_{\ell=1}^{k} (x_{k+1} - x_\ell) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x_1 & (x_n - x_2)(x_n - x_1) & \cdots & \prod_{\ell=1}^{k} (x_n - x_\ell) \end{bmatrix} \in \mathbb{R}^{n \times (k+1)},$$

and

$$H_2^{(k)} = \left[\begin{array}{cccc} 0_{(k+1)\times 1} & 0_{(k+1)\times 1} & \cdots & 0_{(k+1)\times 1} \\ \prod_{\ell=1}^{k}(x_{k+2}-x_{1+\ell}) & 0 & \cdots & 0 \\ \prod_{\ell=1}^{k}(x_{k+3}-x_{1+\ell}) & \prod_{\ell=1}^{k}(x_{k+3}-x_{2+\ell}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \prod_{\ell=1}^{k}(x_n-x_{1+\ell}) & \prod_{\ell=1}^{k}(x_n-x_{2+\ell}) & \cdots & \prod_{\ell=1}^{k}(x_n-x_{n-k-1+\ell}) \end{array}\right] \in \mathbb{R}^{n\times(n-k-1)}.$$

Lemma 8.1 claims that $H^{(0)} = L_n$, the lower triangular matrix of 1s, which can be seen directly by inspection (recalling our convention of defining thee empty product to be 1). The lemma further claims that $H^{(k)}$ can be recursively factorized into the following form:

$$H^{(k)} = H^{(k-1)} \cdot \left[\begin{array}{cc} I_k & 0 \\ 0 & \Delta^{(k)} \end{array}\right] \cdot \left[\begin{array}{cc} I_k & 0 \\ 0 & L_{n-k} \end{array}\right], \tag{8.22}$$

for all $k \geq 1$. We prove the above factorization in this current section. In what follows, we denote the last $n-k-1$ columns of the product (8.22) by $\tilde{M}^{(k)} \in \mathbb{R}^{n\times(n-k-1)}$, and also write

$$\tilde{M}^{(k)} = \left[\begin{array}{c} 0_{(k+1)\times(n-k-1)} \\ \tilde{L}^{(k)}, \end{array}\right],$$

i.e., we use $\tilde{L}^{(k)}$ to denote the lower $(n-k-1)\times(n-k-1)$ submatrix of $\tilde{M}^{(k)}$. To prove the lemma, we show that $\tilde{M}^{(k)}$ is equal to the corresponding block $H_2^{(k)}$, by induction on k. The proof that the first block of $k+1$ columns of the product is equal to $H_1^{(k)}$ follows from the arguments given for the proof of the second block, and therefore we do not explicitly rewrite the proof for this part.

We begin the inductive proof by checking the case $k=1$. Note

$$\tilde{M}^{(1)} = \left[\begin{array}{c} 0_{2\times(n-2)} \\ \tilde{L}^{(1)} \end{array}\right] = \left[\begin{array}{c} 0_{1\times(n-1)} \\ L_{n-1} \end{array}\right] (\Delta^{(k)})^{-1} \left[\begin{array}{c} 0_{1\times(n-2)} \\ L_{n-2} \end{array}\right]$$
$$= \left[\begin{array}{cccc} 0_{2\times 1} & 0_{2\times 1} & \cdots & 0_{2\times 1} \\ x_3-x_2 & 0 & \cdots & 0 \\ x_4-x_2 & x_4-x_3 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_n-x_2 & x_n-x_3 & \cdots & x_n-x_{n-1} \end{array}\right].$$

This gives precisely the last $n-2$ columns of $H^{(1)}$, as defined in (8.4).

Next we verify that if the statement holds for some $k \geq 1$, then it is true for $k+1$. To avoid confusion, we will use i,j as indices $H^{(k+1)}$ and α,θ as indices of $\tilde{L}^{(k+1)}$. The universal rule for the relationship between the two sets of indices is

$$\begin{pmatrix} i \\ j \end{pmatrix} = \begin{pmatrix} \alpha \\ \theta \end{pmatrix} + k + 2.$$

We consider an arbitrary element, $\tilde{L}^{(k+1)}_{\alpha\theta}$. Due to the upper triangular shape of $\tilde{L}^{(k)}$, we have $\tilde{L}^{(k)}_{\alpha\theta} = 0$ if $\alpha < \theta$. For $\alpha \geq \theta$, we plainly calculate, using the inductive hypothesis

$$\begin{aligned}
\tilde{L}^{(k+1)}_{\alpha\theta} &= \sum_{q=1+\theta}^{1+\alpha} \tilde{L}^{(k)}_{1+\alpha,q} \cdot (\Delta^{(k+1)})^{-1}_{qq} \\
&= \sum_{q=1+\theta}^{1+\alpha} \prod_{\ell=1}^{k} (x_{k+2+\alpha} - x_{q+\ell}) \cdot (x_{k+1+q} - x_q) \\
&= \prod_{\ell=1}^{k+1} (x_{k+2+\alpha} - x_{\theta+\ell}) \cdot A = H^{(k)}_{ij} \cdot A,
\end{aligned}$$

where A is the sum of terms that scales each summand to the desired quantity (by multiplying and dividing by missing factors). To complete the inductive proof, it suffices to show that $A = 1$. It turns out that there are two main cases to consider, which we examine below.

Case 1. When $\alpha - \theta \leq k$, the term A can be expressed as

$$\begin{aligned}
A =& \frac{x_{k+1+1+\theta} - x_{1+\theta}}{x_{k+2+\alpha} - x_{1+\theta}} + \frac{(x_{k+1+2+\theta} - x_{2+\theta})(x_{k+2+\alpha} - x_{k+1+1+\theta})}{(x_{k+2+\alpha} - x_{1+\theta})(x_{k+2+\alpha} - x_{2+\theta})} \\
&+ \cdots + \frac{(x_{k+1+\gamma+\theta} - x_{\gamma+\theta})(x_{k+2+\alpha} - x_{k+2+\theta}) \cdots (x_{k+2+\alpha} - x_{k+\gamma+\theta})}{(x_{k+2+\alpha} - x_{1+\theta}) \cdots (x_{k+2+\alpha} - x_{\gamma-1+\theta})(x_{k+2+\alpha} - x_{\gamma+\theta})} \\
&+ \cdots + \frac{(x_{k+1+\alpha} - x_{\alpha})(x_{k+2+\alpha} - x_{k+2+\theta}) \cdots (x_{k+2+\alpha} - x_{k+1+\alpha})}{(x_{k+2+\alpha} - x_{1+\theta}) \cdots (x_{k+2+\alpha} - x_{\alpha-1})(x_{k+2+\alpha} - x_{\alpha})} \\
&+ \frac{\cancel{(x_{k+2+\alpha} - x_{1+\alpha})}(x_{k+2+\alpha} - x_{k+2+\theta}) \cdots (x_{k+2+\alpha} - x_{k+1+\alpha})}{(x_{k+2+\alpha} - x_{1+\theta}) \cdots (x_{k+2+\alpha} - x_{\alpha})\cancel{(x_{k+2+\alpha} - x_{1+\alpha})}}.
\end{aligned}$$

Note that in the last term, the factor $(x_{k+2+\alpha} - x_{1+\alpha})$ in both the denominator and numerator cancels out, leaving the denominator to be the same as the second to last term. Combining the last two terms, we again get a common factor $(x_{k+2+\alpha} - x_{\alpha})$ in denominator and numerator, which cancels out, and makes the denominator of this term the same as that previous term. Continuing in this manner, we can recursively eliminate the terms from last to the first, leaving

$$\frac{\cancel{x_{k+2+\theta}} - x_{1+\theta} + x_{k+2+\alpha} - \cancel{x_{k+2+\theta}}}{x_{k+2+\alpha} - x_{1+\theta}} = 1.$$

In other words, we have shown that $A = 1$.

Case 2. When $\alpha - \theta \geq k + 1$, the denominators in terms of A will remain the same after they reach

$$(x_{k+2+\alpha} - x_{1+\theta}) \cdots (x_{k+2+\alpha} - x_{1+k+\theta}) = \prod_{\ell=1}^{k+1} (x_{k+2+\alpha} - x_{\theta+\ell}) := B.$$

Again, we begin by expressing A explicitly as

$$\begin{aligned}
A =&\frac{x_{k+1+1+\theta} - x_{1+\theta}}{x_{k+2+\alpha} - x_{1+\theta}} + \frac{(x_{k+1+2+\theta} - x_{2+\theta})(x_{k+2+\alpha} - x_{k+1+1+\theta})}{(x_{k+2+\alpha} - x_{1+\theta})(x_{k+2+\alpha} - x_{2+\theta})} \\
&+ \cdots + \frac{(x_{k+1+\gamma+\theta} - x_{\gamma+\theta})(x_{k+2+\alpha} - x_{k+2+\theta})\cdots(x_{k+2+\alpha} - x_{k+\gamma+\theta})}{(x_{k+2+\alpha} - x_{1+\theta})\cdots(x_{k+2+\alpha} - x_{\gamma-1+\theta})(x_{k+2+\alpha} - x_{\gamma+\theta})} \\
&+ \cdots + \frac{(x_{k+1+k+1+\theta} - x_{k+1+\theta})(x_{k+2+\alpha} - x_{k+2+\theta})\cdots(x_{k+2+\alpha} - x_{k+k+1+\theta})}{(x_{k+2+\alpha} - x_{1+\theta})\cdots(x_{k+2+\alpha} - x_{1+k+\theta})} \\
&+ \frac{(x_{k+1+k+2+\theta} - x_{k+2+\theta})(x_{k+2+\alpha} - x_{k+3+\theta})\cdots(x_{k+2+\alpha} - x_{k+k+2+\theta})}{(x_{k+2+\alpha} - x_{1+\theta})\cdots(x_{k+2+\alpha} - x_{1+k+\theta})} \\
&+ \cdots + \frac{(x_{k+1+\alpha} - x_{1+\alpha})(x_{k+2+\alpha} - x_{1+\alpha})\cdots(x_{k+2+\alpha} - x_{k+\alpha})}{(x_{k+2+\alpha} - x_{1+\theta})\cdots(x_{k+2+\alpha} - x_{1+k+\theta})} \\
&+ \frac{(x_{k+1+1+\alpha} - x_{1+\alpha})(x_{k+2+\alpha} - x_{2+\alpha})\cdots(x_{k+2+\alpha} - x_{k+1+\alpha})}{(x_{k+2+\alpha} - x_{1+\theta})\cdots(x_{k+2+\alpha} - x_{1+k+\theta})}.
\end{aligned}$$

Now we divide first factor of the transition term, in the third line above, into two halves by

$$x_{k+1+k+1+\theta} - x_{k+1+\theta} = (x_{k+2+\alpha} - x_{1+k+\theta}) + (x_{k+1+k+1+\theta} - x_{k+2+\alpha}).$$

The first half triggers the recursive reduction on the first k terms exactly as in the first case, so the sum of the first k terms equal to 1 and we get

$$\begin{aligned}
B(A-1) =& -(x_{k+2+\alpha} - x_{k+k+2+\theta})(x_{k+2+\alpha} - x_{k+2+\theta})\cdots(x_{k+2+\alpha} - x_{k+k+1+\theta}) \\
&+ (x_{k+1+k+2+\theta} - x_{k+2+\theta})(x_{k+2+\alpha} - x_{k+3+\theta})\cdots(x_{k+2+\alpha} - x_{k+k+2+\theta}) \\
&+ \cdots + (x_{k+1+\alpha} - x_{1+\alpha})(x_{k+2+\alpha} - x_{1+\alpha})\cdots(x_{k+2+\alpha} - x_{k+\alpha}) \\
&+ (x_{k+1+1+\alpha} - x_{1+\alpha})(x_{k+2+\alpha} - x_{2+\alpha})\cdots(x_{k+2+\alpha} - x_{k+1+\alpha}).
\end{aligned}$$

Now we can do a recursive reduction starting from the first two terms, the sum of which is

$$\begin{aligned}
&\Big[x_{k+1+k+2+\theta} - x_{k+2+\theta} - (x_{k+2+\alpha} - x_{k+2+\theta})\Big](x_{k+2+\alpha} - x_{k+3+\theta})\cdots(x_{k+2+\alpha} - x_{k+k+2+\theta}) \\
=& -(x_{k+2+\alpha} - x_{k+1+k+2+\theta})(x_{k+2+\alpha} - x_{k+3+\theta})\cdots(x_{k+2+\alpha} - x_{k+k+2+\theta})
\end{aligned}$$

This can be combined with the third term in a similar fashion and the recursion continues. At the end, we get

$$\begin{aligned}
B(A-1) =& -(x_{k+2+\alpha} - x_{k+1+\alpha})(x_{k+2+\alpha} - x_{1+\alpha})\cdots(x_{k+2+\alpha} - x_{k+\alpha}) \\
&+ (x_{k+1+1+\alpha} - x_{1+\alpha})(x_{k+2+\alpha} - x_{2+\alpha})\cdots(x_{k+2+\alpha} - x_{k+1+\alpha}) \\
=&\Big[x_{k+1+1+\alpha} - x_{1+\alpha} - (x_{k+2+\alpha} - x_{1+\alpha})\Big](x_{k+2+\alpha} - x_{2+\alpha})\cdots(x_{k+2+\alpha} - x_{k+1+\alpha}) = 0.
\end{aligned}$$

That is, we have shown that $A = 1$.

With $A = 1$ proved between these two cases, we have completed the inductive argument, and hence the proof of the lemma.

8.7.2 Proof of Lemma 8.2 (inverse representation)

We prove Lemma 8.2, which claims that he inverse of falling factorial basis matrix is

$$(H^{(k)})^{-1} = \begin{bmatrix} C \\ \frac{1}{k!} \cdot D^{(k+1)} \end{bmatrix}, \tag{8.23}$$

where $D^{(k+1)}$ is the $(k+1)^{st}$ order discrete difference operator defined in (8.10), and the rows of the matrix $C \in \mathbb{R}^{(k+1)\times n}$ obey $C_1 = e_1$ and

$$C_{i+1} = \left[\frac{1}{i!} \cdot (\Delta^{(i)})^{-1} \cdot D^{(i)}\right]_1, \quad i = 1, \dots k.$$

Again we use induction on k. When $k = 0$, it is easily verified that

$$(H^{(0)})^{-1} = L_n^{-1} = \begin{bmatrix} e_1 \\ D^{(1)} \end{bmatrix} = \begin{bmatrix} e_1 \\ \frac{1}{0!} \cdot D^{(1)} \end{bmatrix}.$$

The rest of the inductive proof is relatively straightforward, following from Lemma 8.1, i.e., from (8.22). Inverting both sides of (8.22) gives

$$\begin{aligned} (H^{(k)})^{-1} &= \begin{bmatrix} I_k & 0 \\ 0 & L_{n-k} \end{bmatrix}^{-1} \cdot \begin{bmatrix} I_k & 0 \\ 0 & \Delta^{(k)} \end{bmatrix}^{-1} \cdot (H^{(k-1)})^{-1} \\ &= \begin{bmatrix} I_k & 0 \\ 0 & L_{n-k}^{-1} \end{bmatrix} \cdot \begin{bmatrix} I_k & 0 \\ 0 & (\Delta^{(k)})^{-1} \end{bmatrix} \cdot (H^{(k-1)})^{-1}. \end{aligned}$$

Now, using that $L_{n-k}^{-1} = \begin{bmatrix} e_1 \\ D^{(1)} \end{bmatrix}$, and assuming that $(H^{(k-1)})^{-1}$ obeys (8.23),

$$\begin{aligned} (H^{(k)})^{-1} &= \begin{bmatrix} I_k & 0 \\ 0 & \begin{bmatrix} e_1 \\ D^{(1)} \end{bmatrix} \end{bmatrix} \cdot \begin{bmatrix} I_k & 0 \\ 0 & (\Delta^{(k)})^{-1} \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ \left[\frac{1}{1!}(\Delta^{(1)})^{-1} D^{(1)}\right]_1 \\ \vdots \\ \left[\frac{1}{(k-1)!}(\Delta^{(k-1)})^{-1} D^{(k-1)}\right]_1 \\ \frac{1}{(k-1)!} \cdot D^{(k)} \end{bmatrix} \\ &= \begin{bmatrix} e_1 \\ \left[\frac{1}{1!}(\Delta^{(1)})^{-1} D^{(1)}\right]_1 \\ \vdots \\ \left[\frac{1}{(k-1)!}(\Delta^{(k-1)})^{-1} D^{(k-1)}\right]_1 \\ \frac{1}{k!} \begin{bmatrix} e_1 \\ D^{(1)} \end{bmatrix} \cdot k(\Delta^{(k)})^{-1} \cdot D^{(k)} \end{bmatrix} = \begin{bmatrix} e_1 \\ \left[\frac{1}{1!}(\Delta^{(1)})^{-1} D^{(1)}\right]_1 \\ \vdots \\ \left[\frac{1}{(k-1)!}(\Delta^{(k-1)})^{-1} D^{(k-1)}\right]_1 \\ \left[\frac{1}{(k)!}(\Delta^{(k)})^{-1} D^{(k)}\right]_1 \\ \frac{1}{k!} \cdot D^{(k+1)} \end{bmatrix} = \begin{bmatrix} C \\ \frac{1}{k!} \cdot D^{(k+1)} \end{bmatrix}, \end{aligned}$$

as desired.

8.7.3 Algorithms for multiplication by $(H^{(k)})^T$ and $[(H^{(k)})^T]^{-1}$

Recall that, given a vector y, we write $y_{a:b}$ to denote its subvector $(y_a, y_{a+1}, \ldots y_b)$, and we write cumsum and diff for the cumulative sum pairwise difference operators. Furthermore, we define flip to be the operator the reverses the order of its input, e.g., $\mathrm{flip}((1,2,3)) = (3,2,1)$, and we write $\circ$ to denote operator composition, e.g., $\mathrm{flip} \circ \mathrm{cumsum}$. The remaining two algorithms from Lemma 8.3 are given below, in Algorithms 9 and 10.

Algorithm 9 Multiplication by $(H^{(k)})^T$

```
Input: Vector to be multiplied $y \in \mathbb{R}^n$, order $k \geq 0$, sorted inputs vector $x \in \mathbb{R}^n$.
Output: $y$ is overwritten by $(H^{(k)})^T y$.
for $i = 0$ to $k$ do
    if $i \neq 0$ then
        $y_{(i+1):n} = y_{(i+1):n} \,./\, (x_{(i+1):n} - x_{1:(n-i)})$.
    end if
    $y_{(i+1):n} = \mathrm{flip} \circ \mathrm{cumsum} \circ \mathrm{flip}(y_{(i+1):n})$.
end for
Return $y$.
```

Algorithm 10 Multiplication by $[(H^{(k)})^T]^{-1}$

```
Input: Vector to be multiplied $y \in \mathbb{R}^n$, order $k \geq 0$, sorted inputs vector $x \in \mathbb{R}^n$.
Output: $y$ is overwritten by $[(H^{(k)})^T]^{-1} y$.
for $i = k$ to $0$ do
    $y_{(i+1):n-1} = \mathrm{flip} \circ \mathrm{diff} \circ \mathrm{flip}(y_{(i+1):n})$.
    if $i \neq 0$ then
        $y_{(i+1):n} = (x_{(i+1):n} - x_{1:(n-i)})^{-1} \,.*\, y_{(i+1):n}$.
    end if
end for
Return $y$.
```

8.7.4 Proof of Lemma 8.4 (proximity to truncated power basis)

Recall that we denote

$$\delta = \max_{i=1,\ldots n} (x_i - x_{i-1}),$$

and write $x_0 = 0$ for notational convenience. Taking the elementwise difference between the falling factorial and truncated power basis matrices, we get

$$H_{ij} - G_{ij} = \begin{cases} 0 & \text{for } i = 1, \ldots n, \ j = 1 \\ \prod_{\ell=1}^{j-1} (x_i - x_\ell) - x_i^{j-1} & \text{for } i > j-1, \ j = 2, \ldots k+1 \\ -x_i^{j-1} & \text{for } i \leq j-1, \ j = 2, \ldots k+1 \\ 0 & \text{for } i \leq j - \lceil k/2 \rceil, \ j \geq k+2 \\ -(x_i - x_{j-\lceil k/2 \rceil})^k & \text{for } j - \lceil k/2 \rceil < i \leq j-1, \ j \geq k+2 \\ \prod_{\ell=1}^{k} (x_i - x_{j-k-1+\ell}) - (x_i - x_{j-\lceil k/2 \rceil})^k & \text{for } i > j-1, \ j \geq k+2. \end{cases} \tag{8.24}$$

In the above, we use $\lceil z \rceil$ to denote the least integer greater than or equal to z (the ceiling function). We will bound the absolute value of each nonzero difference $H_{ij} - G_{ij}$ in (8.24). Starting with the second row,

$$\begin{aligned} \left| \prod_{\ell=1}^{j-1} (x_i - x_\ell) - x_i^{j-1} \right| &\leq x_i^{j-1} - (x_i - x_{j-1})^{j-1} \\ &= x_{j-1} \left[x_i^{j-2} + x_i^{j-3}(x_i - x_{j-1}) + \ldots + x_i (x_i - x_{j-1})^{j-3} + (x_i - x_{j-1})^{j-2} \right] \\ &\leq x_{j-1} \cdot (j-1) \cdot x_i^{j-2} \leq k\delta \cdot k \cdot 1 \leq k^2 \delta. \end{aligned}$$

In the second line above, we used the expansion

$$a^k - b^k = (a-b)(a^{k-1} + a^{k-2}b + \ldots + b^{k-1}), \tag{8.25}$$

and in the third line, we used the fact that $j - 1 \leq k$, so that $x_{j-1} \leq k\delta$, and also $0 \leq x_i \leq 1$. The third row of (8.24) is simpler. Since $0 \leq x_i \leq 1$ and $i \leq j - 1 < k$,

$$|-x_i^{j-1}| \leq x_i \leq k\delta.$$

For the fourth row in (8.24), using the range of i, j, and the fact that $k\delta \leq 1$,

$$|-(x_i - x_{j-\lceil k/2 \rceil})^k| \leq (x_{j-1} - x_{j-\lceil k/2 \rceil})^k \leq (k\delta)^k \leq k\delta.$$

This leaves us to deal with the last row in (8.24). Defining $p = i$, $q = j - (k+1)$, the problem transforms into bounding

$$\prod_{\ell=1}^{k} (x_p - x_{\ell+q}) - (x_p - x_{\lfloor \frac{k+2}{2} \rfloor + q})^k,$$

for any $p = k+2, k+3, \ldots n$, $q = 1, \ldots p-k$, where now $\lfloor z \rfloor$ denotes the greatest integer less than or equal to z (the floor function). We let $\mu_{pq} = x_p - x_{\lfloor \frac{k+2}{2} \rfloor + q}$ and $\eta_q = x_p - x_{q+1} - \mu_{pq}$. Note that η_q is the gap between the maximum multiplicant in the first term above and μ_{pq}. Then

$$\eta_q = x_{\lfloor \frac{k+2}{2} \rfloor + q} - x_{q+1} \leq k\delta.$$

Therefore

$$\begin{aligned}\prod_{\ell=1}^{k}(x_p - x_{\ell+q}) - (x_p - x_{\lfloor\frac{k+2}{2}\rfloor+q})^k &\le (x_p - x_{1+q})^k - \mu_{pq}^k \\ &= (\mu_{pq} + \eta_q)^k - \mu_{pq}^k \\ &= k\delta \cdot \sum_{\ell=0}^{k-1} (\mu_{pq} + \eta_q)^\ell \mu_{pq}^{k-\ell} \\ &\le k^2\delta \cdot (\mu_{pq} + \eta_q)^k \le k^2\delta.\end{aligned}$$

The third line above follows again from the expansion (8.25), and the fact that $\eta_q \le k\delta$. The fourth line uses $\mu_{pq} + \eta_q \ge \mu_{pq}$, and ultimately $\mu_{pq} + \eta_q = x_p - x_{1+q} \in [0,1]$. This completes the proof.

8.7.5 Proof of Theorem 8.5 (trend filtering rate, fixed inputs)

This proof follows the same strategy as the convergence proofs in Tibshirani [221]. Recall that the trend filtering estimate (8.13) can be expressed in terms of the lasso problem (8.14), in that $\hat{\theta} = H^{(k)}\hat{\alpha}$; also consider consider the problem

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \frac{1}{2}\|y - G^{(k)}\theta\|_2^2 + \lambda' \cdot \sum_{j=k+2}^{n} |\theta_j|, \tag{8.26}$$

where $G^{(k)}$ is the truncated power basis matrix of order k. Let $\mu = (f_0(x_1), \ldots f_0(x_n)) \in \mathbb{R}^n$ denote the true function evaluated across the inputs. Then under the assumptions of Theorem 8.5, it is known that

$$\|G^{(k)}\hat{\theta} - \mu\|_2^2 = O_{\mathbb{P}}(n^{-(2k+2)/(2k+3)}),$$

when $\lambda = \Theta(n^{1/(2k+3)})$; see Theorem 10 of Mammen and van de Geer [152]. It now suffices to show that $\|H^{(k)}\hat{\alpha} - G^{(k)}\hat{\theta}\|_2^2 = O_{\mathbb{P}}(n^{-(2k+2)/(2k+3)})$, since

$$\|H^{(k)}\hat{\alpha} - \mu\|_2^2 \le 2\|H^{(k)}\hat{\alpha} - G^{(k)}\hat{\theta}\|_2^2 + 2\|G^{(k)}\hat{\theta} - \mu\|_2^2.$$

For this, we can use the results in Appendix B of Tibshirani [221], specifically Corollary 4 of this work, to argue that we have $\|H^{(k)}\hat{\alpha} - G^{(k)}\hat{\theta}\|_2^2 = O_{\mathbb{P}}(n^{-(2k+2)/(2k+3)})$ as long as $\lambda = (1+\delta)\lambda'$ for any $\delta > 0$, and

$$n^{(2k+2)/(2k+3)} \cdot \max_{i,j=1,\ldots n} |G_{ij}^{(k)} - H_{ij}^{(k)}| \to 0 \quad \text{as } n \to \infty.$$

But by Lemma 8.4, and our condition (8.16) on the inputs, we have $\max_{i,j=1,\ldots n} |G_{ij}^{(k)} - H_{ij}^{(k)}| \le k^2 \log n/n$, which verifies the above, and hence gives the result.

8.7.6 Proof of Lemma 8.6 (maximum gap between random inputs)

Given sorted i.i.d. draws $x_1 \leq \ldots \leq x_n$ from a continuous distribution supported on $[0,1]$, whose density is bounded below by $p_0 > 0$, we consider the maximum gap $\delta = \max_{i=1,\ldots n}(x_i - x_{i-1})$ (recall that we set $x_0 = 0$ for notational convenience). This is a well-studied quantity. In the case of a uniform distribution on $[0,1]$, we know that the spacings vector follows a symmetric Dirichelet distribution, which is equivalent to uniform sampling from an n-simplex, e.g., see David and Nagaraja [60]. Furthermore, the asymptotics of the kth largest gap have also been extensively studied, e.g., in Barbe [13]. Here, we provide a simple finite sample bound on δ, without using distributional or geometric characterizations, but rather a direct argument based on binning.

Consider an arbitrary point x in $[0, 1-\alpha]$. Then the probability that at least one draw from our underlying distribution occurs in $[x, x+\alpha]$ is bounded below by $1-(1-p_0\alpha)^n$. Now divide $[0,1]$ into bins of length α (the last bin can be overlapping with the second to last bin). Note that the event in which there is at least one sample point in each bin implies that the maximum gap δ between adjacent points is less than or equal to 2α. By the union bound, this event occurs with probability at least $1-\lceil \frac{1}{\alpha} \rceil (1-p_0\alpha)^n$.

Let $\alpha = r\log n/(p_0 n)$, and assume n is sufficiently large so that $r \log n/(p_0 n) < 1$. Then we have

$$\begin{aligned}\left\lceil \frac{1}{\alpha} \right\rceil (1-p_0\alpha)^n \leq \Big(\frac{1}{\alpha}+1\Big)(1-p_0\alpha)^n = \frac{p_0 n + r\log n}{r \log n}\Big(1-\frac{r\log n}{n}\Big)^n \\ \leq 2p_0 n \exp(-r\log n) = 2p_0 n^{1-r}.\end{aligned}$$

Plugging in $r = 11$, we get the desired result for $C = 22$, i.e., with probability at least $1-2p_0n^{-10}$, the maximum gap satisfies $\delta \leq 22\log n/(p_0 n)$.

8.7.7 Proof of Corollary 8.7 (trend filtering rate, random inputs)

The proof of this result is entirely analogous to the proof of Theorem 8.5; the only difference is that

$$\max_{i=1,\ldots n-1} (x_{i+1} - x_i) = O_{\mathbb{P}}(\log n/n),$$

(i.e., convergence in probability now), and so accordingly,

$$n^{(2k+2)/(2k+3)} \cdot \max_{i,j=1,\ldots n} |G_{ij}^{(k)} - H_{ij}^{(k)}| \ \overset{p}{\to}\ 0 \quad \text{as } n \to \infty,$$

employing Lemmas 8.4 and 8.6. The same arguments now apply; the stability result in Corollary 4 in Appendix B of Tibshirani [221] must now be applied to random predictor matrices, but this is an extension that is straightforward to verify.

8.8 The higher order KS test

8.8.1 Motivating arguments

As described in the text, the classical KS test is

$$\mathrm{KS}(X_{(m)}, Y_{(n)}) = \max_{z_j \in Z_{(m+n)}} \left| \frac{1}{m} \sum_{i=1}^{m} 1\{x_i \le z_j\} - \frac{1}{n} \sum_{i=1}^{n} 1\{y_i \le z_j\} \right|, \tag{8.27}$$

over samples $X_{(m)} = (x_1, \dots x_m)$ and $Y_{(n)} = (y_1, \dots y_n)$, written in combined form as $Z_{(m+n)} = X_{(m)} \cup Y_{(n)} = (z_1, \dots z_{m+n})$. It is well-known that the above definition is equivalent to

$$\mathrm{KS}(X_{(m)}, Y_{(n)}) = \max_{f\,:\,\mathrm{TV}(f) \le 1} \left| \hat{\mathbb{E}}_{X_{(m)}}[f(X)] - \hat{\mathbb{E}}_{Y_{(n)}}[f(Y)] \right|, \tag{8.28}$$

where we write $\hat{\mathbb{E}}_{X_{(m)}}$ for the empirical expectation under $X_{(m)}$, so $\hat{\mathbb{E}}_{X_{(m)}}[f(X)] = 1/m \sum_{i=1}^{m} f(x_i)$, and similarly for $\hat{\mathbb{E}}_{Y_{(n)}}$. The equivalence between these two definitions follows from the fact that the maximum in (8.28) always occurs at an indicator function, $f(x) = 1\{x \le z_i\}$, for some $i = 1, \dots m+n$.

We now will step through a sequence of motivating arguments that lead to the definition of the higher order KS test in (8.20). The basic idea is to alter the constraint set in (8.28), and consider functions of bounded variation in their kth derivative, for some fixed $k \ge 0$. This gives

$$\max_{f\,:\,\mathrm{TV}(f^{(k)}) \le 1} \left| \hat{\mathbb{E}}_{X_{(m)}}[f(X)] - \hat{\mathbb{E}}_{Y_{(n)}}[f(Y)] \right|. \tag{8.29}$$

Is it possible to compute such a quantity? By a variational result in Mammen and van de Geer [152], the maximum in (8.29) is always achieved by a kth order spline function. In principle, if we knew some finite set T containing the knots of the maximizing spline, then we could restrict our attention to the space of splines with knots in T. However, when $k \ge 2$, such a set T is not generically easy to find, because the knots of the maximizing spline in (8.29) can lie outside of the set of data samples $Z_{(m+n)} = \{z_1, \dots z_{m+1}\}$ [152]. Therefore, we further restrict the functions in consideration in (8.29) to be kth order splines with knots contained in $Z = Z_{(m+n)}$. Letting $\mathcal{S}_Z^{(k)}$ denote the space of such spline functions, we hence examine

$$\max_{f \in \mathcal{S}_Z^{(k)}\,:\,\mathrm{TV}(f^{(k)}) \le 1} \left| \hat{\mathbb{E}}_{X_{(m)}}[f(X)] - \hat{\mathbb{E}}_{Y_{(n)}}[f(Y)] \right|. \tag{8.30}$$

As $\mathcal{S}_Z^{(k)}$ is a finite-dimensional function space (in fact, $(m+n)$-dimensional), we can rewrite (8.30) in a parametric form, similar to (8.27). Let $g_1, \dots g_{m+n}$ denote the kth order truncated power basis with knots over the set of joined data samples Z. Then any function $f \in \mathcal{S}_Z^{(k)}$ with $\mathrm{TV}(f^{(k)}) \le 1$ can be expressed as $f = \sum_{j=1}^{m+n} \alpha_j g_j$, where the coefficients satisfy $\sum_{j=k+2}^{m+n} |\alpha_j| \le 1$. In terms of the evaluations of the function f over $z_1, \dots z_{m+n}$, we have

$$\big(f(z_1), \dots f(z_{m+n})\big) = G^{(k)} \alpha,$$

where $G^{(k)}$ is the truncated power basis matrix, i.e., its columns give the evaluations of $g_1, \dots g_{m+n}$ over the points $z_1, \dots z_{m+n}$. Therefore (8.30) can be re-expressed as

$$\max_{\sum_{j=k+2}^{m+n} |\alpha_j| \leq 1} \left| \frac{1}{m} \mathbb{1}_{X_{(m)}}^T G^{(k)} \alpha - \frac{1}{n} \mathbb{1}_{Y_{(n)}}^T G^{(k)} \alpha \right|. \tag{8.31}$$

Here $\mathbb{1}_{X_{(m)}}$ is an indicator vector of length $m+n$, indicating the membership of each point in the joined sample $Z_{(m+n)}$ to the set $X_{(m)}$. The analogous definition is made for $\mathbb{1}_{Y_{(n)}}$.

Upon inspection, some care must be taken in evaluating the maximum in (8.31). Let us decompose the coefficient vector into blocks as $\alpha = (\alpha_1, \alpha_2)$, where α_1 denotes the first $k+1$ coefficients and α_2 the last $m+n-k-1$. Then the constraint in (8.31) is simply $\|\alpha_2\|_1 \leq 1$, and it is not hard to see that since α_1 is unconstrained, we can choose it to make the criterion in (8.31) arbitrarily large. Therefore, in order to make (8.31) well-defined (finite), we employ the further restriction $\alpha_1 = 0$, yielding

$$\max_{\|\alpha_2\|_1 \leq 1} \left| \frac{1}{m} \mathbb{1}_{X_{(m)}}^T G_2^{(k)} \alpha_2 - \frac{1}{n} \mathbb{1}_{Y_{(n)}}^T G_2^{(k)} \alpha_2 \right|, \tag{8.32}$$

where $G_2^{(k)}$ denotes the last $m-n-k-1$ columns of $G^{(k)}$. A simple duality argument shows that (8.32) can be written in terms of the ℓ_∞ norm, finally giving

$$\mathrm{KS}_G^{(k)}(X_{(m)}, Y_{(n)}) = \left\| (G_2^{(k)})^T \left(\frac{\mathbb{1}_{X_{(m)}}}{m} - \frac{\mathbb{1}_{Y_{(n)}}}{n} \right) \right\|_\infty, \tag{8.33}$$

matching the our definition of the kth order KS test in (8.20). Note that when $k=0$, this reduces to the usual (classic) KS test in (8.27).

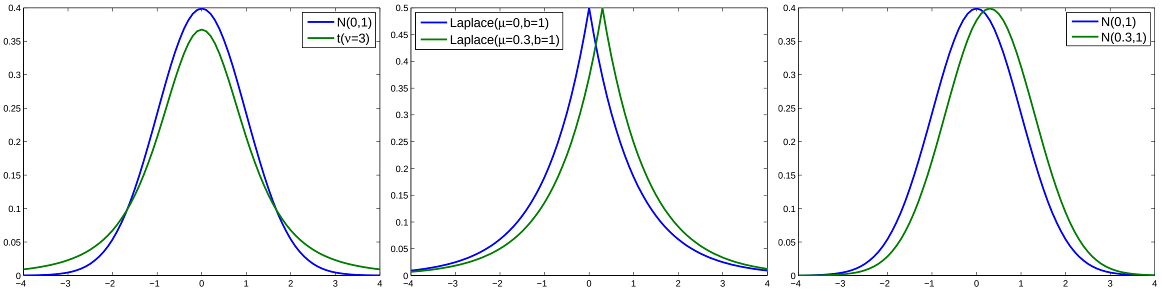
For $k \geq 1$, unlike the usual KS test which requires $O(m+n)$ operations, the kth order KS test in (8.33) requires $O((m+n)^2)$ operations, due to the lower triangular nature of $G^{(k)}$. Armed with our falling factorial basis, we can approximate $\mathrm{KS}_G^{(k)}(X^m, Y^n)$ by

$$\mathrm{KS}_H^{(k)}(X_{(m)}, Y_{(n)}) = \left\| (H_2^{(k)})^T \left(\frac{\mathbb{1}_{X_{(m)}}}{m} - \frac{\mathbb{1}_{Y_{(n)}}}{n} \right) \right\|_\infty, \tag{8.34}$$

where $H^{(k)}$ is the kth order falling factorial basis matrix (and $H_2^{(k)}$ its last $m+n-k-1$ columns) over the points $z_1, \dots z_{m+n}$. After sorting $z_1, \dots z_{m+n}$, the statistic in (8.34) can be computed in $O(k(m+n))$ time; see Algorithm 9, described above in Section 8.7.3.

8.8.2 Additional experiments

In the main text, we presented two numerical experiments, on testing between samples from different distributions P, Q. In the first experiment $P = N(0,1)$ and $Q = t_3$, so the difference between P, Q was mainly in the tails; in the second, $P = \mathrm{Laplace}(0)$ and $Q = \mathrm{Laplace}(0.3)$, and the difference between P, Q was mainly in the centers of the distributions. The first experiment demonstrated that the power of the higher order KS test generally increased as we increased



(a) Experiment 1 (b) Experiment 2 (c) Experiment 3

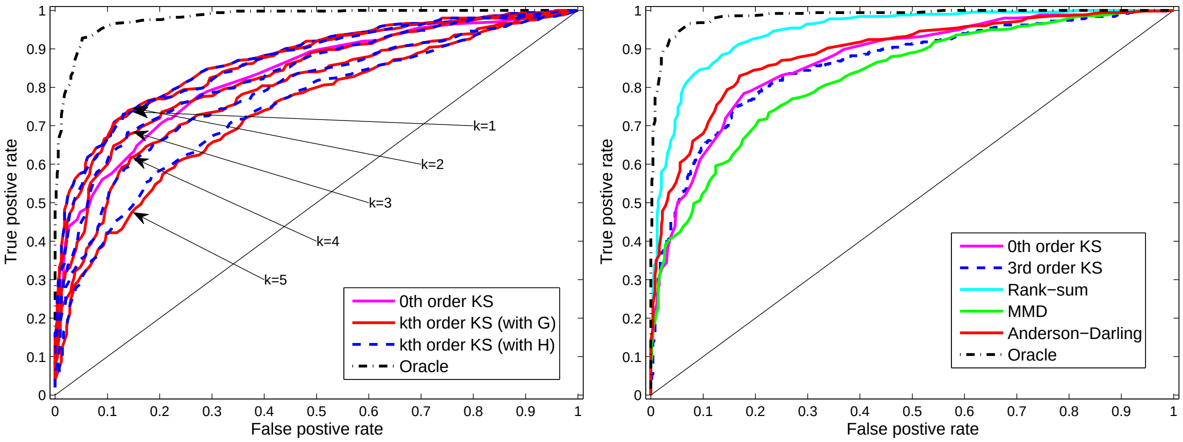
Figure 8.5: An illustration of distribution P vs. Q in our numerical experiments.

the polynomial degree k, the second demonstrated the opposite, i.e., that its power generally decreased for increasing k. Refer back to Figures 8.3 and 8.4 in the main text.

We should note that the first experiment was not carefully crafted in any way; the same performance is seen with a number of similar setups. However, we did have to look carefully to reveal the negative behavior shown in the second experiment. For example, in detecting the difference between mean-shifted standard normals (as opposed to Laplace distributions), the higher order KS tests do not encounter nearly as much difficulty. To demonstrate this, we examine a third experiment here with $P = N(0, 1)$ and $Q = N(0.3, 1)$. Figure 8.5 gives a visual illustration of the distributions across the three experimental setups (the first two considered in the main text, and the third investigated here).

The ROC curves for experiment 3 are given in Figure 8.6. The left panel shows that the test for $k = 1$ improves on the usual test ($k = 0$), even though the difference between the two distributions is mainly near their centers. The right panel shows that the higher order KS tests are competitive with other commonly used nonparametric tests in this setting. The results of this experiment hence suggest that the higher order KS tests provide a utility beyond simply detecting finer tail differences, and the tradeoff induced by varying the polynomial order k is not completely explained as a tradeoff between tail and center sensitivity.

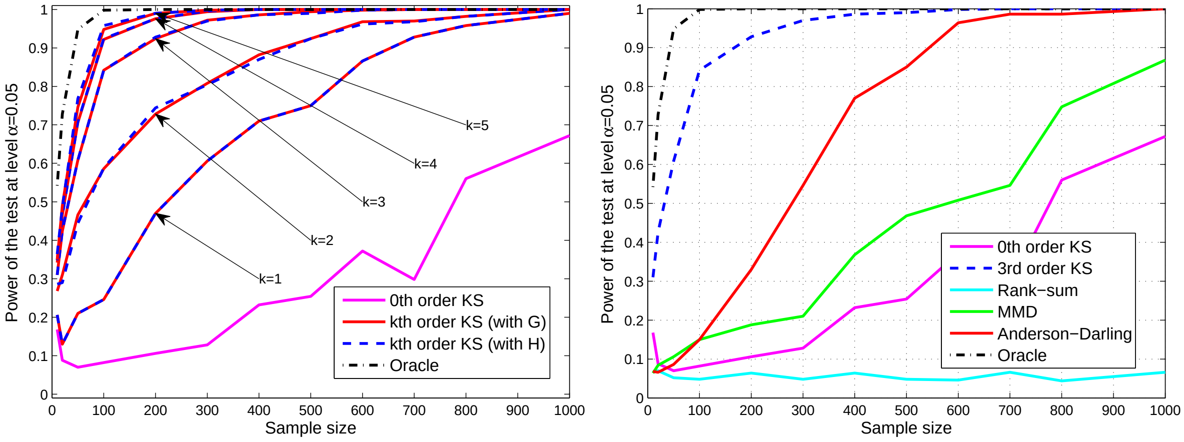
We also study the sample complexity of tests in the three experimental setups. Specifically, over $R = 1000$ repetitions, we find the true positive rate associated with a 0.05 false positive rate, as we let n vary over $10, 20, 50, 100, 200, \dots 1000$. The results for this sample complexity sudy are shown in Figures 8.7, 8.8, and 8.9. We see that the higher order KS tests perform quite favorably the first experimental setup, not so favorably in the second, and somewhere in the middle in the third.



(a) Comparing higher order KS tests

(b) Comparing other tests

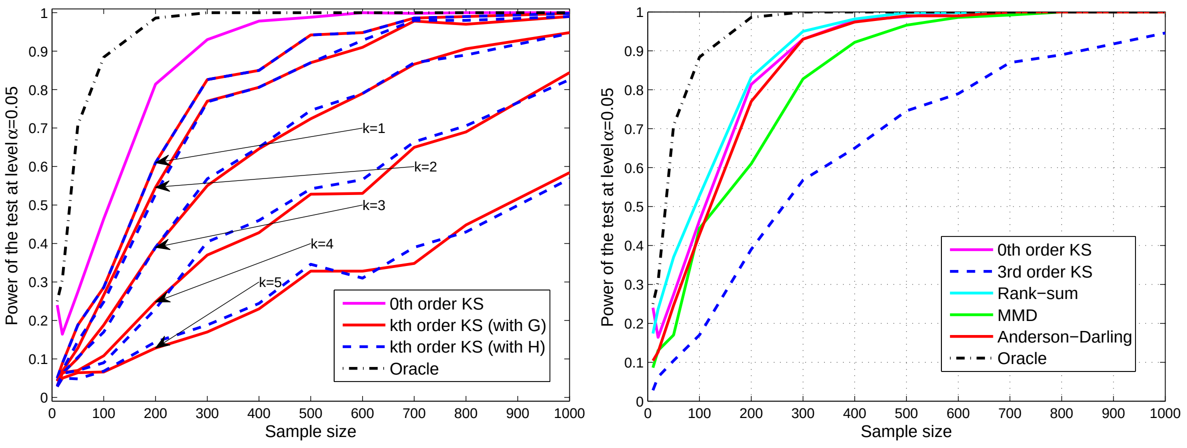
Figure 8.6: ROC curves for experiment 3, normal vs. shifted normal.



(a) Comparing higher order KS tests

(b) Comparing other tests

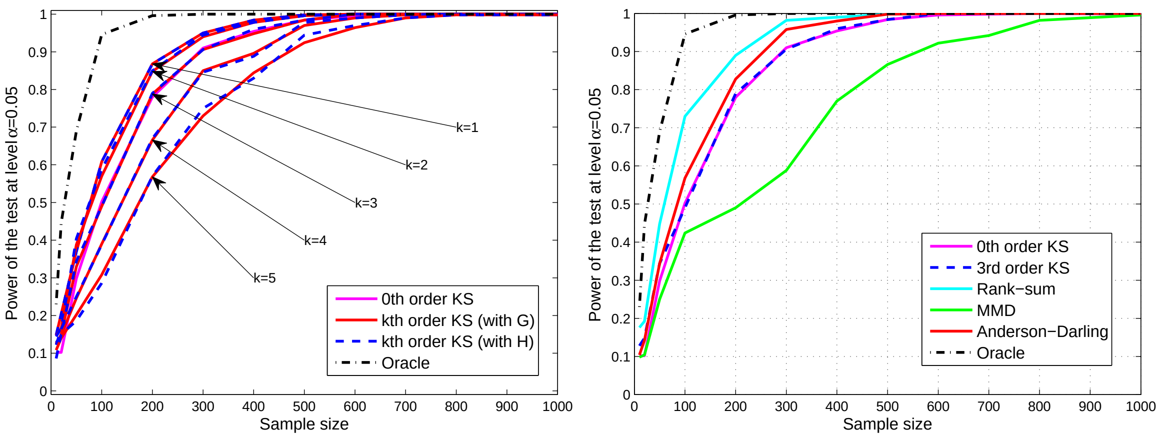
Figure 8.7: Sample complexities at the level $\alpha = 0.05$ for experiment 1, normal vs. t.



(a) Comparing higher order KS tests

(b) Comparing other tests

Figure 8.8: Sample complexities at level $\alpha = 0.05$ in experiment 2, Laplace vs. shifted Laplace.



(a) Comparing higher order KS tests

(b) Comparing other tests

Figure 8.9: Sample complexities at level $\alpha = 0.05$ in experiment 3, normal vs. shifted normal.

Subsequent work and applications

We conclude Part II of the thesis with some pointers to our more recent findings and applications that did not make it to the thesis, and also a few notable related work that builds upon our work.

Chapter 7 is part of a bigger agenda that aims at extending trend filtering to multiple dimensions and understanding the behavior of multivariate trend filtering for signals on grids. The GTF that we described in Chapter 6 is one specific way to construct such an estimator and it is unclear whether it is the right way. The results in Chapter 7 suggest that GTF indeed achieves the optimal rate for the TV-classes with $k = 0$ and all constant d, but $k = 0$ is a very specific case where different ways of defining total variation are the same.

Our new paper [185] addresses the issue by defining discrete higher order total variation classes from the first principle. It identifies subtle issues that causes GTF to be suboptimal even for a discretized Hölder class, due to some boundary artifact of discretization. This also motivated us to consider a new classes of estimators, which we show to achieve the minimax rate using proof techniques from Chapter 6. The results characterizes the minimax rate for $d = 2$ and all constant $k \geq 1$. We summarize the minimax rates that is known so far for TV-classes on grids in Table 8.1. The major open problem right now is to characterize the minimax rate for higher order TV classes for$d \geq 3$.

As we mentioned earlier in Chapter 6, trend filtering-based techniques are easy to adapt to specific application domain since it follows an analysis framework. As a result, we can conduct graph

	$k = 0$ Lower	$k = 0$ Upper	$k \geq 1$ Lower	$k \geq 1$ Upper
$d = 1$	$n^{-2/3}C_n$[70]	$n^{-2/3}C_n$[70, 224]	$n^{-\frac{2k+2}{2k+3}}C_n^{\frac{2}{2k+3}}$[70]	$n^{-\frac{2k+2}{2k+3}}C_n^{\frac{2}{2k+3}}$ [70, 224].
$d = 2$	$n^{-1}C_n\sqrt{\log n}$	$n^{-1}C_n \log n$ [118]	$n^{-\frac{2}{k+2}}C_n^{\frac{2}{k+2}}$	$n^{-\frac{2}{k+2}}C_n^{\frac{2}{k+2}}(\log n)^{\frac{2}{k+2}}$
$d \geq 3$	$n^{-1}C_n\sqrt{\log n}$	$n^{-1}C_n\sqrt{\log n}$[118]	$n^{-\frac{2d}{2k+2+d}}C_n^{\frac{2d}{2k+2+d}}$	?

Table 8.1: Summary of the minimax bounds for GTF on grids. Assuming constant noise. The corresponding canonical rate in corresponding continuous domain TV-class is given in brackets. Results from from Chapter 7 and [185] are highlighted in red and blue respectively. The lower bounds for $k \geq 1$ and $d \geq 3$ do not have a matching upper bound, and in fact, we have reason to believe that they are suboptimal.

transductive trend filtering, generalized linear model trend filtering and so on, by simply modifying the loss function. On the other hand, when there are multiple features in a time varying regression problem, it makes sense to assume the effect of features are additive and they change smoothly over time. Such additive trend filtering is formally studied in [184].

A concrete application of the additive trend filtering is to solve the cybersecurity problem of "attributing hacks". In [148], we derived an additive trend filtering based survival analysis model. The local adaptivity of trend filtering allows sharp changes of certain features' contribution to the hazard function, and therefore allows identifications of major release of certain exploits and subsequent hacker campaigns.

We conclude Part II by mentioning some concurrent and follow-up work in the literature. Padilla et al. [169] realized that a chain-graph is the worst graph and provided a linear-time denoising algorithm by constructing a chain through running DFS over general graphs. El Alaoui et al. [89] picks up the problem of semisupervised learning on graphs with ℓ_p graph-structure regularization, and provided conditions under which the regularizations are asymptotically meaningful. Hutter and Rigollet [118] first realized an $C_n \log n/n$ rate for the $k = 0$ case, and their result motivated us to consider the scaling factors more carefully. Trend filtering on graphs fit naturally into the emerging field signal processing on graphs [186, 198]. [105] investigated the adaptivity of trend filtering to piecewise polynomial signals, with a sparse number of "discrete spline" basis, then the algorithm achieves faster rate. "Discrete splines" [209] are closely related to falling factorials basis that we described in Chapter 8, but do not easily support non-even spacing.

Part III

Minimax theory for modern sequential learning models

Motivation and overview

In Part I and Part II, we have been exclusively considering the batch learning setting, where a data set is given and we are fitting a fixed model class that is specified independent to the data. In practice, however, this ideal scenario of a batch learning setting is seldomly true.

Learning systems are often interactive. In many machine learning systems, e.g., web search, recommendation system, the algorithms being used will largely affect the data being collected. To give a concrete example in web search:

1. My algorithm chooses the top-k search results to show to a user, upon receiving the search query and its corresponding context.
2. The user provides feedbacks through clicks, page flips, or revised search queries.
3. My user will not see anything outside the top-k, therefore, I will never know what would happen if I show them a different set of choices (that comes from a different algorithm).

In the above example, the collected data, e.g., in forms of CONTEXT×ACTION×FEEDBACK, will be specific to the algorithm being used.

Even in cases where the data comes from carefully designed experiments, it is unlikely for the data analyst to only look at the data once. Instead, an experienced statistician or data scientist will first conduct an exploratory data analysis to figure out what model to use. This is also often what is being taught and recommended in modern statistics and machine learning education. Doing exploratory data analysis and trial and errors too extensively, however, could have serious issues. Here we give two examples.

1. An oncologist collects DNA sequences of a few thousands subjects. However, the first couple of attempts to finding features that are associated with cancer do not prove to be statistically significant. The oncologist is not going to let the years of effort in data collection to go in vain, so he decided to work a lot harder, and come up with more features of interest by looking into the data. Luckily after looking at 100 of these, 6 of them have p-value smaller than 0.05.
2. The National Institute of Standard and Technology (NIST) collected a large hand-written digit data set and splitted it into a training set of $60,000$ images and a test set of $10,000$. They also made the data set public. Ever since, thousands of algorithms had been tested on this data set, each potentially depends on the results of previous attempts. Now we have

algorithms that can achieve 99.9% accuracy on the test data set.

The first example clearly reduces the power of a statistical test to null, and the second example might suggest that there are implicit overfitting, and the actual performance of the algorithm achieving near-perfect test error might be a few percentages off on new data.

In this part of the thesis, we study the aforementioned two problems under reasonable mathematical models (and its associated assumptions). In Chapter 9, we address the problem of off-policy evaluation under the contextual bandits model — this is to evaluate a new algorithm when the data sets were collected while running an old algorithm. In Chapter 10, we model the process of choosing what to do with the data set as a partial information game, and design estimators that can be successful despite a sequence of adversarial choices.

In both cases, we find it natural to address the problem in a minimax framework and our main technical contributions are information-theoretical lower bounds. These lower bounds are not necessarily negative results. Instead, they reveal whether the underlying models are reasonable and what we need to assume to further improve upon them.

In the concluding remarks, we will discuss connections of the two models studied and mention some ongoing work in the interactive learning setting that is not included in the thesis.

Chapter 9

Optimal and adaptive off-policy evaluation in contextual bandits

We study the off-policy evaluation problem—estimating the value of a target policy using data collected by another policy—under the contextual bandit model. We consider the general (agnostic) setting without access to a consistent model of rewards and establish a minimax lower bound on the mean squared error (MSE). The bound is matched up to constants by the inverse propensity scoring (IPS) and doubly robust (DR) estimators. This highlights the difficulty of the agnostic contextual setting, in contrast with multi-armed bandits and contextual bandits with access to a consistent reward model, where IPS is suboptimal. We then propose the SWITCH estimator, which can use an existing reward model (not necessarily consistent) to achieve a better bias-variance tradeoff than IPS and DR. We prove an upper bound on its MSE and demonstrate its benefits empirically on a diverse collection of data sets, often outperforming prior work by orders of magnitude.

9.1 Introduction

Contextual bandits refer to a learning setting where the learner repeatedly observes a context, takes an action and observes a reward for the chosen action in the observed context, *but no feedback on any other action*. An example is movie recommendation, where the context describes a user, actions are candidate movies and the reward measures if the user enjoys the recommended movie. The learner produces a policy, meaning a mapping from contexts to actions. A common question in such settings is, given a *target policy*, what is its expected reward? By letting the policy choose actions (e.g., recommend movies to users), we can compute its reward. Such *online evaluation* is typically costly since it exposes users to an untested experimental policy, and does not scale to evaluating many different target policies.

Off-policy evaluation is an alternative paradigm for the same question. Given logs from the existing system, which might be choosing actions according to a very different *logging policy* than the one we seek to evaluate, can we estimate the expected reward of the *target policy*? There are three classes of approaches to address this question: the *direct method* (DM), also known as regression adjustment, *inverse propensity scoring* (IPS) [116] and *doubly robust* (DR) estimators [12, 74, 75, 178].

Our first goal in this chapter is to study the optimality of these three classes of approaches (or lack thereof), and more fundamentally, to quantify the statistical hardness of off-policy evaluation. This problem was previously studied for multi-armed bandits [146] and is related to a large body of work on asymptotically optimal estimators of average treatment effects (ATE) [107, 112, 119, 181], which can be viewed as a special case of off-policy evaluation (see a more detailed exposition in Section 9.9). In both settings, a major underlying assumption is that rewards can be consistently estimated from the features (i.e., covariates) describing contexts and actions, either via a parametric model or non-parametrically. Under such consistency assumptions, it has been shown that DM and/or DR are optimal [119, 146, 181],[1] whereas standard IPS is not [107, 146], but it becomes (asymptotically) optimal when the true propensity scores are replaced by suitable estimates [112].

Unfortunately, consistency of a reward model can be difficult to achieve in practice. Parametric models tend to suffer from a large bias (see, e.g., the empirical evaluation of [74]) and non-parametric models are limited to small dimensions, otherwise non-asymptotic terms become too large (see, e.g., the analysis of non-parametric regression by [29]). Therefore, here we ask: *What can be said about hardness of policy evaluation in the absence of reward-model consistency?*

In this pursuit, we provide the first rate-optimal lower bound on the mean-squared error (MSE) for off-policy evaluation in contextual bandits without consistency assumptions. Our lower bound matches the upper bounds of IPS and DR up to constants, when given a non-degenerate context distributions. This result is in contrast with the suboptimality of IPS under previously studied consistency assumptions, which implies that the two settings are qualitatively different.

Whereas IPS and DR are both minimax optimal, our experiments (similar to prior work) show that IPS is readily outperformed by DR, even when using a simple parametric regression model that is not asymptotically consistent. We attribute this to a lower variance of the DR estimator. We also empirically observe that while DR is generally highly competitive, it is sometimes substantially outperformed by DM. We therefore ask whether it is possible to achieve an even better bias-variance tradeoff than DR. We answer affirmatively and propose a new class of estimators, called the SWITCH estimators, that *adaptively interpolate* between DM and DR (or IPS). We show that SWITCH has MSE no worse than DR (or IPS) in the worst case, but is robust to large importance weights and can achieve a substantially smaller variance than DR or IPS.

We empirically evaluate the SWITCH estimators against a number of strong baselines from prior work, using a previously used experimental setup to simulate contextual bandit problems on real-world multiclass classification data. The results affirm the superior bias-variance tradeoff of SWITCH estimators, with substantial improvements across a number of problems.

[1] The precise assumptions vary for each estimator, and are somewhat weaker for DR than for DM.

In summary, the first part of our paper initiates the study of optimal estimators in a finite-sample setting and without making strong modeling assumptions, while the second part shows how to practically exploit domain knowledge by building better estimators.

9.2 Setup

In contextual bandit problems, the learning agent observes a context x, takes an action a and observes a scalar reward r for the action chosen in the context. Here the context x is a feature vector from some domain $\mathcal{X} \subseteq \mathbb{R}^d$, drawn according to a distribution λ. Actions a are drawn from a finite set $\mathcal{A}$. Rewards r have a distribution conditioned on x and a denoted by $D(r \mid x, a)$. The decision rule of the agent is called a policy, which maps contexts to distributions over actions, allowing for randomization in the action choice. We write $\mu(a \mid x)$ and $\pi(a \mid x)$ to denote the *logging* and *target* policies respectively. Given a policy π, we extend it to a joint distribution over (x, a, r), where $x \sim \lambda$, action $a \sim \pi(a \mid x)$, and $r \sim D(r \mid x, a)$. With this notation, given n i.i.d. samples $(x_i, a_i, r_i) \sim \mu$, we wish to compute the *value of* π:

$$v^\pi = \mathbb{E}_\pi[r] = \mathbb{E}_{x\sim\lambda}\mathbb{E}_{a\sim\pi(\cdot|x)}\mathbb{E}_{r\sim D(\cdot|a,x)}[r]. \tag{9.1}$$

In order to correct for the mismatch in the action distributions under μ and π, it is typical to use *importance weights*, defined as $\rho(x, a) := \pi(a \mid x)/\mu(a \mid x)$. For consistent estimation, it is standard to assume that $\rho(x, a) \neq \infty$, corresponding to *absolute continuity* of π with respect to μ, meaning that whenever $\pi(a \mid x) > 0$, then also $\mu(a \mid x) > 0$. We make this assumption throughout the paper. In the remainder of the setup we present three common estimators of v^π.

The first is the inverse propensity scoring (IPS) estimator [116], defined as

$$\hat{v}^\pi_{\text{IPS}} = \sum_{i=1}^{n} \rho(x_i, a_i) r_i. \tag{9.2}$$

IPS is unbiased and makes no assumptions about how rewards might depend on contexts and actions. When such information is available, it is natural to posit a parametric or non-parametric model of $\mathbb{E}[r \mid x, a]$ and fit it on the logged data to obtain a reward estimator $\hat{r}(x, a)$. Policy evaluation can now simply be performed by scoring π according to $\hat{r}$ as

$$\hat{v}^\pi_{\text{DM}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}} \pi(a \mid x_i)\hat{r}(x_i, a), \tag{9.3}$$

where the DM stands for *direct method* [74], also known as *regression adjustment* or *imputation* [181]. IPS can have a large variance when the target and logging policies differ substantially, and parametric variants of DM can be inconsistent, leading to a large bias. Therefore, both in theory and practice, it is beneficial to combine the approaches into a *doubly robust* estimator [48, 74, 178], such as the following variant,

$$\hat{v}^\pi_{\text{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left[\rho(x_i, a_i)\big(r_i - \hat{r}(x_i, a_i)\big) + \sum_{a\in\mathcal{A}} \pi(a \mid x_i)\hat{r}(x_i, a)\right].$$

Note that IPS is a special case of DR with $\hat{r} \equiv 0$. In the sequel, we mostly focus on IPS and DR, and then suggest how to improve them by further interpolating with DM.

9.3 Limits of Off-policy Evaluation

In this section, we study the off-policy evaluation problem in a minimax setup. After setting up the framework, we present our lower bound and the matching upper bounds for IPS and DR under appropriate conditions.

While minimax optimality is standard in statistical estimations, it is not the only notion of optimality. An alternative framework is that of asymptotic optimality, which establishes Cramer-Rao style bounds on the asymptotic variance of estimators. We use the minimax framework, because it is the most amenable to finite-sample lower bounds, and is complementary to previous asymptotic results, as we discuss after presenting our main results.

9.3.1 Minimax Framework

Off-policy evaluation is a statistical estimation problem, where the goal is to estimate v^π given n i.i.d. samples generated according to a policy μ. We study this problem in a standard minimax framework and seek to answer the following question. What is the smallest MSE that *any* estimator can achieve in the worst case over a large class of contextual bandit problems? As is usual in the minimax setting, we want the class of problems to be rich enough so that the estimation problem is not trivial, and to be small enough so that the lower bounds are not driven by complete pathologies. In our problem, we fix λ, μ and π, and only take worst case over a class of reward distributions. This allows the upper and lower bounds to depend on λ, μ and π, highlighting how these ground-truth parameters influence the problem difficulty. The family of reward distributions $D(r \mid x, a)$ that we study is a natural generalization of the class studied by Li et al. [146] for multi-armed bandits. We assume we are given maps $R_{\max} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}_+$ and $\sigma : \mathcal{X} \times \mathcal{A} \to \mathbb{R}_+$, and define the class of reward distributions $\mathcal{R}(\sigma, R_{\max})$ as[2]

$$\begin{aligned}\mathcal{R}(\sigma, R_{\max}) := \Big\{ D(r|x,a) :\ & 0 \le \mathbb{E}_D[r|x,a] \le R_{\max}(x,a) \\ & \text{and } \mathrm{Var}_D[r|x,a] \le \sigma^2(x,a) \text{ for all } x, a \Big\}.\end{aligned}$$

Note that σ and $R_{\max}$ are allowed to change over contexts and actions. Formally, an estimator is any function $\hat{v} : (\mathcal{X} \times \mathcal{A} \times \mathbb{R})^n \to \mathbb{R}$ that takes n data points collected by μ and outputs an estimate of v^π. The *minimax risk* of off-policy evaluation over the class $\mathcal{R}(\sigma, R_{\max})$, denoted by $R_n(\pi; \lambda, \mu, \sigma, R_{\max})$, is defined as

$$\inf_{\hat{v}} \sup_{D(r|x,a) \in \mathcal{R}(\sigma, R_{\max})} \mathbb{E}\left[(\hat{v} - v^\pi)^2\right]. \tag{9.4}$$

[2] Technically, the inequalities in the definition of $\mathcal{R}(\sigma, R_{\max})$ need to hold almost surely with $x \sim \lambda$ and $a \sim \mu(\cdot \mid x)$.

Recall that the expectation is taken over the n samples drawn from μ, along with any randomness in the estimator. The main goal of this section is to obtain a lower bound on the minimax risk. To state our bound, recall that $\rho(x, a) = \pi(a \mid x)/\mu(a \mid x) < \infty$ is an importance weight at (x, a). We make the following technical assumption on our problem instances, described by tuples of the form $(\pi, \lambda, \mu, \sigma, R_{\max})$:

Assumption 9.1. *There exists $\epsilon > 0$ such that $\mathbb{E}_\mu\left[(\rho\sigma)^{2+\epsilon}\right]$ and $\mathbb{E}_\mu\left[(\rho R_{\max})^{2+\epsilon}\right]$ are finite.*

This assumption is only a slight strengthening of the assumption that $\mathbb{E}_\mu[(\rho\sigma)^2]$ and $\mathbb{E}_\mu[(\rho R_{\max})^2]$ be finite, which is required for consistency of IPS (see, e.g., [75]). Our assumption holds for instance when the context space is finite, because then both ρ and $R_{\max}$ are bounded.

9.3.2 Minimax Lower Bound for Off-policy Evaluation

With the minimax setup in place, we now give our main lower bound on the minimax risk for off-policy evaluation and discuss its consequences. Our bound depends on a parameter $\gamma \in [0, 1]$ and a derived indicator random variable $\xi_\gamma(x, a) := \mathbf{1}(\mu(x, a) \le \gamma)$, which separates out the pairs (x, a) that appear "frequently" under μ.[3] As we will see, the "frequent" pairs (x, a) (where $\xi_\gamma = 0$) correspond to the intrinsically realizable part of the problem, where consistent reward models can be constructed. The "infrequent" pairs (where $\xi_\gamma = 1$) constitute the part that is non-realizable in the worst-case. When $\mathcal{X} \subseteq \mathbb{R}^d$ and λ is continuous with respect to the Lebesgue measure, then $\xi_\gamma(x, a) = 1$ for all $\gamma \in [0, 1]$, so the problem is non-realizable everywhere in the worst-case. Our result uses the following problem-dependent constant (defined with the convention $0/0 = 0$):

$$C_\gamma := 2^{2+\epsilon} \max\left\{\frac{\mathbb{E}_\mu[(\rho\sigma)^{2+\epsilon}]^2}{\mathbb{E}_\mu[(\rho\sigma)^2]^{2+\epsilon}}, \frac{\mathbb{E}_\mu[\xi_\gamma(\rho R_{\max})^{2+\epsilon}]^2}{\mathbb{E}_\mu[\xi_\gamma(\rho R_{\max})^2]^{2+\epsilon}}\right\}.$$

Theorem 9.2. *Assume that a problem instance satisfies Assumption 9.1 with some $\epsilon > 0$. Then for any $\gamma \in [0, 1]$ and any $n \ge \max\left\{16C_\gamma^{1/\epsilon}, 2C_\gamma^{2/\epsilon}\mathbb{E}_\mu[\sigma^2/R_{\max}^2]\right\}$, the minimax risk $R_n(\pi; \lambda, \mu, \sigma, R_{\max})$ satisfies the lower bound*

$$\frac{\mathbb{E}_\mu\left[\rho^2\sigma^2\right] + \mathbb{E}_\mu\left[\xi_\gamma\rho^2 R_{\max}^2\right]\Big(1 - 350n\gamma\log(5/\gamma)\Big)}{700n}.$$

The bound holds for every $\gamma \in [0, 1]$, and we can take the maximum over γ. In particular, we get the following simple corollary under continuous context distributions.

Corollary 9.3. *Under conditions of Theorem 9.2, assume further that λ has a density relative to Lebesgue measure. Then*

$$R_n(\pi; \lambda, \mu, \sigma, R_{\max}) \ge \frac{\mathbb{E}_\mu\left[\rho^2\sigma^2\right] + \mathbb{E}_\mu\left[\rho^2 R_{\max}^2\right]}{700n}.$$

[3] Formally, $\mu(x, a)$ corresponds to $\mu(\,\{(x, a)\}\,)$, i.e., the measure under μ of the set $\{(x, a)\}$. For example, when λ is a continuous distribution then $\mu(x, a) = 0$ everywhere.

If λ is a mixture of a density and point masses, then $\gamma = 0$ will exclude the point masses from the second term of the lower bound. In general, choosing $\gamma = \mathcal{O}\big(1/(n \log n)\big)$ excludes the contexts likely to appear multiple times, and ensures that the second term in Theorem 9.2 remains non-trivial (when $\mu(x, a) \leq \gamma$ with positive probability).

Before sketching the proof of Theorem 9.2, we discuss its preconditions and implications.

Preconditions of the theorem: The theorem assumes the existence of a (problem-dependent) constant C_γ which depends on the constant γ and various moments of the importance-weighted rewards. When $R_{\max}$ and σ are bounded (a common situation), C_γ measures how heavy-tailed the importance weights are. Note that $C_\gamma < \infty$ for all $\gamma \in [0, 1]$ whenever Assumption 9.1 holds, and so the condition on n in Theorem 9.2 is eventually satisfied as long as the random variable $\sigma/R_{\max}$ has a bounded second moment. This is quite reasonable since in typical applications the *a priori* bound on expected rewards is on the same order or larger than the *a priori* bound on the reward noise. For the remainder of the discussion, we assume that n is appropriately large so the preconditions of the theorem hold.

Comparison with upper bounds: The setting of Corollary 9.3 is typical of many contextual bandit applications. In this setting both IPS and DR achieve the minimax risk up to a multiplicative constant. Let $r^*(x, a) := \mathbb{E}[r \mid x, a]$. Recall that DR is using an estimator $\hat{r}(x, a)$ of $r^*(x, a)$, and IPS can be viewed as a special case of DR with $\hat{r} \equiv 0$. By Lemma 3.3(i) of Dudík et al. [75], the MSE of DR is

$$\mathbb{E}[(\hat{v}^\pi_{\text{DR}} - v^\pi)^2] = \frac{1}{n}\Big(\mathbb{E}_\mu[\rho^2\sigma^2] + \text{Var}_{x\sim D}\mathbb{E}_{a\sim\mu(\cdot|x)}[\rho r^*] + \mathbb{E}_{x\sim D}\text{Var}_{a\sim\mu(\cdot|x)}[\rho(\hat{r} - r^*)]\Big).$$

Note that $0 \leq r^* \leq R_{\max}$, so if the estimator $\hat{r}$ also satisfies $0 \leq \hat{r} \leq R_{\max}$, we obtain that the risk of DR (with IPS as a special case) is at most $\mathcal{O}\big(\frac{1}{n}(\mathbb{E}_\mu[\rho^2\sigma^2] + \mathbb{E}_\mu[\rho^2 R^2_{\max}])\big)$. This means that IPS and DR are unimprovable, in the worst case, beyond constant factors. Another implication is that the lower bound of Corollary 9.3 is sharp, and the minimax risk is precisely $\Theta\big(\frac{1}{n}(\mathbb{E}_\mu[\rho^2\sigma^2] + \mathbb{E}_\mu[\rho^2 R^2_{\max}])\big)$. While IPS and DR exhibit the same minimax rates, Eq. (**??**) also immediately shows that DR will be better than IPS whenever $\hat{r}$ is even moderately good (better than $\hat{r} \equiv 0$).

Comparison with asymptotic optimality results: As discussed in Section 9.1, previous work on optimal off-policy evaluation, specifically the average treatment estimation, assumes that it is possible to consistently estimate $r^*(x, a) = \mathbb{E}[r \mid x, a]$. Under such an assumption it is possible to (asymptotically) match the risk of DR with the perfect reward estimator $\hat{r} \equiv r^\star$, and this is the best possible asymptotic risk [107]. This optimal risk is $\frac{1}{n}\big(\mathbb{E}_\mu[\rho^2\sigma^2] + \text{Var}_{x\sim D}\mathbb{E}_\pi[r^* \mid x]\big)$, corresponding to the first two terms of Eq. (**??**), with no dependence on $R_{\max}$. Several estimators achieve this risk, *including the multiplicative constant*, under various consistency assumptions [107, 112, 119, 181]. Note that this is strictly below our lower bound for continuous λ. That is, consistency assumptions yield a better asymptotic risk than possible in the agnostic setting. The gap in constants between our upper and lower bounds is due to the finite-sample setting, where lower-order terms cannot be ignored, but have to be explicitly bounded. Indeed, apart from the result of Li et al. [146], discussed below, ours is the first finite-sample lower bound for off-policy evaluation.

Comparison with multi-armed bandits: For multi-armed bandits, equivalent to contextual bandits with a single context, Li et al. [146] show that the minimax risk equals $\Theta(\mathbb{E}_\mu[\rho^2\sigma^2]/n)$ and is achieved, e.g., by DM, whereas IPS is suboptimal. They also obtain a similar result for contextual bandits, assuming that each context appears with a large-enough probability to estimate its associated rewards by empirical averages (amounting to realizability). While we obtain a larger lower bound, this is not a contradiction, because we allow arbitrarily small probabilities of individual contexts and even continuous distributions, where the probability of any single context is zero.

On a closer inspection, the first term of our bound in Theorem 9.2 coincides with the lower bound of Li et al. [146] (up to constants). The second term (optimized over γ) is non-zero only if there are contexts with small probabilities relative to the number of samples. In multi-armed bandits, we recover the bound of Li et al. [146]. When the context distribution is continuous, or the probability of seeing repeated contexts in a data set of size n is small, we get the minimax optimality of IPS.

One of our key contributions is to highlight this *agnostic contextual* regime where IPS is optimal. In the *non-contextual* regime, where each context appears frequently, the rewards for each context-action pair can be consistently estimated by empirical averages. Similarly, the asymptotic results discussed earlier focus on a setting where rewards can be consistently estimated thanks to parametric assumptions or smoothness (for non-parametric estimation), with the goal of asymptotic efficiency. Our work complements that line of research. In many practical situations, we wish to evaluate policies on high-dimensional context spaces, where the consistent estimation of rewards is not a feasible option. In other words, the agnostic contextual regime dominates.

The distinction between the contextual and non-contextual regime is also present in our proof, which combines a non-contextual lower bound due to the reward noise, similar to the analysis of Li et al. [146], and an additional bound arising for non-degenerate context distributions. This latter result is a key technical novelty of our paper.

Proof sketch: We only sketch some of the main ideas here and defer the full proof to Section 9.7.1. For simplicity, we discuss the case where λ is a continuous distribution. We consider two separate problem instances corresponding to the two terms in Theorem 9.2. The first part is relatively straightforward and reduces the problem to Gaussian mean estimation. We focus on the second part which depends on $R_{\max}$. Our construction defines a prior over the reward distributions, $D(r \mid x, a)$. Given any (x, a), a problem instance is given by

$$\mathbb{E}[r \mid x, a] = \eta(x, a) = \begin{cases} R_{\max}(x, a) & \text{w.p. } \theta(x, a), \\ 0 & \text{w.p. } 1 - \theta(x, a), \end{cases}$$

for $\theta(x, a)$ to be appropriately chosen. Once η is drawn, we consider a problem instance defined by η where the rewards are deterministic and the only randomness is in the contexts. In order to lower bound the MSE across all problems, it suffices to lower bound $\mathbb{E}_\theta[\mathrm{MSE}_\eta(\hat{v})]$. That is, we can compute the MSE of an estimator for each individual η, and take expectation of the MSEs under the prior prescribed by θ. If the expectation is large, we know that there is a problem instance where the estimator incurs a large MSE.

A key insight in our proof is that this expectation can be lower bounded by $\mathrm{MSE}_{\mathbb{E}_\theta[\eta(x,a)]}(\hat{v})$, corresponding to the MSE of a single problem instance with the actual *rewards*, rather than $\eta(x,a)$, drawn according to θ and with the mean reward function $\mathbb{E}_\theta[\eta(x,a)]$. This is powerful, since this new problem instance has stochastic rewards, just like Gaussian mean estimation, and is amenable to standard techniques. The lower bound by $\mathrm{MSE}_{\mathbb{E}_\theta[\eta(x,a)]}(\hat{v})$ is only valid when the context distribution λ is rich enough (e.g., continuous). In that case, our reasoning shows that with enough randomness in the context distribution, a problem with even a deterministic reward function is extremely challenging.

9.4 Incorporating Reward Models

As discussed in the previous section, it is generally possible to beat our minimax bound when consistent reward models exist. We also argued that even in the absence of a consistent model, when DR and IPS both achieve optimal risk rates, the performance of DR on finite samples will be better than IPS as long as the reward model is even moderately good (see Eq. **??**). However, under a large reward noise σ, DR may still suffer from high variance when the importance weights are large, even when given a perfect reward model. In this section, we derive a class of estimators that leverage reward models to directly address this source of high variance, in a manner very different from the standard DR approach.

9.4.1 The SWITCH Estimators

Our starting point is the observation that insistence on maintaining unbiasedness puts the DR estimator at one extreme end of the bias-variance tradeoff. Prior works have considered ideas such as truncating the rewards or importance weights when the importance weights are large (see, e.g., Bottou et al. 37), which can dramatically reduce the variance at the cost of a little bias. We take the intuition a step further and propose to estimate the rewards for actions by two distinct strategies, based on whether they have a large or a small importance weight in a given context. When importance weights are small, we continue to use our favorite unbiased estimators, but switch to directly applying the (potentially biased) reward model on actions with large importance weights. Here, "small" and "large" are defined via a *threshold parameter* τ. Varying this parameter between 0 and ∞ leads to a family of estimators which we call the SWITCH estimators as they switch between an agnostic approach (such as DR or IPS) and the direct method.

We now formalize this intuition, and begin by decomposing v^π according to importance weights:

$$\begin{aligned}\mathbb{E}_\pi[r] &= \mathbb{E}_\pi[r\mathbf{1}(\rho \le \tau)] + \mathbb{E}_\pi[r\mathbf{1}(\rho > \tau)] \\ &= \mathbb{E}_\mu[\rho r\mathbf{1}(\rho \le \tau)] + \mathbb{E}_{x\sim\lambda}\Big[\sum_{a\in\mathcal{A}} \mathbb{E}_D[r \mid x,a]\,\pi(a \mid x)\,\mathbf{1}(\rho(x,a)>\tau)\Big].\end{aligned}$$

Conceptually, we split our problem into two. The first problem always has small importance weights, so we can use unbiased estimators such as IPS or DR. The second problem, where importance weights are large, is addressed by DM. Writing this out leads to the following estimator:

$$\hat{v}_{\text{SWITCH}} = \frac{1}{n}\sum_{i=1}^{n}\left[r_i\rho_i\mathbf{1}(\rho_i \le \tau)\right] + \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}}\hat{r}(x_i,a)\pi(a \mid x_i)\mathbf{1}(\rho(x_i,a) > \tau). \tag{9.5}$$

Note that the above estimator specifically uses IPS on the first part of the problem. When DR is used instead of IPS, we refer to the resulting estimator as SWITCH-DR. The reward model used within the DR part of the SWITCH-DR estimator can be the same or different from the reward model used to impute rewards in the second part. We next present a bound on the MSE of the SWITCH estimator using IPS. A similar bound holds for SWITCH-DR.

Theorem 9.4. *Let $\epsilon(a,x) := \hat{r}(a,x) - \mathbb{E}[r|a,x]$ be the bias of $\hat{r}$ and assume $\hat{r}(x,a) \in [0, R_{\max}(x,a)]$ almost surely. Then for $\hat{v}_{\text{SWITCH}}$, with $\tau > 0$, the MSE is at most*

$$\frac{2}{n}\left\{\mathbb{E}_{\mu}\left[\left(\sigma^2 + R_{\max}^2\right)\rho^2\mathbf{1}(\rho \le \tau)\right] + \mathbb{E}_{\pi}\left[R_{\max}^2\mathbf{1}(\rho > \tau)\right]\right\} + \mathbb{E}_{\pi}\left[\epsilon\mathbf{1}(\rho > \tau)\right]^2.$$

The proposed estimator interpolates between DM and IPS. For $\tau = 0$, SWITCH coincides with DM, while $\tau \to \infty$ yields IPS. Consequently, SWITCH estimator is minimax optimal when τ is appropriately chosen. However, unlike IPS and DR, the SWITCH and SWITCH-DR estimators are by design more robust to large (or heavy-tailed) importance weights. Several estimators related to SWITCH have been previously studied:

1. Bottou et al. [37] consider a special case of SWITCH with $\hat{r} \equiv 0$, meaning that all the actions with large importance weights are eliminated from IPS. We refer to this method as *Trimmed IPS*.
2. Thomas and Brunskill [218] study an estimator similar to SWITCH in the more general setting of reinforcement learning. Their *MAGIC* estimator can be seen as using several candidate thresholds τ and then evaluating the policy by a weighted sum of the estimators corresponding to each τ. Similar to our approach of automatically determining τ, they determine the weighting of estimators via optimization (as we discuss below).

9.4.2 Automatic Parameter Tuning

So far we have discussed the properties of the SWITCH estimators assuming that the parameter τ is chosen well. Our goal is to obtain the best of IPS and DM, but a poor choice of τ might easily give us the worst of the two estimators. Therefore, a method for selecting τ plays an essential role. A natural criterion would be to pick τ that minimizes the MSE of the resulting estimator. Since we do not know the precise MSE (as v^{π} is unknown), an alternative is to minimize its data-dependent estimate. Recalling that the MSE can be written as the sum of variance and squared bias, we estimate and bound the terms individually.

Recall that we are working with a data set (x_i, a_i, r_i) and $\rho_i := \pi(a_i \mid x_i)/\mu(a_i \mid x_i)$. Using this data, it is straightforward to estimate the variance of the SWITCH estimator. Let $Y_i(\tau)$ denote the estimated value that π obtains on the data point x_i according to the SWITCH estimator with the threshold τ, that is

$$Y_i(\tau) := r_i \rho_i \mathbf{1}(\rho_i \leq \tau) + \sum_{a \in \mathcal{A}} \hat{r}(x_i, a) \pi(a \mid x_i) \mathbf{1}(\rho(x_i, a) > \tau),$$

and $\bar{Y}(\tau) := \frac{1}{n} \sum_{i=1}^{n} Y_i(\tau)$. Since $\hat{v}_{\text{SWITCH}} = \bar{Y}(\tau)$ and the x_i are i.i.d., the variance can be estimated as

$$\operatorname{Var}(\bar{Y}(\tau)) \approx \frac{1}{n^2} \sum_{i=1}^{n} (Y_i(\tau) - \bar{Y}(\tau))^2 =: \widehat{\operatorname{Var}}_\tau, \tag{9.6}$$

where the approximation above is clearly consistent since the random variables Y_i are appropriately bounded as long as the rewards are bounded, because the importance weights are capped at the threshold τ.

Next we turn to the bias term. For understanding bias, we look at the MSE bound in Theorem 9.4, and observe that the last term in that theorem is precisely the squared bias. Rather than using a direct bias estimate, which would require knowledge of the error in $\hat{r}$, we will upper bound this term. We assume that the function $R_{\max}(x, a)$ is known. This is not limiting since in most practical applications an *a priori* bound on the rewards is known. Then we can upper bound the squared bias as

$$\mathbb{E}_\pi \big[\epsilon \mathbf{1}(\rho > \tau)\big]^2 \leq \mathbb{E}_\pi \big[R_{\max} \mathbf{1}(\rho > \tau)\big]^2.$$

Replacing the expectation with an average, we obtain

$$\widehat{\operatorname{Bias}}_\tau^2 := \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\pi \big[R_{\max} \mathbf{1}(\rho > \tau) \bigm| x_i\big]\right]^2.$$

With these estimates, we pick the threshold $\widehat{\tau}$ by optimizing the sum of estimated variance and the upper bound on bias,

$$\widehat{\tau} := \operatorname*{argmin}_\tau \widehat{\operatorname{Var}}_\tau + \widehat{\operatorname{Bias}}_\tau^2. \tag{9.7}$$

Our upper bound on the bias is rather conservative, as it upper bounds the error of DM at the largest possible value for every data point. This has the effect of favoring the use of the unbiased part in SWITCH whenever possible, unless the variance would overwhelm even an arbitrarily biased DM. This conservative choice, however, immediately implies the minimax optimality of the SWITCH estimator using $\widehat{\tau}$, because the incurred bias is no more than our upper bound, and it is incurred only when the minimax optimal IPS estimator would be suffering an even larger variance.

Our automatic tuning is related to the MAGIC estimator of Thomas and Brunskill [218]. The key differences are that we pick only one threshold τ, while they combine the estimates with many different τs using a weighting function. They pick this weighting function by optimizing a bias-variance tradeoff, but with significantly different bias and variance estimators. In our experiments, the automatic tuning using Eq. (9.7) generally works better than MAGIC.

9.5 Experiments

We next empirically evaluate the proposed SWITCH estimators on the 10 UCI data sets previously used for off-policy evaluation [74]. We convert the multi-class classification problem to contextual bandits by treating the labels as actions for a policy μ, and recording the reward of 1 if the correct label is chosen, and 0 otherwise.

In addition to this *deterministic* reward model, we also consider a *noisy* reward model for each data set, which reveals the correct reward with probability 0.5 and outputs a random coin toss otherwise. Theoretically, this should lead to bigger σ^2 and larger variance in all estimators. In both reward models, $R_{\max} \equiv 1$ is a valid bound.

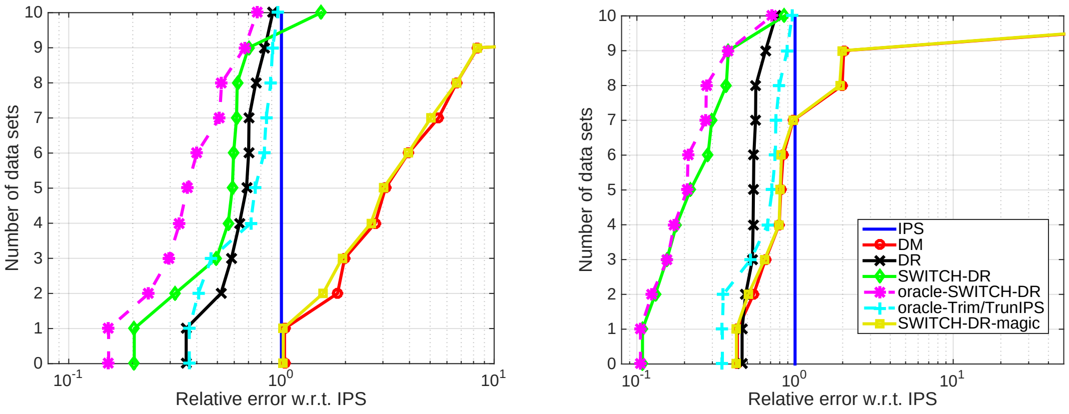
The target policy π is the deterministic decision of a logistic regression classifier learned on the multi-class data, while the logging policy μ samples according to the probability estimates of a logistic model learned on a covariate-shifted version of the data. The covariate shift is obtained as in prior work [74, 103].

In each data set with n examples, we treat the uniform distribution over the data set itself as a surrogate of the population distribution so that we know the ground truth of the rewards. Then, in the simulator, we randomly draw i.i.d. data sets of size $100, 200, 500, 1000, 2000, 5000, 10000, \ldots$ until reaching n, with 500 different repetitions of each size. We estimate MSE of each estimator by taking the empirical average of the squared error over the 500 replicates; note that we can calculate the squared error exactly, because we know v^π. For some of the methods, e.g., IPS and DR, the MSE can have a very large variance due to the potentially large importance weights. This leads to very large error bars if we estimate their MSE even with 500 replicates. To circumvent this issue, we report a clipped version of the MSE that truncates the squared error to 1, namely $\mathrm{MSE} = \mathbb{E}[(\hat{v} - v^\pi)^2 \wedge 1]$. This allows us to get valid confidence intervals for our empirical estimates of this quantity. Note that this does not change the MSE estimate of our approach at all, but is significantly more favorable towards IPS and DR. In this section, whenever we refer to "MSE", we are referring to this truncated version.

We compare SWITCH and SWITCH-DR against the following baselines: 1. *IPS*; 2. *DM trained via logistic regression*; 3. *DR*; 4. *Truncated and Reweighted IPS (TrunIPS)*; and 5. *Trimmed IPS (TrimIPS)*.

In DM, we train $\hat{r}$ and then evaluate the policy on the same contextual bandit data set. Following Dudík et al. [74], DR is constructed by randomly splitting the contextual bandit data into two folds, estimating $\hat{r}$ on one fold, and then evaluating π on the other fold and vice versa, obtaining two estimates. The final estimate is the average of the two. TrunIPS is a variant of IPS, where importance weights are capped at a threshold τ and then renormalized to sum to one [see, e.g., 27]. TrimIPS is a special case of SWITCH due to Bottou et al. [37] described earlier, where $\hat{r} \equiv 0$.

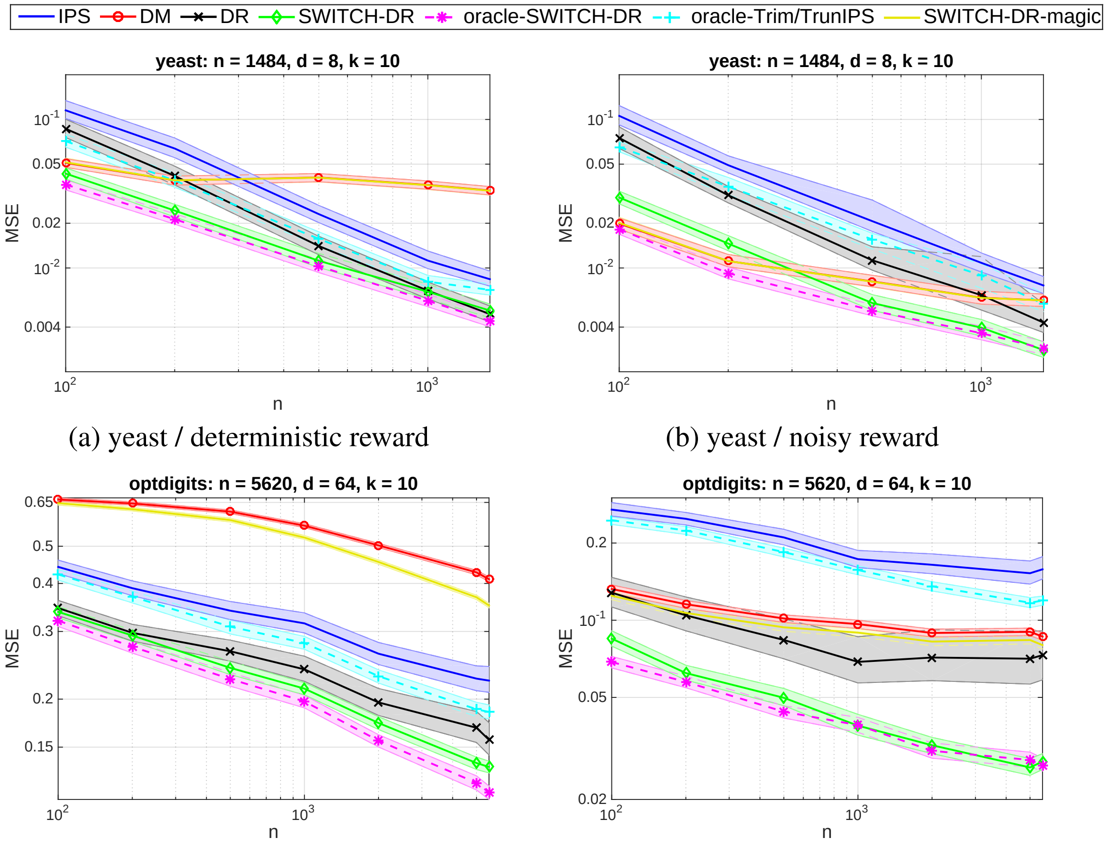
For SWITCH and SWITCH-DR as well as TrunIPS and TrimIPS we select the parameter τ by our automatic tuning from Section 9.4.2. To evaluate our tuning approach, we also include the results for the τ tuned optimally in hindsight, which we refer to as the *oracle* setting, and also show



(a) Deterministic reward

(b) Noisy reward

Figure 9.1: The number of UCI data sets where each method achieves at least a given Rel. MSE. On the left, the UCI labels are used as is; on the right, label noise is added. Curves towards top-left achieve smaller MSE in more cases. Methods in dashed lines are "cheating" by choosing the threshold τ to optimize test MSE. SWITCH-DR outperforms baselines and our tuning of τ is not too far from the best possible. Each data set uses an n which is the size of the data set, drawn via bootstrap sampling and results are averaged over 500 trials.



(a) yeast / deterministic reward

(b) yeast / noisy reward

(c) optdigits / deterministic reward

(d) optdigits / noisy reward

Figure 9.2: MSE of different methods as a function of input data size. *Top:* optdigits data set. *Bottom:* yeast data set.

results obtained by the multi-threshold MAGIC approach. In all these approaches we optimize among 21 possible thresholds, from an exponential grid between the smallest and the largest importance weight observed in the data, considering all actions in each observed context.

In order to stay comparable across data sets and data sizes, our performance measure is the relative MSE with respect to the IPS. Thus, for each estimator $\hat{v}$, we calculate

$$\text{Rel. MSE}(\hat{v}) = \frac{\text{MSE}(\hat{v})}{\text{MSE}(\hat{v}_{\text{IPS}})}.$$

The results are summarized in Figure 9.1, plotting the number of data sets where each method achieves at least a given relative MSE.[4] Thus, methods that achieve smaller MSE across more data sets are towards the top-left corner of the plot, and a larger area under the curve indicates better performance. Some of the differences in MSE are several orders of magnitude large since the relative MSE is shown on the logaritmic scale. As we see, SWITCH-DR dominates all baselines and our empirical tuning of τ is not too far from the best possible. The automatic tuning by MAGIC tends to revert to DM, because its bias estimate is too optimistic and so DM is preferred whenever IPS or DR have some significant variance. The gains of SWITCH-DR are even greater in the noisy-reward setting, where we add label noise to UCI data.

In Figure 9.2, we illustrate the convergence of MSE as n increases. We select two data sets and show how SWITCH-DR performs against baselines in two typical cases: (i) when the direct method works well initially but is outperformed by IPS and DR as n gets large, and (ii) when the direct method works poorly. In the first case, SWITCH-DR outperforms both DM and IPS, while DR improves over IPS only moderately. In the second case, SWITCH-DR performs about as well as IPS and DR despite a poor performance of DM. In all cases, SWITCH-DR is robust to additional noise in the reward, while IPS and DR suffer from higher variance. Results for the remaining data sets are in Section 9.8.

9.6 Conclusion

In this chapter we have carried out minimax analysis of off-policy evaluation in contextual bandits and showed that IPS and DR are minimax optimal in the worst-case, when no consistent reward model is available. This result complements existing asymptotic theory with assumptions on reward models, and highlights the differences between agnostic and consistent settings. Practically, the result further motivates the importance of using side information, possibly by modeling rewards directly, especially when importance weights are too large. Given this observation, we propose a new class of estimators called SWITCH that can be used to combine any importance weighting estimators, including IPS and DR, with DM. The estimators adaptively switch between DM when the importance weights are large and either IPS or DR when the importance weights are small. We show that the new estimators have favorable theoretical properties and also work

[4] For clarity, we have excluded SWITCH, which significantly outperforms IPS, but is dominated by SWITCH-DR. Similarly, we only report the better of oracle-TrimIPS and oracle-TrunIPS.

well on real-world data. Many interesting directions remain open for future work, including high-probability upper bounds on the finite-sample MSE of SWITCH estimators, as well as sharper finite-sample lower bounds under realistic assumptions on the reward model.

9.7 Proofs

The proof is organized as follows. In Section 9.7.1, 9.7.2 and 9.7.3, we provide detailed proofs of the theoretical results in the paper. In Section 9.8, we provide additional figures for the experiments described in Section 9.5.

9.7.1 Proof of Theorem 9.2

In this section we prove the minimax bound of Theorem 9.2. The result is obtained by combining the following two lower bounds:

Theorem 9.5 (Lower bound 1)**.** *For each problem instance such that* $\mathbb{E}_\mu[\rho^2\sigma^2] < \infty$*, we have*

$$R_n(\pi;\lambda,\mu,\sigma,R_{\max}) \geq \frac{\mathbb{E}_\mu[\rho^2\sigma^2]}{32en}\left[1 - \frac{\mathbb{E}_\mu\left[\rho^2\sigma^2\mathbf{1}\Big(\rho\sigma^2 > R_{\max}\sqrt{n\mathbb{E}_\mu[\rho^2\sigma^2]/2}\Big)\right]}{\mathbb{E}_\mu[\rho^2\sigma^2]}\right]^2 .$$

Theorem 9.6 (Lower bound 2)**.** *Assume that* $\mathbb{E}_\mu[\rho^2 R_{\max}^2] < \infty$*, and we are given* $\gamma \in [0,1]$ *and* $\delta \in (0,1]$*. Write* $\xi := \xi_\gamma$ *and* $\gamma' := \max\{\gamma,\delta\}$*. Then there exist functions* $\hat{R}(x,a)$ *and* $\hat{\rho}(x,a)$ *such that*

$$\hat{R}^2(x,a) \leq R_{\max}^2(x,a) \leq (1+\delta)\hat{R}^2(x,a)\ , \qquad \hat{\rho}^2(x,a) \leq \rho^2(x,a) \leq (1+\delta)\hat{\rho}^2(x,a)$$

and the following lower bound holds:

$$\begin{aligned}&R_n(\pi;\lambda,\mu,\sigma,R_{\max})\\ &\geq \frac{\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]}{32en}\left[1 - \frac{\mathbb{E}_\mu\left[\xi\hat{\rho}^2\hat{R}^2\mathbf{1}\Big(\xi\hat{\rho}\hat{R} > \sqrt{n\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]/16}\Big)\right]}{\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]}\right]^2 - \gamma'\log\big(5/\gamma'\big)(1+\delta)\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]\ .\end{aligned}$$

The reason for introducing γ' in Theorem 9.6 is to allow $\gamma = 0$, which is an important special case of the theorem. Otherwise, we could just assume $0 < \delta \leq \gamma$. The first bound captures the intrinsic difficulty due to the variance of reward, and is present even in a vanilla multi-armed bandit problem without contexts. The second result shows the additional dependence on $R_{\max}^2$, even when $\sigma \equiv 0$, whenever the distribution λ is not too degenerate, and captures the additional difficulty of the contextual bandit problem. We next show how these two lower bounds yield Theorem 9.2 and then return to their proofs.

Proof of Theorem 9.2. Throughout the theorem we write $\xi := \xi_\gamma$. We begin by simplifying the two lower bounds. Assume that Assumption 9.1 holds with ϵ. This also means that $\mathbb{E}_\mu[\xi(\rho R_{\max})^{2+\epsilon}]$ is finite as well as $\mathbb{E}_\mu[\xi(\rho R_{\max})^2]$ is finite and either both of them are zero or both of them are non-zero. Similarly, $\mathbb{E}_\mu[(\rho\sigma)^{2+\epsilon}]$ and $\mathbb{E}_\mu[(\rho\sigma)^2]$ are both finite and either both of them are zero or both of them are non-zero, so C_γ is a finite constant. Let $p = 1 + \epsilon/2$ and $q = 1 + 2/\epsilon$, i.e., $1/p + 1/q = 1$. Further, let $\hat{R}$ and $\hat{\rho}$ be the functions from Theorem 9.6. Then the definition of C_γ means that

$$\begin{aligned}
C_\gamma^{1/(\epsilon q)} = C_\gamma^{1/(2+\epsilon)} &= 2 \cdot \max\left\{ \frac{\mathbb{E}_\mu\left[\xi(\rho^2 R_{\max}^2)^{\frac{2+\epsilon}{2}}\right]^{\frac{2}{2+\epsilon}}}{\mathbb{E}_\mu\left[\xi\rho^2 R_{\max}^2\right]}, \frac{\mathbb{E}_\mu\left[(\rho^2\sigma^2)^{\frac{2+\epsilon}{2}}\right]^{\frac{2}{2+\epsilon}}}{\mathbb{E}_\mu\left[\rho^2\sigma^2\right]} \right\} \\
&= 2 \cdot \max\left\{ \frac{\mathbb{E}_\mu\left[\xi(\rho^2 R_{\max}^2)^p\right]^{1/p}}{\mathbb{E}_\mu\left[\xi\rho^2 R_{\max}^2\right]}, \frac{\mathbb{E}_\mu\left[(\rho^2\sigma^2)^p\right]^{1/p}}{\mathbb{E}_\mu\left[\rho^2\sigma^2\right]} \right\} \\
&\geq 2 \cdot \max\left\{ \frac{\mathbb{E}_\mu\left[\xi(\hat{\rho}^2 \hat{R}^2)^p\right]^{1/p}}{\mathbb{E}_\mu\left[\xi\rho^2 R_{\max}^2\right]}, \frac{\mathbb{E}_\mu\left[(\rho^2\sigma^2)^p\right]^{1/p}}{\mathbb{E}_\mu\left[\rho^2\sigma^2\right]} \right\} \quad , \qquad (9.8)
\end{aligned}$$

and recall that we assume that

$$n \geq \max\left\{16 C_\gamma^{1/\epsilon},\ 2C_\gamma^{2/\epsilon}\mathbb{E}_\mu[\sigma^2/R_{\max}^2]\right\} \quad . \qquad (9.9)$$

First, we simplify the correction term in the lower bound of Theorem 9.5. Using Hölder's inequality and Eq. (9.8), we have

$$\begin{aligned}
&\mathbb{E}_\mu\left[\rho^2\sigma^2 \mathbf{1}\left(\rho\sigma^2 > R_{\max}\sqrt{n\mathbb{E}_\mu[\rho^2\sigma^2]/2}\right)\right] \\
&\quad \leq \mathbb{E}_\mu\left[\left(\rho^2\sigma^2\right)^p\right]^{1/p} \cdot \mathbb{P}_\mu\left[\rho\sigma^2 > R_{\max}\sqrt{n\mathbb{E}_\mu[\rho^2\sigma^2]/2}\right]^{1/q} \\
&\quad \leq \frac{1}{2}\mathbb{E}_\mu[\rho^2\sigma^2] \cdot C_\gamma^{1/(\epsilon q)} \cdot \mathbb{P}_\mu\left[\rho\sigma^2/R_{\max} > \sqrt{n\mathbb{E}_\mu[\rho^2\sigma^2]/2}\right]^{1/q} .
\end{aligned}$$

We further invoke Markov's inequality, Cauchy-Schwartz inequality, and Eq. (9.9) in the following three steps to simplify this event as

$$\begin{aligned}
&\leq \frac{1}{2}\mathbb{E}_\mu[\rho^2\sigma^2] \cdot C_\gamma^{1/(\epsilon q)} \cdot \left(\frac{\mathbb{E}_\mu\left[\rho\sigma \cdot (\sigma/R_{\max})\right]}{\sqrt{n\mathbb{E}_\mu[\rho^2\sigma^2]/2}}\right)^{1/q} \\
&\leq \frac{1}{2}\mathbb{E}_\mu[\rho^2\sigma^2] \cdot C_\gamma^{1/(\epsilon q)} \cdot \left(\frac{\sqrt{\mathbb{E}_\mu[\rho^2\sigma^2]} \cdot \sqrt{\mathbb{E}_\mu[\sigma^2/R_{\max}^2]}}{\sqrt{n\mathbb{E}_\mu[\rho^2\sigma^2]/2}}\right)^{1/q} \\
&= \frac{1}{2}\mathbb{E}_\mu[\rho^2\sigma^2] \cdot \left(C_\gamma^{2/\epsilon} \cdot \frac{2\mathbb{E}_\mu[\sigma^2/R_{\max}^2]}{n}\right)^{1/2q} \leq \frac{1}{2}\mathbb{E}_\mu[\rho^2\sigma^2] \quad . \qquad (9.10)
\end{aligned}$$

For the correction term in Theorem 9.6, we similarly have

$$\begin{aligned}
&\mathbb{E}_\mu\left[\xi\hat{\rho}^2\hat{R}^2\mathbf{1}\left(\xi\hat{\rho}\hat{R} > \sqrt{n\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]/16}\right)\right] \\
&\quad\le \mathbb{E}_\mu\Big[\big(\xi\hat{\rho}^2\hat{R}^2\big)^p\Big]^{1/p}\cdot\mathbb{P}_\mu\Big[\xi\hat{\rho}\hat{R} > \sqrt{n\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]/16}\Big]^{1/q} \\
&\quad\le \frac{1}{2}\mathbb{E}_\mu[\xi\rho^2R_{\max}^2]\cdot C_\gamma^{1/(\epsilon q)}\cdot\mathbb{P}_\mu\Big[\xi\hat{\rho}^2\hat{R}^2 > n\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]/16\Big]^{1/q},
\end{aligned}$$

so that Markov's inequality and Eq. (9.9) further yield

$$\begin{aligned}
&\le \frac{1}{2}\mathbb{E}_\mu[\xi\rho^2R_{\max}^2]\cdot C_\gamma^{1/(\epsilon q)}\cdot\left(\frac{\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]}{n\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]/16}\right)^{1/q} \\
&= \frac{1}{2}\mathbb{E}_\mu[\xi\rho^2R_{\max}^2]\cdot\left(C_\gamma^{1/\epsilon}\cdot\frac{16}{n}\right)^{1/q} \le \frac{1}{2}\mathbb{E}_\mu[\xi\rho^2R_{\max}^2] \le \frac{(1+\delta)^2}{2}\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]\ . \qquad (9.11)
\end{aligned}$$

Using Eq. (9.10), the bound of Theorem 9.5 simplifies as

$$\begin{aligned}
&R_n(\pi;\lambda,\mu,\sigma,R_{\max}) \\
&\quad\ge \frac{\mathbb{E}_\mu[\rho^2\sigma^2]}{32en}\left[1-\frac{\mathbb{E}_\mu\left[\rho^2\sigma^2\mathbf{1}\left(\rho\sigma^2 > R_{\max}\sqrt{n\mathbb{E}_\mu[\rho^2\sigma^2]/2}\right)\right]}{\mathbb{E}_\mu[\rho^2\sigma^2]}\right]^2 \\
&\quad\ge \frac{\mathbb{E}_\mu[\rho^2\sigma^2]}{32en}\left(1-\frac{1}{2}\right)^2 = \frac{\mathbb{E}_\mu[\rho^2\sigma^2]}{128en}\ . \qquad (9.12)
\end{aligned}$$

Similarly, by Eq. (9.11), Theorem 9.6 simplifies as

$$\begin{aligned}
&R_n(\pi;\lambda,\mu,\sigma,R_{\max}) \\
&\ge \frac{\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]}{32en}\left[1-\frac{\mathbb{E}_\mu\left[\xi\hat{\rho}^2\hat{R}^2\mathbf{1}\Big(\xi\hat{\rho}\hat{R} > \sqrt{n\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]/16}\Big)\right]}{\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]}\right]^2 - \gamma'\log(5/\gamma')(1+\delta)\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2] \\
&\ge \frac{\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]}{32en}\left[1-\frac{(1+\delta)^2}{2}\right]^2 - \gamma'\log(5/\gamma')(1+\delta)\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2] \\
&= \frac{\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]}{128en}\big(1-2\delta-\delta^2\big)^2 - \gamma'\log(5/\gamma')(1+\delta)\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2] \\
&\ge \frac{\mathbb{E}_\mu[\xi\rho^2R_{\max}^2]}{128en}\frac{\big(1-2\delta-\delta^2\big)^2}{(1+\delta)^2} - \gamma'\log(5/\gamma')(1+\delta)\mathbb{E}_\mu[\xi\rho^2R_{\max}^2]\ .
\end{aligned}$$

Since this bound is valid for all $\delta > 0$, taking $\delta \to 0$, we obtain

$$R_n(\pi;\lambda,\mu,\sigma,R_{\max}) \ge \frac{\mathbb{E}_\mu[\xi\rho^2R_{\max}^2]}{128en} - \gamma\log(5/\gamma)\mathbb{E}_\mu[\xi\rho^2R_{\max}^2]\ .$$

Combining this bound with Eq. (9.12) yields

$$\begin{aligned}
&R_n(\pi; \lambda, \mu, \sigma, R_{\max}) \\
&\quad \geq \frac{1}{2} \cdot \frac{\mathbb{E}_\mu[\rho^2\sigma^2]}{128en} + \frac{1}{2} \cdot \frac{\mathbb{E}_\mu[\xi\rho^2 R_{\max}^2]}{128en} - \frac{1}{2} \cdot \gamma \log(5/\gamma) \mathbb{E}_\mu[\xi\rho^2 R_{\max}^2] \\
&\quad \geq \frac{\mathbb{E}_\mu[\rho^2\sigma^2]}{700n} + \frac{\mathbb{E}_\mu[\xi\rho^2 R_{\max}^2]}{700n} - \frac{1}{2} \cdot \gamma \log(5/\gamma) \mathbb{E}_\mu[\xi\rho^2 R_{\max}^2] \\
&\quad = \frac{1}{700n} \Big[\mathbb{E}_\mu[\rho^2\sigma^2] + \mathbb{E}_\mu[\xi\rho^2 R_{\max}^2] \Big(1 - 350n\gamma \log(5/\gamma)\Big)\Big] \ . \qquad \square
\end{aligned}$$

It remains to prove Theorems 9.5 and 9.6. They are both proved by a reduction to hypothesis testing, and invoke Le Cam's argument to lower-bound the error in this testing problem. As in most arguments of this nature, the key contribution lies in the construction of an appropriate testing problem that leads to the desired lower bounds. Before proving the theorems, we recall the basic result of Le Cam which underlies our proofs. We point the reader to the excellent exposition of Lafferty et al. [136, Section 36.4] on more details about Le Cam's argument.

Theorem 9.7 (Le Cam's method, [136, Theorem 36.8])**.** *Let $\mathcal{P}$ be a set of distributions, let $X_1, \ldots, X_n$ be an i.i.d. sample from some $P \in \mathcal{P}$, let $\theta(P)$ be any function of $P \in \mathcal{P}$, let $\hat{\theta}(X_1, \ldots, X_n)$ be an estimator, and d be a metric. For any pair $P_0, P_1 \in \mathcal{P}$,*

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\hat{\theta}, \theta(P))] \geq \frac{\Delta}{8} e^{-nD_{\mathrm{KL}}(P_0 \| P_1)} \tag{9.13}$$

where $\Delta = d(\theta(P_0), \theta(P_1))$, and $D_{\mathrm{KL}}(P_0 \| P_1) = \int \log(dP_0/dP_1)dP_0$ is the KL-divergence.

While the proofs of the two theorems share a lot of similarities, they have to use reductions to slightly different testing problems given the different mean and variance constraints in the two results. We begin with the proof of Theorem 9.5, which has a simpler construction.

Proof of Theorem 9.5

The basic idea of this proof is to reduce the problem of policy evaluation to that of Gaussian mean estimation where there is a mean associated with each x, a pair. We now describe our construction.

Creating a family of problems Since we aim to show a lower bound on the hardness of policy evaluation in general, it suffices to show a particular family of hard problem instances, such that every estimator requires the stated number of samples on at least one of the problems in this family. Recall that our minimax setup assumes that π, μ and λ are fixed and the only aspect of the problem which we can design is the conditional reward distribution $D(r \mid x, a)$. For Theorem 9.5, this choice is further constrained to satisfy $\mathbb{E}[r \mid x, a] \leq R_{\max}(x, a)$ and $\mathrm{Var}(r \mid x, a) \leq \sigma^2(x, a)$. In order to describe our construction, it will be convenient to define the shorthand $\mathbb{E}[r \mid x, a] = \eta(x, a)$. We will identify a problem in our family with the function $\eta(x, a)$

as that will be the only changing element in our problems. For a chosen η, the policy evaluation question boils down to estimating $v_\eta^\pi = \mathbb{E}[r(x,a)]$, where the contexts x are chosen according to λ, actions are drawn from $\pi(x,a)$ and the reward distribution $D_\eta(r \mid x,a)$ is a normal distribution with mean $\eta(x,a)$ and variance $\sigma^2(x,a)$

$$D_\eta(r \mid x,a) = \mathcal{N}(\eta(x,a),\, \sigma^2(x,a)).$$

Clearly this choice meets the variance constraint by construction, and satisfies the upper bound so long as $\eta(x,a) \le R_{\max}(x,a)$ almost surely. Since the evaluation policy π is fixed throughout, we will drop the superscript and use v_η to denote v_η^π in the remainder of the proofs. With some abuse of notation, we also use $\mathbb{E}_\eta[\cdot]$ to denote expectations where contexts and actions are drawn based on the fixed choices λ and μ corresponding to our data generating distribution, and the rewards drawn from η. We further use P_η to denote this entire joint distribution over (x,a,r) triples.

Given this family of problem instances, it is easy to see that for any pair of η_1, η_2 which are both pointwise upper bounded by $R_{\max}$, we have the lower bound:

$$R_n(\lambda,\pi,\mu,\sigma^2,R_{\max}) \ge \inf_{\hat{v}} \max_{\eta\in\eta_1,\eta_2} \mathbb{E}_\eta\Big[\underbrace{(\hat{v}-v_\eta)^2}_{\ell_\eta(\hat{v})}\Big],$$

where we have introduced the shorthand $\ell_\eta(\hat{v})$ to denote the squared error of $\hat{v}$ to v_η. For a parameter $\epsilon > 0$ to be chosen later, we can further lower bound this risk for a fixed $\hat{v}$ as

$$\begin{aligned} R_n(\hat{v}) &\ge \max_{\eta\in\eta_1,\eta_2} \mathbb{E}_\eta[\ell_\eta(\hat{v})] \ge \max_{\eta\in\eta_1,\eta_2} \epsilon\mathbb{P}_\eta(\ell_\eta \ge \epsilon) \\ &\ge \frac{\epsilon}{2}\Big[\mathbb{P}_{\eta_1}(\ell_{\eta_1}(\hat{v}) \ge \epsilon) + \mathbb{P}_{\eta_2}(\ell_{\eta_2}(\hat{v}) \ge \epsilon)\Big], \end{aligned} \tag{9.14}$$

where the last inequality lower bounds the maximum by the average. So far we have been working with an estimation problem. We next describe how to reduce this to a hypothesis testing problem.

Reduction to hypothesis testing For turning our estimation problem into a testing problem, the idea is to identify a pair η_1, η_2 such that they are far enough from each other so that any estimator which gets a small estimation loss can essentially identify whether the data generating distribution corresponds to P_{η_1} or P_{η_2}. In order to do this, we take any estimator $\hat{v}$ and identify a corresponding test statistic which maps $\hat{v}$ into one of η_1, η_2. The way to do this is essentially identified in Eq. (9.14), and we describe it next.

Note that since we are constructing a hypothesis test for a specific pair of distributions P_{η_1} and P_{η_2}, it is reasonable to consider test statistics which have knowledge of η_1 and η_2, and hence the corresponding distributions. Consequently, these tests also know the true policy values v_{η_1} and v_{η_2} and the only uncertainty is which of them gave rise to the observed data samples. Therefore, for any estimator $\hat{v}$, we can a associate a statistic $\phi(\hat{v}) = \operatorname{argmin}_\eta \{\ell_{\eta_1}(\hat{v}), \ell_{\eta_2}(\hat{v})\}$.

Given this hypothesis test, we are interested in its error rate $\mathbb{P}_\eta(\phi(\hat{v}) \neq \eta)$. We first relate the estimation error of $\hat{v}$ to the error rate of the test. Suppose for now that

$$\ell_{\eta_1}(\hat{v}) + \ell_{\eta_1}(\hat{v}) \geq 2\epsilon, \tag{9.15}$$

so that at least one of the losses is at least ϵ. Suppose that the data comes from η_1. Then if $\ell_{\eta_1}(\hat{v}) < \epsilon$, we know that the test is correct, because by Eq. (9.15) the other loss is greater than ϵ, and therefore $\phi(\hat{v}) = \eta_1$. This means that the error under η_1 can only occur if $\ell_{\eta_1}(\hat{v}) \geq \epsilon$. Similarly, the error under η_2 can only occur if $\ell_{\eta_2}(\hat{v}) \geq \epsilon$, so the test error can be bounded as

$$\begin{aligned}\max_{\eta \in \eta_1, \eta_2} \mathbb{P}_\eta(\phi(\hat{v}) \neq \eta) &\leq \mathbb{P}_{\eta_1}(\phi(\hat{v}) \neq \eta_1) + \mathbb{P}_{\eta_2}(\phi(\hat{v}) \neq \eta_2) \\ &\leq \mathbb{P}_{\eta_1}(\ell_{\eta_1}(\hat{v}) \geq \epsilon) + \mathbb{P}_{\eta_2}(\ell_{\eta_2}(\hat{v}) \geq \epsilon) \\ &\leq \frac{2}{\epsilon} R_n(\hat{v}), \end{aligned} \tag{9.16}$$

where the final inequality uses our earlier lower bound in Eq. (9.14).

To finish connecting our the estimation problem to testing, it remains to establish our earlier supposition (9.15). Assume for now that η_1 and η_2 are chosen such that

$$\left(v_{\eta_1} - v_{\eta_2}\right)^2 \geq 4\epsilon. \tag{9.17}$$

Then an application of the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ yields

$$4\epsilon \leq \left(v_{\eta_1} - v_{\eta_2}\right)^2 \leq 2(\hat{v} - v_{\eta_1})^2 + 2(\hat{v} - v_{\eta_2})^2 = 2\ell_{\eta_1}(\hat{v}) + 2\ell_{\eta_2}(\hat{v}),$$

which yields the posited bound (9.14).

Invoking Le Cam's argument So far we have identified a hypothesis testing problem and a test statistic whose error is upper bounded in terms of the minimax risk of our problem. In order to complete the proof, we now place a lower bound on the error of this test statistic. Recall the result of Le Cam (9.13), which places an upper bound on the attainable error in any testing problem. In our setting, this translates to

$$\max_{\eta \in \eta_1, \eta_2} \mathbb{P}_\eta(\phi(\hat{v}) \neq \eta) \geq \frac{1}{8} e^{-n D_{\mathrm{KL}}(P_{\eta_1} \,||\, P_{\eta_2})}.$$

Since the distribution of the rewards is a spherical Gaussian, the KL-divergence is given by the squared distance between the means, scaled by the variance, that is

$$D_{\mathrm{KL}}(P_{\eta_1} \,||\, P_{\eta_2}) = \mathbb{E}\left[\frac{(\eta_1(x,a) - \eta_2(x,a))^2}{2\sigma^2(x,a)}\right],$$

where we recall that the contexts and actions are drawn from λ and μ respectively. Since we would like the probability of error in the test to be a constant, it suffices to choose η_1 and η_2 such that

$$\mathbb{E}\left[\frac{(\eta_1(x,a) - \eta_2(x,a))^2}{2\sigma^2(x,a)}\right] \leq \frac{1}{n}. \tag{9.18}$$

Picking the parameters So far, we have not made any concrete choices for η_1 and η_2, apart from some constraints which we have introduced along the way. Note that we have the constraints (9.17) and (9.18) which try to ensure that η_1 and η_2 are not too close that an estimator does not have to identify the true parameter, or too far that the testing problem becomes trivial. Additionally, we have the upper and lower bounds of 0 and $R_{\max}$ on η_1 and η_2. In order to reason about these constraints, it is convenient to set $\eta_2 \equiv 0$, and pick $\eta_1(x,a) = \eta_1(x,a) - \eta_2(x,a) = \Delta(x,a)$. We now write all our constraints in terms of Δ.

Note that v_{η_2} is now 0, so that the first constraint (9.17) is equivalent to

$$v_{\eta_1} = \mathbb{E}_{\eta_1}[\rho(x,a)r(x,a)] = \mathbb{E}_{\Delta}[\rho(x,a)r(x,a)] \geq 2\sqrt{\epsilon},$$

where the importance weighting function ρ is introduced since P_{η_1} is based on choosing actions according to μ and we seek to evaluate π. The second constraint (9.18) is also straightforward

$$\mathbb{E}\left[\frac{\Delta^2}{2\sigma^2}\right] \leq \frac{1}{n}.$$

Finally, the bound $R_{\max}$ and non-negativity of η_1 and η_2 are enforced by requiring $0 \leq \Delta(x,a) \leq R_{\max}(x,a)$ almost surely.

The minimax lower bound is then obtained by the largest ϵ in the constraint (9.17) such that the other two constraints can be satisfied. This gives rise to the following variational characterization of the minimax lower bound:

$$\begin{aligned} \max_{\Delta} \quad & \epsilon \\ \text{such that} \quad & \mathbb{E}_{\Delta}[\rho(x,a)r(x,a)] \geq 2\sqrt{\epsilon}, && (9.19) \\ & \mathbb{E}\left[\frac{\Delta^2}{2\sigma^2}\right] \leq \frac{1}{n}, && (9.20) \\ & 0 \leq \Delta(x,a) \leq R_{\max}(x,a). && (9.21) \end{aligned}$$

Instead of finding the optimal solution, we just exhibit a feasible setting of Δ here. We set

$$\Delta = \min\left\{\frac{\alpha\sigma^2\rho}{\mathbb{E}_\mu[\rho^2\sigma^2]}, R_{\max}\right\}, \quad \text{where} \quad \alpha = \sqrt{\frac{2\mathbb{E}_\mu[\rho^2\sigma^2]}{n}}. \tag{9.22}$$

This setting satisfies the bounds (9.21) by construction. A quick substitution also verifies that the constraint (9.20) is satisfied. Consequently, it suffices to set ϵ to the value attained in the constraint (9.19). Substituting the value of Δ in the constraint, we see that

$$\begin{aligned} \mathbb{E}_{\Delta}[\rho(x,a)r(x,a)] &= \mathbb{E}_{x\sim\lambda, a\sim\mu}[\rho(x,a)\Delta(x,a)] \\ &\geq \mathbb{E}_{x\sim\lambda, a\sim\mu}\left[\rho\frac{\alpha\sigma^2\rho}{\mathbb{E}_\mu[\rho^2\sigma^2]}\mathbf{1}\Big(\rho\sigma^2\alpha \leq R_{\max}\mathbb{E}_\mu[\rho^2\sigma^2]\Big)\right] \\ &= \alpha\left(1 - \frac{\mathbb{E}_\mu\left[\rho^2\sigma^2\mathbf{1}\left(\rho\sigma^2\alpha > R_{\max}\mathbb{E}_\mu[\rho^2\sigma^2]\right)\right]}{\mathbb{E}_\mu[\rho^2\sigma^2]}\right) \\ &=: 2\sqrt{\epsilon}. \end{aligned}$$

Putting all the foregoing bounds together, we obtain that for all estimators $\hat{v}$

$$\begin{aligned}
R_n(\hat{v}) &\geq \frac{\epsilon}{2} \cdot \Big(\max_{\eta \in \eta_1, \eta_2} \mathbb{P}_\eta(\phi(\hat{v}) \neq \eta) \Big) \\
&\geq \frac{\epsilon}{2} \cdot \frac{1}{8} e^{-n D_{\mathrm{KL}}(P_{\eta_1} \,\|\, P_{\eta_2})} \\
&\geq \frac{\epsilon}{2} \cdot \frac{1}{8e} = \frac{\epsilon}{16e} \\
&= \frac{1}{16e} \cdot \frac{\alpha^2}{4} \left(1 - \frac{\mathbb{E}_\mu \left[\rho^2 \sigma^2 \mathbf{1}\big(\rho\sigma^2 > R_{\max} \mathbb{E}_\mu[\rho^2\sigma^2]/\alpha\big) \right]}{\mathbb{E}_\mu[\rho^2 \sigma^2]} \right)^2 \\
&= \frac{\mathbb{E}_\mu[\rho^2\sigma^2]}{32en} \left(1 - \frac{\mathbb{E}_\mu \left[\rho^2 \sigma^2 \mathbf{1}\big(\rho\sigma^2 > R_{\max} \sqrt{n \mathbb{E}_\mu[\rho^2\sigma^2]/2}\big) \right]}{\mathbb{E}_\mu[\rho^2 \sigma^2]} \right)^2 .
\end{aligned}$$

Proof of Theorem 9.6

We now give the proof of Theorem 9.6. While it shares a lot of reasoning with the proof of Theorem 9.5, it has one crucial difference. In Theorem 9.5, there is a non-trivial noise in the reward function, unlike in Theorem 9.6. This allowed the proof to work with just two candidate mean-reward functions, since any realization in the data is corrupted with noise. However, in the absence of added noise, the task of mean identification becomes rather trivial: an estimator can just check whether η_1 or η_2 matches the observations exactly.

To prevent such a strategy, we instead construct a richer family of reward functions. Instead of merely two mean rewards, our construction will involve a randomized design of the expected reward function from an appropriate prior distribution. The draw of the mean reward from a prior will essentially generate noise even though any given problem is noiseless. The construction will also highlight the crucial sources of difference between the contextual and multi-armed bandit problems, since the arguments here rely on having access to a rich context distribution, by which we mean distribution that puts non-trivial probability on many contexts. In the absence of this property, the bound of Theorem 9.6 becomes weaker.

Creating a family of problems Our family of problems will be parametrized by the two reals δ and γ from the statement of the theorem. Our construction begins with a discretization step at the resolution δ, whose goal is to create a countable partition of the set of pairs $\mathcal{X} \times \mathcal{A}$. If sets $\mathcal{X}$ and $\mathcal{A}$ are countable or finite, this step is vacuous, but if the sets of contexts or actions have continuous parts, this step is required.

First, let $\mu(x, a)$ denote the joint probability measure obtained by first drawing $x \sim \lambda$ and then $a \sim \mu(\cdot \mid x)$. In Lemma 9.8, we show that $\mathcal{X} \times \mathcal{A}$ can be split into countably many disjoint sets B_i, $\biguplus_{i \in \mathcal{I}} B_i = \mathcal{X} \times \mathcal{A}$, such that the following conditions are satisfied:

- Each $i \in \mathcal{I}$ is associated with numbers $R_i \geq 0$, $\rho_i \geq 0$ and $\xi_i \in \{0,1\}$ such that

  $$R^2_{\max}(x,a) \in [R_i^2,\, (1{+}\delta)R_i^2] \;\;,\quad \rho^2(x,a) \in [\rho_i^2,\, (1{+}\delta)\rho_i^2] \;\;,\quad \xi_\gamma(x,a) = \xi_i \quad \text{for all } (x,a) \in B_i.$$

- Each B_i either satisfies $\mu(B_i) \leq \delta$ or consists of a single pair (x_i, a_i).

The numbers R_i and ρ_i will be exactly $\hat{R}(x,a)$ and $\hat{\rho}(x,a)$ from the theorem statement.

As before, we parametrize the family of reward distributions in terms of the mean reward function $\eta(x,a)$. However, now $\eta(x,a)$ is itself a random variable, which is drawn from a prior distribution. The reward function $\eta(x,a)$ will be constant on each B_i, and its value on B_i, written as $\eta(i)$, will be drawn from a scaled Bernoulli, parametrized by a prior function $\theta(i)$ as follows:

$$\eta(i) = \begin{cases} \xi_i R_i & \text{with probability } \theta(i), \\ 0 & \text{with probability } 1-\theta(i). \end{cases} \tag{9.23}$$

We now set $D_\eta(r \mid x,a) = \eta(i)$ whenever $(x,a) \in B_i$. This clearly satisfies the constraints on the mean since $0 \leq \mathbb{E}[r \mid x,a] \leq R_i \leq R_{\max}(x,a)$ from the property of the partition, and also $\text{Var}(r \mid x,a) = 0$ as per the setting of Theorem 9.6. The goal of an estimator is to take n samples generated by drawing $x \sim \lambda$, $a \mid x \sim \mu$ and $r \mid x,a \sim D_\eta$, and output an estimate $\hat{v}$ such that $\mathbb{E}_\eta[(\hat{v} - v^\pi_\eta)^2]$ is small. We recall our earlier shorthand v_η to denote the value of π under the reward distribution generated by η. For showing a lower bound on this quantity, it is clearly sufficient to pick any prior distribution governed by a parameter θ, as in Eq. (9.23), and lower bound the expectation $\mathbb{E}_\theta\left[\mathbb{E}_\eta[(\hat{v} - v_\eta)^2 \mid \eta]\right]$. If this expectation is large for some estimator $\hat{v}$, then there must be some realization η, which induces a large error least one function $\eta(x,a)$ which induces a large error $\mathbb{E}_\eta[(\hat{v} - v_\eta)^2 \mid \eta]$, as desired. Consequently, we focus in the proof on lower bounding the expectation $\mathbb{E}_\theta[\cdot]$. This expectation can be decomposed with the use of the inequality $a^2 \geq (a+b)^2/2 - b^2$ as follows:

$$\mathbb{E}_\theta\Big[\mathbb{E}_\eta[(\hat{v} - v_\eta)^2 \mid \eta]\Big] \geq \frac{1}{2}\mathbb{E}_\theta\Big[\mathbb{E}_\eta[(\hat{v} - \mathbb{E}_\theta[v_\eta])^2 \mid \eta]\Big] - \mathbb{E}_\theta\Big[(v_\eta - \mathbb{E}_\theta[v_\eta])^2\Big].$$

Taking the worst case over all problems in the above inequality, we obtain

$$\begin{aligned} \sup_\eta \mathbb{E}_\eta[(\hat{v} - v_\eta)^2] &\geq \sup_\theta \mathbb{E}_\theta\Big[\mathbb{E}_\eta[(\hat{v} - v_\eta)^2 \mid \eta]\Big] \\ &\geq \underbrace{\sup_\theta \frac{1}{2}\mathbb{E}_\theta\Big[\mathbb{E}_\eta[(\hat{v} - \mathbb{E}_\theta[v_\eta])^2 \mid \eta]\Big]}_{\mathcal{T}_1} - \underbrace{\sup_\theta \mathbb{E}_\theta\Big[(v_\eta - \mathbb{E}_\theta[v_\eta])^2\Big]}_{\mathcal{T}_2}. \end{aligned} \tag{9.24}$$

This decomposition says that the expected MSE of an estimator in estimating v_η can be related to the MSE of the same estimator in estimating the quantity $\mathbb{E}_\theta[v_\eta]$, as long as the variance of the quantity v_η under the distribution generated by θ is not too large. This is a very important observation, since we can now choose to instead study the MSE of an estimator in estimating $\mathbb{E}_\theta[v_\eta]$ as captured by $\mathcal{T}_1$. Unlike the distribution D_η which is degenerate, this problem has a non-trivial noise arising from the randomized draw of η according to θ. Thus we can use similar techniques as the proof of Theorem 9.5, albeit where the reward distribution is a scaled Bernoulli instead of Gaussian. For now, we focus on controlling $\mathcal{T}_1$, and $\mathcal{T}_2$ will be handled later.

In order to bound $\mathcal{T}_1$, we will consider two carefully designed choices θ_1 and θ_2 to induce two different problem instances and show that $\mathcal{T}_1$ is large for *any estimator* under one of the two parameters. In doing this, it will be convenient to use the additional shorthand $\ell_\theta(\hat{v}) = (\hat{v} - \mathbb{E}_\theta[v_\eta])^2$. Proceeding as in the proof of Theorem 9.5, we have

$$\begin{aligned}\mathcal{T}_1 &= \frac{1}{2}\sup_\theta \mathbb{E}_\theta\Big[\mathbb{E}_\eta[(\hat{v} - \mathbb{E}_\theta[v_\eta])^2 \mid \eta]\Big] = \frac{1}{2}\sup_\theta \mathbb{E}_\theta\Big[\mathbb{E}_\eta[\ell_\theta(\hat{v}) \mid \eta]\Big]\\ &\geq \frac{\epsilon}{2}\sup_\theta \mathbb{P}_\theta\left(\ell_\theta(\hat{v}) \geq \epsilon\right) \geq \frac{\epsilon}{2}\max_{\theta \in \theta_1, \theta_2} \mathbb{P}_\theta\left(\ell_\theta(\hat{v}) \geq \epsilon\right)\\ &\geq \frac{\epsilon}{4}\Big[\mathbb{P}_{\theta_1}\left(\ell_{\theta_1}(\hat{v}) \geq \epsilon\right) + \mathbb{P}_{\theta_2}\left(\ell_{\theta_2}(\hat{v}) \geq \epsilon\right)\Big].\end{aligned}$$

Reduction to hypothesis testing As in the proof of Theorem 9.5, we now reduce the estimation problem into a hypothesis test for whether the data is generated according to the parameter θ_1 or θ_2. The arguments here are similar to the earlier proof, so we will be terser in this presentation.

As before, our hypothesis test has entire knowledge of D_η as well as θ_1 and θ_2. Consequently, we construct a test based on picking θ_1 whenever $\ell_{\theta_1}(\hat{v}) \leq \ell_{\theta_2}(\hat{v})$. As before, we will ensure that $|\mathbb{E}_{\theta_1}[v_\eta] - \mathbb{E}_{\theta_2}[v_\eta]| \geq 2\sqrt{\epsilon}$ so that for any estimator $\hat{v}$, we have

$$\ell_{\theta_1}(\hat{v}) + \ell_{\theta_2}(\hat{v}) \geq 2\epsilon.$$

Under this assumption, we can similarly conclude that the error of our hypothesis test is at most

$$\mathbb{P}_{\theta_1}\left(\ell_{\theta_1}(\hat{v}) \geq \epsilon\right) + \mathbb{P}_{\theta_2}\left(\ell_{\theta_2}(\hat{v}) \geq \epsilon\right).$$

Invoking Le Cam's argument Once again, we can lower bound the error rate of our test by invoking the result of Le Cam. This requires an upper bound on the KL-divergence $D_{\mathrm{KL}}(P_{\theta_1} \| P_{\theta_2})$. The only difference from our earlier argument is that these distributions are now Bernoulli instead of Gaussian, based on the construction in Eq. (9.23). More formally, we have

$$\begin{aligned}D_{\mathrm{KL}}(P_{\theta_1} \| P_{\theta_2}) &= \sum_{i \in \mathcal{I}} \sum_{r \in \{0, xi_i R_i\}} \log\left(\frac{p(r; \theta_1(i))}{p(r; \theta_2(i))}\right) p(r; \theta_1(i))\mu(B_i)\\ &= \mathbb{E}_\mu\Big[\,\xi_i D_{\mathrm{KL}}\big(\text{Ber}(\theta_1(i)) \;\big\|\; \text{Ber}(\theta_2(i))\,\big)\Big], \qquad (9.25)\end{aligned}$$

where i is treated as a random variable under μ, and ξ_i is included, because the two distributions assign $r = 0$ with probability one if $\xi_i = 0$.

Picking the parameters It remains to carefully choose θ_1 and θ_2. We define $\theta_2(i) \equiv 0.5$, and let $\theta_1(i) = \theta_2(i) + \Delta_i$, where Δ_i will be chosen to satisfy certain constraints as before. Then, by Lemma 9.10, the KL divergence in Eq. (9.25) can be bounded as

$$D_{\mathrm{KL}}(P_{\theta_1} \| P_{\theta_2}) \leq \frac{1}{4}\mathbb{E}_\mu\left[\xi_i \Delta_i^2\right].$$

It remains to choose Δ_i. Following a similar logic as before, we seek to find a good feasible solution of the maximization problem

$$\begin{aligned} \max_{\Delta} \quad & \epsilon \\ \text{such that} \quad & \mathbb{E}_\mu\big[\rho(x,a)\xi_i\Delta_i R_i\big] \geq 2\sqrt{\epsilon}, && (9.26) \\ & \frac{1}{4}\mathbb{E}_\mu\left[\xi_i\Delta_i^2\right] \leq \frac{1}{n}, && (9.27) \\ & 0 \leq \Delta_i \leq 0.5. && (9.28) \end{aligned}$$

For some $\alpha > 0$ to be determined shortly, we set

$$\Delta_i = \min\left\{\frac{\xi_i\rho_i R_i\alpha}{\mathbb{E}_\mu[\xi_i\rho_i^2 R_i^2]},\, 0.5\right\}.$$

The bound constraint (9.28) is satisfied by construction and we set $\alpha = \sqrt{4\mathbb{E}_\mu[\xi_i\rho_i^2R_i^2]/n}$ to satisfy the constraint (9.27). To obtain a feasible choice of ϵ, we bound $\mathbb{E}_\mu[\rho(x,a)\xi_i\Delta_iR_i]$ as follows:

$$\begin{aligned} \mathbb{E}_\mu\big[\rho(x,a)\xi_i\Delta_iR_i\big] &\geq \mathbb{E}_\mu\big[\xi_i\rho_i\Delta_iR_i\big] \\ &\geq \mathbb{E}_\mu\left[\frac{\xi_i\rho_i^2R_i^2\alpha\mathbf{1}\big(\xi_i\rho_iR_i \leq \mathbb{E}_\mu[\xi_i\rho_i^2R_i^2]/2\alpha\big)}{\mathbb{E}_\mu[\xi_i\rho_i^2R_i^2]}\right] \\ &= \alpha\left(1 - \frac{\mathbb{E}_\mu\left[\xi_i\rho_i^2R_i^2\mathbf{1}(\xi_i\rho_iR_i > \mathbb{E}_\mu[\xi_i\rho_i^2R_i^2]/2\alpha)\right]}{\mathbb{E}_\mu[\xi_i\rho_i^2R_i^2]}\right) \\ &=: 2\sqrt{\epsilon}. \end{aligned}$$

Collecting our arguments so far, we have established that

$$\begin{aligned} \mathcal{T}_1 &\geq \frac{\epsilon}{4}\cdot\Big(\mathbb{P}_{\theta_1}\left(\ell_{\theta_1}(\hat{v}) \geq \epsilon\right) + \mathbb{P}_{\theta_2}\left(\ell_{\theta_2}(\hat{v}) \geq \epsilon\right)\Big) \\ &\geq \frac{\epsilon}{4}\cdot\frac{1}{8}e^{-nD_{\mathrm{KL}}(P_{\theta_1}\|P_{\theta_2})} \\ &\geq \frac{\epsilon}{4}\cdot\frac{1}{8e} = \frac{\epsilon}{32e} \\ &= \frac{1}{32e}\cdot\frac{\alpha^2}{4}\left(1 - \frac{\mathbb{E}_\mu\left[\xi_i\rho_i^2R_i^2\mathbf{1}(\xi_i\rho_iR_i > \mathbb{E}_\mu[\xi_i\rho_i^2R_i^2]/2\alpha)\right]}{\mathbb{E}_\mu[\xi_i\rho_i^2R_i^2]}\right)^2 \\ &= \frac{\mathbb{E}_\mu[\xi_i\rho_i^2R_i^2]}{32en}\left(1 - \frac{\mathbb{E}_\mu\left[\xi_i\rho_i^2R_i^2\mathbf{1}\left(\xi_i\rho_iR_i > \sqrt{n\mathbb{E}_\mu[\xi_i\rho_i^2R_i^2]/16}\right)\right]}{\mathbb{E}_\mu[\xi_i\rho_i^2R_i^2]}\right)^2. \end{aligned}$$

In order to complete the proof, we need to further upper bound $\mathcal{T}_2$ in the decomposition (9.24).

Bounding $\mathcal{T}_2$ We need to bound the supremum over all priors θ. Consider an arbitrary prior θ and assume that η is drawn according to Eq. (9.23). To bound $\mathbb{E}_\theta\big[(v_\eta - \mathbb{E}_\theta[v_\eta])^2\big]$, we view $(v_\eta - \mathbb{E}_\theta[v_\eta])^2$ as a random variable under θ and bound it using Hoeffding's inequality.

We begin by bounding its range. From the definition of η and v_η,

$$0 \le v_\eta \le \mathbb{E}_\pi[\xi_i R_i] = \mathbb{E}_\mu[\rho(x,a)\xi_i R_i] \le (1+\delta)^{1/2}\mathbb{E}_\mu[\xi_i \rho_i R_i] \ \ ,$$

so also $0 \le \mathbb{E}_\theta[v_\eta] \le (1+\delta)^{1/2}\mathbb{E}_\mu[\xi_i\rho_i R_i]$. Hence, $|v_\eta - \mathbb{E}_\theta[v_\eta]| \le (1+\delta)^{1/2}\mathbb{E}_\mu[\xi_i\rho_i R_i]$, and we obtain the bound

$$(v_\eta - \mathbb{E}_\theta[v_\eta])^2 \le (1+\delta)(\mathbb{E}_\mu[\xi_i\rho_i R_i])^2 \le (1+\delta)\mathbb{E}_\mu[\xi_i\rho_i^2 R_i^2]. \tag{9.29}$$

The proof proceeds by applying Hoeffding's inequality to control the probability that $(v_\eta - \mathbb{E}_\theta[v_\eta])^2 \ge t^2$ for a suitable t. Then we can, with high probability, use the bound $(v_\eta - \mathbb{E}_\theta[v_\eta])^2 \ge t^2$, and with the remaining small probability apply the bound of Eq. (9.29).

To apply Hoeffding's inequality, we write v_η explicitly as

$$v_\eta = \sum_{i\in\mathcal{I}} \mu(B_i)\rho_i'\eta_i =: \sum_{i\in\mathcal{I}} Y_i$$

where $\rho_i' := \mathbb{E}_\mu[\rho(x,a) \mid (x,a) \in B_i]$. Thus, v_η can be written as a sum of countably many independent variables, but we can only apply Hoeffding's inequality to their finite subset. Note that the variables Y_i are non-negative and upper-bounded by a summable series, namely $Y_i \le \mu(B_i)\rho_i' R_i$, where the summability follows because $\mathbb{E}_\mu[\rho R_{\max}] \le 1 + \mathbb{E}_\mu[\rho^2 R_{\max}^2] < \infty$. This means that for any $\delta_0 > 0$, we can choose a finite set $\mathcal{I}_0$ such that $\sum_{i\notin\mathcal{I}_0} Y_i \le \delta_0$. We will determine the sufficiently small value of δ_0 later; for now, consider the corresponding set $\mathcal{I}_0$ and define an auxiliary variable

$$v_\eta' := \sum_{i\in\mathcal{I}_0} Y_i \ \ ,$$

which by construction satisfies $v_\eta' \le v_\eta \le v_\eta' + \delta_0$. Note that the summands Y_i can be bounded as

$$0 \le Y_i \le \xi_i\rho_i' R_i \mu(B_i) \le \xi_i(1+\delta)^{1/2}\rho_i R_i \sqrt{\mu(B_i)}\sqrt{\gamma'}$$

because $\rho_i' \le (1+\delta)^{1/2}\rho_i$ and $\xi_i\mu(B_i) \le \xi_i\sqrt{\mu(B_i)}\sqrt{\gamma'}$, because $\xi_i = 0$ whenever $\mu(B_i) > \max\{\gamma,\delta\} = \gamma'$. By Hoeffding's inequality, we thus have

$$\begin{aligned}\mathbb{P}(|v_\eta' - \mathbb{E}_\theta v_\eta'| \ge t) &\le 2\exp\left\{-\frac{2t^2}{(1+\delta)\sum_{i\in\mathcal{I}_0}\xi_i\rho_i^2R_i^2\mu(B_i)\gamma'}\right\}\\ &\le 2\exp\left\{-\frac{2t^2}{(1+\delta)\gamma'\mathbb{E}_\mu[\xi_i\rho_i^2R_i^2]}\right\}.\end{aligned}$$

Now take $t = \sqrt{\gamma'\log(4/\gamma')(1+\delta)\mathbb{E}_\mu[\xi_i\rho_i^2R_i^2]/2}$ in the above bound, which yields

$$\mathbb{P}\Big[(v_\eta' - \mathbb{E}_\theta v_\eta')^2 \ge t^2\Big] = \mathbb{P}\Big[|v_\eta' - \mathbb{E}_\theta v_\eta'| \ge t\Big] \le \frac{\gamma'}{2} \ \ .$$

Now, we can go back to analyzing v_η. We set δ_0 sufficiently small, so

$$t + \delta_0 \le \sqrt{\gamma' \log(5/\gamma')(1+\delta)\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]/2}.$$

Thus, using Eq. (9.29), we have

$$\begin{aligned}
\mathbb{E}_\theta\left[(v_\eta - \mathbb{E}_\theta v_\eta)^2\right] &\le (t+\delta_0)^2 \cdot \mathbb{P}\Big[(v_\eta - \mathbb{E}_\theta v_\eta)^2 < (t+\delta_0)^2\Big] \\
&\qquad + (1+\delta)\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2] \cdot \mathbb{P}\Big[(v_\eta - \mathbb{E}_\theta v_\eta)^2 \ge (t+\delta_0)^2\Big] \\
&\le (t+\delta_0)^2 + (1+\delta)\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2] \cdot \mathbb{P}\Big[(v'_\eta - \mathbb{E}_\theta v'_\eta)^2 \ge t^2\Big] \\
&= \frac{\gamma' \log(5/\gamma')(1+\delta)\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]}{2} + (1+\delta)\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2] \cdot \frac{\gamma'}{2} \\
&\le \gamma' \log(5/\gamma')(1+\delta)\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]\ .
\end{aligned}$$

Combining this bound with the bound on $\mathcal{T}_1$ yields the theorem.

Lemma 9.8. *Let $\mathcal{Z} := \mathcal{X} \times \mathcal{A}$ be a subset of $\mathbb{R}^d$, let μ be a probability measure on $\mathcal{Z}$ and $R_{\max}$ and ρ be non-negative measurable functions on $\mathcal{Z}$. Given $\gamma \in [0,1]$, define a random variable $\xi_\gamma(z) := \mathbf{1}(\mu(z) \le \gamma)$. Then for any $\delta \in (0,1]$, there exists a countable index set $\mathcal{I}$ and disjoint sets $B_i \subseteq \mathcal{Z}$ alongside non-negative reals R_i, ρ_i and $\xi_i \in \{0,1\}$ such that the following conditions hold:*

- *Sets B_i form a partition of $\mathcal{Z}$, i.e., $\mathcal{Z} = \uplus_{i \in \mathcal{I}} B_i$.*
- *Reals R_i and ρ_i approximate $R_{\max}$ and ρ, and ξ_i equals ξ_γ as follows:*

$$R_{\max}^2(z) \in [R_i^2,\, (1+\delta)R_i^2]\ , \quad \rho^2(z) \in [\rho_i^2,\, (1+\delta)\rho_i^2]\ , \quad \xi_\gamma(z) = \xi_i \quad \textit{for all } z \in B_i.$$

- *Each set B_i either satisfies $\mu(B_i) \le \delta$ or consists of a single $z \in \mathcal{Z}$.*

Proof. Let $\mathcal{Z} := \mathcal{X} \times \mathcal{A}$. We begin our construction by separating out atoms, i.e., the elements $z \in \mathcal{Z}$ such that $\mu(z) > 0$. Specifically, we write $\mathcal{Z} = \mathcal{Z}^{\text{na}} \uplus \mathcal{Z}^{\text{a}}$ where $\mathcal{Z}^{\text{a}}$ consists of atoms and $\mathcal{Z}^{\text{na}}$ of all non-atoms. The set $\mathcal{Z}^{\text{a}}$ is either finite or countably infinite, so $\mathcal{Z}^{\text{na}}$ is measurable.

By a theorem of Sierpiński [199], since μ does not have any atoms on $\mathcal{Z}^{\text{na}}$, it must be continuous on $\mathcal{Z}^{\text{na}}$ in the sense that if A is a measurable subset of $\mathcal{Z}^{\text{na}}$ with $\mu(A) = a$ then for any $b \in [0,a]$, there exists a measurable set $B \subseteq A$ such that $\mu(B) = b$. This means that we can decompose $\mathcal{Z}^{\text{na}}$ into $N := \lceil 1/\delta \rceil$ sets $\mathcal{Z}_1^{\text{na}}, \mathcal{Z}_2^{\text{na}}, \dots, \mathcal{Z}_N^{\text{na}}$ such that each has a measure at most δ and $\mathcal{Z}^{\text{na}} = \biguplus_{j=1}^N \mathcal{Z}_j^{\text{na}}$.

We next ensure the approximation properties for $R_{\max}$ and ρ. We begin by a countable decomposition of non-negative reals. We consider the countable index set $\mathcal{J} := \mathbb{Z} \cup \{-\infty\}$ and define the sequence $a_j := (1+\delta)^{j/2}$, for $j \in \mathbb{Z}$. Positive reals can then be decomposed into the following intervals indexed by $\mathcal{J}$:

$$I_{-\infty} := \{0\}\ , \qquad I_j := (a_j, a_{j+1}] \quad \text{for } j \in \mathbb{Z}.$$

It will also be convenient to set $a_{-\infty} := 0$. Thus, the construction of I_j guarantees that for all $j \in \mathcal{J}$ and all $t \in I_j$ we have $a_j^2 \le t^2 \le (1+\delta)a_j^2$.

The desired partition, with the index set $\mathcal{I} = \mathcal{Z}^{\mathrm{a}} \cup [N] \times \mathcal{J}^2$, is as follows:

$$\begin{aligned}
&\text{for } i = z \in \mathcal{Z}^{\mathrm{a}}: && B_i \coloneqq \{z\},\ R_i \coloneqq R_{\max}(z),\ \rho_i \coloneqq \rho(z),\ \xi_i \coloneqq \xi_\gamma(z);\\
&\text{for } i = (j, j_R, j_\rho) \in [N] \times \mathcal{J}^2: && B_i \coloneqq \mathcal{Z}_j^{\mathrm{na}} \cap R_{\max}^{-1}(I_{j_R}) \cap \rho^{-1}(I_{j_\rho}),\\
& && R_i \coloneqq a_{j_R},\ \rho_i \coloneqq a_{j_\rho},\ \xi_i \coloneqq 1. \qquad \square
\end{aligned}$$

9.7.2 Proof of Theorem 9.4

Let $A_x \coloneqq \{a \in \mathcal{A} :\ \rho(x,a) \le \tau\}$. For brevity, we write $A_i \coloneqq A_{x_i}$. We decompose the mean squared error into the squared bias and variance and control each term separately,

$$\mathrm{MSE}(\hat{v}_{\text{SWITCH}}) = \left|\mathbb{E}[\hat{v}_{\text{SWITCH}}] - v^\pi\right|^2 + \mathrm{Var}[\hat{v}_{\text{SWITCH}}].$$

We first calculate the bias. Note that bias is incurred only in the terms that fall in A_x^c, so

$$\begin{aligned}
\mathbb{E}[\hat{v}_{\text{SWITCH}}] - v^\pi &= \mathbb{E}\left[\sum_{a \in A_x^c} \hat{r}(x,a)\pi(a|x)\right] - \mathbb{E}\left[\sum_{a \in A_x^c} \mathbb{E}[r|x,a]\,\pi(a|x)\right]\\
&= \mathbb{E}_\pi\Big[\big(\hat{r}(x,a) - \mathbb{E}[r|x,a]\big)\,\mathbf{1}(a \in A_x^c)\Big]\\
&= \mathbb{E}_\pi\big[\,\epsilon(x,a)\,\mathbf{1}(\rho > \tau)\,\big]
\end{aligned}$$

where we recall that $\epsilon(x,a) = \hat{r}(x,a) - \mathbb{E}[r|x,a]$.

Next we upper bound the variance. Note that the variance contributions from the IPS part and the DM part are not independent, since the indicators $\rho(x_i,a) > \tau$ and $\rho(x_i,a) \le \tau$ are mutually exclusive. To simplify the analysis, we use the following inequality that holds for any random variable X and Y:

$$\mathrm{Var}(X+Y) \le 2\mathrm{Var}(X) + 2\mathrm{Var}(Y).$$

This allows us to calculate the variance of each part separately.

$$\begin{aligned}
\mathrm{Var}[\hat{v}_{\text{SWITCH}}] &\le 2\,\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n [r_i\rho_i\mathbf{1}(a_i \in A_i)]\right] + 2\,\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n\sum_{a\in\mathcal{A}} \hat{r}(x_i,a)\pi(a|x_i)\mathbf{1}(a \in A_i^c)\right]\\
=&\frac{2}{n}\,\mathrm{Var}_\mu\big[r\rho\mathbf{1}(a \in A_x)\big] + \frac{2}{n}\,\mathrm{Var}\left[\sum_{a\in A_x^c}\hat{r}(x,a)\pi(a|x)\right]\\
=&\frac{2}{n}\,\mathbb{E}_\mu\mathrm{Var}\big[r\rho\mathbf{1}(a \in A_x)\,\big|\,x,a\big] + \frac{2}{n}\,\mathrm{Var}_\mu\mathbb{E}\big[r\rho\mathbf{1}(a \in A_x)\,\big|\,x,a\big] + \frac{2}{n}\,\mathrm{Var}\left[\sum_{a\in A_x^c}\hat{r}(x,a)\pi(a|x)\right]\\
\le&\frac{2}{n}\,\mathbb{E}_\mu\mathrm{Var}\big[r\rho\mathbf{1}(a \in A_x)\,\big|\,x,a\big] + \frac{2}{n}\,\mathbb{E}_\mu\Big[\mathbb{E}[r\rho\mathbf{1}(a \in A_x)\,|\,x,a]^2\Big] + \frac{2}{n}\,\mathbb{E}\left[\Big(\sum_{a\in A_x^c}\hat{r}(x,a)\pi(a|x)\Big)^2\right]\\
\le&\frac{2}{n}\,\mathbb{E}_\mu\big[\sigma^2\rho^2\mathbf{1}(a \in A_x)\big] + \frac{2}{n}\,\mathbb{E}_\mu\big[R_{\max}^2\rho^2\mathbf{1}(a \in A_x)\big] + \frac{2}{n}\,\mathbb{E}\left[\Big(\sum_{a\in A_x^c}\hat{r}(x,a)\pi(a|x)\Big)^2\right].
\end{aligned}$$

To complete the proof, note that the last term is further upper bounded using Jensen's inequality as

$$\begin{aligned}
\mathbb{E}\Big[\Big(\sum_{a\in A_x^c}\hat{r}(x,a)\pi(a|x)\Big)^2\Big] &= \mathbb{E}\Big[\Big(\sum_{a\in A_x^c}\pi(a|x)\Big)^2\Big(\sum_{a\in A_x^c}\frac{\hat{r}(x,a)\pi(a|x)}{\sum_{a\in A_x^c}\pi(a|x)}\Big)^2\Big]\\
&\le \mathbb{E}\Big[\Big(\sum_{a\in A_x^c}\pi(a|x)\Big)\Big(\sum_{a\in A_x^c}\hat{r}(x,a)^2\pi(a|x)\Big)\Big]\\
&\le \mathbb{E}_\pi\big[R_{\max}^2\mathbf{1}(\rho>\tau)\big],
\end{aligned}$$

where the final inequality uses $\sum_{a\in A_x^c}\pi(a|x)\le 1$ and $\hat{r}(x,a)\in[0,R_{\max}(x,a)]$ almost surely. Combining the bias and variance bounds, we get the stated MSE upper bound. □

9.7.3 Utility Lemmas

Lemma 9.9 ([113, Theorem 2]). *Let $X_i\in[a_i,b_i]$ and $X_1,...,X_n$ are drawn independently. Then the empirical mean $\bar{X}=\frac{1}{n}(X_1+...+X_n)$ obeys*

$$\mathbb{P}(|\bar{X}-\mathbb{E}[\bar{X}]|\ge t)\le 2e^{-\frac{2n^2t^2}{\sum_{i=1}^n(b_i-a_i)^2}}.$$

Lemma 9.10 (Bernoulli KL-divergence). *For $0<p,q<1$, we have*

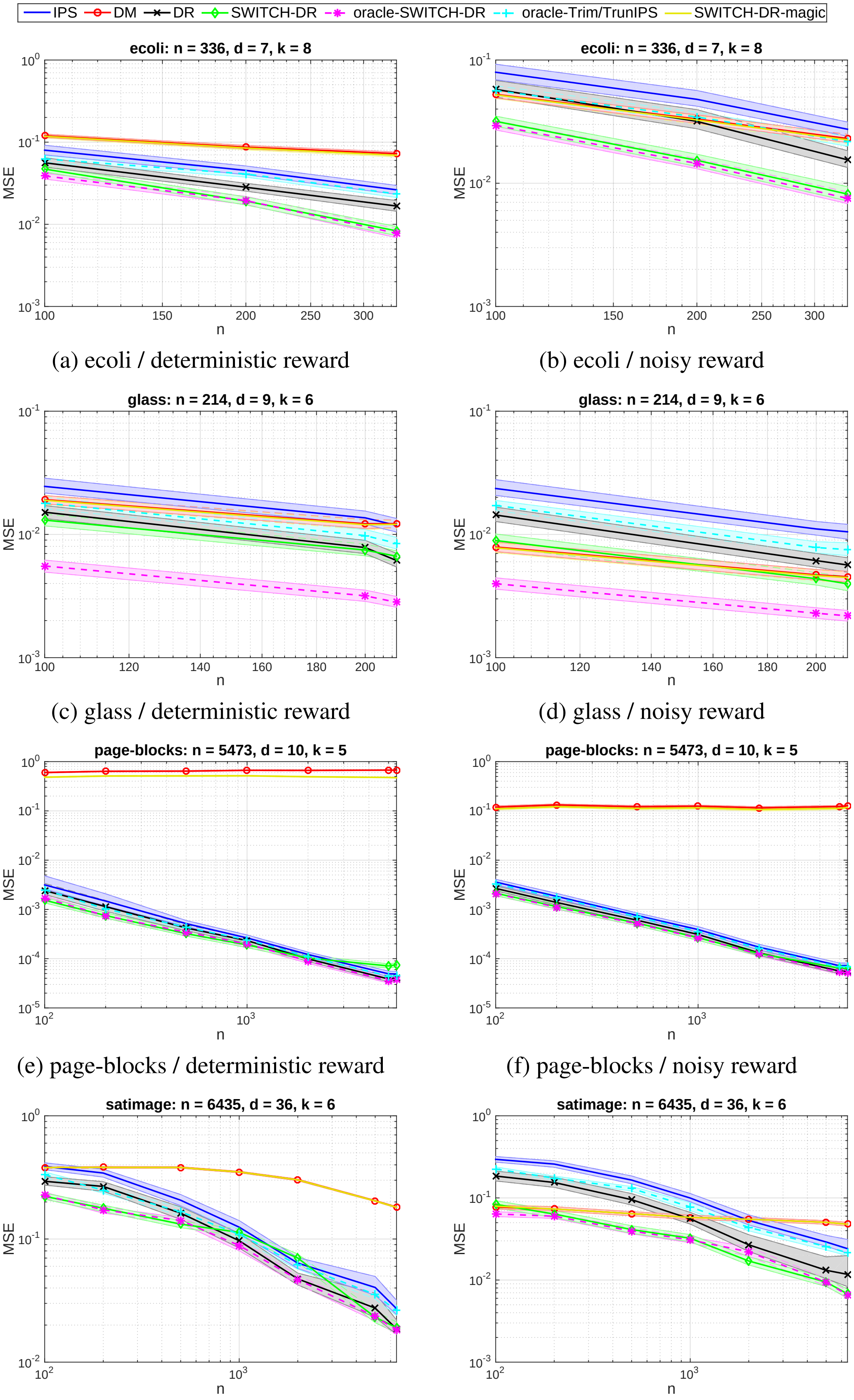
$$D_{\mathrm{KL}}(\mathrm{Ber}(p)\|\mathrm{Ber}(q))\le(p-q)^2(\frac{1}{q}+\frac{1}{1-q}).$$

Proof.

$$\begin{aligned}
D_{\mathrm{KL}}(\mathrm{Ber}(p)\|\mathrm{Ber}(q)) &= p\log\left(\frac{p}{q}\right)+(1-p)\log\left(\frac{1-p}{1-q}\right)\\
&\le p\frac{p-q}{q}+(1-p)\frac{q-p}{1-q}=\frac{(p-q)^2}{q}+(p-q)+\frac{(p-q)^2}{1-q}+(q-p)\\
&=(p-q)^2\left(\frac{1}{q}+\frac{1}{1-q}\right).
\end{aligned}$$

□

9.8 Additional Figures from the Experiments



(a) ecoli / deterministic reward

(b) ecoli / noisy reward

(c) glass / deterministic reward

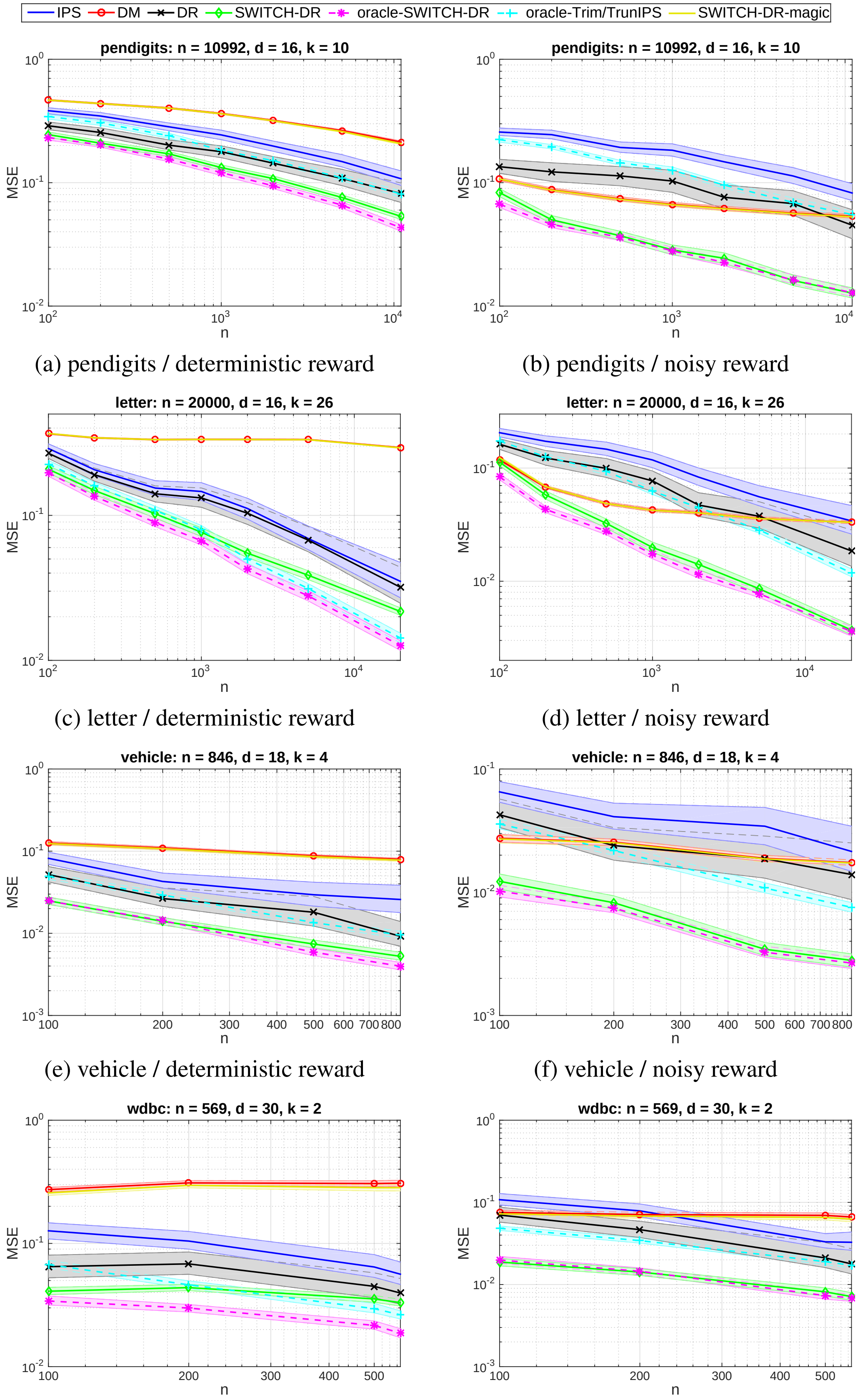
(d) glass / noisy reward

(e) page-blocks / deterministic reward

(f) page-blocks / noisy reward

(g) satimage / deterministic reward

(h) satimage / noisy reward



(a) pendigits / deterministic reward

(b) pendigits / noisy reward

(c) letter / deterministic reward

(d) letter / noisy reward

(e) vehicle / deterministic reward

(f) vehicle / noisy reward

(g) wdbc / deterministic reward

(h) wdbc / noisy reward

9.9 Relationship to causal effect estimation and its asymptotic optimality theory in the econometrics

Off-policy evaluation in the bandit literature and mean-effect estimation causal inference are mathematically equivalent problems but are studied under vastly different assumptions due to the cultural differences in the two communities. In this section, we connect the two problems and illustrate how our results fit into the existing literature on causal effect estimation.

We start by explaining the differences in the notations and how the two problems are equivalent. In causal inference literature, the "action" a is denoted as T for "treatment". The "treatment" is often binary, i.e., $T \in \{0, 1\}$. The "context" x is called "covariates" in the regression sense. The base policy $\mu(a = 1|x)$ is therefore the treatment probability $p(T = 1|X)$ (or just $p(X)$) for the patient with covariate X. The reward random variable $r(x, a)$ is simply the response/outcome variable Y in the "average treatment effect on the treated" (ATT) problem and an affine transformation of the response variable Y in the "average treatment effect" (ATE) problem. Specifically, the value of a policy $v^\pi = \mathbb{E}_\pi r(x, a)$ is the same as ATT when taking $\pi = [1, 0]$, $r = Y$; and it is the same as ATE when taking $\pi = [0.5, 0.5]$ and $r(x, a = 1) = 2Y(x, a = 1)$ and $r(x, a = 0) = -2Y(x, a = 0)$.

Therefore our results for estimating policy values can be directly applied to the causal inference problems of estimating ATE and ATT. We restrict our discussion to ATE estimation but everything can be trivially stated for ATT estimation.

There are important differences too. In off-policy evaluation, $\mu(\cdot|x)$ is often known, but in causal inference, $\mu(\cdot|x)$ needs to be estimated from data unless it is a controlled experiment. More critically, in off-policy evaluation, we do not make assumptions on what $\mu(\cdot|x)$ and $\mathbb{E}(r|x, a)$ could be, while realizable assumptions on these quantities are often needed in the econometrics literature.

In the remainder of the section, we describe the asymptotic optimality theory, in particular, the "semi-parametric efficiency" lower bound and the kind of estimators that matches the lower bound, and then highlight the gap to our minimax lower bound in Theorem 9.2. Our key conclusions are that

1. the asymptotic efficiency bound implies a strictly smaller rate than our lower bound;
2. all known estimators that achieves the asymptotic efficiency use an assumption that the expected reward function is within a realizable class of functions that is sufficiently small;
3. and our lower bound implies that the rate of asymptotic efficiency *cannot* be achieved in general (without such realizable assumptions).

In addition, we will address the following two related questions:

1. Is knowing propensities helpful in ATE estimation?
2. What are the assumptions hidden in estimators achieving asymptotic efficiency? In particular, do they suffer from the curse of dimensionality in finite sample?

Hopefully, the discussion will shed a light on how our bound expands the understanding of the causal inference problem as a byproduct.

9.9.1 Asymptotic efficiency, efficient estimators and realizability assumption

We start by describing the asymptotic (semi-parametric) efficiency lower bound due to Hahn [107].

Theorem 9.11 (Theorem 1 of [107] in our notation)**.** *No regular (local asymptotically normal) estimator sequence $\hat{v}_n$ of ATE has a (scaled) asymptotic variance $\lim_{n\to\infty}\mathrm{Var}[\sqrt{n}\hat{v}_n]$ that is smaller than*

$$\mathbb{E}_{x\sim\mathcal{D}}\left\{\frac{\mathrm{Var}(r|x,a=1)}{4\mu(a=1|x)}+\frac{\mathrm{Var}(r|x,a=0)}{4(1-\mu(a=1|x))}\right.$$
$$\left.+\left(\frac{\mathbb{E}(r|x,a=1)-\mathbb{E}(r|x,a=0)}{2}-\frac{\mathbb{E}(r|a=1)-\mathbb{E}(r|a=0)}{2}\right)^2\right\}.$$

Using the importance weight $\rho(x,a)=0.5/\mu(x,a)$, the above bound can be rewritten as

$$\mathbb{E}_{x\sim\mathcal{D}}\left\{\mathbb{E}_\mu[\rho^2\mathrm{Var}(r|x,a)|x]\right\}+\mathrm{Var}_{x\sim\mathcal{D}}\left\{\mathbb{E}_\mu[\rho r|x]\right\}. \tag{9.30}$$

The lower bound can be obtained asymptotically by various estimators under different regularity assumptions. These are mostly plug-in estimators of form

$$\hat{v}_{\text{Hahn}}=\frac{1}{n}\sum_{i=1}^n\frac{\hat{\mathbb{E}}[r\mathbf{I}(a=1)|x_i,a=1]}{\hat{\mu}(a=1|x_i)}-\frac{\hat{\mathbb{E}}[r\mathbf{I}(a=0)|x_i,a=1]}{1-\hat{\mu}(a=1|x_i)}.$$

$$\hat{v}_{\text{Hirano}}=\frac{1}{n}\sum_{i=1}^n\frac{r_i\mathbf{I}(a=1)}{\hat{\mu}(a=1|x_i)}-\frac{r_i\mathbf{I}(a=0)}{1-\hat{\mu}(a=1|x_i)}.$$

$$\hat{v}_{\text{Imbens}}=\frac{1}{n}\sum_{i=1}^n\hat{r}(x_i,a=1)-\hat{r}(x_i,a=0).$$

where $\hat{\cdot}$ is used to denote (nonparametric) estimate of these quantities. Note that the $\hat{v}_{\text{Imbens}}$ is simply the direct method, and $\hat{v}_{\text{Hirano}}$ is IPS but with estimated propensity scores. What's common about these estimators is that they all require assumptions on these functions of interest so $\hat{\cdot}$ would be consistent with a sufficiently fast rate, which often requires stronger smoothness assumption as the dimensionality of x gets larger. We describe the flavor of this type of results only for $\hat{v}_{\text{Hirano}}$ below and refer interested readers to a recent paper with a survey of these estimators [181] and their drawbacks.

Theorem 9.12 (Theorem 1 of [112])**.** *Assume ignorabiltiy, a.k.a, unconfoundedness of treatment assignment — $a\perp[r(x,a=1),r(x,a=0)]|x$ — and the following regularity conditions.*

(i) the support of x is $[0,1]^d$

(ii) *the density of x is bounded and bounded away from* 0.

(iii) $\mathbb{E}(r^2|a=1) < \infty$, $\mathbb{E}(r^2|a=0) < \infty$

(iv) $\mathbb{E}[r|x, a=0]$ *and* $\mathbb{E}[r|x, a=1]$ *are continuously differentiable for all* x.

(v) $\mu(a=1|x)$ *as a function of* x *is continuously differentiable of all order* $s \geq 7d$.

(vi) $\mu(a=1|x)$ *is bounded away from* 0 *and* 1 *for all* x.

Then there is an estimator $\hat{\mu}(a=1|x)$ *that estimates* $\mu(a=1|x)$ *for every* $x = x_1, ..., x_n$ *under which,* $\hat{v}_{\mathit{Hirano}}$ *attains the lower bound* (9.30) *at the limit of* $n \to \infty$.

We note that the assumptions, especially (v), are very strong and they hide the exponential dimension dependence by making stronger and stronger smoothness assumptions as dimension gets larger. Also, we note that (iv) is a realizable assumption that constrains $\mathbb{E}[r|x, a=0]$ and $\mathbb{E}[r|x, a=1]$ to be Hölder class functions.

One interesting, but curious observation by Hahn [107] is that knowing the true propensity scores (primarily $\mu(a|x)$, as $\pi(a|x)$ is known in the mean-effect estimation) is not useful in efficient estimation. In fact, it could even be harmful. It's shown in Hahn [107] and Hirano et al. [112] that using the true propensities in $\hat{v}_{\text{Hahn}}$ or $\hat{v}_{\text{Hirano}}$ (that would simply be IPS!) will lead to an estimator that is no longer asymptotically efficient.

Remark 9.13 (k-arm bandit setting)**.** *It is useful to point out another example of this seemingly magical fact that using a noisy estimate is asymptotically better than using the exact quantity. In Li et al. [146], it is shown that the DM estimator can in fact be rewritten into IPS with estimated propensities, therefore in this simple setting at least,* $\hat{v}_{\mathit{Hahn}}$ *and* $\hat{v}_{\mathit{Hirano}}$ *are equivalent.*

The k-arm bandit example suggests that the regularity assumptions needed for nonparametric estimation might be just one example of assuming a realizable class of functions such that we can learn $\mathbb{E}(r|x, a)$ as n gets large. Our Theorem 9.2 clearly indicates that when we do not make such assumption, it is information-theoretically impossible to achieve the semiparametric efficiency.

To reinforce this point of view, we now compare (9.30) to the error bound of IPS, to our lower bound in Corollary 9.3 and to the upper bound achieved by the doubly robust estimator with oracle $\mathbb{E}(r|x, a)$.

Remark 9.14 (Comparison to IPS upper bound and Corollary 9.3.)**.** *First of all,* (9.30) *is strictly smaller than the variance of IPS (therefore also smaller than our lower bound in Corollary 9.3, which matches IPS variance up to a universal constant for the hardest problem in the class when* x *is drawn from a density). Recall that the variance of IPS can be decomposed into*

$$\mathbb{E}_\mu[\rho^2 \mathrm{Var}(r|x, a)] + \mathrm{Var}_\mu\left[\rho \mathbb{E}(r|x, a)\right].$$

The first term matches exactly with the first term in (9.30)*, while the second term is strictly bigger than the second term in* (9.30)*, because*

$$\mathrm{Var}_{(x,a)\sim\mu}[\rho \mathbb{E}(r|x, a)] = \mathrm{Var}_{x\sim\mathcal{D}} \mathbb{E}_\mu(\rho r|x) + \mathbb{E}_{a\sim\mu} \mathrm{Var}_{x\sim\mathcal{D}}\left[\mathbb{E}(\rho r|x, a)|a\right] \geq \mathrm{Var}_{x\sim\mathcal{D}} \mathbb{E}_\mu(\rho r|x).$$

They differ by a positive term $\mathbb{E}_{a\sim\mu}\mathrm{Var}_{x\sim\mathcal{D}}\left[\mathbb{E}_\mu(\rho^2 r|x,a)|a\right]$ *which is not in* (9.30)*, and this term cannot be bounded by a constant multiple of either* $\mathrm{Var}_{x\sim\mathcal{D}}\mathbb{E}_\mu(\rho r|x)$ *or* $\mathbb{E}_\mu[\rho^2\mathrm{Var}(r|x,a)]$.
Remark 9.15 (Comparison to our lower bound.)**.** *Assume the context distribution is a probability density, our minimax lower bound is different from* (9.30) *in several ways. Ours is a finite sample bound and applies to all estimators and to every n (not just LAN estimators). Because we do not restrict the class of estimators, the bound is necessarily a maximum over a class of problems rather than a per-instance variance lower bound like* (9.30)*. In order to make them comparable, we ignore the constant factor, take supremum of* (9.30) *over the class* $\mathcal{R}(\sigma, R_{\max})$*, which gives us*

$$\mathbb{E}_\mu[\rho^2\sigma^2] + \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_\mu[\rho R_{\max}|x]^2\right].$$

The second term is smaller than our bound in Corollary 9.3 by a Jensen's inequality.
Remark 9.16 (Comparison to Oracle DR.)**.** *By Lemma 3.1(i) of Dudík et al. [75], DR with a perfect oracle has variance*

$$\mathrm{MSE}(\hat{v}_{\mathrm{Oracle-DR}}) = \frac{1}{n}\mathbb{E}_\mu\rho^2\mathrm{Var}(r|x,a)^2 + \frac{1}{n}\mathrm{Var}\left[\sum_{a\in\mathcal{A}}\pi(a|x)\mathbb{E}(r|x,a)\right].$$

This exactly matches the Cramer-Rao lower bound (9.30)*, although it uses oracle information. In other word, the asymptotic efficiency [107] is effectively a measure of how well an estimator performs relative to the oracle doubly-robust estimator.*

These observations suggest that the realizability assumption is critical. On the one extreme, when the regression function is known or can be accurately estimated from the data, then it is possible to achieve a faster rate than the minimax lower bound. One the other extreme, when the regression function is unknown and we do not assume it can be accurately estimated, then the IPS approach that makes no assumptions on the regression functions become optimal.

Specifically, our results imply that Assumption (iv) in Theorem 9.12 cannot be removed, otherwise no estimator could match the rate in the bound of (9.30), let alone achieving asymptotic efficiency, because otherwise it would violate our lower bound.

9.9.2 Curse of dimensionality and the benefits of knowing the propensity scores

Recall that the aforementioned estimators all make use of nonparametric estimators of either $\mathbb{E}(r|x, a = 0$ or $1)$ or $\mu(a = 1|x)$ (or both) as subroutines. Therefore, in order to achieve asymptotic efficiency, the rate to estimate these quantities need to be sufficiently fast (typically requiring the sup-norm convergence rate being $o_p(n^{-1/4})$). This requires stronger smoothness assumptions of $\mathbb{E}(r|x, a = 0$ or $1)$ or $\mu(a = 1|x)$ (or both) in order to achieve asymptotic efficiency and a much larger number of samples before the asymptotics become a reasonable assumption. If $x \in [0, 1]^d$ and $\mathbb{E}[r|x, a]$ is (α)th differentiable in argument x for both actions, then the standard minimax rate of estimating this function in sup-norm is on the order of $n^{-\frac{\alpha}{2\alpha+d}}$. In order word, the $n^{-1/4}$ rate requires that $\alpha > d/2$ and as d gets larger the smoothness assumption gets more restrictive. We encourage readers to checkout the detailed discussion in Rothe [181].

Rothe [181] also considered a setting more related to us where the propensities $\mu(a|x)$ are known and showed that a doubly robust estimator using an oversmoothed local linear regression estimator of $\mathbb{E}[r|x, a]$ is able to achieve asymptotic efficiency without requiring $\mathbb{E}(r|x, a = 0$ or $1)$ to be smoother as dimension increases. We translate the result into our notations as follows.

Theorem 9.17 (Theorem 2 in [181] with $l = 0$)**.** *Assume ignorabiltiy $a \perp [r(x, a = 1), r(x, a = 0)]|x$ and the regularity conditions (i)(ii)(iv)(vi) of Theorem 9.12, replace (iii) and (v) with*

(iii) $\mathbb{E}(|r|^{2+\delta}|x, a = 1) < \infty$, $\mathbb{E}(|r|^{2+\delta}|x, a = 0) < \infty$ *for a constant $\delta > 0$ for all x.*

(v) $\mu(a|x)$ *is known for all x and $a \in \{0, 1\}$.*

(Note that (iii) is slightly stronger and (v) gives more power to the estimator but does not require any regularity assumptions on μ.) Then there exists an estimator $\hat{v}$ such that

$$\sqrt{n}(\hat{v} - \hat{v}_{\mathrm{Oracle-DR}}) = \tilde{O}_{\mathbb{P}}(n^{-\min\{\frac{2}{3d}, \frac{1}{2+d}\}}).$$

In conclusion, Hahn [107]'s observation that it is not useful to know the propensity scores is not technically true unless we make higher order smoothness assumption with order linear in dimension d. Knowing the propensity scores allow Rothe [181] to attain asymptotic efficiency (9.30) under only differentiability assumption on expected reward function. The result hence applies to all constant d.

However, the finite sample performance of the proposed estimator remains to suffer from"curse-of-dimensionality" in the supposedly lower order terms. By the results in Bertin et al. [29], the exact constant of the minimax risk in sup-norm for the class of functions with derivatives bounded by L is

$$\sigma^{\frac{2}{2+d}} L^{\frac{d}{2+d}} \left[\frac{(d-1)!(d^2 + d^3)}{2^d}\right]^{\frac{2}{2+d}}.$$

This suggests that the asymptotic approximation is only meaningful when n gets as large as $d!$, which can quickly become unrealistic beyond $d < 10$ in practice.

Chapter 10

Optimal Gaussian adaptive data analysis

In modern data analysis, data analysts often have many rounds of interaction with a data set before deciding on a model to fit, or have a list of scientific questions to answer with the data. The process is often sequential and adaptive, and can lead to significant bias in the subsequent inferences. In this chapter, we propose a game-theoretic minimax framework to formally study this problem under a model called "sequential selective estimation", where in each round the data analyst estimates a parameter that is chosen based on the revealed estimates in previous rounds. Assuming Gaussianity of data, we establish the first sharp minimax lower bound on the squared error in the order of $O(\sqrt{k}\sigma^2)$, where k is the number of sequentially chosen quantities to estimate, and σ^2 is the ordinary signal-to-noise ratio for a single parameter estimate (often on the order of $1/n$ in the i.i.d. setting). Our lower bound is based on the construction of an approximately least favorable adversary who picks a sequence of parameters that are most likely to be affected by overfitting. The key technical component of the lower bound proof is a reduction to finding the convoluting distribution that optimally obfuscates the sign of a Gaussian signal. Our lower bound construction also reveals a transparent and elementary proof of the matching upper bound as an alternative approach to [183], who used information-theoretic tools to provide the same upper bound. We believe that the proposed framework opens up opportunities to obtain insights for many other settings of adaptive data analysis, which would extend the idea to more practical realms.

10.1 Introduction

In most traditional statistical data analysis, the validity of inference requires the inference procedure (for example, the null hypothesis and test statistic to be considered) to be specified before looking at the data. In modern scientific and engineering research with large-scale data and complex hypotheses, it is more natural to choose models and inference tasks in a sequential and data-driven manner. For example, one may want to fit a second model to the data after seeing that the first model did not fit well; or to test significance of the variables chosen by a variable selection procedure. If traditional frequentist inference procedures are applied to these adaptively chosen tasks, the validity are often questionable due to overfitting or what is known as "Researcher Degree of Freedom" [49, 138, 200].

In this chapter, we consider the problem of "sequential selective estimation", which is designed to appropriately quantify the amount of additional error due to the selection procedure. It is obvious that this additional error is determined by how the parameters to be estimated are chosen. For example, if the parameters to be estimated are determined before seeing the data, then there is no additional error. On the other hand, as we well see in this chapter, there exists selection methods that can incur substantial bias in the parameter estimate if traditional methods are used. It is impractical to know what kind of selection criterion a data analyst will use, so we consider a worst-case scenario and study the problem from a minimax perspective:

What is the worst-case estimation error among all possible selection methods?

Formally, let $\{\theta_t : t \in T\}$ be a collection of parameters, and X the observed data, given loss function $L(\cdot,\cdot)$, we would like to find out the following minimax risk of selective sequential estimation.

$$\min_{\hat{\theta}} \max_{\mathcal{D}} \max_{T_1,...,T_k} \max_{i=1:k} \mathbb{E}\left[L(\hat{\theta}_{T_i}, \theta_{T_i})\right] ,$$

where the minimization is taken over all possible estimators $\hat{\theta}$, $T_1, ..., T_k$ is a sequence of adaptively chosen parameter indices from an index set $\mathcal{T}$, and $\mathcal{D}$ is the distribution of X whose unknown parameter includes θ. Here the expectation takes into account both the randomness of X and $T_1, ..., T_k$. Our results can be extended to any fixed distribution $\mathcal{D}$, but with some mild restrictions to the estimators.

We focus on the joint Gaussian model, in which we assume that under the distribution of the data set:

$$\theta_t = \mathbb{E}(\phi_t(X)) , \ \forall\, t \in \mathcal{T}, \quad \text{and} \ \ (\phi_t(X) : t \in \mathcal{T}) \ \text{ is a Gaussian process} .$$

Example 10.1 (Sequential estimation of projections)**.** *Let $X \in \mathbb{R}^d$, drawn from a multivariate normal distribution. Moreover, assume each column of X is appropriately normalized so that the marginal variances are equal to σ^2/n. Here*

$$\mathcal{T} = \left\{t \in \mathbb{R}^d : \|t\|_2 \leq 1\right\}$$

is the class of all unit vectors and $\phi_t(X) = \langle t, X\rangle$. It is clear that for any $t \in \mathcal{T}$, $\mathrm{Var}(\phi_t(X)) \leq \frac{\sigma^2}{n}$. Also, for any fixed finite subset of $\{t_1, ..., t_k\} \subset \mathcal{T}$, $\phi_{t_1,...,t_k}(X)$ is a multivariate normal distribution.

Example 10.2 (Sequential model selection for linear regression)**.** *Let $X \in \mathbb{R}^{n\times d}$ be a fixed design matrix, response vector $Y \sim \mathcal{N}(\mu, \sigma^2 I)$. Choose any feature subset $t \subset [d]$ and fit a linear regression model, and then*

$$\hat{\theta}_t = \phi_t(X, Y) = (X_t^T X_t)^{-1} X_t^T Y.$$

Example 10.3 (Hyper-parameter tuning with Bayesian optimization)**.** *Let $t \in \mathcal{T}$ be an index to a tuple of d hyperparameters, and $\mathcal{T}$ could be $[0, 1]^d$. Let X be a hold-out dataset. The objective of* *hyperparameter tuning is to $t \in \mathcal{T}$ such that $\mathbb{E}R(h(t), X)$ is minimized for some risk functional R, where the expectation is taken over X and the randomness in learning a model h on a fixed training data set. We can only observe a random variable $\hat{R}(h(t), X)$ with mean $\mathbb{E}R(h(t), X)$. These correspond to ϕ_t and μ_t in our model. In Bayesian optimization, the indices $T_1, T_2, ..., T_k$ are chosen sequentially, where for each i, T_i is either a deterministic or randomized function of $(T_1, ..., T_{i-1}, \hat{R}(h(T_1), X), ..., \hat{R}(h(T_{i-1}), X))$. It is often assumed that the black box function $\hat{R}(h(t), X)$ of t is a Gaussian process. This exactly matches our assumption.*

In our sequential selective estimation framework, natural plug-in estimators such as $\hat{\theta}_t = \phi_t(X)$ will work poorly because each round reveals too much information about the realized data set too quickly, and there may exist a new $t \in \mathcal{T}$ such that $\phi_t(X)$ significantly deviates from its mean. In order to avoid such issues, many estimators for this problem often involve additional post-randomization of the plug-in estimator [see e.g. 18, 85, 87, 183]. The main idea is that if the parameters are estimated in a way such that they provide little information about the details of the dataset, it is unlikely for the subsequent selections to overfit. A good example is to make the estimators differentially private [85, 87]. These approaches work directly with information-theoretic quantities, so they are applicable to any selection procedure so long as it is fed with only sufficiently perturbed releases. These methods tend to be overly conservative, and it is unclear whether the large random noises added for strong data privacy protection is also necessary for the conceptually easier sequential selective estimation problem. Also, despite some study in lower bounds [109, 211] in related settings, it is not well understood whether the existing estimators are optimal.

Summary of results. In this chapter, we prove the following (informally):

Theorem 10.4 (Main result I (a minimax lower bound))**.** *There is no estimator $\hat{\phi}$ that can estimate the mean of any k sequentially chosen parameters uniformly better than $\Omega(\sqrt{k}\sigma^2)$ for all distributions under the jointly-Gaussian model when $|\mathcal{T}| = \Omega(2^k)$.*

Theorem 10.5 (Main result II (a per-instance lower bound))**.** *There is no estimator $\hat{\phi}$ that is "natural" (which includes all noise adding estimators), that can estimate the mean of any k sequentially chosen parameters uniformly better than $\Omega(\sqrt{k}\sigma^2)$ for any fixed distribution, provided that $\phi_{\mathcal{T}}$ is sufficiently "rich".*

We did not use standard information-theory tools to obtain these results. For the best of our knowledge, there are no standard information-theoretical proof techniques directly applicable to our problem. Instead, we resort to a novel constructive proof, which reduces the problem to finding an optimal noise-adding procedure against a sequence of optimal binary classifiers.

These lower bounds are achieved (up to a constant factor) by the Gaussian noise-adding estimator due to [183] of the following form:

$$\hat{\theta}_{T_i} = \phi_{T_i}(X) + k^{1/4}\sigma Z_i, \text{ for } i = 1, ..., k,$$

where $Z_i \sim \mathcal{N}(0, 1)$, independent of everything else. This indicates that Gaussian noise adding cannot be significantly improved.

We also provide an alternative (and much more transparent) analysis for the Gaussian model that does not use mutual information as in Russo and Zou [183]. This leads to an improved constant in the upper bound.

The results suggest that, when the set of potential parameters is rich enough and the plug-in estimators are Gaussian, independent Gaussian noise adding is minimax optimal up to a constant. We show that for k-step adaptive data analysis, the smallest worst-case amplification factor of the squared estimation error that can be achieved by any (possibly adaptive) releasing procedures is $\sqrt{k}$. Here term "worst-case" refers to any possible adaptive parameter selection mechanism. Our motivation here has been driven substantially by work on relaxations of the method of differential privacy, and its primary mechanism of protection through additive noise. We return to that link at the end of Section 10.2.

Related work The problem of estimating sequentially selected parameters has been studied under the name "adaptive data analysis" or "preventing false recoveries" in the computer science community [See, e.g., 18, 85, 87, 109, 183, 211]. The most commonly used setting assumes the selection mechanism can adaptively choose any low-sensitivity query (i.e. plug-in estimate). For k-step adaptive data analysis, Dwork et al. [87] produces the first sample complexity upper bound for Laplace noise adding in the order of $\tilde{O}(\sqrt{k}/\epsilon^{2.5})$, where ϵ is the target error level, defined as the largest absolute error over all k steps. Here the sample complexity is the number of iid data points needed for k-step adaptive data analysis to achieve a target error level with high probability. Bassily et al. [18] improves the bound to $\tilde{O}(\sqrt{k}/\epsilon^{2})$ and extends to approximate differential privacy, as well as convex optimization queries. The factor $\sqrt{k}$ is shown to be optimal [109, 211] for polynomial time algorithms or any algorithms if the dimension of $\mathcal{X}$ is sufficiently large, but the optimal dependence on ϵ remains open.

Our results apply to a different setting studied in Russo and Zou [183, Proposition 9], where the plug-in estimates are assumed to be jointly Gaussian. This is neither stronger nor weaker than the requirement of low-sensitivity as shown in Figure 10.1. We also define the risk differently as the maximum expected squared error. Due to these differences, our bounds are only loosely comparable to those in Bassily et al. [18], Dwork et al. [87], Steinke and Ullman [211] in terms of the maximum expected absolute error — a middle ground that both our bounds and theirs imply. In particular, our upper bound is on the same order as Bassily et al. [18] and Russo and Zou [183] with a constant improvement over Russo and Zou [183] due to a more direct proof.

Our lower bound (when taking $\sigma^2 = 1/n$) becomes $\Omega(k^{1/4}/n^{1/2})$, which is substantially larger than the best available lower bound in Steinke and Ullman [211] that translates into $\Omega(\min\{k^{1/2}/n, 1\})$. However, this is not a fair comparison as the settings are not quite the same. We will clarify this point after we present our results and discuss implications of our work in the setting of Steinke and Ullman [211] and vice versa in Section 10.4. Both our lower bound and that in Steinke and Ullman [211] apply to the more general jointly-subgaussian setting, but it remains an open problem to find an algorithm that matches the lower bound in this more general regime.

When additional assumptions are made, e.g., when we assume $\mathcal{X}$ is finite and low-dimensional, then one can improve the dependence on k exponentially with a computationally inefficient

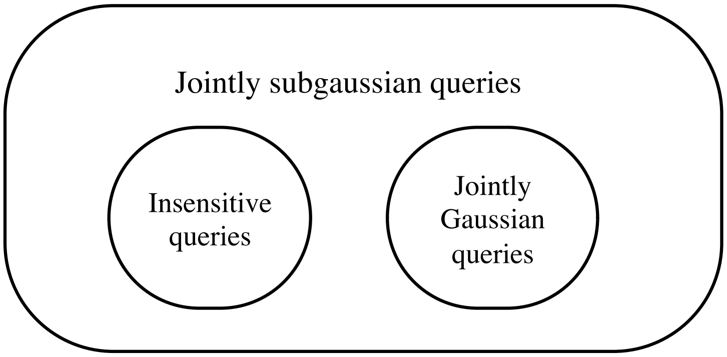


Figure 10.1: Relationship of the class of queries.

algorithm [18, 87].

The problem of valid inference for data-dependent tasks has been studied through a different perspective in the statistics community. Fithian et al. [95], Lockhart et al. [149], Taylor and Tibshirani [215], Taylor et al. [216] and others developed a series of "selective inference" methods that work with specific variable selection tools (e.g., Lasso) and adjust the confidence intervals or p-values accordingly based on the selections such that they have the exact or asymptotically correct frequentist coverage. There are several major differences between this framework and the adaptive data analysis framework. First, selective inference essentially considers two-step problems, where the variables are selected in the first step, and their significance are tested in the second step. Second, these methods are *passive observers* in that they release the variable selection result without randomization, but only adjust the inference in the second step to correct the selection bias. In other word, the goal of selective inference is to produce valid (but potentially very large) confidence intervals for even highly biased statistical quantities, whereas the goal of adaptive data analysis is to prevent "bad" statistical inference tasks from being selected in the first place.

10.2 A minimax framework

In this section we formulate the minimax problem of sequential selective estimation in a game-theoretic framework. The game-theoretic approach we take is to cover all possible ways to choose the parameters to estimate in the sequence. The game goes as follows:

1. **Basic context.** Both players, the data analyst (i.e. the player) and the adversary, are given a collection of parameters $\{\theta_t : t \in \mathcal{T}\}$, with each $\theta_t \in \Theta_t \subseteq \mathbb{R}$.
2. **The game procedure.** The game consists of k steps. The ith step works as follows ($1 \le i \le k$):
    (a) The adversary announces a $T_i \in \mathcal{T}$, which can be a possibly randomized function of the released estimates up to time $i-1$.

(b) After seeing T_i, the player releases an estimate of θ_{T_i}, denoted by A_i, which can be a possibly randomized function of the data X, and the historical parameter indices $T_1, ..., T_{i-1}$.

3. **Assumptions and remarks.**

(a) The data X is drawn such that $\mathbb{E}\phi_t(X) = \theta_t$ for all $t \in \mathcal{T}$. The functions $\phi_t : \mathcal{X} \mapsto \Theta_t$ are known to both sides of the game.

(b) The joint distribution of $\{\phi_t(X) : t \in \mathcal{T}\}$ belongs to $\mathbb{D}$, a family of distributions on $\prod_{t\in\mathcal{T}} \Theta_t$ that is known to both sides.

(c) The number of steps, k, in the sequential selective estimation procedure is known to both sides.

(d) The player designs a k-step strategy $(\mathcal{A}_i : 1 \le i \le k)$ where $\mathcal{A}_i : (X, T_1, ..., T_i, Z_i) \mapsto A_i \in \Theta_{T_i}$ where Z_i represents a fresh random variable.

(e) The adversary chooses parameters to estimate using a k-step strategy $(\mathcal{W}_i : 1 \le i \le k)$ where $\mathcal{W}_i : (T_1, ..., T_{i-1}, A_1, ..., A_{i-1}, R_i) \mapsto T_i \in \mathcal{T}$, where R_i is another fresh random variable used in $\mathcal{W}_i$. Note that $\mathcal{W}_i$ may depend on any deterministic information, such as the distribution of $\{\phi_t(X) : t \in \mathcal{T}\}$. But $\mathcal{W}$ does not have access to the data X.

4. **Loss function.**

$$\max_{1\le i\le k} \mathbb{E}(A_i - \theta_{T_i})^2 .$$

The player wants to minimize the loss function while the adversary want to maximize it.

The problem of interest is to identify the optimal "player" strategy (estimators) that will ensure that if the game is repeatedly played, the expected loss is minimized. Specifically, we consider the case when $\diamond$ is the "max" operator and the loss function is the standard mean square loss. The corresponding minimax risk is

$$\mathcal{R}(k, \mathbb{A}_{1:k}, \mathbb{W}_{1:k}, \mathbb{D}) := \inf_{\mathcal{A}_{1:k}\in\mathbb{A}_{1:k}} \sup_{\mathcal{D}\in\mathbb{D}} \sup_{\mathcal{W}_{1:k}\in\mathbb{W}_{1:k}} \max_{i\in[k]} \mathbb{E}(A_i - \mu_{T_i})^2.$$

We consider $\mathbb{D} = \mathbb{D}(\theta, \mathcal{T}, \sigma^2)$ to be the collection of all Gaussian processes indexed by $\mathcal{T}$ with mean $(\theta_t : t \in \mathcal{T})$ and bounded marginal variance:

$$\mathrm{Var}(\phi_t) \le \sigma^2 , \ \forall\, t \in \mathcal{T} .$$

We would like to point out that this is a challenging partial information game where the adversary knows $\mathcal{D}$, the distribution of $\phi_{\mathcal{T}}$ but does not know the specific data set drawn from that distribution. The player does not know $\mathcal{D}$ (and in particular, the mean of $\phi_{\mathcal{T}}$, which are what he would like to estimate in the first place) and his strategy is expected to work for all $\mathcal{D} \in \mathbb{D}$ and all adversarial strategy $\mathcal{W}_{1:k}$. This quantification of the amount of information available to both the "player" and the "adversary" is important because

(a) if the “player” knows the distribution, choosing $A_i = \mu_{T_i}$ independent to the data results in a minimax risk of 0,

(b) if the “adversary” knows the data and the player does not know the distribution, then the problem becomes the standard uniform convergence problem of characterizing

$$\inf_{\mathcal{A}_1} \sup_{\mathcal{D}\in\mathbb{D}} \mathbb{E}\left[\sup_{t\in\mathcal{T}} \mathbb{E}[(A_1 - \mu_t)^2 | \phi_{\mathcal{T}}]\right]$$

because even before the game start, $\mathcal{W}_1$ would have already chosen the worst t for what $\mathcal{A}_1$ to output in this specific realization. This is the central problem in empirical process theory and statistical learning theory, and has been studied extensively for many years. However, if $\mathcal{T}$ has unbounded “entropy”, the minimax risk could be ∞.

The main objective of this chapter is to understand what the minimax rate (minimax risk up to a universal constant) for interesting and practically relevant choices of $\mathbb{A}_{1:k}$, $\mathbb{W}_{1:k}$ and $\mathbb{D}$. In most of our results, both $\mathbb{A}_{1:k}$ and $\mathcal{W}_{1:k}$ are defined in an information-theoretical fashion because we only restrict them to be functions of everything they can see at a certain point of the game. We also consider smaller classes of $\mathbb{A}_{1:k}$ when we aim to get a much stronger per-instance lower bound where $\mathcal{D}$ is fixed and known to both sides of the game. In this sense, such restrictions are necessary.

Lastly, we acknowledge that this minimax framework has been studied by Steinke and Ullman [211] under the name “adaptive data analysis”, where both computational and information-theoretic lower bounds were established when $\mathcal{T}$ is the class of all statistical queries and $\Phi_{\mathcal{T}}$ is generated by an arbitrary iid distribution defined on $\{0,1\}^d$. For the best of our knowledge, this current chapter is the first concrete treatment of the problem when $\Phi_{\mathcal{T}}$ is jointly Gaussian (or a Gaussian process when $|\mathcal{T}| = \infty$), which we believe is instrumental in understanding the implication of adaptive data analysis in statistical theory and practice.

10.3 Results

In this section, we will work out the upper and lower bounds for the minimax risk.

10.3.1 An explicit adversary in the Gaussian model

Before we derive the minimax lower bound, we first provide an alternative proof to the mutual information-based argument of Russo and Zou [183] using only elementary arguments. This allows us to understand intuitively what an “adversary” would do, knowing that the “player” will just be adding Gaussian noise.

We first present results for a simpler version of the adaptive data analysis that has only one step adaptivity, where the adversary chooses the queries by explicitly maximizing the selection bias in the form of a conditional expectation. Building upon this result, we extend the argument to form an explicit upper bound for the minimax risk in the k-step setting. The proof provides intuition for constructing the minimax lower bound presented in Section 10.3.2.

1-step adaptivity under Gaussian additive noise

Our first result applies to the case when $(T_1, ..., T_{k-1})$ take an arbitrary fixed vector $(t_1, t_2, ..., t_{k-1})$. For each $i = 1, ..., k$, we will choose release protocol $\mathcal{A}_i$ to be such that $A_i = \phi_{t_i}(X) + Z_i$ where $Z_i \sim \mathcal{N}(0, w^2)$ is a freshly drawn normal random variable.

After observing the realized values of $A_1, ..., A_{k-1}$, the adversary samples T_k from a distribution $\mathcal{P}$ on $\mathcal{T}$, with $\mathcal{P}$ depending on $(T_{1:k-1}, A_{1:k-1})$. Again, we emphasize that T_k belongs to the class

$$\mathcal{T} = \{t : |\phi_t(X) \sim \mathcal{N}(\mu_t, \sigma_t^2), \sigma_t^2 \leq \sigma^2\}.$$

The key idea is that the choice of T_k boils down to choosing a covariance vector $\Sigma_{k,1:k-1}$ with previously selected $\phi_t, ..., \phi_{t_{k-1}}$. The following result constructs the least favorable choice of T_k.

We use $\Sigma_{1:i}$ to denote the covariance matrix of $\phi_{t_{1:i}}$, and $\Sigma_{j,1:i}$ the covariance vector between ϕ_{t_j} and $\phi_{t_{1:i}}$.

Theorem 10.6 (1-step adaptivity). *Let $t_1, ..., t_{k-1} \in \mathcal{T}$. Moreover, let observation noise $Z_i \sim \mathcal{N}(0, w^2)$ and T_k generated by any adaptive selection protocol. Then the squared bias*

$$\sup_{t_1,...,t_{k-1} \in \mathcal{T}} [\mathbb{E}(\phi_{T_k} - \mu_{T_k})]^2 \leq \frac{(k-1)\sigma^4}{w^2}.$$

Moreover, if $w^2 = \sqrt{k-1}\sigma^2$, the square error of the estimate

$$\mathbb{E}(A_k - \mu_{T_k})^2 \leq (2\sqrt{k-1} + 1)\sigma^2.$$

Proof sketch. The proof relies on the law of total expectation that expands the bias into

$$\begin{aligned}\mathbb{E}(\phi_{T_k} - \mu_{T_k}) &= \mathbb{E}_{A_{1:k-1}} \mathbb{E}_{T_k | A_{1:k-1}} \mathbb{E}_{\phi_{T_k} | T_k, A_{1:k-1}} (\phi_{T_k} - \mu_{T_k}) \\ &\leq \mathbb{E}_{A_{1:k-1}} \sup_{t_k \in \mathcal{T}} \mathbb{E}_{\phi_{t_k} | A_{1:k-1}} (\phi_{t_k} - \mu_{t_k}).\end{aligned}$$

Since $A_{1:k-1} = \phi_{t_{1:k-1}} + Z_{1:k-1}$ and ϕ_{t_k} are jointly normal, we can explicitly write down the conditional expectation

$$\mathbb{E}(\phi_{t_k} - \mu_{t_k} \mid A_{1:k-1}) = \Sigma_{k,1:k-1}^T (\Sigma_{1:k-1} + w^2 I_{k-1})^{-1} (\phi_{t_{1:k-1}} + Z_{1:k-1} - \mu_{t_{1:k-1}}).$$

Finding the supremum of $t_k \in \mathcal{T}$ reduces to finding the maximum over the covariance $\mathbb{E}(\phi_{t_k} - \mu_{t_k})(\phi_{t_{1:k-1}} - \mu_{t_{1:k-1}}) = \Sigma_{k,1:k-1} =: v$ and variance $\mathrm{Var}(\phi_{t_k}) =: w^2$. These quantities cannot be arbitrary, since $w^2 \leq \sigma^2$ and the covariance matrix $\Sigma_{1:k}$ need to be positive definite. Under these constraints, this optimization for is a quadratically constrained linear optimization and we can write down optimal solution in closed form. It remains to evaluate the outer most supremum over $t_1, ..., t_{k-1} \in \mathcal{T}$ which is standard calculations. Details are left to the full proof in Appendix 10.6.1. □

Remark 10.7. *The bound in Theorem 10.6 is tight for the Gaussian noise-adding algorithm because if we take $t_1, ..., t_k$ such that $\phi_{t_{1:k-1}} \sim \mathcal{N}(0, \sigma^2 I)$, then expected squared bias $\geq \frac{(k-1)\sigma^4}{w^2+\sigma^2}$. When $w^2 \gg \sigma^2$, this nearly attains the upper bound (up to a multiplicative factor of $\frac{w^2+\sigma^2}{w^2}$).*

k-Step Adaptive Data Analysis: Upper Bound

Now we extend the above argument to k-step adaptive data analysis.

Theorem 10.8 (upper bound for k-step adaptivity)**.** *Let the distribution of data X and class of functions $\mathcal{T}$ obey our assumptions. Now let $T_1, ..., T_k$ be random variables drawn by any (potentially randomized) adaptive procedure that chooses $T_i \in \mathcal{T}$ based on outputs of actively perturbed statistics $A_1 \sim \mathcal{N}(\phi_{T_1}, w_1^2), ..., A_{i-1} \sim \mathcal{N}(\phi_{T_{i-1}}, w_{i-1}^2)$. Then for any integer k, the square bias of $\phi_{T_{i-1}}$ obeys*

$$|\mathbb{E}\phi_{T_k} - \mu_{T_k}|^2 \leq \sigma^4 \sum_{i=1}^{k-1} \left(\frac{1}{w_i^2} + \frac{\sigma^2}{w_i^4} \right) ,$$

where the expectation is taken over the both the randomness of X, the randomness in the adaptive choice of $(T_1, ..., T_k)$, and the randomness of perturbation used in $(A_1, ..., A_k)$. Furthermore, by taking $w_i^2 = \sqrt{k-1}\sigma^2$ for all $i < k$ and $w_k = 0$ we have

$$\max_{i=1,..,k} \mathbb{E}\, |A_i - \mu_{T_i}|^2 \leq 2(\sqrt{k-1} + 1)\sigma^2 \,.$$

The k-step adaptive analysis upper bound is on the same order as in Theorem 10.6 where we only allow one step adaptivity.

Proof Sketch. Similar to the previous theorem, we use the law of total expectation to expand the expectation. It is more involved in that we need to expand $\mathbb{E}\, |\phi_{T_k} - \mu_{T_k}|^2$ recursively into

$$\mathbb{E}_{A_1}\mathbb{E}_{T_2|A_1}\mathbb{E}_{A_2|T_{1:2},A_1}...\mathbb{E}_{T_{k-1}|A_{1:k-2}}\mathbb{E}_{A_{k-1}|T_{1:k-1},A_{1:k-2}}\mathbb{E}_{T_k|T_{1:k-1},A_{1:k-1}}\mathbb{E}_{\phi_{T_k}|T_{1:k},A_{1:k-1}} |\phi_{T_k} - \mu_{T_k}|^2 \,.$$

An upper bound can be obtained by replacing all $\mathbb{E}_{T_i|A_{1:i-1}}$ from a specific selection rule $\mathcal{W}_i$ with a supremum over $\mathcal{T}$. By bias and variance decomposition, it can be shown that the dominating term is the squared bias. Using the same argument as in Theorem 10.6, we can write down the conditional bias and maximize it explicitly, which gives us $\mathbb{E}(\phi_{T_k} - \mu_{T_k}) \leq \sqrt{\sigma^2 \mathbb{E} \boldsymbol{f}_{k-1}}$ where

$$\boldsymbol{f}_{k-1} := (A_{1:k-1} - \mu_{T_{1:k-1}})^T (\Sigma_{1:k-1} + W_{1:k-1})^{-1} \Sigma_{1:k-1} (\Sigma_{1:k-1} + W_{1:k-1})^{-1} (A_{1:k-1} - \mu_{T_{1:k-1}}).$$

In the above expression, $W_{1:k-1}$ is the (diagonal) covariance matrix of the noise we add in the first $k-1$ iterations. Note that $\boldsymbol{f}_{i-1}$ can be defined for any $1 \leq i \leq k$.

It turns out that we can "peel off" the supremum one at a time from the inner most conditional expectation all the way to the first one like an "onion", using the following formula (details in Lemma 10.15)

$$\sup_{t_{i-1} \in \mathcal{T}} \mathbb{E}_{A_{i-1}|T_{1:i-2},t_{i-1},A_{1:i-2}} \boldsymbol{f}_{i-1} \leq \boldsymbol{f}_{i-2} + \frac{\sigma^2}{w_i^2} + \frac{\sigma^4}{w_i^4}.$$

Applying this recursively and summing up the residuals gives us the bound of bias for any k. We invite readers to check out the detailed proof in Appendix 10.6.2. $\square$

Remark 10.9. *Since* 1*-step adaptivity is a special case of* k*-step adaptivity, the choice of* $t_1, ..., t_k$ *that nearly attains the risk bound also applies here.*

Note that the bound in Theorem 10.8 implies a expected absolute error upper bound on the same order as the result in Russo and Zou [183, Proposition 9]. Our result is slightly stronger as we bound the square error. That said, Proposition 5 and arguments in the proof of Proposition 9 in Russo and Zou [183] can be combined to obtain the same square error bound. Our proof leads to a sharper constant and transparent understanding of the least favorable adaptive selection protocol. Despite the differences in the settings, we remark that our bound is also on the same order as Bassily et al. [18], modulo that Bassily et al. [18] also has strong concentration — a characteristic that follows from McDiarmid's inequality under the low-sensitivity assumption. Our proof technique is arguably more direct, as we do not rely on differential privacy to control the generalization error.

10.3.2 Minimax lower bound

In this section, we prove a minimax lower bound that certifies that the Gaussian noise-adding attains the optimal rate up to a small constant factor in both Gaussian model (and hence trivially the subgaussian model). The proof will be constructive in that we will construct a strengthened version of the "adversary" that beats not only Gaussian noise adding, but any estimators.

Theorem 10.10. *Let* $\mathcal{D}(\sigma^2, N) = \{\mathcal{N}(\mu, \Sigma) | \Sigma_{i,i} \leq \sigma^2$ *for all* $i = 1, ..., N\}$. *If* $|\mathcal{T}| \geq 2^{k-1} + k - 1$, *then*

$$\inf_{\mathcal{A}_{1:k}} \sup_{P_{\phi_{\mathcal{T}}} \in \mathcal{D}(\sigma^2, |\mathcal{T}|)} \sup_{\mathcal{W}_{1:k}} \max_{i \in [k]} \mathbb{E}(A_i - \mu_{T_i})^2 = \Omega(\sqrt{k-1}\sigma^2)$$

This bound implies that the Gaussian-noise adding approach attains the minimax risk up to a universal constant. We now provide a sketch of the proof and highlight the key technical contributions. Detailed proof is presented in a later section.

Proof Sketch. The idea of the proof is that we construct a specific selection rule $\mathcal{W}_{1:k}$ and a nearly least favorable prior distribution of the parameters of the jointly Gaussian distribution of $\phi_{\mathcal{T}}$, and then find a tight Bayes risk lower bound.

Constructing a nearly least-favorable prior, adversary pair. Inspired by our upper bound proof in Section 10.3.1, we choose $\mathcal{W}_{1:k}$ that "explores" in the first $k-1$ rounds to learn as much about the data as possible, then "exploits" in the last round by finding a $t_k \in \mathcal{T}$ such that the realized value of ϕ_{t_k} deviates significantly from its mean μ_{t_k}. This is jointly achieved by a carefully chosen prior on the covariance matrix Σ and choices made by $\mathcal{W}_{1:k}$.

In the "exploration" phase, $\mathcal{W}_{1:k-1}$ chooses $T_{1:k-1} = t_{1:k-1}$, and the prior distribution ensures that $\Sigma_{1:k-1} = \sigma^2 I_{k-1}$, i.e., the first $k-1$ statistics are mutually independent. The cardinality constraint ensures that besides $t_1, ..., t_k$, there are 2^{k-1} remaining element in $\mathcal{T}$ such that we can dedicate one for every element $s \in \{-1, 1\}^{k-1}$; and we choose a prior distribution on Σ such that $t(s)$ obeys that

$$\begin{aligned}&\mathbb{P}(\mathrm{Cov}[\phi_{t(s)}, \phi_{t_1}] = \sigma^2 s_1, ..., \mathrm{Cov}[\phi_{t(s)}, \phi_{t_{k-1}}] = \sigma^2 s_{k-1})\\=&\mathbb{P}(\mathrm{Cov}[\phi_{t(s)}, \phi_{t_1}] = -\sigma^2 s_1, ..., \mathrm{Cov}[\phi_{t(s)}, \phi_{t_{k-1}}] = -\sigma^2 s_{k-1}) = 0.5.\end{aligned}$$

In the "exploitation" phase, the adversary $\mathcal{W}_k$ then uses the Bayes classifier to predict the signs of $\phi_{t_i} - \mu_{t_i}$ for each i. This is the likelihood ratio test which output 1 if

$$\log \frac{p(A_1, ..., A_{k-1} | \phi_{t_i} - \mu_{t_i} > 0)}{p(A_1, ..., A_{k-1} | \phi_{t_i} - \mu_{t_i} < 0)} \geq 0$$

and -1 otherwise. Let the predicted sign vector be $\hat{s}$, $\mathcal{W}_k$ simply chooses $t_k = t(\hat{s})$. This implies that

$$\mathbb{E}[(\phi_{t_k} - \mu_{t_k})(\phi_{t_{1:k-1}} - \mu_{t_{1:k-1}})] = \frac{\hat{s}\sigma^2}{\sqrt{k-1}} \text{ or } -\frac{\hat{s}\sigma^2}{\sqrt{k-1}}$$

depending on the random prior.

Intuitively, the adversary tries to maximize the bias in the last step as much as she can. Once $\phi_{t_{1:k-1}}$ is realized, the plan is to choose ϕ_{t_k} with maximally allowed correlation with all previously released queries, where the signs of correlations are chosen such that all bias caused by correlation will have the same sign. This is the "secret sauce" that the optimal adversary used in our upper bound proof of Theorem 10.6.

As we will see, the prior distribution on Σ plays an important role that ensures that the direction of the bias remains agnostic to the "player". Lastly, we choose a prior distribution for $\mu_{\mathcal{T}}$ that is independently and uniformly distributed for each $t \in \mathcal{T}$. We will supply details of these choices later.

Concretely, denote the minimax risk by $\mathcal{R}(k, \sigma^2)$, we use expectation to lower bound the maximum

$$\begin{aligned}\mathcal{R}(k, \sigma^2) &= \inf_{\mathcal{A}_{1:k}} \sup_{P_{\phi_{\mathcal{T}}} \in \mathcal{D}(\sigma^2, |\mathcal{T}|)} \sup_{\mathcal{W}_{1:k}} \max_{i \in [k]} \mathbb{E}(A_i - \mu_{T_i})^2\\&\geq \inf_{\mathcal{A}_{1:k}} \mathbb{E}_{\mu,\Sigma} \max_{i \in [k]} \mathbb{E}[(A_i - \mu_{T_i})^2 | \mu, \Sigma]\\&\geq \tau \wedge \inf_{\mathcal{A}_{1:k}: \mathbb{E}[(A_i - \mu_{T_i})^2] \leq \tau\ \forall i \in [k-1]} \mathbb{E}[(A_k - \mu_{T_k})^2]. \qquad (10.1)\end{aligned}$$

where $\cdot \wedge \cdot$ denotes the minimum of the two sides. The third line is true because if $\tau \geq \mathcal{R}(k, \sigma^2)$, the lower bound holds trivially. If $\tau > \mathcal{R}(k, \sigma^2)$, then the restriction is without loss of generality. This allows us to converted the problem to a constrained form with an unspecified constant τ, so that we can focus on dealing with the last round first.

Play optimally against the adversary: in the kth round. By Bayesian decision theory, the optimal Bayes estimator is the conditional expectation, namely,

$$A_k^* = \mathbb{E}[\mu_{T_k} | \phi_{\mathcal{T}}, T_{1:k}, A_{1:k-1}].$$

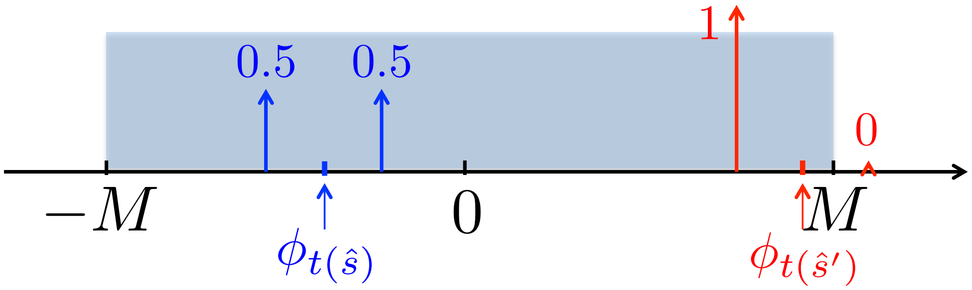


Figure 10.2: Illustration of the two cases of the posterior distribution $\mathbb{P}(\mu_{t(\hat{s})}|\phi_{\mathcal{T}}, \mu_{t_{1:k-1}})$.

This is somewhat hard to evaluate since taking $T_k = t_k$ reveals $\hat{s}$ to the player, which encodes the following message: The "adversary" who knows $\mu_{t_{1:k-1}}$ uses the optimal test to infer that $\mathrm{sign}(\phi_{t_{1:k-1}} - \mu_{t_{1:k-1}})$ is $\hat{s}$. Using this info, the the "player" now knows a little bit more about where $\mu_{t_{1:k-1}}$. Also, note that $\phi_{\mathcal{T}}$ is the common child of $\mu_{t_{1:k-1}}$ and μ_{t_k} so knowledge about $\mu_{t_{1:k-1}}$) could propagate over to μ_{t_k}. For a detailed graphical model representation of the dependency structure, we invite readers to check out Figure 10.3 in the proof.

A cute trick that we used to get around this issue is to consider an estimator $\mathcal{A}_k$ that is more powerful. In particular, we give $\mathcal{A}_k$ the privilege of seeing $\mu_{t_{1:k-1}}$, which d-separates the information flow from T_k to $\mu_{t(s)}$ for all sign vector s, including $t(\hat{s})$. This allows us to analytically write down the posterior distribution of $\mu_{t(s)}$ for any fixed s.

A side effect is that the player now knows exactly how much the bias is after combining the new knowledge with the old. In particular, the player now knows that

$$b(\hat{s}) = \hat{s}^T(\phi_{t_{1:k-1}} - \mu_{t_{1:k-1}})/\sqrt{k-1}.$$

What fortunately remains unknown (or at least unknown most of the times) is whether $\mu_{t(\hat{s})}$ is on the left or on the right of $\phi_{t(s)}$.

Under the uniform distribution $[-M', M']$, the posterior distribution of $\mu_{t(\hat{s})}$ has an interesting form — basically, if $\phi_{t(s)}$ happens to be in $[-M'+|b(\hat{s})|, M'-|b(\hat{s})|]$ then the posterior distribution assigns 0.5 probability to each hypothesis, therefore the posterior mean is simply $\phi_{t(\hat{s})}$ and the estimator completely consumes a square loss of $|b(\hat{s})|^2$. Otherwise, the player can denoise and estimate $\mu_{t(\hat{s})}$ perfectly. This is illustrated in Figure **??**.

Fortunately, we can choose a sequence of priors such that $M' \to \infty$ and the probability that the latter event occurs converges to 0.

This allows us to write

$$\begin{aligned}
\mathcal{R}(k,\sigma^2) &\geq \tau \wedge \inf_{\mathcal{A}_{1:k-1}:\mathbb{E}[(A_i-\mu_{T_i})^2]\leq\tau\ \forall i\in[k-1]} \mathbb{E}\left[\big(\frac{\langle \hat{s}, \phi_{t_{1:k-1}} - \mu_{t_{1:k-1}}\rangle}{\sqrt{k-1}}\big)^2 \right] \\
&\geq \tau \wedge \inf_{\mathcal{A}_{1:k-1}:\mathbb{E}[(A_i-\mu_{T_i})^2]\leq\tau\ \forall i\in[k-1]} \left(\frac{\mathbb{E}\left[\langle \hat{s}, \phi_{t_{1:k-1}} - \mu_{t_{1:k-1}}\rangle\right]}{\sqrt{k-1}} \right)^2 \\
&= \tau \wedge \inf_{\mathcal{A}_{1:k-1}:\mathbb{E}[(A_i-\mu_{T_i})^2]\leq\tau\ \forall i\in[k-1]} \frac{1}{k-1} \left(\sum_{i=1}^{k-1} \mathbb{E}\left[\hat{s}_i(\phi_{T_i} - \mu_{T_i})\right] \right)^2 \qquad (10.2)
\end{aligned}$$

Play optimally against the adversary: in the first 1th to $(k-1)$th rounds. We will now show that for a sequence of increasingly flat prior on $\mu_{t_{1:k-1}}$, it becomes without loss of generality to consider only $\mathcal{A}_{1:k-1}$ that adds zero-mean independent noise.

First of all, since the in (10.2) depends only on the overall expectation, it is not hard to show that for every dependent estimators $\mathcal{A}_i$ that chooses A_i as a function of all past observations, we can construct an alternative $\tilde{\mathcal{A}}_i$ that behaves the same for as $\mathcal{A}_i$ but depends only on ϕ_{t_i}. Note that this is equivalent to adding noise Z_i to ϕ_{t_i} where Z_i's distribution could be a function that depends on ϕ_{t_i}.

This reduces the problem to solving

$$\inf_{\mathcal{Z}:\mathbb{E}(\phi_i+Z_i-\mu_i)^2\leq\tau} \max_{\hat{s}_i} \mathbb{E}[\hat{s}_i(\phi_i - \mu_i)]$$

where $\hat{s}_i$ is any mapping from $\phi_i + Z_i, \mu$ to $\{-1, 1\}$.

Note that it is intuitively quite pointless for the noise Z_i to depends on ϕ_{t_i} since it does not reveal much information about μ_{t_i}, and it is unclear how depending on ϕ_{t_i} can prevent $\hat{s}_i$ to estimates the sign of $\phi_{t_i} - \mu_{t_i}$. However, formally showing that one can get rid of those possibility is tricky.

The redeeming property is that we know what $\hat{s}_i$ is (a likelihood ratio test) and can analytically write down the expression of $\max_{\hat{s}_i} \mathbb{E}[\hat{s}_i(\phi_i - \mu_i)]$ as a double integral for any choice of $\mathcal{Z}_i$. In addition, we used a "smoothing" technique with a wide uniform prior on μ_i with width $2M$. By exchanging the order of the integral, we are able to construct a $\tilde{\mathcal{Z}}_i$ that does not depend on ϕ_i and differs to the expectation using $\mathcal{Z}_i$ only by a factor of σ^2/M. This suggests that as $M \to \infty$, we can safely restrict our attention to independent noise adding.

Lastly, removing the mean from an independent noise Z_i does not change the objective and $\mathbb{E}(\phi_i + Z_i - \mu_i)^2$ only decreases, so it suffices to consider zero-mean independent noise adding. In other word, we have

$$\mathcal{R}(k,\sigma^2) \geq \tau \wedge \left(\sqrt{k-1} \inf_{Z\sim q \text{ s.t. } \mathbb{E}[Z]=0, \mathrm{Var}[Z]\leq\tau-\sigma^2} \frac{1}{2} \mathbb{E}_{x\sim\mathcal{N}(0,\sigma^2)} \left[|x| \| q(\cdot + |x|) - q(\cdot - |x|) \|_1 \right] \right)^2. \qquad (10.3)$$

Near optimal noise to obfuscate the sign of a Gaussian signal. It turns out that the above problem can be formulated as a variational convex optimization problem. While it is not possible to solve it analytically, we were able to construct a nearly optimal feasible dual functions, and the corresponding dual objective value is a lower bound of the optimal value by duality. The result, summarized in Lemma 10.16, implies that if $\tau \geq \sigma^2$,

$$\inf_{Z\sim q \text{ s.t. } \mathbb{E}[Z]=0, \mathrm{Var}[Z]\leq \tau-\sigma^2} \frac{1}{2}\mathbb{E}_{x\sim\mathcal{N}(0,\sigma^2)}\left[|x|\|q(\cdot+|x|)-q(\cdot-|x|)\|_1\right] \geq \frac{\sigma^2}{2\sqrt{3}\tau^{1/2}}.$$

Finally, we balance the error in by choosing $\tau = \sqrt{k-1}\sigma^2$, which ultimately gives a lower bound of the minimax risk as claimed. □

With a slight modification of the argument, we can also obtain a stronger and more explicit bound for the setting in Example 10.1.

Corollary 10.11. *Consider $X \sim \mathcal{N}(\theta, \sigma^2 I)$ and $X' = v(X-\theta)+\theta'$. Note that (X, X') are jointly normal with a covariance matrix $\begin{bmatrix}\sigma^2 I & v\sigma^2 I\\ v\sigma^2 I & \sigma^2 v^2 I\end{bmatrix}$. Let $\mathcal{T} = \mathcal{S}^{2d-1}$. Then for all $k = 1, ..., d+1$, we have*

$$\inf_{\mathcal{A}_{1:k}} \sup_{\theta_1,\theta_2,v\in\{-1,1\}} \sup_{\mathcal{W}_{1:k}} \max_{i\in[k]} \mathbb{E}[(A_i - \langle (X,X'), T_i\rangle)^2] = \Omega(\sqrt{k-1}\sigma^2).$$

This corollary suggests that the lower bound holds even if the covariance matrix is known to the player for up to one unknown bit, there are no methods that can perform better than simple noise adding.

The proof follows closely the arguments of the proof of Theorem 10.10, but differs in some of the details. Most importantly, $\phi_{\mathcal{I}^c}$ are now projections of X', so their means are now coupled, so we cannot define a separate prior for each t that is not the first independent statistics. For this reason, we needed to use a Gaussian prior instead of the uniform prior for θ'. We defer detailed proofs to the appendix.

10.3.3 Per-instance lower bound

The lower bound can also be stated in a per-distribution fashion, if we restrict our attention to a class of natural adaptive estimators that depend only on ϕ_{T_i} and the entire history, namely,

$$A_i = f(\phi_{T_i}, \phi_{T_1}, ..., \phi_{T_{i-1}}, A_1, ..., A_{i-1}, T_1, ..., T_{i-1}).$$

for some measurable function f. This includes noise-adding, shrinkage and so on.

We consider this is natural because $\phi_{T_{1:i}}$ is a sufficient statistic of $\mu_{T_{1:i}}$. Also, it is not clear whether estimators that infers $\mu_{T_{1:i}}$ through choices of $T_{1:i}$ assuming that the adversary actually knows $\mu_{T_{1:i}}$ are of any practical interest. Moreover, some form of restriction on the estimator class is required because in the setting with a fixed data distribution, knowledge T_i "gives away" μ_{T_i} and allows the trivial solution of $A_i := \mu_{T_i}$.

An alternative way of viewing this is that, while it is allowed for the estimator to make use the fixed distribution, it is required for it to use it "discreetly" and "pretend" that it only got the estimator from the data alone. In other word, this estimator cannot "cheat" too blatantly, such that it would appear unnatural.

The result requires the distribution to obey the following "richness" assumption.

(C) Within $\mathcal{T}$, we can define random choices of $T_1, ..., T_{k-1}$ such that the induced distributions $\mu_{T_1}, ..., \mu_{T_{k-1}} \sim$ Uniform$(-M, M)$ for some $M = \Omega(k^{1/4}\sigma)$, and for every $t_1, ..., t_{k-1}$ in the range, $\phi_{t_1} \perp\!\!\!\perp ... \perp\!\!\!\perp \phi_{t_{k-1}}$. Also, there always exists T_k such that for every t_k in its range, the correlation between ϕ_{t_k} and $\phi_{t_1,...,t_{k-1}}$ is $\sigma^2 s$ for any sign vector s of length $k-1$, and in addition, the induced distribution μ_{T_k} is uniformly distributed on $[-M', M']$ for $M' = \Omega(k\sigma)$.

In the following, we explicitly construct a distribution satisfying the above condition.

Example 10.12 (Projections of multivariate Gaussian distribution)**.** *Define a $2k$ dimensional multivariate Gaussian distribution as follows:*

$$X_1, X_3, ..., X_{2k-1} \sim \mathcal{N}(-M, \sigma^2/2)$$

$$X_2, X_4, ..., X_{2k} \sim \mathcal{N}(M, \sigma^2/2)$$

$$\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_3, ..., X_4) = ... = \mathrm{Cov}(X_{2k-1}, X_{2k}) = \sigma^2/2;$$

Recall that

$$\mathcal{T} = \left\{t \in \mathbb{R}^{2k} \middle| \|t\|_2 = 1\right\}$$

and

$$\phi_t = t^T x$$

We can work out that

$$\max_t \mathrm{Var}(t^T x) = \|\Sigma\|_2 = \sigma^2$$

We can define T_1 that interpolates between X_1 and X_2 such that μ_{T_1} has arbitrary distribution supported on $[-M, M]$. Now, for each sign vector s, we can choose the range of T_k to be

$$\left\{\sqrt{[t_k]_{2i-1}^2 + [t_k]_{2i}^2} = s_i/\sqrt{k-1} \textit{ for } i = 1, ..., k-1 \textit{ and } 0 \textit{ otherwise.}\right\}$$

The degree of freedom that we leave in the construction allows us to set μ_{T_k} to be anywhere between $[-M, M]$.

Theorem 10.13. *For each fixed distribution $P_{\phi_{\mathcal{T}}}$ satisfying (A), (B) and (C), the minimax risk for natural estimators obeys*

$$\inf_{\mathcal{A}_{1:k} \in \mathbb{A}_{1:k}} \sup_{\mathcal{W}_{1:k}} \max_{i \in [k]} \mathbb{E}(A_i - \mu_{T_i})^2 = \Omega(\sqrt{k-1}\sigma^2).$$

The proof follows the same proof techniques that we used to establish Theorem 10.10, except that the randomization and smoothing operations are now simulated by the adversary choosing randomized $T_1, ..., T_k$. The results suggest that if we allow sequential selection of what to estimate, natural estimators has an error growing in $\sqrt{k-1}$ not only in the worst case, but essentially for every distribution that has enough independent randomness and correlations in it, even if we restrict the statistics of interest to be among linear projections of a Gaussian vector.

10.4 Discussion and open problems

Richness, dimension and uniform convergence. As we point out in Assumption (C), the lower bound applies for each given k when there is a sufficiently rich class of functions $\mathcal{T}$ that satisfy Assumption (A) and (B). So what happens to the minimax risk when $\mathcal{T}$ is not sufficiently rich?

For instance, if $\mathcal{T}$ contains only a finite number of functions, or is a class of smooth functions that has slowly growing metric entropy, then standard uniform convergence argument implies that for sufficiently large k, the minimax risk will no longer be proportional to $\sqrt{k}$, but rather become some quantity independent of k. In addition, these rates can be achieved without randomization, which essentially deems the whole discussion of adaptive data analysis meaningless.

This picture is more intricate than just the two extremes. For any finite k, there is often a large space between small function classes that has uniform convergence (Uniform Glivenko-Cantelli), and big function classes with minimax risk growing in the order of $O(\sqrt{k})$. For instance, adaptive data analysis can be meaningful even for a finite class of functions, if $\sqrt{k} \ll \log|\mathcal{T}|$ but $k \simeq \log|\mathcal{T}|$, then we gain orders of magnitude improvements through this upper bound in the adaptive data analysis. Moreover, for any fixed k, the lower bound also holds if $\mathcal{T}$ sufficiently rich. This richness can often be measured in terms of dimension or $\log$-cardinality.

To be more concrete on the above discussion, we revisit the example of sequential estimation of the projections of multivariate Gaussian vector as described in Example 10.1.

First off, if k is very large, namely, when $k \gg d^2$, we know for sure that the upper bound is no longer tight since by standard uniform convergence, we can get

$$\mathbb{E} \sup_{t \in \mathcal{T}} |\phi_t(X)|^2 = O\left(d\sigma^2\right). \tag{10.4}$$

This suggests that the adaptive data analysis model is only meaningful when $k = o(d^2)$, in which case the Gaussian noise adding algorithm achieves a square error of $O(\sqrt{k}\sigma^2) = o(d\sigma^2)$.

On the other hand, our lower bounds imply that this is the optimal risk when $k = O(d)$, because our construction requires at least $k-1$ independent statistics, which is not possible for the problem with $d = O(\sqrt{k})$. This brings up the following open problem:

- Can we improve over Gaussian noise adding algorithm in the sequential estimation of Gaussian projections when $d < k < d^2$?

It is unclear what answers to this question might be. A negative answer would also reveal that the uniform convergence based bound is optimal for $k = \Omega(d^2)$. A positive answer would suggest that there might be a way to systematically obtain smaller estimation error than the uniform convergence bound even if $k \gg d^2$.

Our conjecture is that Gaussian noise adding is actually optimal up to $k = d^2$ and beyond d^2, uniform convergence bound is tight. We are able to prove this for a restricted setting of independent noise-adding estimators but a lower bound for arbitrary estimators remains extremely challenging and we will leave the concrete treatment of the problem to future work.

Gaussian sequence model, iid data and comparison to Steinke and Ullman [211]. So far we have been focusing on Gaussian sequence model with one observation.

A major open problem is therefore:

- Can we do better when we have iid data?

In the standard statistical estimation problem, extension to multiple observations are easy, since the KL-divergence of P^n and Q^n is simply $nD_{KL}(P\|Q)$ and the standard proof techniques like Le Cam's method and Fano's method follow straightforwardly. However, the problem becomes much trickier when we allow sequential selection, where the main workhorse we used relies on delicate construction of a strong adversary.

The iid setting makes it harder to prove the lower bound because it effectively allows more flexibility in designing estimators. In our framework, this can be modeled as additional equivalence and independence structures in $\mathcal{T}$ that are known and can be exploited by the player. These structures allow the player to find t' and use $\phi_{t'}$ (instead of ϕ_t) to estimate μ_t, and trade-off between bias (which could be large due to sequential selection) and variance (which could be large if data splitting is used excessively).

In a related setting studied in [211], where the adversary sequentially chooses statistical queries on an iid data set of sample size n over $\{0,1\}^d$ with a sufficiently large $d \geq k^2$, it was shown that there exists a special iid distribution that comes out of the fingerprinting codes in traitor-tracing schemes, under which it is possible to reconstruct a large percentage of the entire data set after using sufficiently accurate answers to most of k statistical queries, thus allowing the subsequent statistic to have a bias of magnitude $k^{1/2}/n$. This is still far from the upper bound, which is on the order of $k^{1/4}/\sqrt{n}$. It remains open whether the gap can be closed.

Also, we note that their technique is very specific to the distribution constructed using fingerprinting codes in which identifying the data sets correspond to identifying subsets of a known finite population in which each element is highly unique (a high dimensional data set). It is unclear how or whether their arguments could be modified to prove a lower bound in the jointly Gaussian setting.

Our conjecture is that Steinke and Ullman [211]'s lower bound is not tight, and a square error of order $\sqrt{k-1}\sigma^2/n$ should be the minimax rate for the iid setting too.

Data splitting based estimators. To illustrate why we believe in such a pessimistic conjecture, when there had been algorithms proposed that make clever use of a “holdout” data set to prevent the type of adversary that “explore” first and “exploit” only once.

In particular, the algorithm Thresholdout proposed by Dwork et al. [85] reveals half of the data set (the training set) to the data analyst (the “adversary”) and makes use of the other half of the data set (the holdout set) to estimates a statistic only when the estimates on the two halves differ significantly. Of course, in this case the adversary does not need to “explore” the training set and can easily force the Thresholdout estimator to makes use of the holdout set in every iterations, so with k rounds, our constructed “adversary” implies that it could lead to an estimation error of $\sqrt{k-1}\sigma^2/n$ in the last round.

Another contender is the EffectiveRounds algorithm proposed in Dwork et al. [87], which effectively split the training set in to q equal parts and each with $n/(2q)$ data points. Basically, the algorithm uses only one of the splits to do estimations until the answer differs significantly from the holdout set, in which case the answer from the holdout set is used and that split is thrown away. Subsequent questions are answered by the next split in the remaining $q-1$ and we start from scratch. The thing is that, since no randomization is used, this algorithms rely completely on data splitting and therefore for sequentially chosen queries, in other word, it only takes $O(q)$ steps to break all q training data sets. In other word, the square error risk achieved by this algorithm is only $k\sigma^2/n$.

There are many other ways to use data splitting when designing estimators to handle the easy cases when the adversary only sporadically “exploit”s, but it is unclear whether they can achieve the same worst-case risk that is achievable by directly adding noise to the statistics evaluated using the full data set. For instance, if we split the data set into q equal parts and each time we only use one of them to estimate statistics, then when $k < q$ such estimators necessarily will suffer from an error of $q\sigma^2/n$ even for non-adaptively chosen queries, while the naive noise adding approach that does not split the data could achieve $\sqrt{k-1}\sigma^2/n$.

An intriguing open question is that:

- Does data splitting actually help to improve the estimation error for $k > q^2$?

While we do not know the answer in general, as decision rules on, let us first consider a simplified problem where we do know the answer. Let us restrict the player’s strategy to the following model $\mathbb{A}_{1:k}^{split}$:

1. The player is informed of the adversary’s strategy $\mathcal{W}_1, ..., \mathcal{W}_k$ which in each round, chooses T_i using $T_{1:i-1}$, $A_{1:i-1}$.
2. The player randomly splits the data into q equal parts with disjoint sets of indices $\mathcal{I}_1, ..., \mathcal{I}_q$ satisfying $|\mathcal{I}_1| = ... = |\mathcal{I}_q| = n/q$ (for simplicity, assume q and n/q are integers). Let the averages of each partition be $\bar{X}^1, ..., \bar{X}^q$.
3. In Round i, upon receiving T_i, the player will answer by choosing (a possibly random) $J_i \in [q]$ and answer $\langle \bar{X}^{J_i}, T_i \rangle + Z$ where Z is an arbitrary “fresh” random variable generated independently to all states of the game, and J_i is determined using $J_{1:i-1}, T_{1:i}, A_{1:i-1}$, and

the oracle loss $\mathbb{E}[\langle \bar{X}^j - \theta, T_i\rangle^2 | T_{1:i}, J_{1:i-1}]$ for any $j \in [q]$.

Note that this model implies that J_i cannot depend on the data set $X_1, X_2, ..., X_n$. Yet, we think this is a reasonable model because the player now have access ot the oracle loss associated with each choice, the data distribution, the adversary's strategy and not having to reveal $J_1, ..., J_k$ to the adversary.

We note that the key idea in both Thresholdout and EffectiveRounds is to detect whether the query is "adaptive" through inconsistency in the estimates from different data split. Under this simplified model, the algorithm can know perfectly whether a query is (harmfully) "adaptive" and in fact can calculate the exact bias and variance associated with each choice $j \in [q]$.

Proposition 10.14. *In the problem of sequential selective estimation of Gaussian projections with iid data set of size n. Assume $d > k - 1$ and $X_1, ..., X_n \sim \mathcal{N}(\theta, \sigma^2 I)$. There is a fixed sequence of adversary strategy $\mathcal{W}_{1:k}$ that, such that for any $\theta \in \mathbb{R}^d$*

$$\inf_{\mathcal{A}_{1:k} \in \mathbb{A}^{split}_{1:k}} \mathbb{E} \max_{i \in [k]} [(A_i - \langle T_i, \theta\rangle)^2] = \Omega(\frac{\sqrt{k-1}\sigma^2}{n}).$$

The proof, which we present in details later in Appendix 10.8, constructs an explicit adversary that "explore"s and "exploit" at the same time using different dimensions of the projection T_i, which forces each data partition to answer not more than $(k-1)/q$ statistical queries with added noise having a variance at least $\Omega(\sqrt{k-1}\sigma^2/n)$. The key observation in the proof is that since each partition now has an intrinsic variance of order $q\sigma^2/n$, we only need a $1/q$ fraction of the dimensions in the specified projections to blow up the bias if that partition happens to be used.

This lower bound implies that even if the minimax risk can be improved in the iid setting, it probably cannot be achieved by variants of Thresholdout and EffectiveRounds.

10.5 Conclusion

In this chapter, we presented a minimax framework for adaptive data analysis and derived the first minimax lower bound for the joint-Gaussian setting that matches the upper bound in Russo and Zou [183] up to constant. We also presented an elementary proof for the same upper bound for the Gaussian noise-adding procedure and obtained an improved constant for the squared error metric. Our results reveal that an approximate least favorable adversary that "explores" for $k-1$ steps and then maximizes the bias in the last step by choosing a statistical quantity that is simultaneously correlated with all previous chosen statistical quantities with the signs of the covariance vector decided by an optimal classifier.

In the discussion, we pointed out the implicit dependency of the lower bound on the richness of the problem class. Through an illustrative example, we discussed the intriguing regime that interpolates the risk bounds via adaptive data analysis and uniform convergence. Lastly, we discussed challenges in extending the minimax analysis to the iid setting and illustrated why data-splitting based methods probably will not improve the minimax risk., do not work better relative to those that do not split the data set.

Future work includes finding the upper bound for the jointly-subgaussian case, strengthening the lower bound to the cases with an iid data set, bridging the dimensionality gap from between uniform convergence and adaptive data analysis, as well as extending the minimax framework for more practical regimes.

10.6 Upper bound proofs

10.6.1 Proof of Theorem 10.6

Proof. In this proof, we denote short hand $\phi := \phi_{t_{1:k-1}}$, $\mu := \mu_{t_{1:k-1}}$ and $Z := Z_{1:k-1}$. We first bound the bias term using law of total expectations

$$\mathbb{E}_{T_k,\phi_{T_k}}(\phi_{T_k}-\mu_{T_k}) = \mathbb{E}_{\phi,Z}\mathbb{E}_{T_k\sim\mathcal{W}(\phi+Z)}\mathbb{E}_{\phi_{T_k}}(\phi_{T_k}-\mu_{T_k}\big|T_k,\phi+Z) \leq \mathbb{E}_{\phi,Z}\sup_{t_k\in\mathcal{T}}\mathbb{E}_{\phi_{t_k}}(\phi_{t_k}-\mu_{t_k}\big|\phi+Z)$$

By Gaussianity, the conditional expectation is

$$\mathbb{E}_{\phi_{t_k}}(\phi_{t_k}-\mu_{t_k}\big|\phi+Z) = \Sigma_{k,1:k-1}^T(\Sigma_{1:k-1}+w^2I)^{-1}(\phi+Z-\mu)$$

What is the worst $\Sigma_{1:k-1,k}$ to use? If unconstrained the above bias can go to infinity. The choice of $\Sigma_{1:k-1,k}$ must obey that after we augment $\Sigma_{1:k-1}$ with it Σ remains positive semidefinite. In general, suppose we have block matrix $M = \begin{bmatrix}\lambda & v^T\\ v & \Sigma\end{bmatrix}$ where Σ is positive definite, then

$$M \succ 0 \Leftrightarrow \det(M) = \det(A)\det(A - v\lambda^{-1}v^T) > 0.$$

which tells us that the optimal solution to

$$\left\{\max_{v,\lambda}\langle v,x\rangle\Big|M\succ 0, \lambda\leq\sigma^2\right\}$$

is to take $\lambda = \sigma^2$ and the problem translates into

$$\left\{\max_{v}\langle v,x\rangle\Big|vv^T\prec\sigma^2\Sigma\right\} = \left\{\max_{v}\langle v,x\rangle\Big|\sqrt{v^T\Sigma^{-1}v}<\sigma\right\} = \|v^*\|_{\Sigma^{-1}}\|x\|_{\Sigma} = \sigma\|x\|_{\Sigma}$$

where $\|\cdot\|_\Sigma$ is the dual norm of $\|\cdot\|_{\Sigma^{-1}}$.

Take $\Sigma = \Sigma_{1:k-1}$ and we use the above argument can work out $\Sigma_{k,1:k-1}$ that attains the supremum and the corresponding conditional bias is

$$\sigma\|(\Sigma_{1:k-1}+w^2I)^{-1}(\phi+Z-\mu)\|_{\Sigma_{1:k-1}}.$$

Since $(\cdot)^2$ is monotonically increasing on $\mathbb{R}_+$, the supremum of the square conditional bias is attained by the same choice of t_k and

$$\begin{aligned}
&\mathbb{E}_{\phi,Z}\sigma^2(\phi+Z-\mu)^T(\Sigma_{1:k-1}+w^2I)^{-1}\Sigma_{1:k-1}(\Sigma_{1:k-1}+w^2I)^{-1}(\phi+Z-\mu)\\
=&\sigma^2\mathrm{tr}\left\{\mathbb{E}_{\phi,Z}\left[(\phi+Z-\mu)(\phi+Z-\mu)^T\right](\Sigma_{1:k-1}+w^2I)^{-1}\Sigma_{1:k-1}(\Sigma_{1:k-1}+w^2I)^{-1}\right\}\\
=&\sigma^2\mathrm{tr}\left\{\Sigma_{1:k-1}(\Sigma_{1:k-1}+w^2I)^{-1}\right\}\\
=&\sigma^2\sum_{i=1}^{k-1}\frac{\lambda_i}{\lambda_i+w^2}\le\frac{\sigma^2}{w^2}\sum_{i=1}^{k-1}\lambda_i=\frac{(k-1)\sigma^4}{w^2}. \qquad (10.5)
\end{aligned}$$

where λ_i are the eigenvalues of $\Sigma_{1:k-1}$. The last line diagonalizes $\Sigma_{1:k-1}$ and uses the fact that trace operator is unitary invariant. Note that the inequality is sharp up to a small constant as for the case when $\lambda_i=\sigma^2$ for all i, the quantity is equal to $\frac{(k-1)\sigma^2}{\sigma^2+w^2}$. By Jensen's inequality, the upper bound for the expected conditional bias in (10.5) is also upper bounds the bias.

The proof of the second claim is simply decomposing the square error of $A_k=\phi_{T_k}+Z_k$ into square bias and variance, and upper bound each term by choosing $w=\sqrt{k-1}\sigma^2$.

The bias of A_k is the same as that of ϕ_{T_k} in (10.5). The variance obeys

$$\begin{aligned}
&\Sigma_k+w^2-\mathbb{E}_{\phi,Z}\sup_{\phi_k}\Sigma_{k,1:k-1}^T(\Sigma_{1:k-1}+w^2I_{1:k-1})^{-1}\Sigma_{k,1:k-1}\\
\le&\Sigma_k+w^2\le\sigma^2+w^2\le(\sqrt{k-1}+1)\sigma^2
\end{aligned}$$

Adding the two upper bounds give us the right form. □

10.6.2 Proof of Theorem 10.8

Proof. We first control the bias. As we worked out in the proof of Theorem 10.6,

$$\begin{aligned}
\mathbb{E}\left(\phi_{T_k}-\mu_{T_k}\right)=&\mathbb{E}_{T_{1:k-1},A_{1:k-1}}\mathbb{E}_{T_k\sim\mathcal{W}_k|T_{1:k-1},A_{1:k-1}}\left(\mathbb{E}_{\phi_{T_k}|T_{1:k},A_{1:k-1}}\phi_{T_k}-\mu_{T_k}\right)\\
\le&\mathbb{E}_{T_{1:k-1},A_{1:k-1}}\sup_{t_k\in\mathcal{T}}\left(\mathbb{E}_{\phi_{t_k}|T_{1:k-1},t_k,A_{1:k-1}}\phi_{t_k}-\mu_{T_k}\right)\\
=&\mathbb{E}_{T_{1:k-1},A_{1:k-1}}\sqrt{\sigma^2\boldsymbol{f}_{k-1}}\le\sqrt{\sigma^2\mathbb{E}_{T_{1:k-1},A_{1:k-1}}\boldsymbol{f}_{k-1}}, \qquad (10.6)
\end{aligned}$$

where for simplicity, we denote

$$\boldsymbol{f}_{k-1}:=(A_{1:k-1}-\mu_{T_{1:k-1}})^T(\Sigma_{1:k-1}+W_{1:k-1})^{-1}\Sigma_{1:k-1}(\Sigma_{1:k-1}+W_{1:k-1})^{-1}(A_{1:k-1}-\mu_{T_{1:k-1}}),$$

and the last step follows from the Jensen's inequality on the concave function $\sqrt{\cdot}$. Note that every variable in $\boldsymbol{f}_{k-1}$ is a random variable.

We will further expand the above expectation into a sequence of expectations and recursively evaluate the conditional expectation and then taking supremum of $\boldsymbol{f}_{k-1}$.

$$\begin{aligned}
\mathbb{E}\boldsymbol{f}_{k-1}&=\mathbb{E}_{T_{1:k-2},A_{1:k-2}}\mathbb{E}_{T_{k-1}|T_{1:k-2},A_{1:k-2}}\mathbb{E}_{A_{k-1}|T_{1:k-1},A_{1:k-2}}\boldsymbol{f}_{k-1}\\
&\le\mathbb{E}_{T_{1:k-2},A_{1:k-2}}\sup_{\phi_{t_{k-1}}\in\mathcal{T}}\mathbb{E}_{A_{k-1}|T_{1:k-2},t_{k-1},A_{1:k-2}}\boldsymbol{f}_{k-1}
\end{aligned}$$

It turns out that we can neatly express the conditional expectation in a closed form as a function of $\Sigma_{1:k-1}$ and the diagonal covariance of the added noise $W_{1:k-1}$.

Lemma 10.15. *Denote $w^2 = W_{k-1}, W = W_{1:k-2}$ and $\Sigma = \Sigma_{1:k-2}, v = \Sigma_{k-1,1:k-2}, \lambda = \Sigma_{k-1}$ such that*

$$W_{1:k-1} = \begin{bmatrix} W & 0 \\ 0 & w^2 \end{bmatrix}, \qquad \Sigma_{1:k-1} = \begin{bmatrix} \Sigma & v \\ v^T & \lambda \end{bmatrix}.$$

In addition, $\Omega := (\Sigma + W)^{-1}\Sigma(\Sigma + W)^{-1}$. We have

$$\mathbb{E}_{A_{k-1}|T_{1:k-2},t_{k-1},A_{1:k-2}} \boldsymbol{f}_{k-1} = \boldsymbol{f}_{k-2} + \frac{(\lambda + v^T\Omega v - v^T(\Sigma + W)^{-1}v)(\lambda + w^2)}{(\lambda + w^2 - v^T(\Sigma + W)^{-1}v)^2}. \tag{10.7}$$

In order to not interrupt the flow of the arguments, we defer the proof of Lemma 10.15 to the appendix.

With this parametric form, the supremum can be rewritten as

$$\begin{aligned} &\sup_{T_{k-1}\in\mathcal{T}} \mathbb{E}_{A_{k-1}|T_{1:k-2},t_{k-1},A_{1:k-2}} \boldsymbol{f}_{k-1} \\ =&\boldsymbol{f}_{k-2} + \max_{\lambda \le \sigma^2, \Sigma_{1:k-1}\succeq 0} \frac{(\lambda + v^T\Omega v - v^T(\Sigma + W)^{-1}v)(\lambda + w^2)}{\left(\lambda + w^2 - v^T(\Sigma + W)^{-1}v\right)^2}. \end{aligned} \tag{10.8}$$

As in the previous calculations, the semidefinite constraint requires that $v^T\Sigma^{-1}v \le \lambda$, also $(\Sigma + W)^{-1} \preceq \Sigma^{-1}$, therefore for any v,

$$v^T\Omega v = v^T(\Sigma + W)^{-1}\Sigma(\Sigma + W)^{-1}v \le v^T(\Sigma + W)^{-1}v \le v^T\Sigma^{-1}v \le \lambda.$$

Substitute into (10.8), we get an upper bound of the supremum

$$\sup_{T_{k-1}\in\mathcal{T}} \mathbb{E}_{A_{k-1}|T_{1:k-2},t_{k-1},A_{1:k-2}} \boldsymbol{f}_{k-1} \le \boldsymbol{f}_{k-2} + \max_{\lambda\le\sigma^2} \frac{\lambda(\lambda + w^2)}{w^4} \le \boldsymbol{f}_{k-2} + \frac{\sigma^2}{w^2} + \frac{\sigma^4}{w^4}.$$

Recursively evaluating and upper bounding the supremum until the base case

$$\sup_{t_1\in\mathcal{T}} \mathbb{E}_{A_1|t_1} \boldsymbol{f}_1 = \max_{\sigma_1^2\le\sigma^2} \frac{\mathbb{E}(A_1 - \mu_{t_1})^2\sigma_1^2}{(\sigma_1^2 + w_1^2)} = \max_{\sigma_1^2\le\sigma^2} \frac{\sigma_1^2}{\sigma_1^2 + w_1^2} \le \frac{\sigma^2}{w_1^2} + \frac{\sigma^4}{w_1^4},$$

we end up with

$$\mathbb{E}\boldsymbol{f}_{k-1} = \sigma^2 \sum_{i=1}^{k-1} \left(\frac{1}{w_i^2} + \frac{\sigma^2}{w_i^4}\right),$$

and substitute into (10.6), we obtain an upper bound for the bias

$$\mathbb{E}\left(\phi_{T_k} - \mu_{T_k}\right) \le \sqrt{\sigma^4 \sum_{i=1}^{k-1} \frac{1}{w_i^2} + \frac{\sigma^2}{w_i^4}}.$$

This gives the desired bound for the first claim.

The variance can be easily evaluated using the conditional variance.

$$\begin{aligned}\mathrm{Var}(\phi_k) &= \mathbb{E}_{T_{1:k-1},A_{1:k-1}}\mathbb{E}_{T_k|T_{1:k-1},A_{1:k-1}}\mathbb{E}_{\phi_{T_k}|T_{1:k},A_{1:k-1}}\left(\phi_{T_k} - \mathbb{E}_{\phi_{T_k}}\right)^2 \\ &= \mathbb{E}_{T_{1:k-1},A_{1:k-1}}\mathbb{E}_{T_k|T_{1:k-1},A_{1:k-1}}(\sigma_i^2 - \Sigma_{k,1:k-1}\Sigma_{1:k-1}\Sigma_{1:k-1,k}) \leq \sigma^2.\end{aligned}$$

Combining the bounds for bias and variance, we upper bound the mean square error by

$$\sigma^4 \sum_{i=1}^{k-1}\left(\frac{1}{w_i^2} + \frac{\sigma^2}{w_i^4}\right) + \sigma^2.$$

The second claim follows directly by taking $w_i^2 = \sqrt{k-1}\sigma^2$ for each $i = 1, .., k-1$, and add the variance of the additional noise Z_k. □

10.6.3 Proof of Lemma 10.15

Proof. We prove by a direct calculation. First of all, we invert $(\Sigma_{1:k-1} + W_{1:k-1})$ in block form and use Sherman-Morrison formula on the first principle minor:

$$(\Sigma_{1:k-1} + W_{1:k-1})^{-1} = \begin{bmatrix}\Sigma + W & v \\ v^T & \lambda + w^2\end{bmatrix}^{-1}$$

$$= \begin{bmatrix}\left(\Sigma + W - \frac{vv^T}{\lambda+w}\right)^{-1} & -(\Sigma+W)^{-1}v\left(\lambda + w^2 - v^T(\Sigma+W)^{-1}v\right)^{-1} \\ -\left(\lambda + w^2 - v^T(\Sigma+W)^{-1}v\right)^{-1}v^T(\Sigma+W)^{-1} & \left(\lambda + w^2 - v^T(\Sigma+W)^{-1}v\right)^{-1}\end{bmatrix}$$

$$= \begin{bmatrix}(\Sigma+W)^{-1} + \alpha(\Sigma+W)^{-1}vv^T(\Sigma+W)^{-1} & -\alpha(\Sigma+W)^{-1}v \\ -\alpha v^T(\Sigma+W)^{-1} & \alpha\end{bmatrix}$$

where we denote $\alpha := \left(\lambda + w^2 - v^T(\Sigma+W)^{-1}v\right)^{-1}$.

For any symmetric block matrices

$$\begin{bmatrix}X & Y \\ Y^T & Z\end{bmatrix}\begin{bmatrix}A & B \\ B^T & C\end{bmatrix}\begin{bmatrix}X & Y \\ Y^T & Z\end{bmatrix}$$
$$= \begin{bmatrix}XAX + XBY^T + Y^TBX + YCY^T & XAY + XBZ + YB^TY + YCZ \\ Y^TAX + Y^TBY^T + ZB^TX + ZCY^T & Y^TAY + Y^TBZ + ZB^TY + ZCZ\end{bmatrix}$$

For the special case here let $A_{1:k-2} - \mu_{T_{1:k-2}} =: x$ and $A_{k-1} - \mu_{T_{k-1}} =: y$,

$$\boldsymbol{f}_{k-1} = \begin{bmatrix}x^T & y^T\end{bmatrix}\begin{bmatrix}F_{11} & F_{12} \\ F_{21} & F_{22}\end{bmatrix}\begin{bmatrix}x \\ y\end{bmatrix} = x^TF_{11}x + 2x^TF_{12}y + y^TF_{22}y$$

for some blocks $F_{\cdot,\cdot}$.

Further denote $b := v^T(\Sigma + W)^{-1}x$, $\Omega := (\Sigma + W)^{-1}\Sigma(\Sigma + W)^{-1}$,

$$\begin{cases} x^T F_{11} x = & x^T\Omega x + 2\alpha b(x^T\Omega v) + \alpha^2 b^2 (v^T \Omega v) - 2\alpha b^2 - 2\alpha^2 b^2 [v^T(\Sigma+W)^{-1}v] + \lambda\alpha^2 b^2. \\ 2x^T F_{12} y = & -2\alpha b(x^T\Omega v) - 2\alpha^2 b^2(v^T\Omega v) + 2\alpha b^2 + 4\alpha^2 b^2[v^T(\Sigma+W)^{-1}v] - 2\lambda\alpha^2 b^2 \\ y^T F_{22} y = & \lambda\alpha^2 b^2 + \alpha^2 b^2(v^T\Omega v) - 2\alpha^2 b^2[v^T(\Sigma+W)^{-1}v] \\ & +\alpha^2\left[\lambda + v^T\Omega v - v^T(\Sigma+W)^{-1}v\right](\lambda+w^2) \end{cases}$$

It's easy to check that $x^T\Omega x = \boldsymbol{f}_{k-2}$ and almost everything cancels out when we sum the three terms up. All that remains gives exactly what we claim to be true. □

10.7 Lower bound proofs

We use $H_i = [\mathcal{Z}_{1:i-1}, T_{1:i-1}, A_{1:i-1}]$ to denote the shared knowledge up to step i by both the player and adversary, and we use $\bar{H}_i = [H_i, Z_{1:i-1}]$ to denote the entire history and the data, which further includes information known to the player but not known to the adversary.

The proof constructs one specific sequence of $\mathcal{W}_{1:k}$ and show that, no estimators $\mathcal{A}_{1:k}$ can achieve a better risk than $\Omega(\sqrt{k}\sigma^2)$.

We discuss the two different settings that we considered, where the constructions are related but slightly different:

(a) when we considers the standard minimax risk, where we maximize over the class of all distributions, as well as the selection procedure. Note that here the only assumption we need is $|\mathcal{T}| > 2^{k-1} + k - 1$.

(b) when we consider a fixed distribution satisfying richness assumptions and maximize over only the selection procedure (restricting the estimators to only natural ones).

10.7.1 Part (a) minimax lower bound

Constructing a specific adversary strategy.

In the first scenario, the adversary simply chooses $T_1 = 1, ..., T_{k-1} = k - 1$, and T_k will be a deterministic map of H_k to range $k, ..., k + 2^{k-1} - 1$. The remainder of $\mathcal{T}$ hence becomes unimportant and we can specify it in an arbitrary way that does not depend on the first $k + 2^{k-1} - 1$ elements of the set. Without loss of generality, we assume $|\mathcal{T}| = k + 2^{k-1} - 1$ from here onwards until we discuss Scenario (b).

For ease of presentation, define operator $\mathsf{dec2bin} : \{0, 1, ..., 2^{k-1}\} \to \{-1, 1\}^{k-1}$, which convert any nonnegative integer smaller than 2^{k-1} to its binary number representation in the form of a sign vector. For example, $\mathsf{dec2bin}(1) = [-1, -1, ..., -1, 1]$ and $\mathsf{dec2bin}(5) = [-1, -1, ..., -1, 1, -1, 1]$.

Similarly, the inverse operator $\mathsf{bin2dec} : \{-1,1\}^{k-1} \to \{0,1,...,2^{k-1}\}$ maps a binary number's sign vector representation to the corresponding integer.

Specifically, the adversary takes $\mathcal{W}_k$ to be the following algorithm:

1. Observing $A_1, A_2, ..., A_{k-1}$ and $\mu_1, ..., \mu_{k-1}$, $\mathcal{W}_k$ uses an optimal classifier to infer the signs of $\phi_{T_i} - \mu_i$. Specifically. likelihood ratio test that output $\hat{s}_i = 1$ if
$$\log \frac{\mathbb{P}(A_1, ..., A_{k-1} | \phi_{T_i} - \mu_{T_i} > 0)}{\mathbb{P}(A_1, ..., A_{k-1} | \phi_{T_i} - \mu_{T_i} < 0)} \ge 0$$
and $\hat{s}_i = -1$ otherwise. Let the estimated sign vector be $\hat{s}$.
2. Choose $T_k = k + \mathsf{bin2dec}(\hat{s})$.

Let the above strategy be $\tilde{\mathcal{W}}_{1:k}$. Moreover, define the subset of estimators such that $\sup_{\mu,\Sigma} \max_{\mathcal{W}_{1:k}} \max_{i \in [k-1]} \mathbb{E}(A_i$
$\mu_{T_i})^2 \le \tau$. Denote this set by $\mathbb{A}^{\tau}_{1:k-1}$. The minimax risk $\mathcal{R}(k, \sigma^2)$ satisfies:

$$\begin{aligned} \mathcal{R}(k,\sigma^2) &\ge \tau \wedge \min_{\mathcal{A}_{1:k-1} \in \mathbb{A}^{\tau}_{1:k-1}, \mathcal{A}_k} \max_{\mu,\Sigma} \max_{\mathcal{W}_{1:k}} \mathbb{E}_{\mathcal{A}_{1:k}, \mathcal{W}_{1:k}} (A_k - \mu_{T_k})^2 \\ &\ge \tau \wedge \min_{\mathcal{A}_{1:k-1} \in \mathbb{A}^{\tau}_{1:k-1}, \mathcal{A}_k} \mathbb{E}_{\mu,\Sigma} \mathbb{E}_{\mathcal{A}_{1:k}, \tilde{\mathcal{W}}_{1:k}} (A_k - \mu_{T_k})^2, \end{aligned} \tag{10.9}$$

The first inequality is true for all τ because if $\tau \le \mathcal{R}(k,\sigma^2)$ the bound holds trivially and if $\tau > \mathcal{R}(k,\sigma^2)$ then the minimax estimator is in $\mathbb{A}^{\tau}_{1:k-1}$. We will choose τ later to maximize the bound. In the second line we used a Bayes risk to lower bound the minimax risk. Note that the above inequality holds for the Bayes risk evaluated under *any* prior distribution of $\mu_{\mathcal{T}}$ and $\Sigma_{\mathcal{T},\mathcal{T}}$.

Constructing a nearly least favorable prior.

To approximately maximize the lower bound, we will consider the following prior distribution.

$$\mu_t \sim \text{Unif}(-M_t, M_t)$$

for each $t \in \mathcal{T}$ independently. For $t = 1, ..., k-1$, $M_t = M$ and for $t = k, ..., k + 2^{k-1} - 1$, $M_t = M'$. Parameter M and M' will be specified later.

Let $\Sigma_{1:k-1,1:k-1} = \sigma^2 I_{k-1}$. Each of the remaining 2^{k-1} statistics has marginal variance also fixed at σ^2, but they are all correlated with the first $k-1$ statistics. Specifically,

$$\text{Cov}(\phi_{k+j}, \phi_{1:k-1}) = \frac{\sigma^2 \nu_{k+j}}{\sqrt{k-1}} \cdot \mathsf{dec2bin}(j).$$

where independently for every $j = 0, ..., 2^{k-1} - 1$,

$$\mathbb{P}(\nu_{k+j} = -1) = \mathbb{P}(\nu_{k+j} = 1) = 0.5.$$

Note that here the prior distribution of $\Sigma_{\mathcal{T},\mathcal{T}}$ is induced by the distribution of the random sign vector $\nu = \{\nu_k, ..., \nu_{k+2^{k-1}-1}\}$.

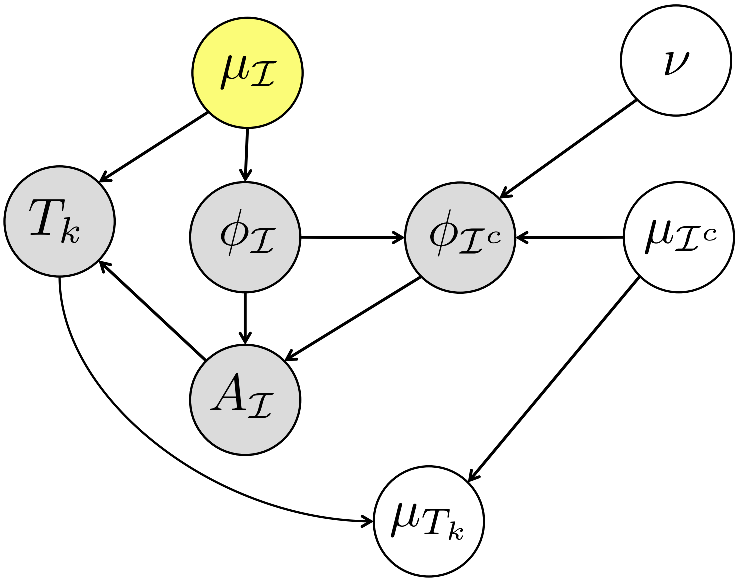


Figure 10.3: A directed acyclic graphical representation of the joint distribution. The nodes shaded in gray are random variables observed by the estimators. $\mu_\mathcal{I}$ (shaded in yellow) is not known to the "player", but we deliberately give it away to the "player"[1] to break the dependence between $A_\mathcal{I}$ and μ_{T_k} so that the calculations become tractable.

Optimal player strategy in the last step.

We will now work out the optimal estimator in the kth step under the above fixed adversary strategy and the prior distribution.

The conditional independence structure of the variables are summarized in Figure 10.3. Note that we can factorize the joint distribution in a different way. For clarity in notation, denote index set $\mathcal{I} = \{1, ..., k-1\}$ and $\mathcal{I}^c = \{k, ..., k+2^{k-1}-1\}$.

$$\mathbb{E}(A_k - \mu_{T_k})^2 = \mathbb{E}_{\mu_\mathcal{I}}\mathbb{E}_{\phi_\mathcal{I}|\mu_\mathcal{I}}\mathbb{E}_{\phi_{\mathcal{I}^c}|\mu_\mathcal{I},\phi_\mathcal{I}}\mathbb{E}_{A_\mathcal{I}}\mathbb{E}_{A_\mathcal{I}|\mu_\mathcal{I},\phi_\mathcal{T}}\mathbb{E}_{T_k|A_\mathcal{I},\mu_\mathcal{I}}\mathbb{E}[(A_k - \mu_{T_k})^2|T_k, \mu_\mathcal{I}, \phi_\mathcal{T}, A_\mathcal{I}].$$

The final expectation is taken over $P(\mu_{I^c}, \nu|T_k, \mu_{1:k-1}, \phi_\mathcal{T}, A_\mathcal{I})$. By the conditional independences, we can drop the dependence on $A_\mathcal{I}$ and T_k:

$$P(\mu_{I^c}, \nu|T_k, \mu_\mathcal{I}, \phi_\mathcal{T}, A_\mathcal{I}) = P(\mu_{I^c}, \nu|\mu_\mathcal{I}, \phi_\mathcal{T})$$

Also, for any $i, j \in \mathcal{I}^c$,

$$(\mu_i, \nu_i) \perp\!\!\!\perp (\mu_i, \nu_j) \mid \mu_\mathcal{I}, \phi_\mathcal{T}$$

So for each $T_k = i \in \mathcal{I}^c$, so we can consider a separate univariate estimation problem independently for each T_k chosen.

[1] This makes the "player" stronger, the minimax risk smaller, therefore a lower bound of the easier setting for the "player" is a lower bound of the original setting.

Moreover, by Bayes rule,

$$\begin{aligned}P(\mu_{I^c}, \nu|\mu_{\mathcal{I}}, \phi_{\mathcal{T}}) \propto P(\phi_{\mathcal{T}}|\mu_{\mathcal{I}^c}, \nu, \mu_{\mathcal{I}})\pi(\mu_{I^c}, \nu|\mu_{1:k-1}) &= P(\phi_{\mathcal{T}}|\mu_{\mathcal{I}^c}, \nu, \mu_{\mathcal{I}})\pi(\mu_{I^c}, \nu) \\ &= P(\phi_{\mathcal{I}}|\mu_{\mathcal{I}})P(\phi_{\mathcal{I}^c}|\mu_{\mathcal{I}^c}, \nu, \phi_{\mathcal{I}}, \mu_{\mathcal{I}})\pi(\mu_{I^c})\pi(\nu).\end{aligned}$$

For $i \in \mathcal{I}^c$, denote $b_i := \frac{1}{\sqrt{k-1}} s_i^T(\phi_{\mathcal{I}} - \mu_{\mathcal{I}})$, where $s_i = \mathsf{dec2bin}(i-k)$. The posterior distribution obeys that

$$\begin{cases}\mathbb{P}\left(\mu_i = \phi_i + b_i, \nu_i = 1|\mu_{\mathcal{I}}, \phi_{\mathcal{T}}\right) = 1 & \text{if } \phi_i - b_i < -M' \\ \mathbb{P}\left(\mu_i = \phi_i - b_i, \nu_i = -1|\mu_{\mathcal{I}}, \phi_{\mathcal{T}}\right) = 1 & \text{if } \phi_i + b_i > M' \\ \begin{bmatrix}\mathbb{P}\left(\mu_i = \phi_i - b_i, \nu_i = -1|\mu_{\mathcal{I}}, \phi_{\mathcal{T}}\right) = 0.5 \\ \mathbb{P}\left(\mu_i = \phi_i + b_i, \nu_i = 1|\mu_{\mathcal{I}}, \phi_{\mathcal{T}}\right) = 0.5.\end{bmatrix} & \text{otherwise.}\end{cases}$$

This defines a prior distribution for the problem of estimating μ_{T_k}, when we condition on $T_k, \mu_{\mathcal{I}}, \phi_{\mathcal{T}}, A_{\mathcal{I}}$. The corresponding optimal estimator A_k (even if we give $\mu_{\mathcal{I}}$ as an auxiliary input to make it more powerful), is simply the Bayes estimator, which for squared error loss function, is simply the posterior mean $\mathbb{E}(\mu_{T_k}|T_k, \mu_{1:k-1}, \phi_{\mathcal{T}}, A_{1:k-1}) = \phi_{T_k}$ [see e.g. 143, Corollary 4.1.2], namely,

$$A_k = \begin{cases}\phi_{T_k} + b_{T_k} & \text{if } \phi_{T_k} - b_{T_k} < -M' \\ \phi_{T_k} - b_{T_k} & \text{if } \phi_{T_k} + b_{T_k} > M' \\ \phi_{T_k} & \text{if } -M' + b_{T_k} < \phi_{T_k} < M' - b_{T_k},\end{cases}$$

and the corresponding mean square error is 0 if the first two events happen and is $b_{T_k}^2$ otherwise.

Of course as M gets large, the probability that we run into the first two events are smaller and at the limit of $M' \to \infty$, the first two events occur with probability 0. We now formalize this idea.

$$\begin{aligned}&\mathbb{P}(\phi_{T_k} - b_{T_k} < -M') = \mathbb{P}(\nu_{T_k} = -1, \mu_{T_k} < -M' + b_{T_k}) \\ \leq &\mathbb{P}(\mu_{T_k} < -M' + b_{T_k}) = \mathbb{P}(\mu_{T_k} < -M' + b_{T_k}, b_{T_k} > h) + \mathbb{P}(\mu_{T_k} < -M' + b_{T_k}, b_{T_k} < h) \\ \leq &\mathbb{P}(b_{T_k} > h) + \mathbb{P}(\mu_{T_k} < -M' + h, b_{T_k} < t) \leq \mathbb{P}(b_{T_k} > h) + \mathbb{P}(\mu_{T_k} < -M' + h)\end{aligned}$$

where the inequality holds for every fixed $h > 0$. Similarly,

$$\mathbb{P}(\phi_{T_k} + b_{T_k} > M') \leq \mathbb{P}(b_{T_k} > h) + \mathbb{P}(\mu_{T_k} > M' - h).$$

Note that

$$\mathbb{P}(\mu_{T_k} > M' - h) = \mathbb{P}(\mu_{T_k} < -M' + h) = h/(2M').$$

Also by Markov's inequality

$$\mathbb{P}(b_{T_k} > h) \leq \frac{\mathbb{E}[b_{T_k}]}{h} \leq \frac{\mathbb{E}[\max_{i \in \mathcal{I}^c} b_i]}{h} = \frac{1}{h}\sqrt{\frac{2(k-1)}{\pi}}\sigma.$$

Therefore

$$\mathbb{P}(\phi_{T_k} - b_{T_k} < -M' \text{ or } \phi_{T_k} + b_{T_k} > M') \leq \frac{h}{M'} + \frac{1}{h}\sqrt{\frac{2(k-1)}{\pi}}\sigma.$$

As $M' \to \infty$, take $h = \sqrt{M'}$, then $\mathbb{P}(\phi_{T_k} - b_{T_k} < -M' \text{ or } \phi_{T_k} + b_{T_k} > M') \to 0$. Also, it can be shown that for each $\mathcal{A}_{1:k-1} \in \mathbb{A}^{\tau}_{1:k-1}$, by dominated convergence theorem

$$\lim_{M' \to \infty} \mathbb{E}(b^2_{T_k} | - M' + b_{T_k} \leq \phi_{T_k} \leq M' - b_{T_k}) = \mathbb{E}(b^2_{T_k}).$$

In summary, as $M' \to \infty$, we get

$$\begin{aligned}
\mathcal{R}(k, \sigma^2) \geq & \tau \wedge \min_{\mathcal{A}_{1:k-1} \in \mathbb{A}^{\tau}_{1:k-1}} \mathbb{E}_{\mu_{1:k-1}} \mathbb{E}_{\mathcal{A}_{1:k-1}} b^2_{T_k} \\
= & \tau \wedge \min_{\mathcal{A}_{1:k-1} \in \mathbb{A}^{\tau}_{1:k-1}} \mathbb{E}_{\mu_{1:k-1}} \mathbb{E}_{\bar{H}_k} \left[\sum_{i=1}^{k-1} \frac{\hat{s}_i(\phi_{T_i} - \mu_{T_i})}{\sqrt{k-1}} \right]^2 \\
\geq & \tau \wedge \min_{\mathcal{A}_{1:k-1} \in \mathbb{A}^{\tau}_{1:k-1}} \left[\sum_{i=1}^{k-1} \frac{\mathbb{E}_{\mu_{1:k-1}} \mathbb{E}_{\bar{H}_k} \hat{s}_i(\phi_{T_i} - \mu_{T_i})}{\sqrt{k-1}} \right]^2 \qquad (10.10)
\end{aligned}$$

where the last inequality follows from Jensen's inequality.

Let $\hat{s}$ be a function $\mathbb{R}^{2(k-1)} \to \{-1, 1\}^{k-1}$, that takes $A_{1:k-1}, \mu_{1:k-1}$ as input. Denote

$$Q(\mathcal{A}_{1:k-1}, \hat{s}) := \left[\sum_{i=1}^{k-1} \frac{\mathbb{E}_{\mu_{1:k-1}} \mathbb{E}_{\bar{H}_k} \hat{s}_i(\phi_{T_i} - \mu_{T_i})}{\sqrt{k-1}} \right]^2.$$

By the optimality of the previously defined adversary that chooses $\hat{s}_i$ to be the likelihood ratio test, we can rewrite the expression in (10.10) as a variational form

$$\mathcal{R}(k, \sigma^2) \geq \tau \wedge \min_{\mathcal{A}_{1:k-1} \in \mathbb{A}^{\tau}_{1:k-1}} \max_{\hat{s}} Q(\mathcal{A}_{1:k-1}, \hat{s}). \qquad (10.11)$$

What we have achieved so far is a reduction of the problem into a simper minimax problem, where we know the optimal $\hat{s}$ analytically.

It remains to work out the optimal player strategy for the first $k-1$ steps. This is still very challenging as $\mathcal{A}_{1:k-1}$ could have complex dependency structures. We now show that under the specific prior, it suffices for us to consider an even simpler class of player strategies.

Reduction to independent noise adding.

To begin with, define $Z_i = A_i - \phi_{T_i}$. Since Z_i is allowed to depend on everything that A_i depends on including T_i and ϕ_{T_i}, $\mathcal{A}_{1:k-1}$ is equivalent to $\phi_{T_{1:k-1}} + \mathcal{Z}_{1:k-1}$. Without loss of generality, this converts the problem of finding the optimal estimators to the equivalent problem of finding the optimal *dependent* noise adding procedures. We overload the score function Q that takes $\mathcal{Z}_{1:k-1}$ as an input to mean $Q(\phi_{T_{1:k-1}} + \mathcal{Z}_{1:k-1}, \hat{s})$.

Now we will show that it suffices to consider only independent noise adding. The argument is broken into two steps. In Step 1, we show that it suffices to consider Z_i that depends only on ϕ_{T_i} but not everything else. Then in Step 2, we will further show that it suffices to consider independent noise adding that does not even depend on ϕ_{T_i}.

Step 1: We will show that for every feasible $\mathcal{Z}_{1:k-1}$ obeying $\mathbb{E}(\phi_{T_i} + Z_i - \mu_{T_i})^2 \leq \tau$, there is another feasible randomization strategy $\tilde{\mathcal{Z}}_{1:k-1}$ that adds Z_i such that $Z_i \perp\!\!\!\perp$ Everything $\mid \phi_{T_i}$ and $\max_{\hat{s}} Q(\tilde{\mathcal{Z}}_{1:k-1}, \hat{s}) \leq \max_{\hat{s}} Q(\mathcal{Z}_{1:k-1}, \hat{s})$.

Take $\tilde{\mathcal{Z}}_{1:k-1}$ such that $Z_i|\phi_{T_i}$ is identically distributed to that of $\mathcal{Z}_{1:k-1}$ and $Z_i \perp\!\!\!\perp$ Everything $\mid \phi_{T_i}$. This can be obtained by simply marginalizing over all other variables that $\mathcal{Z}_{1:k-1}$ depends on.

Clearly, $\tilde{\mathcal{Z}}_{1:k-1}$ is feasible. It is easy to see that the bias remains unchanged. And the following expansion shows that the variance also remains unchanged.

$$\begin{aligned}\mathrm{Var}(\phi_{T_i} + Z_i - \mu_{T_i}) &= \mathbb{E}\mathrm{Var}(\phi_{T_i} + Z_i - \mu_{T_i}|\phi_{T_i}, T_i) + \mathrm{Var}\mathbb{E}(\phi_{T_i} + Z_i - \mu_{T_i}|\phi_{T_i}, T_i) \\ &= \mathbb{E}\mathrm{Var}(Z_i|\phi_{T_i}) + \mathrm{Var}(\phi_{T_i} - \mu_i + \mathbb{E}(Z_i|\phi_{T_i})).\end{aligned}$$

Now define $\hat{s}' \in \{f : \mathbb{R} \times \mathbb{R} \to \{-1, 1\}\}$ that takes in only A_i and μ_{T_i} instead of all $A_1, ..., A_{k-1}$, $\mu_{T_1,...,T_{k-1}}$.

$$\max_{\hat{s}} Q(\tilde{\mathcal{Z}}_{1:k-1}, \hat{s}) = \max_{\hat{s}'} Q(\tilde{\mathcal{Z}}_{1:k-1}, \hat{s}') = \max_{\hat{s}'} Q(\mathcal{Z}_{1:k-1}, \hat{s}') \leq \max_{\hat{s}} Q(\mathcal{Z}_{1:k-1}, \hat{s})$$

The first equal sign is true because $A_1 \perp\!\!\!\perp A_2 \perp\!\!\!\perp ... \perp\!\!\!\perp A_{k-1}$ regardless of any form of conditioning. The second equal sign is true because $\hat{s}'$ uses only A_i to estimate the sign of $\phi_{T_i} - \mu_{T_i}$ and the distribution of $A_i|\phi_{T_i}, T_i$ remains unchanged by definition of $\tilde{Z}_{1:k-1}$. Then the last inequality holds because $\hat{s}$ is more powerful than $\hat{s}'$.

Step 2: Intuitively, if we are given a perturbation Z that depends on ϕ_{T_i}, it reveals information about ϕ_{T_i} that can be used to the adversary's advantage. We will now formalize the intuition. Thanks to Step 1, we can now consider each summand of $Q(\mathcal{Z}_{1:k-1}, \hat{s})$ separately, that is:

$$\min_{\mathcal{A}_{1:k-1} \in \mathbb{A}^{\tau}_{1:k-1}} \max_{\hat{s}} Q(\mathcal{A}_{1:k-1}, \hat{s}) = \left\{ \frac{1}{\sqrt{k-1}} \sum_{i=1}^{k-1} \min_{\mathcal{Z}_i \in \mathbb{Z}_i^{\tau}} \max_{\hat{s}'_i} \mathbb{E}_{\mu_i} \mathbb{E}_{\phi_i, Z_i} \hat{s}'_i(\phi_i - \mu_i) \right\}^2 =: \left\{ \frac{\sum_{i=1}^{k-1} \tilde{F}_i}{\sqrt{k-1}} \right\}^2 \tag{10.12}$$

and we denote

$$\tilde{F}_i = \min_{\mathcal{Z}_i \in \mathbb{Z}_i^{\tau}} \max_{\hat{s}'_i} \mathbb{E}_{\mu_i} \mathbb{E}_{\phi_i} \hat{s}'_i(\phi_i - \mu_i)$$

where the expectation over μ_{t_i} is defined as Unif$(-M, M)$. We will further drop the subscript i, t_i and the prime in $\hat{s}'$ for notational simplicity. Now let $p_a(\cdot)$ be the probability density of $\mathcal{Z}$

when $\phi = a$. Denote $x := |\phi - \mu|$. We know that x follows the half-normal distribution. By the decision rules of $\hat{s}$ as defined earlier, for each $\mathcal{Z}$, we can write we get:

$$\begin{aligned}
&\max_{\hat{s}} \mathbb{E}_\mu \mathbb{E}_\phi \hat{s}(\phi - \mu) \\
\geq &\int_{-M}^{M} \left\{ \int_0^\infty x \frac{1}{2} \left[\int |p_{\mu-x}(t+x) - p_{\mu+x}(t-x)| dt \right] \frac{\sqrt{2}}{\sqrt{\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx \right\} \frac{1}{2M} d\mu \\
= &\frac{1}{2} \int_0^\infty x \left\{ \int_{-M}^{M} \left[\int |p_{\mu-x}(t+x) - p_{\mu+x}(t-x)| dt \right] \frac{\sqrt{2}}{\sqrt{\pi}\sigma} \frac{1}{2M} d\mu \right\} e^{-\frac{x^2}{2\sigma^2}} dx \\
\geq &\frac{1}{2} \int_0^\infty x \left[\int \left| \int_{-M}^{M} \left(p_{\mu-x}(t+x) - p_{\mu+x}(t-x) \right) \frac{1}{2M} d\mu \right| dt \right] \frac{\sqrt{2}}{\sqrt{\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx.
\end{aligned}$$

where the last line follows from Jensen's inequality. Denote $\int |f(t)| dt$ by L1-norm $\|f\|_1$, the above expression can be rewritten compactly as

$$\begin{aligned}
&\frac{1}{2} \mathbb{E}_x \left[x \| \mathbb{E}_\mu (p_{\mu-x}(\cdot + x) - p_{\mu+x}(\cdot - x)) \|_1 \right] \\
= &\frac{1}{2} \mathbb{E}_x \left[x \| \mathbb{E}_\mu (p_{\mu-x}(\cdot + x) - p_\mu(\cdot + x) + p_\mu(\cdot + x) - p_\mu(\cdot - x) + p_\mu(\cdot - x) - p_{\mu+x}(\cdot - x)) \|_1 \right] \\
\geq &\frac{1}{2} \mathbb{E}_x \left[x \| \mathbb{E}_\mu p_\mu(\cdot + x) - \mathbb{E}_\mu p_\mu(\cdot - x) \|_1 \right] \\
&- \frac{1}{2} \mathbb{E}_x \left[x \| \mathbb{E}_\mu p_{\mu-x}(\cdot + x) - \mathbb{E}_\mu p_\mu(\cdot + x) \|_1 \right] - \frac{1}{2} \mathbb{E}_x \left[x \| \mathbb{E}_\mu p_\mu(\cdot - x) - \mathbb{E}_\mu p_{\mu+x}(\cdot - x)) \|_1 \right]
\end{aligned}$$

Recall that μ is a uniform distribution from $[-M, M]$. By Holder's inequality,

$$\left\| \mathbb{E}_\mu (p_{\mu-x}(\cdot)) - \mathbb{E}_\mu (p_\mu(\cdot)) \right\|_1 \leq \frac{x}{M}.$$

It follows that

$$\mathbb{E}_x \left[x \left\| \mathbb{E}_\mu p_{\mu-x}(\cdot + x) - \mathbb{E}_\mu p_\mu(\cdot + x) \right\|_1 \right] \leq \frac{1}{M} \mathbb{E}_x \left[x^2 \right] \leq \frac{\sigma^2}{M}.$$

This ensures that the two negative terms can both be arbitrarily small as we take $M \to \infty$.

Now, let $\mathcal{Z}'$ be the "smoothed" distribution such that Z has distribution $q(t) = \mathbb{E}_\mu p_\mu(t)$. Clearly, $\mathcal{Z}'$ is feasible, it does not depend on ϕ and

$$\mathbb{E}_x \left[x \| \mathbb{E}_\mu p_\mu(\cdot + x) - \mathbb{E}_\mu p_\mu(\cdot - x) \|_1 \right] = \mathbb{E}_x \left[x \| q(\cdot + x) - q(\cdot - x) \|_1 \right].$$

Since for each $\mathcal{Z}$ we can construct this $\mathcal{Z}'$ that does not depend on ϕ and has the same objective value, therefore it suffices to consider only independent noise adding. Lastly, we can also offset $\mathcal{Z}'$ such that the noise it adds is zero-mean without affecting the feasibility and the objective value.

To put things together, we get

$$\tilde{F} \geq \min_{q \text{ s.t. } \mathbb{E}Z = 0, \text{Var}(Z) \leq \tau} \underbrace{\frac{1}{2} \mathbb{E}_x \left[x \left\| q(\cdot + x) - q(\cdot - x) \right\|_1 \right]}_{(\dagger)} - \frac{\sigma^2}{M} \tag{10.13}$$

where the minimization is over the distribution q of the zero-mean independent noise.

Near optimal obfuscation of a Bayes classifier.

It remains to lower bound $(\dagger)$ using Lemma 10.16 following lemma (which we defer the proof to later)

Lemma 10.16. *Let signal random variable X and noise random variable Z be such that $X \sim \mathcal{N}(0, \sigma^2)$, $\mathbb{E}Z = 0$, $\mathrm{Var}(Z) \leq w^2$. Let Y be the half-normal distribution with parameter σ^2. Let $\hat{s}$ be the optimal Bayes classifier of $\mathrm{sign}(X)$ by observing $X + Z$, then*

$$\mathbb{E}_{X,Z}(\hat{s}X) = \frac{1}{2}\mathbb{E}_Y\left[Y \left\|q(\cdot + Y) - q(\cdot - Y)\right\|_1\right]$$

and

$$\mathbb{E}_{X,Z}(\hat{s}X) \geq \begin{cases} \frac{\sigma^2}{\sqrt{3}w} - \frac{\sigma^4}{2\sqrt{3}w^3} & \textit{when } w^2 \geq \sigma^2, \\ \frac{\sigma}{2\sqrt{3}} & \textit{when } w^2 < \sigma^2. \end{cases} \tag{10.14}$$

Moreover,

$$\lim_{\frac{\sigma^2}{w} \to 0} \frac{w}{\sigma^2}\mathbb{E}_{X,Z}(\hat{s}X) \geq \frac{1}{\sqrt{3}}$$

with equal sign attained by uniform distribution $U([-\sqrt{3}w, \sqrt{3}w])$.

Recall that we can take τ to be any value. For $\tau \geq \sigma^2$, (10.14) implies that

$$(\dagger) \geq \frac{\sigma^2}{2\sqrt{3\tau}}. \tag{10.15}$$

Putting everything together

When $k = 1$, there is no dependence in T and ϕ_T. The minimax risk of estimating a normal mean suggests that $F_k \geq \sigma^2$.

When $k > 1$, $\tau = \sqrt{k-1}\sigma^2 \geq \sigma^2$, combine (10.11), (10.12), (10.13), (10.15), we get

$$\begin{aligned} \mathcal{R}(k, \sigma^2) &\geq \left[\frac{1}{\sqrt{k-1}}\sum_{i=1}^{k-1}\left(\frac{\sigma^2}{2\sqrt{3\tau}} + \frac{\sigma^2}{M}\right)\right]^2 \\ &\geq \left[\frac{\sqrt{(k-1)}\sigma^2}{12} + \frac{(k-1)\sigma^4}{M^2} + \frac{2(k-1)^{3/4}\sigma^3}{\sqrt{3}M}\right]. \end{aligned}$$

Note that this bound holds for all $M > 0$. Take $M \to \infty$, we get

$$\mathcal{R}(k, \sigma^2) \geq \frac{\sqrt{k-1}\sigma^2}{12}.$$

In addition, as $k \to \infty$ This completes the proof.

Lower bound for the sequential projections of Gaussian vectors

Proof of Corollary 10.11. We will mostly reuse arguments in the sections before, but also make important changes to address issues specific to this problem.

For $i = 1, ..., k-1$, the adversary will choose $t_i = e_i$ the standard basis for "exploration". Then in the kth step, the adversary will "exploit" by estimating the signs of $X - \theta_1$ using the likelihood ratio test to obtain $\hat{s}$ and then choose $T_k = [0_d; \hat{s}/\sqrt{k-1}; 0_{d-k+1}]$.

We will also consider the following prior distribution:

$$\theta_1 \sim \text{Unif}(-M, M), \quad \theta_2 \sim \mathcal{N}(0, M^2 I), \quad \mathbb{P}(v=1) = \mathbb{P}(v=-1) = 0.5.$$

Constraining $\mathcal{A}_{1:k-1}$ to outputting an $A_{1:k-1}$ such that $\max_{i\in[k-1]} \mathbb{E}[(A_i - \langle T_i, \theta\rangle)^2 \leq w^2]$, if w^2 is greater than the minimax risk, then this restriction is without loss of generality; otherwise we can lower bound the minimax risk using

$$\begin{aligned}&\inf_{\mathcal{A}_{1:k}} \sup_{\theta_1,\theta_2,v\in\{-1,1\}} \sup_{\mathcal{W}_{1:k}} \max_{i\in[k]} \mathbb{E}[(A_i - \langle (X,X'), T_i\rangle)^2] \\ \geq &\min\{w^2, \underbrace{\inf_{\mathcal{A}_{1:k-1}\in\mathbb{A}^w_{1:k-1},\mathcal{A}_k} \sup_{\theta_1,\theta_2,v\in\{-1,1\}} \sup_{\mathcal{W}_{1:k}} \mathbb{E}[(A_k - \langle (X,X'), T_k\rangle)^2]}_{(*)}\}\end{aligned}$$

We will choose w^2 later and focus on lower bounding the second term by choosing the specific adversary and a specific prior, which gives us

$$(*) \geq \inf_{\mathcal{A}_{1:k-1}\in\mathbb{A}^w_{1:k-1},\mathcal{A}_k} \mathbb{E}_{\theta_1,\theta_2,v\in\{-1,1\}} \mathbb{E}[(A_k - \langle (X,X'), T_k\rangle)^2]$$

Also, we reveal θ_1 to $\mathcal{A}_k$ (to make it an even stronger estimator) to break the dependence of $\mu_k = \langle T_k, (X, X')\rangle$ to A_k, so we can more easily calculate the posterior distribution.

First check that

$$\mu_k = \frac{\langle \hat{s}_{1:k-1}, \theta'\rangle}{\sqrt{k-1}} = \frac{\langle \hat{s}_{1:k-1}, X'\rangle}{\sqrt{k-1}} - v\frac{\langle \hat{s}_{1:k-1}, X-\theta_1\rangle}{\sqrt{k-1}} = \phi_{T_k} - vb$$

where we denote $b := \frac{\langle \hat{s}_{1:k-1}, X-\theta_1\rangle}{\sqrt{k-1}}$. From the equation, it is clear that is not known to the player is v.

It follows that the posterior distribution

$$\mathbb{P}(\mu_k | X, X', A_{1:k-1}, T_{1:k-1}, T_k, \theta_1) = \mathbb{P}(\mu_k | b, \phi_{T_k}) = \sum_{v\in\{-1,1\}} \mathbb{P}(\mu_k, v | b, \phi_{T_k})$$

where

$$\mathbb{P}(\mu_k, v | b, \phi_{T_k}) \propto p(b) p(\phi_{T_k} | v, \mu_k) \pi(v) \pi(\mu_k) = \begin{cases} 0.5\pi(\mu_k = \phi_{T_k} - b | \phi_{T_k}) & \text{if } v = 1 \text{ and } \mu_k = \phi_{T_k} - b \\ 0.5\pi(\mu_k = \phi_{T_k} + b | \phi_{T_k}) & \text{if } v = -1 \text{ and } \mu_k = \phi_{T_k} + b \end{cases}$$

Using conjugacy of the prior, we have $\mu_k|\phi_{T_k} \sim \mathcal{N}\left(\frac{M^2\phi_{T_k}}{\sigma^2+M^2}, (1/M^2 + 1/\sigma^2)^{-1}\right)$.

$$\frac{\pi(\mu_k = \phi_{T_k} - b|\phi_{T_k})}{\pi(\mu_k = \phi_{T_k} + b|\phi_{T_k})} = e^{-(\frac{\sigma^2\phi_{T_k}}{\sigma^2+M^2}+b)^2/(2\sigma^2)+(\frac{\sigma^2\phi_{T_k}}{\sigma^2+M^2}-b)^2/(2\sigma^2)} = e^{-\frac{2b\sigma^2\phi_{T_k}}{(\sigma^2+M^2)\sigma^2}}$$

Note that $\phi_{T_k} = O_{\mathbb{P}}(M)$ and $b = O_{\mathbb{P}}(\sqrt{k-1}\sigma)$, so as $M \to \infty$

$$\mathbb{E}(\mu_k|X, X', A_{1:k-1}, T_{1:k}, \theta_1) \to \frac{\langle \hat{s}, X' \rangle}{k-1}.$$

More formally, the optimal $A_k = \mathbb{E}(\mu_k|X, X', A_{1:k-1}, T_{1:k})$ obeys that

$$\begin{aligned}\mathbb{E}[(A_k - \mu_k)^2] \geq& \mathbb{E}\left[(\mathbb{E}(\mu_k|X, X', A_{1:k-1}, T_{1:k}, \theta_1) - \mu_k)^2\right] \\ &\geq \Big(1 - \mathbb{P}(|\phi_{T_k}| > M\log(M)) - \mathbb{P}(|b| > \sqrt{k-1}\sigma\log(M))\Big) \\ &\quad \mathbb{E}\left[b^2 \middle| |b| < \sqrt{k-1}\sigma\log(M), |\phi_{T_k}| < M\log(M)\right]\end{aligned}$$

Take $\lim_{M\to\infty}$ on both sides, we get

$$\lim_{M\to\infty} \mathbb{E}[(A_k - \mu_k)^2] = \mathbb{E}[b^2] = \frac{\mathbb{E}[(\langle \hat{s}, X' - \theta_2 \rangle)^2]}{k-1} = \frac{\mathbb{E}[(\langle \hat{s}, X - \theta_1 \rangle)^2]}{k-1}$$

The problem thus reduces to:

$$(*) \geq \inf_{\mathcal{A}_{1:k-1} \in \mathbb{A}^w_{1:k-1}} \frac{\mathbb{E}[(\langle \hat{s}, X - \theta_1 \rangle)^2]}{k-1} \geq \inf_{\mathcal{A}_{1:k-1} \in \mathbb{A}^w_{1:k-1}} \frac{(\mathbb{E}[\langle \hat{s}, X - \theta_1 \rangle])^2}{k-1}$$

This gets us to (10.11). The remainder of the proof is exactly the same as that of Theorem 10.10 as in Section 10.7.1, which involves reduction to zero-mean independent noise adding using a "smoothing trick" via the uniform prior distribution of θ_1, and then invoking Lemma 10.16. Then the proof is complete by choosing $w = \sqrt{k-1}\sigma$. □

10.7.2 Part (b) fixed distribution

In this section, we present a stronger lower bound construction that works for a class of fixed distribution. The difficulty is that now the estimator also knows the underlying distribution, so we need to somehow prevent the triviality that $A_{1:k} = \mu_{T_{1:k}}$ for any chosen $T_{1:k}$. For this matter, we restrict our estimator A_i to be "natural", which at step i uses only $\phi_{T_{1:i}}, A_{1:i-1}, T_{1:k-1}$ as well as the distributional parameters $\mu_{T_{1:i-1}}, \Sigma_{T_{1:i-1}}$ of previously answered statistics.

The idea is construct adversary that chooses randomized $T_1, ..., T_k$, to simulate the same randomization as in the previous section, which is possible under the richness assumption.

Specifically, let $\mathcal{W}_{1:k-1}$ pick randomized $T_{1:k-1}$ such that $\phi_{T_{1:k-1}}(X) \sim \mathcal{N}(\mu_{T_{1:k-1}}, \sigma^2 I)$, namely, $\phi_{T_{1:k-1}}$ are independent and their mean distributed uniformly between $[-M, M]$.

And we take $\mathcal{W}_k$ to be the following algorithm:

1. Observing $A_1, A_2, ..., A_{k-1}$ and $\mu_1, ..., \mu_{k-1}$, $\mathcal{W}_k$ uses an optimal classifier to infer the signs of $\phi_{T_i} - \mu_i$. Specifically. likelihood ratio test that output $\hat{s}_i = 1$ if

$$\log \frac{\mathbb{P}(A_1, ..., A_{k-1} | \phi_{T_i} - \mu_{T_i} > 0)}{\mathbb{P}(A_1, ..., A_{k-1} | \phi_{T_i} - \mu_{T_i} < 0)} \geq 0$$

   and $\hat{s}_i = -1$ otherwise. Let the estimated sign vector be $\hat{s}$.

2. Then $\mathcal{W}_k$ chooses T_k such that $\mathrm{Var}(\phi_{t_k}) = \sigma^2$ and the covariance vector with known realized $T_{1:k-1}$ being $v = \frac{\sigma^2}{\sqrt{k-1}} \hat{s}$ with probability 0.5 and $-v$ with probability the other 0.5. At the same time the distribution of μ_{T_k} induced by the distribution of T_k is uniform on $[-M', M']$.

   Hence from the player's perspective, it is impossible to offset the bias in ways that depend only on ϕ_{T_k}, and we will show that the optimal strategy for the player is to just output ϕ_{T_k}.

The argument to show that $A_k = \phi_{T_k}$ is somewhat simpler than previously, because now T_k is chosen conditioned the first $k-1$ rounds, so we do not have to calculate the posterior distribution of the model parameters, but rather can specify the posterior prior distributions through T_k directly.

$$\mathbb{E}(A_k - \mu_{T_k})^2 = \mathbb{E}_{\bar{H}_k} \mathbb{E}((A_k - \mu_{T_k})^2 | \bar{H}_k)$$

Now the second part becomes an estimation problem with a known prior and Bayes-estimator (the posterior mean) is optimal. In other word, the Bayes estimator

$$A_k^* = \mathbb{E}(\mu_{T_k} | \phi_{T_k}, \bar{H}_k).$$

$$p(\mu_{T_k} | \phi_{T_k}, \bar{H}_k) \propto p(\phi_{T_k} | \mu_{T_k}, \bar{H}_k) p(\mu_{T_k} | \bar{H}_k)$$

Let $b := |\mathbb{E}(\phi_{T_k} - \mu_{T_k} | \bar{H}_k)|$. If $\mu_{T_k} \in [-M + b, M - b]$

$$\mathbb{P}(\phi_{T_k} = \mu_{T_k} - b | \mu_{T_k}, \bar{H}_k) = \mathbb{P}(\phi_{T_k} = \mu_{T_k} + b | \mu_{T_k}, \bar{H}_k) = 0.5$$

and

$$p(\mu_{T_k} | \bar{H}_k) = p(\mu_{T_k} | \bar{H}_k) = 1/(2M).$$

and the corresponding posterior mean is ϕ_{T_k}. When $\mu_{T_k} \notin [-M' + b, M' - b]$, which happens with probability b/M' conditioned on $\bar{H}_k$, we can simply lower bound the risk by 0. This gives us

$$\begin{aligned}
\mathbb{E}(A_k - \mu_{T_k})^2 &= \mathbb{E}_{\bar{H}_k} \mathbb{E}((A_k - \mu_{T_k})^2 | \bar{H}_k) \\
&\geq \mathbb{E}_{\bar{H}_k} (1 - b/M') \mathbb{E}((\phi_{T_k} - \mu_{T_k})^2 | \bar{H}_k) = \mathbb{E}_{\bar{H}_k} (1 - b/M') b^2 \\
&= \mathbb{E}_{\bar{H}_k} b^2 - \mathbb{E}_{\bar{H}_k} b^3 / M'.
\end{aligned}$$

Note that

$$b = \left| \frac{1}{\sqrt{k-1}} \hat{s}^T (\phi_{T_{1:k-1}} - \mu_{T_{1:k-1}}) \right| = \frac{1}{\sqrt{k-1}} \sum_{i=1}^{k-1} \hat{s}_i (\phi_{T_i} - \mu_{T_i}).$$

This is the mean of $k-1$ i.i.d. half-normal distribution. Since half-normal is sub-gaussian, therefore by the moment bound of subgaussian, we get $\mathbb{E}b^3 \leq O[(k-1)^{3/2}\sigma^3]$, therefore it suffices to take $M' = \Omega(k\sigma)$ to match the conjectured minimax risk of $O(\sqrt{k}\sigma^2)$.

Since we can simulate a uniform distribution of μ_{T_i} in $[-M, M]$ by choosing T_i, using exactly the same argument as previously, we can show that it suffices to consider only independent noise adding to $\phi_{T_{1:k-1}}$ in the case with fixed distributions.

10.7.3 Proof of Lemma 10.16

Proof. $\mathbb{E}\hat{s}X$, the absolute margin risk of a classifier $\hat{s}$, is a function of the noise distribution p. For example, if $p = 0$, $\hat{s} \equiv \mathrm{sign}(X)$ therefore $\mathbb{E}\hat{s}X = \mathbb{E}|X|$, if p is normal with variance $w^2 \to \infty$, $\hat{s}$ is independent to the signs of X, therefore this quantity converges to 0. The specific shape of p also matters, e.g., adding Bernoulli noise with $w^2 \to \infty$ yields $\mathbb{E}\hat{s}X = \mathbb{E}|X|$. The idea of the proof is to formulate an optimization problem that minimizes $\mathbb{E}\hat{s}X$ over the class of all p and then try to solve it analytically.

To begin with, we first express $\mathbb{E}\hat{s}X$ as the L_1 norm of a linear transformation of p. Decompose X into sign s and magnitude t, where $\mathbb{P}(s=1) = \mathbb{P}(s=-1) = 0.5$ and t is drawn from a half-normal distribution which we denote by q.

$$\begin{aligned}
&\mathbb{E}_Z\mathbb{E}_X\hat{s}X = \mathbb{E}_Z\mathbb{E}_s\mathbb{E}_t ts\hat{s}\\
=&\mathbb{E}_Z \sum_{s\in\{-1,1\}} 0.5\mathbb{E}_t ts \;\mathrm{sign}[\mathbb{E}_{t'}\mathbb{P}(A|s'>0,t') - \mathbb{E}_{t'}\mathbb{P}(A|s'<0,t') > 0]\\
=&0.5\mathbb{E}_Z\mathbb{E}_t t1_{\{t+Z\in E_1\}} - 0.5\mathbb{E}_Z\mathbb{E}_t t1_{\{t+Z\in E_2\}} - 0.5\mathbb{E}_Z\mathbb{E}_t t1_{\{-t+Z\in E_1\}} + 0.5\mathbb{E}_Z\mathbb{E}_t t1_{\{-t+Z\in E_2\}}\\
=&0.5\mathbb{E}_t\mathbb{E}_{A|t,s=1} t1_{\{A\in E_1\}} - 0.5\mathbb{E}_t\mathbb{E}_{A|t,s=1} t1_{\{A_1\in E_2\}} - 0.5\mathbb{E}_t\mathbb{E}_{A|t,s=-1} t1_{\{A\in E_1\}} + 0.5\mathbb{E}_t\mathbb{E}_{A|t,s=-1} t1_{\{A\in E_2\}}\\
=&0.5\int_z\left[\int_t t(p(z-t)-p(z+t))q(t)dt\right]_+ dz + 0.5\int_z\left[\int_t t(-p(z-t)+p(z+t))q(t)dt\right]_+ dz\\
=&0.5\int_z\left|\int_t t(p(z-t)-p(z+t))q(t)dt\right| dz = 0.5\int_z \left|\mathbb{E}_t t(p(z-t)-p(z+t))\right| dz \qquad (10.16)
\end{aligned}$$

where in Line 3 and 4, we use E_1 to denote the event of A such that $\mathrm{sign}[\mathbb{E}_{t'}\mathbb{P}(A|s'>0,t') - \mathbb{E}_{t'}\mathbb{P}(A|s'<0,t') > 0] = 1$ and $E_2 = E_1^c$. Note that E_1 and E_2 are events in the σ-field of observation A induced only by the σ-field of Z (since X is integrated out).

Consider the following variational optimization problem over distribution p that is 0-mean and has variance bounded by w^2.

$$\begin{aligned}
&\min_p \int |\mathbb{E}_t tp(x+t) - \mathbb{E}_t tp(x-t)|dx\\
&\text{s.t. } p \text{ is a probability distribution defined on } \mathbb{R},\\
&\qquad \mathrm{Var}(Z) \leq w^2, \mathbb{E}(Z) = 0 \text{ for } Z \sim p.
\end{aligned} \qquad (10.17)$$

where t distributes as half-normal distribution with parameter σ.

Define operator A such that $Ap = \int_t t[p(x+t) - p(x-t)]\frac{\sqrt{2}}{\sigma\sqrt{\pi}}e^{-\frac{t}{2\sigma^2}}dt$. The objective can be rewritten as $\|Ap\|_1$. A is a linear operator, all constraints are affine in p, therefore this is a convex optimization problem, which we rewrite in standard form below:

$$
\begin{aligned}
&\min_{\boldsymbol{p}} \|A\boldsymbol{p}\|_1 \\
&\text{s.t. } \langle \boldsymbol{x}^2, \boldsymbol{p}\rangle \leq w^2, \quad -\boldsymbol{p} \leq 0 \\
&\qquad \langle \mathbf{1}, \boldsymbol{p}\rangle = 1, \quad \langle \boldsymbol{x}, \boldsymbol{p}\rangle = 0.
\end{aligned}
\tag{10.18}
$$

The Lagrangian and the corresponding dual problem are

$$
L(\boldsymbol{p}, u_1, \boldsymbol{u}_2, v_1, v_2) = \|A\boldsymbol{p}\|_1 + u_1(\langle \boldsymbol{x}^2, \boldsymbol{p}\rangle - w^2) - \langle \boldsymbol{u}_2, \boldsymbol{p}\rangle + v_1(1 - \langle \mathbf{1}, \boldsymbol{p}\rangle) + v_2\langle \boldsymbol{x}, \boldsymbol{p}\rangle,
$$

$$
\begin{aligned}
&\max_{u1, \boldsymbol{u}_2, v_1, v_2, C} - u_1 w^2 + v_1 \\
&\quad \text{s.t. } \|A^{-1}(-u_1\boldsymbol{x}^2 + \boldsymbol{u}_2 + v_1\mathbf{1} - v_2\boldsymbol{x}) + C\|_\infty \leq 1 \\
&\qquad \boldsymbol{u}_2 \geq \mathbf{0}, \quad u_1 \geq 0.
\end{aligned}
\tag{10.19}
$$

and by the definition of the Lagrange dual, the corresponding dual objective value for any feasible dual variables will be a lower bound of the primal optimal solution, and our proof involves constructing one "nearly optimal" feasible dual solution. In the derivations below, please refer to Figure 10.4 for illustrations.

From Figure 10.4(a), we can see that the linear operator A is closely related to the differentiation operator. Correspondingly, A^{-1} is closely related to indefinite integral operator. Using the moment properties of the half-normal distribution and simple calculus, we derive a few properties about A and A^{-1} when applied to polynomials (see the derivation in the next section).

$$
\begin{aligned}
&A\mathbf{1} = \mathbf{0}, && \\
&A\boldsymbol{x} = 2\sigma^2\mathbf{1}, && A^{-1}\mathbf{1} = \frac{1}{2\sigma^2}\boldsymbol{x} + C, \\
&A\boldsymbol{x}^2 = 4\sigma^2\boldsymbol{x}, && A^{-1}\boldsymbol{x} = \frac{1}{4\sigma^2}\boldsymbol{x}^2 + C, \\
&A\boldsymbol{x}^3 = 6\sigma^2\boldsymbol{x}^2 + 6\sigma^4\mathbf{1}, && A^{-1}\boldsymbol{x}^2 = \frac{1}{6\sigma^2}\boldsymbol{x}^3 - \frac{1}{2}\boldsymbol{x} + C.
\end{aligned}
$$

where C is an arbitrary constant. It follows that

$$
\begin{aligned}
&A^{-1}(-u_1\boldsymbol{x}^2 + \boldsymbol{u}_2 + v_1\mathbf{1} - v_2\boldsymbol{x}) \\
&= -\frac{1}{6\sigma^2}u_1\boldsymbol{x}^3 + \frac{-1}{4\sigma^2}v_2\boldsymbol{x}^2 + \left(\frac{1}{2}u_1 + \frac{1}{2\sigma^2}v_1\right)\boldsymbol{x} + A^{-1}\boldsymbol{u}_2 + C. \\
&= \int(-\nu_0\boldsymbol{x}^2 + \nu_1\boldsymbol{x} + \nu_2)dx + A^{-1}\boldsymbol{u}_2 + C.
\end{aligned}
\tag{10.20}
$$

where $\nu_0 = \frac{1}{2\sigma^2}u_1, \nu_1 = -\frac{1}{2\sigma^2}v_2, \nu_2 = \frac{1}{2}u_1 + \frac{1}{2\sigma^2}v_1$. The only restriction of ν_0 is non-negativity, and ν_1 and ν_2 can be arbitrary due to the flexibility of of v_1 and v_2.

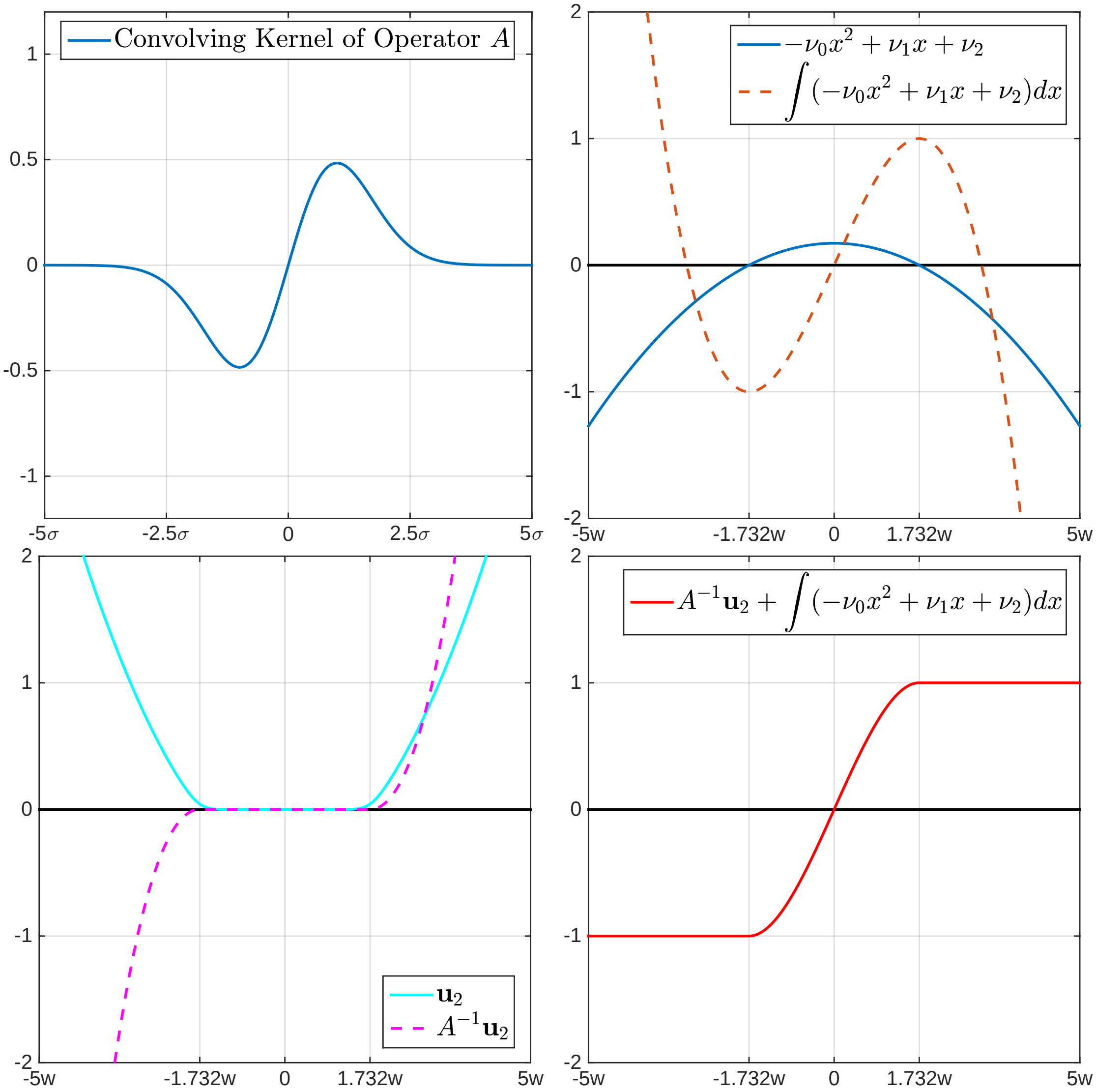


Figure 10.4: Illustrations of our construction of the dual functions. (a) illustrates the operator A, which is essentially a convolution with the shown kernel. (b) shows our constructions of quadratic function f_{ν_0,ν_1,ν_2} and its indefinite integral. (c) shows our construction of $g = A^{-1}\boldsymbol{u}_2$ and the corresponding nonnegative dual function $\boldsymbol{u}_2$. (d) illustrates the that the the ℓ_∞-norm constraint is satisfied.

Let $f_{\nu_0,\nu_1,\nu_2}(x) = -\nu_0 x^2 + \nu_1 x + \nu_2$. Take ν_0, ν_1, ν_2 such that $f_{\nu_0,\nu_1,\nu_2}(x) \geq 0$ between $[-\sqrt{3}w, \sqrt{3}w]$, and

$$\int_{-\sqrt{3}w}^{\sqrt{3}w} (-\nu_0 x^2 + \nu_1 x + \nu_2) dx = 2.$$

The coefficients that satisfy these constraints are

$$\nu_0 = \frac{1}{2\sqrt{3}w^3}, \quad \nu_1 = 0, \quad \nu_2 = \frac{\sqrt{3}}{2w},$$

which correspond to

$$u_1 = \frac{\sigma^2}{\sqrt{3}w^3}, \quad v_2 = 0, \quad v_1 = \frac{\sqrt{3}\sigma^3}{w} - \frac{\sigma^4}{\sqrt{3}w^3}. \tag{10.21}$$

Check that these are feasible in (10.19).

Moreover,

$$F_{\nu_0,\nu_1,\nu_2}(x) := \int f_{\nu_0,\nu_1,\nu_2}(x) dx = -\frac{x^3}{6\sqrt{3}w^3} + \frac{\sqrt{3}x}{2w}.$$

Define function g, where

$$g(x) = \begin{cases} F_{\nu_0,\nu_1,\nu_2}(x) + 1 & \text{when } x \leq -\sqrt{3}w \\ F_{\nu_0,\nu_1,\nu_2}(x) - 1 & \text{when } x \geq \sqrt{3}w \\ 0 & \text{otherwise.} \end{cases}$$

g is a monotonically increasing function, therefore taking $\boldsymbol{u}_2 = Ag$ obeys $\boldsymbol{u}_2 \geq 0$. Check that

$$\int f_{\nu_0,\nu_1,\nu_2}(x) dx + A^{-1}\boldsymbol{u}_2 = -\frac{x^3}{6\sqrt{3}} + \frac{\sqrt{3}x}{2w} + g(x) = \begin{cases} -1 & \text{when } x \leq -\sqrt{3}w \\ -\frac{x^3}{6\sqrt{3}w^3} + \frac{\sqrt{3}x}{2w} & \text{when } -\sqrt{3}w \leq x \leq \sqrt{3}w \\ 1 & \text{when } x \geq \sqrt{3}w \end{cases},$$

therefore obeys the first constraint in (10.19). Together with (10.21), we form $(u_1, \boldsymbol{u}_2, v_1, v_2)$ which is a feasible dual solution and the primal optimal solution p^* obeys

$$\|Dp^*\|_1 \geq \frac{2\sigma^2}{\sqrt{3}w} - \frac{\sigma^4}{\sqrt{3}w^3}.$$

This bound is sharp when $w \gg \sigma$, but becomes meaningless when $w^2 < \sigma^2$. We note that $\|Dp^*\|_1$ is a monotonically decreasing function in w^2, therefore the case when $w^2 = \sigma^2$ gives a lower bound for the case when $w^2 \leq \sigma^2$, therefore for any w, we can write

$$\|Ap^*\|_1 \geq \begin{cases} \frac{2\sigma^2}{\sqrt{3}w} - \frac{\sigma^4}{\sqrt{3}w^3} & \text{when } w^2 \geq \sigma^2, \\ \frac{\sigma}{\sqrt{3}} & \text{when } w^2 < \sigma^2. \end{cases}$$

Combine with (10.16), we get our first claim.

Now we move on to work on the second claim where $\sigma^2/w \to 0$. This is equivalent to solving the problem when w is fixed and $\sigma \to 0$, since we can rescale the real line accordingly. As $\sigma \to 0$, $\frac{A}{2\sigma^2}$ converges to $\frac{\partial(\cdot)}{\partial x}$. We divide the objective of (10.19) by $2\sigma^2$. At the limit, the KKT condition of (10.19) becomes

$$\begin{Bmatrix} \int(-u_1\boldsymbol{x}^2 + \boldsymbol{u}_2 + v_1\mathbf{1} - v_2\boldsymbol{x})dx + C \in \partial\|\cdot\|_1(\partial_x\boldsymbol{p}), \\ u_1 \geq 0, \boldsymbol{u}_2 \geq 0, \\ \boldsymbol{p} \text{ is a zero-mean distribution,} \\ u_1(\langle \boldsymbol{x}^2, \boldsymbol{p}\rangle - w^2) = 0, \\ \boldsymbol{p}(x)\boldsymbol{u}_2(x) = 0 \text{ for every } x \in \mathbb{R}, \end{Bmatrix}$$

where the subgradient of the ℓ_1-norm is

$$\partial\|\cdot\|_1(\partial_x\boldsymbol{p}) = \begin{cases} -1 & \text{when } (\partial_x\boldsymbol{p})(x) < 0 \\ 1 & \text{when } (\partial_x\boldsymbol{p})(x) > 0 \\ [-1,1] & \text{Otherwise.} \end{cases}$$

Now we will construct a set of dual variables $(u_1, \boldsymbol{u}_2, v_1, v_2)$ so that they satisfy the KKT condition with

$$\boldsymbol{p}(x) = \begin{cases} \frac{1}{2\sqrt{3}w} & \text{when } x \in [-\sqrt{3}w, \sqrt{3}w] \\ 0 & \text{otherwise.} \end{cases}$$

First of all, $p(x)$ is a valid zero-mean distribution and $\langle \boldsymbol{x}^2, p\rangle = w^2$.

$$\partial_x\boldsymbol{p}(x) = \begin{cases} -\infty & \text{when } x = -\sqrt{3}w \\ +\infty & \text{when } x = \sqrt{3}w \\ 0 & \text{otherwise.} \end{cases}$$

Now consider the range $x \in [-\sqrt{3}w, \sqrt{3}w]$, where $\boldsymbol{u}_2(x) = 0$. $f_{u_1,v_1,v_2}(x) = -u_1\boldsymbol{x}^2 + v_1\mathbf{1} - v_2\boldsymbol{x}$ is the standard form of a quadratic function, and by $u_1 \geq 0$, this is a concave quadratic function. As we did earlier, we choose the parameter of this quadratic function such that

$$\begin{cases} f_{u_1,v_1,v_2}(-\sqrt{3}w) = 0, \\ f_{u_1,v_1,v_2}(\sqrt{3}w) = 0, \\ \int_{-\sqrt{3}w}^{\sqrt{3}w}(-u_1\boldsymbol{x}^2 + v_1\mathbf{1} - v_2\boldsymbol{x})dx = 2. \end{cases}$$

This is always feasible because as u_1 goes from 0 to ∞, the area under the curve also continuously and monotonically increases to ∞. Now, choosing $C = 1$ ensures that we have $-1 \leq (\int f_{u_1,v_1,v_2}(x)dx + C) \leq 1$, $f_{u_1,v_1,v_2}(-\sqrt{3}w) = -1$ and $f_{u_1,v_1,v_2}(\sqrt{3}w) = 1$.

When we have anything outside $[-\sqrt{3}w, \sqrt{3}w]$, $f_{u_1,v_1,v_2}(x) \leq 0$ and taking $u_2(x) = -f_{u_1,v_1,v_2}(x)$ allows function $\boldsymbol{u}_2 + f_{u_1,v_1,v_2}$ to stay at 0, which checks he stationarity condition. Therefore, the given dual variables certify that the proposed uniform $\boldsymbol{p}$ is optimal. The objective value $\lim_{\frac{\sigma^2}{w}\to 0}\frac{w}{\sigma^2}\int|\mathbb{E}_t t(p(x+t)-p(x-t))|dx=\frac{2}{\sqrt{3}}$. The proof is complete by substituting the quantity into (10.16) (divide by 2). □

10.8 Proof of the bound for data splitting-based estimators

Proof of Proposition 10.14. The proof is organized as follows. We will first describe the adversary $\mathcal{W}_{1:k}$ we construct then lower bound the risk by searching the optimal player strategy under the above model.

Constructing $\mathcal{W}_{1:k}$ Now we are ready construct a specific adversary. Let e_i be the standard basis of $\mathbb{R}^d$. The adversary chooses $T_1=e_1$ and for $i>1$

$$T_i=\frac{T_i^{\text{explore}}}{\sqrt{2}}+v\frac{T_i^{\text{exploit}}}{\sqrt{2}},$$

where $T_i^{\text{explore}}=e_i$ is the "exploration" term and $T_i^{\text{exploit}}=\frac{\hat{s}_{1:i-1}}{\sqrt{i-1}}$ is the "exploitation" term and $\mathbb{P}(v=-1)=\mathbb{P}(v=1)=0.5$.

Just like before, $\hat{s}_i$ is the output of an optimal classifier of the sign of

$$\left\langle\bar{X}^{J_i},e_i\right\rangle-\theta_i:=\bar{X}_i^{J_i}-\theta_i \tag{10.22}$$

using the observation

$$A_i=\frac{1}{\sqrt{2}}\bar{X}_i^{J_i}+\frac{1}{\sqrt{2}}\langle\bar{X}^{J_i},T_i^{\text{exploit}}\rangle+Z_i.$$

In other word, the adversary chooses

$$\hat{s}_i=\begin{cases}1 & \text{if } \mathbb{P}\left(A_i\middle|\bar{X}_i^{J_i}-\theta_i\geq 0\right)\geq\mathbb{P}\left(A_i\middle|\bar{X}_i^{J_i}-\theta_i<0\right)\\ -1 & \text{otherwise}\end{cases}$$

Lastly, in round k, the adversary chooses $T_k=T_k^{\text{exploit}}=\hat{s}_{1:k-1}/\sqrt{k-1}$.

Debiasing and estimating $\hat{s}_i$ using A_i. The adversary sees A_i and converts it to

$$\begin{aligned}\tilde{A}_i:=&\sqrt{2}\left\{A_i-\langle T_i,\theta\rangle-\mathbb{E}Z_i-\mathbb{E}\left[\frac{1}{\sqrt{2}}\langle T_i^{\text{exploit}},\bar{X}^{J_i}-\theta\rangle\right]\right\}\\=&\bar{X}_i^{J_i}-\theta_i+\sqrt{2}(Z_i-\mathbb{E}Z_i)+\langle T_i^{\text{exploit}},\bar{X}^{J_i}-\theta\rangle-\mathbb{E}\left[\langle T_i^{\text{exploit}},\bar{X}^{J_i}-\theta\rangle\right]\end{aligned}$$

By the independence of Z_i to everything,

$$\mathbb{E}[\tilde{A}_i|\bar{X}_i^{J_i}-\theta_i]=\bar{X}_i^{J_i}-\theta_i+\mathbb{E}\left[\langle T_i^{\text{exploit}},\bar{X}^{J_i}-\theta\rangle\middle|\bar{X}_i^{J_i}-\theta_i\right]-\mathbb{E}\left[\langle T_i^{\text{exploit}},\bar{X}^{J_i}-\theta\rangle\right].$$

In addition, since $J_1, ..., J_k$ does not depends on X, and that for every j,

$$\begin{aligned}\mathbb{E}\left[\langle T_i^{\text{exploit}}, \bar{X}^{J_i} - \theta\rangle \Big| \bar{X}_i^{J_i} - \theta_i\right] &= \mathbb{E}\left[\mathbb{E}\left[\langle T_i^{\text{exploit}}, \bar{X}^{J_i} - \theta\rangle \Big| \bar{X}_i^{J_i} - \theta_i, J_i\right] \Big| \bar{X}_i^{J_i} - \theta_i\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\langle T_i^{\text{exploit}}, \bar{X}^{J_i} - \theta\rangle \Big| J_i\right] \Big| \bar{X}_i^{J_i} - \theta_i\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\langle T_i^{\text{exploit}}, \bar{X}^{J_i} - \theta\rangle \Big| J_i\right]\right] = \mathbb{E}\left[\langle T_i^{\text{exploit}}, \bar{X}^{J_i} - \theta\rangle\right]\end{aligned}$$

the second line holds because $\bar{X}_i^j \perp\!\!\!\perp \langle \cdot, \bar{X}_{1:i-1}^j\rangle$ for every j and the third line holds because J_i is independent to $\bar{X}_i - \theta$. This suggests that the adversary can use $\tilde{A}_i$ as an unbiased estimate of $\bar{X}_i^{J_i} - \theta_i$, despite not knowing $J_{1:i}$. The conditional distribution $\tilde{A}_i | \bar{X}_i^{J_i} - \theta_i$ could be complex due to the choice of J_i, but is known to the adversary, and it satisfies that

$$\mathrm{Var}[\tilde{A}_i | \bar{X}_i^{J_i} - \theta_i] = \mathrm{Var}(\langle T_i^{\text{exploit}}, \bar{X}^{J_i} - \theta\rangle | \bar{X}_i^{J_i} - \theta_i) + 2\mathrm{Var}(Z_i)$$

where we can use the same arguments above to get

$$\begin{aligned}&\mathrm{Var}(\langle T_i^{\text{exploit}}, \bar{X}^{J_i} - \theta\rangle | \bar{X}_i^{J_i} - \theta_i) \\ =&\mathbb{E}\left[\mathrm{Var}(\langle T_i^{\text{exploit}}, \bar{X}^{J_i} - \theta\rangle | \bar{X}_i^{J_i} - \theta_i, J_i) \Big| \bar{X}_i^{J_i} - \theta_i\right] + \mathrm{Var}\left[\mathbb{E}[\langle T_i^{\text{exploit}}, \bar{X}^{J_i} - \theta\rangle | \bar{X}_i^{J_i} - \theta_i, J_i] | \bar{X}_i^{J_i} - \theta_i\right] \\ =&\mathbb{E}\left[\mathrm{Var}(\langle T_i^{\text{exploit}}, \bar{X}^{J_i} - \theta\rangle | J_i) \Big|\right] + \mathrm{Var}\left[\mathbb{E}[\langle T_i^{\text{exploit}}, \bar{X}^{J_i} - \theta\rangle | J_i]\right] \\ =&\mathrm{Var}(\langle T_i^{\text{exploit}}, \bar{X}^{J_i} - \theta\rangle).\end{aligned}$$

Lower bounding the bias from each i We will now consider a few restriction of the player's class of strategies that apply without loss of generality. First of all, it suffices to consider the case when $q < \sqrt{k-1}$, because otherwise the MSE at round 1 is bigger than $\sqrt{k-1}\sigma^2/n$, and the proof is complete. Secondly, by the definition of A_i, If $\mathrm{Var}(Z_i) + \frac{1}{2}\mathrm{Var}(\langle T_i^{\text{exploit}}, \bar{X}^{J_i} - \theta\rangle) > \sqrt{k-1}\sigma^2/n$, then the MSE at round i is necessarily greater than $\sqrt{k-1}\sigma^2/n$ and the proof is complete. So we only have to consider the case when J_i and Z obeys that

$$\mathrm{Var}[\tilde{A}_i | \bar{X}_i^{J_i} - \theta_i] \le 2\sqrt{k-1}\sigma^2/n.$$

Therefore, we can apply our main Lemma 10.16 to get for all $i = 1, ..., k-1$

$$\lim_{M\to\infty} \mathbb{E}[\hat{s}_i(\bar{X}_i^{J_i} - \theta_i)] \ge \frac{q\sigma^2}{n}\frac{1}{2\sqrt{3}}\frac{\sqrt{n}}{\sqrt{2}(k-1)^{1/4}\sigma} = \frac{q\sigma}{\sqrt{24}(k-1)^{1/4}\sqrt{n}}.$$

Overall bound. By law of total expectation and repeated application of Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}[\max_i \langle A_i - \theta, T_i\rangle^2] =&\mathbb{E}\left[\mathbb{E}[\max_i \langle A_i - \theta, T_i\rangle^2 | J_1, ..., J_{k-1}]\right] \\ \ge&\mathbb{E}\left[\max_i \mathbb{E}[\langle A_i - \theta, T_i\rangle^2 | J_1, ..., J_{k-1}]\right] \ge \mathbb{E}\left[\max_i \mathbb{E}[\langle A_i - \theta, T_i\rangle | J_1, ..., J_{k-1}]^2\right]\end{aligned} \tag{10.23}$$

Since J is independent to the data, conditioning on these events do not change the distribution of X,

Define $m_i^j := |\{\ell \in [i-1] | J_\ell = j\}|$, that is, the number of times data split j was used at iteration i.

$$\mathbb{E}[A_i - \langle \theta, T_i\rangle | J_1, ..., J_{k-1}] \geq \min_j \mathbb{E}[\langle \bar{X}^j - \theta, T_k\rangle | J_1, ..., J_{k-1}] \geq \left[\min_j m_k^j\right]^2 \frac{1}{k-1} \frac{q^2\sigma^2}{24(k-1)^{1/2}n}.$$

It follows that if we apply law of total expectation on the event E such that that $\min_j m_k^j] \geq \frac{k-1}{q}$

$$\begin{aligned}(10.23) =& \mathbb{P}(E)\mathbb{E}\left[\max_i \mathbb{E}[A_i - \langle\theta, T_i\rangle | J_1, ..., J_{k-1}, E]^2 \middle| E\right] \\ &+ \mathbb{P}(E^c)\mathbb{E}\left[\max_i \mathbb{E}[A_i - \langle\theta, T_i\rangle | J_1, ..., J_{k-1}, E^c]^2 \middle| E^c\right]. \end{aligned} \tag{10.24}$$

When $\min_j m_k^j \geq \frac{k-1}{q}$, then the loss from the kth round is clearly $\sqrt{k-1}C\sigma^2/(24n)$. On the other hand, if it is the complement event where $\min_j m_k^j < \frac{k-1}{q}$, then we know

$$\sum_{j=1}^q m_k^j - \min_j m_k^j \geq k - 1 - \frac{k-1}{q},$$

which implies by the pigeon hole principle that

$$\max_j m_k^j \geq \frac{k-1}{q-1} - \frac{k-1}{q(q-1)} \geq \frac{k-1}{q}.$$

As a result, under event E^c, there exists $j \in [q], i \in [k-1]$, such that $J_i = j$ and $m_i^j = \frac{k-1}{q} - 1$ and thus

$$\mathbb{E}[A_i - \langle\theta, T_i\rangle | J_1, ..., J_{k-1}, E^c]^2 \geq (\frac{k-1}{q} - 1)^2 \frac{1}{2(k-1)} \frac{q^2\sigma^2}{24(k-1)^{1/2}n} = \Omega(\sqrt{k-1}\sigma^2/n),$$

where we need $k \geq 2q + 1$. The proof is complete by substituting lower bounds under both events into (10.24). □

10.9 Derivation of the simple properties of A and A^{-1}

The raw moments of the half-normal distributions are:

$$\mu_1 = \frac{\sqrt{2}\sigma}{\sqrt{\pi}}, \qquad \mu_2 = \sigma^2, \qquad \mu_3 = \frac{2\sqrt{2}\sigma^3}{\sqrt{\pi}}, \qquad \mu_4 = 3\sigma^4.$$

Let q be the half normal density. We start with the forward operator A on polynomials.

$$\begin{aligned}
A\mathbf{1} \quad &= \int_0^\infty |t|q(t)dt - \int_{-\infty}^0 |t|q(t)dt = 0.\\
A\boldsymbol{x} \quad &= \int_0^\infty (x+t)|t|q(t)dt - \int_{-\infty}^0 (x-t)|t|q(t)dt = 2\int_0^\infty t^2 q(t)dt = 2\mu_2.\\
A\boldsymbol{x}^2 \quad &= \int_0^\infty (x+t)^2|t|q(t)dt - \int_{-\infty}^0 (x-t)^2|t|q(t)dt\\
&= \int_0^\infty (x^2+2xt+t^2)|t|q(t)dt - \int_0^\infty (x^2-2xt+t^2)|t|q(t)dt\\
&=4x\int_0^\infty t^2 q(t) = 4x\mu_2.\\
A\boldsymbol{x}^3 \quad &= \int_0^\infty (x+t)^3|t|q(t)dt - \int_{-\infty}^0 (x-t)^3|t|q(t)dt\\
&= \int_0^\infty (x^3+3x^2t+3xt^2+t^3)|t|q(t)dt - \int_0^\infty (x^3-3x^2t+3xt^2-t^3)|t|q(t)dt\\
&=6x^2\int_0^\infty t^2q(t)dt + 2\int_0^\infty t^3 q(t)dt = 6x\mu_2 + 2\mu_4.
\end{aligned}$$

The inverse operator A^{-1} on $\mathbf{1}, \boldsymbol{x}$ and $\boldsymbol{x}^2$ are obtained by simply applying A^{-1} on both sides and rearrange the terms.

Subsequent work and applications

In Part III of the thesis, we described two sequential interactive learning problems that are well-motivated by the fundamental challenges of the big data era. We formulate both problems in a minimax framework and studied their optimal max-risk. The first, the off-policy evaluation problem in contextual bandits, is in fact a pure statistical estimation problem closely related to the causal inference literature. The second, the sequential selective estimation problem, which is inspired by the related problem of adaptive data analysis, is in fact a very delicate partial information game, which we analyzed directly.

In this section, we discuss a novel connection between bandits and sequential selective estimation in their special cases. In particular, this is between linear bandits[2] and sequential selective estimations of Gaussian projection in Example 10.1.

We will first describe the stochastic linear bandits problem in a notation that will be easy for us to draw the connection to our model in Chapter 10. In all stochastic bandits problems, there is an agent interacting with the world by choosing an action given a context, and then the agent receives a reward for the action taken. The job of the agent is to minimize a notion of "regret" after repeatedly taking actions and learn from the outcome.

Specifically, for linear bandits, the agent receives a class of actions $\mathcal{T}_i \subset \mathbb{R}^d$ in round i and chooses an action (or a treatment) given by a vector $T_i \in \mathcal{T}_i$. When an action T_i is chosen at time i, it incurs a reward

$$R_i = \langle T_i, \theta_* \rangle + Z_i$$

where Z_i is assumed to be σ^2-subgaussian. Note that the data set is in forms of triplets

$$(\mathcal{T}_i, T_i, R_i) \text{ for } i = 1, ..., k.$$

At the end of the day, the agent wants to come up with a strategy which minimizes the regret

$$\mathcal{R}(k) = \mathbb{E}\left[\sum_{i=1}^{k} \max_{t \in \mathcal{T}_i} \langle t, \theta_* \rangle - \sum_{i=1}^{k} R_i\right].$$

At a glance it is unclear why this is related to either contextual bandits or sequential adaptive estimation. We will now clarify.

When $\mathcal{T}_i$ is fixed and include only $e_1, ..., e_d$, then this reduces to d-arm bandits. $\mathcal{T}_i$ can also be used to encode a subset of contextual bandit model where the expected reward is linear in the context feature vector x_i and a coefficient vector that is pre-defined for every action. Specifically, in this case for m-actions, $\mathcal{T}_i$ contains m possible actions and they are $x_i \otimes \{e_1, ..., e_m\}$ (where $\otimes$ denotes is Kronecker product). The corresponding θ_* is partitioned into chunks that correspond to each action.

The connection to sequential adaptive estimation is more interesting. We claim that linear bandits with $\mathcal{T}_i \equiv \mathcal{T} = \{t \in \mathbb{R}^d | \|t\|_2 \leq 1\}$ and $\theta_* = X - \theta$, is equivalent to a version of the sequential selective estimation problem with a cumulative (rather than maximum) loss function, where the player is required to use to

$$A_i = \langle X, T_i \rangle + Z_i$$

where Z_i is subgaussian.

To see this, let us put ourselves in the shoes of the adversary in sequential selective estimation. Think about the linear bandits agent who chooses actions as the adversary in sequential selective estimations. In every iteration, the agent (the adversary) chooses t_i and receives $A_i = \langle XT_i \rangle + Z_i$ from the player. The agent then modify it into a "reward":

$$R_i := A_i - \langle \theta T_i \rangle = \langle X - \theta, T_i \rangle + Z.$$

At the end of the day, the job of the adversary in sequential selective estimation is to maximize $\max_i \mathbb{E}[|R_i|^2]$. We could of course choose a different loss function that compute the summation over rounds rather than maximization. Also, since the player only allowed to add noise, without loss of generality, we can assume that R_i is maximized. In other word, the goal of the adversary is to maximize

$$\sum_i \mathbb{E} R_i.$$

This is equivalent to minimize the regret

$$\max_{t \in \mathcal{T}} \mathbb{E}[\langle X - \theta, t \rangle] - \sum_i \mathbb{E}[R_i]$$

A subtle difference here is that in sequential selective estimations, we are measuring the average reward over a "prior" distribution of X, which happens to be $\mathcal{N}(\theta, \sigma^2 I)$.

It is interesting to see how upper bounds and lower bounds of our problem imply for linear bandits and how results in linear bandits can apply to our setting. In particular, since this is a specific case of linear bandits, our upper bound (e.g.,Theorem 10.8) implies a lower bound for linear bandit problems in general. Our lower bound (e.g., Theorem 10.10 and 10.13) on the other hand, can be used as a performance guarantee for a specific subset of linear bandits problems where the player who comes up with answers in stateful and adversarial nature.

We leave details to formalize these implications as future work.

Chapter 11

Conclusion and discussion

In this thesis, I presented theoretical contributions to a number of new models in the broad area of statistical machine learning. These include differentially private machine learning (Part I), locally adaptive nonparametric regression (Part II), off-policy evaluation as well as adaptive data analysis (Part III).

While these problems seem unrelated to each other, having studied all of them, it comes to our realization that on the technical level they all share very similar behaviors and the interplay of these areas could lead to fruitful new theoretical understanding and algorithmic development. In the following, we conclude the thesis with some meta-level observations in statistical learning research and three potential ways the problems in the three parts can meet each other.

11.1 Meta-level take-home messages

Sharing ideas across a diverse research community. Statistical machine learning is a very diverse field in the intersection of statistics, optimization, information theory as well as theoretical computer science. Each field represents a long history of research work with a rich associated literature. The author of this thesis benefits greatly from attending conferences such as ICML and NIPS where a variety of ideas from different branches of mathematics are brought together.

Many of the results in the thesis come from a cross-fertilization of concepts and techniques in different research communities. For instance, the initial idea of Chapter 2 and 4 comes from a realization that differential privacy is closely related to the algorithmic stability in learning theory [38]. Similarly, the connection between the Bayesian posterior distribution and exponential mechanism [156] allows us to connect a vast statistical literature on posterior consistency and Bernstein Von Mises theorems that can be used to analyze the utility of exponential mechanisms.

It is also possible that some of the connections and implications only emerge in a much later phase of a research project. Our results on trend filtering on graphs in Chapter 6 is developed under the primary motivation of extending univariate trend filtering. It comes as a pleasant surprise that the signal processing communities have been independently developing a suite of tools for Fourier and wavelet transforms on graphs (see, e.g.,[198]). Also, spectral graph theory and the corresponding electrical network interpretation of graphs gave us a big "wow" and helped us to identify the key graph-theoretical quantities that can help us obtain strong error bounds.

In addition, the connections between off-policy evaluation in contextual bandits and asymptotic optimality theory in causal inference reveal several new understanding in both fields (see the discussion in the end of Chapter 9); and the sequential selective estimation problem we presented in Chapter 10 aims at explaining the implication of randomization used for adaptive data analysis to the statistics community.

We need more of such fusions and more communications with people who are familiar with the literature in the past 100 years. It is possible that many of the technical challenges that we are facing now are challenges that researchers in a completely different field have already put a lot of thoughts on for decades. As, Isaac Newton put it, only when we "stand on the shoulder of a giant", can we see further and address bigger problems.

On the theory and practice split. Another underlying thread of this thesis is to bridge the theory and practice. This is the underlying motivation for our work in Chapter 3, 4 and 5, and it also motivates our optimality theory presented in Chapter 7 to justify the use of the popular total variation denoising method in image and video denoising.

Nevertheless, bridging the theory and practice is rather challenging to do in general. Theoretical machine learning research has long been regarded as providing general guidelines rather than meaningful bounds that one can use in practice. This is not surprising as many common techniques being used such as concentration inequalities, union bounds and the empirical process theory-style of uniform convergence arguments are at best tight up to a constant.

Theoretical research, however, do not have to be disconnected to practice. This is at least not the case in the old days. Classic statistical theory is filled with tight analysis and exact (asymptotic) optimality results. Classical coding theory ensures that we make use of just enough number of bits to send a message over a noisy channel. The optimal control theory allows us to design controllers that are optimally robust to model-misspecification. From this regard, it is hard to argue that tight theoretical analysis cannot be done. It just takes more time and more careful calculations.

That said, there probably always will be a gap that we cannot hope to close. Theoretical research is limited by the available proof techniques and complexity of the model that we are able to reason formally. Often, this means that we have to decide on whether we want to design a provably optimal algorithm on a wrong model or adopt a completely empirical approach on a somewhat "less wrong" model. The deep learning community took the latter approach and seems to be doing fine so far. Differential privacy, on the other hand, is a special case because the privacy guarantee can only be proven in theory. More fine-grained privacy analysis such as those with per-instance differential privacy as we described in Chapter 5 is much needed.

Lastly, I find it important to choose a practical algorithm to work on in the first place. The practical performance of algorithms can differ significantly, even if they admit the same theoretical rate (see e.g., the experimental comparison of trend filtering and wavelet smoothing in Chapter 6). In the hindsight, trend filtering is unique because it directly penalizes the L1-norm of an approximation of the total variation operator, while wavelet smoothing needs to explicitly construct a set of basis functions that could have small artifacts in themselves due to discretization. In addition, somewhat circularly, I know that trend filtering will work well in practice, because it has already been shown to work in practical examples [129, 182] at least in restricted settings.

A positive view of minimax lower bounds. The last take-home-message is my two cents on lower bounds. Lower bounds, also known as no-free-lunch theorems, are often regarded as negative results, as they certify that no algorithms can solve a problem better beyond a fundamental limit. However, there is also a positive way of looking at them.

First, it is perhaps more important to realize what a minimax lower bound does *not* say. Lower bounds, in most cases are only true in a minimax sense. This means that it does not prevent a specific problem that we see to be easily solvable. Also, an algorithm being minimax, does not necessarily mean that it cannot be significantly improved. There could even be another algorithm that strictly improves a minimax optimal algorithm in every instance, e.g., MLE and Stein's shrinkage in the problem of estimating a Gaussian mean when dimension is larger than 3 [210].

Second, when a lower bound appears large, it often suggests that we should revise the class of problems under consideration. The constructions that lead to the lower bound could offer interesting insight on why the problem class is hard and how to redefine the class so that one can avoid those issues. In our study of private learnability in Chapter 2, it is clear from Beimel et al. [19], Chaudhuri and Hsu [54]'s threshold-function construction that the impossibility can be sidestepped by restricting our attention to only Lipschitz loss functions or smooth class of distributions. In our study of off-policy evaluation in Chapter 9, it will be a much less interesting set of results if we stop at a minimax analysis that shows that IPS is minimax optimal, as then we will not be able to come up with the "SWITCH" estimator that behaves adaptively and orders-of-magnitude better than IPS in practice.

In my opinion, finding a minimax lower bound and an algorithm that matches the minimax rate is not the end. It is also the beginning of the new pursuit of an adaptive and practical algorithm.

11.2 Future directions

Statistical estimation for denoising differential privacy. We start by inspecting the connections of statistical estimation and differential privacy. First of all, both statistics and differential privacy deal with probabilities, but in opposite ways. Statistical estimation address the problem of making probabilistic inferences on the some fixed underlying parameter, while differential privacy is about simulating the randomness that prevents the a classifier from distinguishing between two point hypothesis for sure, which implies a lower bound on the estimation error as well. An interesting idea is to investigate how much can statistical estimation theory help in improving the utility of differentially privately released statistics through a "denoising" post-processing algorithm.

A key drawback of statistic estimation is that it often needs to make assumptions on the distribution that generates the data, which may or may not be true in practice. However, when the noise is simulated by ourselves (to achieve differential privacy), we have full control over its distribution of the noise, and we could have additional information about the parameter to estimate released differentially privately. This is especially helpful for high-dimensional problems, where knowing the sparsity level or the ℓ_1-norm bound could allow us to exploit many of the machinery developed in the past two decades in high-dimensional statistics.

Another interesting direction is to consider the Bayesian estimation setting. More concretely, if the data follows the following generative process

$$X \sim P(X|\theta), \theta \sim \pi$$

for some prior π that is public knowledge and $\phi(X)$ is the sufficient statistics . Then we know that the optimal Bayes estimator of θ in square error is the posterior mean estimator

$$\hat{\theta} = \mathbb{E}(\theta|X) = \mathbb{E}(\theta|\phi(X)).$$

Suppose we all we see is differentially private release of the data, or the differentially private release Y of the sufficient statistics $\phi(X)$, then effectively the generative process becomes

$$Y \sim P'(Y|\theta), \theta \sim \pi$$

where $P'(Y|\theta) = \int Q(Y|X)P(X|\theta)dX$, and the optimal postprocessing estimator is

$$\mathbb{E}(\theta|Y).$$

Lastly, when we consider the statistical estimation problem and differential privacy jointly, the effective amount of noise we see is the combination of that noise from DP and that intrinsic randomness in data generation. In such cases, it makes sense to "denoise" more aggressively and it will be interesting to investigate how much data efficiency we lose. In Chapter 5, we provided an example of this in estimating linear regression coefficients.

Privacy and sequential learning. One interesting property of differential privacy is that it allows graceful composition over a sequence of differentially private algorithms each chosen as a function of the entire history. This enables the application of differential privacy to control false discoveries in science [87] and effectively motivated our work on sequential selective estimation in Chapter 10.

The implication of such adaptive composition in other sequential learning models are not clear. This is of particular interest because in reinforcement learning (RL), bandits and online learning, randomization is used for a variety of other reasons. Specifically, randomization is used for exploration in RL and bandits and is used hedging in adversarial online learning.

Each round of Follow-The-Perturbed-Leader (FTPL), is essentially objective perturbation as studied in [128]. Also, Exp3 algorithm for adversarial bandits, the Hedge algorithm for adversarial online learning as well as the Thompson sampling algorithm for RL and bandits can in fact be treated as a sequence of OPS algorithm that we presented in Chapter 3.

In many of these algorithms, especially those in the adversarial setting, randomization is used to ensure certain notion of stability. As we know, differential privacy implies almost all kinds of stability and it will be interesting to check whether we can obtain a regret bound directly using the composition approach of differential privacy.

Online trend filtering? So far, we have talked about the marriage of privacy and statistical estimation, and that of privacy and sequential learning. How about combining trend filtering with sequential learning?

There could be many ways of setting up the problem, but one reasonable model to consider is the following.

Suppose we observe a sequence of noisy observations $y_1, y_2, ..., y_n$ one by one, and we know that $y_i \sim \mathcal{N}(f(i/n), \sigma^2)$ for $i = 1, ..., n$ for f such that the $(k+1)$th derivative of f has total variation bounded by C_n. For each $i = 1, ..., n$, can we come up with $\hat{\theta}_i$ using C_n, σ^2 and $y_1, ..., y_{i-1}$ such that at the end of the day,

$$\sup_{f \in \mathrm{TV}_k(C_n)} \frac{1}{n} \sum_{i=1}^{n} (\hat{\theta}_i - f(i/n))^2 = o(1).$$

Clearly, this is a problem that is strictly harder than univariate trend filtering, as we are only required to provide an answer after observing $(y_1, ..., y_n)$ all together. A natural question is that is there a price to pay in the sequential setting.

The case for $k = 0$ is a special case of the non-stationary stochastic optimization result in [31] with loss function being $(\hat{\theta}_i - \theta_i)^2$ and noisy gradient oracle. A noisy gradient oracle is equivalent to observing y_i because the gradient of the above loss function is $2(\hat{\theta}_i - \theta_i)$ and $2(\hat{\theta}_i - y_i)$ is an unbiased estimate of the gradient. Moreover, for bounded function values, the bounded nonstationary variation is also equivalent to a total variation bound.

For $C_n = 1$, [31]'s result translates into an upper bound of $O(n^{-1/2})$, achieved via a online gradient descent algorithm with scheduled restarts. This is the same as the minimax linear rate in the batch setting, and much slower than the $O(n^{-2/3})$ rate that can be obtained in the batch setting.

It remains open to see whether the $O(n^{-2/3})$ rate can be achieved in the online setting.

Bibliography

[1] Martín Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security (CCS-16)*, pages 308–318. ACM, 2016. 5.1

[2] Naoki Abe, Alan W Biermann, and Philip M Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003. 10

[3] Robert Acar and Curtis R. Vogel. Analysis of total variation penalty methods. *Inverse Problems*, 10:1217–1229, 1994. 7.1

[4] Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. In *International Conference on Machine Learning (ICML-12)*, 2012. 3.1, 3.2, 3.4.2, 3.4.2, 3.8.2, 5

[5] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014, 2008. 1

[6] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. In *Advances in Neural Information Processing Systems (NIPS-09)*, pages 33–40, 2009. 3.1

[7] Hirotugu Akaike. Likelihood of a model and information criteria. *Journal of econometrics*, 16(1):3–14, 1981. 4.3.3

[8] Yasemin Altun and Alex Smola. Unifying divergence minimization and statistical inference via convex duality. In *International Conference on Computational Learning Theory*, pages 139–153. Springer, 2006. 4.3.3

[9] Theodore Anderson and Donald Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954. 8.1, 8.5.2

[10] Rie Ando and Tong Zhang. Learning on graph with Laplacian regularization. *Advances in Neural Information Processing Systems (NIPS-06)*, 9, 2006. 7.1

[11] David Applegate and Ravi Kannan. Sampling and integration of near log-concave functions. In *Proceedings of 23rd ACM Symposium on Theory of Computing*, pages 156–163, 1991. 2.5.2, 3.3.3

[12] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. 9.1

[13] Philippe Barbe. Limiting distribution of the maximal spacing when the density function admits a positive minimum. *Statistics & Probability Letters*, 14(1):53–60, 1992. 8.7.6

[14] Rina Foygel Barber and John C Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*, 2014. 4.1, 4.3, 4.5, **??**, **??**

[15] Alvaro Barbero and Suvrit Sra. Fast Newton-type methods for total variation regularization. In *International Conference on Machine Learning (ICML-11)*, volume 28, pages 313–320, 2011. 6.4.2

[16] Alvero Barbero and Suvrit Sra. Modular proximal optimization for multidimensional total-variation regularization. arXiv: 1411.0589, 2014. 6.4.1, 7.1, 7.1

[17] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization, revisited. *rem*, 3:17, 2014. 2.1, 2.3, 2.3.2, 2.5.2, 3.1, 3.3.1, 3.3.3, 3.4, 3.4.1, 3.4.1, 3.4.3, 3.6

[18] Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *ACM SIGACT Symposium on Theory of Computing (STOC-16)*, pages 1046–1059. ACM, 2016. 2.1, 4.1, 4.1, 4.5, 4.5, 10.1, 10.1, 10.1, 10.3.1

[19] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Characterizing the sample complexity of private learners. In *Conference on Innovations in Theoretical Computer Science*, pages 97–110. ACM, 2013. 2.1, 2.1, 2.3.3, 2.23, 2.4.3, 11.1

[20] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 363–378. Springer, 2013. 2.1

[21] Amos Beimel, Hai Brenner, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. *Machine learning*, 94 (3):401–437, 2014. 2.2.2, 2.8.1, 3.4, 3.13

[22] Mikhail Belkin and Partha Niyogi. Using manifold structure for partially labelled classification. *Advances in Neural Information Processing Systems (NIPS-02)*, 15, 2002. 7.1

[23] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. 7.1

[24] Mikhail Belkin and Partha Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 56(1–3):209–239, 2004. 7.1

[25] Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. *Conference on Learning Theory (COLT-05)*, 18, 2005. 6, 7.1

[26] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. On manifold regularization. *International Conference on Artificial Intelligence and Statistics*, 8, 2005. 7.1

[27] Oliver Bembom and Mark J van der Laan. Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. 2008. 9.5

[28] Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

4.3.3

[29] Karine Bertin et al. Asymptotically exact minimax estimation in sup-norm for anisotropic Hölder classes. *Bernoulli*, 10(5):873–888, 2004. 9.1, 9.9.2

[30] Dimitri P Bertsekas. Projected Newton methods for optimization problems with simple constraints. *SIAM Journal on Control and Optimization*, 20(2):221–246, 1982. 6.4.2

[31] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015. 11.2

[32] William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural computation*, 13(11):2409–2463, 2001. 3.3.1

[33] Lucien Birge and Pascal Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, 2001. 7.3.2, 7.7.2, 7.13

[34] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. 1, 3.1

[35] Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. The johnson-lindenstrauss transform itself preserves differential privacy. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 410–419. IEEE, 2012. 5.5

[36] Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning (ICML-15)*, pages 1006–1014, 2015. 2.1

[37] Léon Bottou, Jonas Peters, Joaquin Quinonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y Simard, and Ed Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research*, 14(1):3207–3260, 2013. 9.4.1, 1, 9.5

[38] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002. 2.2.3, 4.1, 11.1

[39] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004. 6.4.1

[40] Kristian Bredies, Karl Kunisch, and Thomas Pock. Total generalized variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526, 2010. 6.2.4

[41] Jonathan B Buckheit and David L Donoho. *Wavelab and reproducible research*. Springer, 1995. 8.2.3

[42] Peter Buhlmann and Sara van de Geer. *Statistics for High-Dimensional Data*. Springer, Berlin, 2011. 6.6.1, 6.6.3

[43] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016. 5.1, 5.2.3, **??**

[44] Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *IEEE Symposium on Foundations of Computer*

Science (FOCS-15), pages 634–649. IEEE, 2015. 3

[45] Mark Mar Bun. *New Separations in the Complexity of Differential Privacy*. PhD thesis, Harvard University Cambridge, Massachusetts, 2016. 3

[46] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005. II

[47] Bernd Carl. Metric entropy of convex hulls in Hilbert spaces. *Bulletin of the London Mathematical Society*, 29(04):452–458, 1997. 6.9.11

[48] Claes M Cassel, Carl E Särndal, and Jan H Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3): 615–620, 1976. 9.2

[49] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11: 2079–2107, 2010. 10.1

[50] Antonin Chambolle and Jerome Darbon. On total variation minimization and surface evolution using parametric maximum flows. *International Journal of Computer Vision*, 84: 288–307, 2009. 7.1

[51] Antonin Chambolle and Pierre-Louis Lions. Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76(2):167–188, 1997. 6.4.1, 7.1

[52] Philip K Chan and Salvatore J Stolfo. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In *KDD*, volume 1998, pages 164–168, 1998. 1

[53] Samprit Chatterjee and Ali S Hadi. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, pages 379–393, 1986. 5.3

[54] Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *COLT*, volume 19, pages 155–186, 2011. 2.1, 2.1, 2.3, 2.3.3, 2.3.3, 2.12, 2.4.1, 2.4.2, 2.24, 2.6, 2.7, 11.1

[55] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011. I, 2.1, 2.3, 2.4, 3.1, 3.3.1, 3.3.4, 3.5, 3.6, 5.1

[56] Tianqi Chen, Emily B. Fox, and Carlos Guestrin. Stochastic Gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning (ICML'14)*, 2014. 3.1, 3.2, 3.4.2, 3.4.2, 3.4.2, 5

[57] Fan Chung and Mary Radcliffe. On the spectra of general random graphs. *The Electronic Journal of Combinatorics*, 18(1), 2011. 6.9.3

[58] Ronald Coiman and Mauro Maggioni. Diffusion wavelets. *Applied and Computational Harmonic Analysis*, 21(2):53–94, 2006. 6.8.1, 6.8.2

[59] Samuel Conte and Carl de Boor. *Elementary Numerical Analysis: An Algorithmic Approach*. McGraw-Hill, New York, 1980. International Series in Pure and Applied Mathematics. 6.9.7, 7.1, 7.8

[60] Herbert Aron David and Haikady Navada Nagaraja. *Order Statistics*. Wiley, Hoboken, 1970. 8.7.6

[61] Pierpaolo De Blasi and Stephen G Walker. Bayesian asymptotics with misspecified models. *Statistica Sinica*, 23:169–187, 2013. 3.3.2

[62] Carl de Boor. *A Practical Guide to Splines*. Springer, New York, 1978. 8.1

[63] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990. 1

[64] Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin IP Rubinstein. Robust and private bayesian inference. In *Algorithmic Learning Theory*, pages 291–305. Springer, 2014. 3.1, 1, 3.6, 5.4, 5.6

[65] Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D Skeel, and Hartmut Neven. Bayesian sampling using stochastic gradient thermostats. In *Advances in Neural Information Processing Systems*, pages 3203–3211, 2014. 3.1, 3.2, 3.4.2, 3.4.2, 5

[66] David Dobson and Fadil Santosa. Recovery of blocky images from noisy and blurred data. *SIAM Journal on Applied Mathematics*, 56(4):1181–1198, 1996. 7.1

[67] David Donoho and Iain Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994. 7.1

[68] David Donoho and Iain Johnstone. Minimax risk over ℓ_p-balls for ℓ_q-error. *Probability Theory and Related Fields*, 99(2):277–303, 1994. 7.7.2

[69] David Donoho and Iain Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995. 6.8.1

[70] David Donoho and Iain Johnstone. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26(8):879–921, 1998. II, 7, 7.1, 7.1, 7.1, 7.2, 7.2, 7.5, 7.5, **??**, **??**, **??**, **??**

[71] David Donoho, Richard Liu, and Brenda MacGibbon. Minimax risk over hyperrectangles, and implications. *Annals of Statistics*, 18(3):1416–1437, 1990. 7.1, 7.3.3, 7.4, 7.7.3, 7.7.6

[72] Harish Doraiswamy, Nivan Ferreira, Theodoros Damoulas, Juliana Freire, and Claudio Silva. Using topological analysis to support event-guided exploration in urban data. *Visualization and Computer Graphics, IEEE Transactions on*, PP(99), 2014. 6.5.3

[73] Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, and David P Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13(Dec):3475–3506, 2012. 5.3

[74] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning (ICML-11)*, 2011. 9.1, 9.2, 9.5

[75] Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014. 9.1, 9.3.1, 9.3.2, 9.16

[76] George T. Duncan, Mark Elliot, and Juan Jose Salazar Gonzalez. *Statistical Confidentiality: Principles and Practice*. Springer, 2011. I

[77] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming*, pages 1–12. Springer, 2006. I, 2.2, 2.2.2, 3.1

[78] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *ACM Symposium on Theory of Computing*, pages 371–380. ACM, 2009. I

[79] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *ACM Symposium on Theory of Computing (STOC-09)*, pages 371–380. ACM, 2009. 3.3.1, 5.1, 5.5

[80] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2013. I, 3.2.1, 3.11, 3.14, 5.2, 5.2, 5.2.1

[81] Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016. 5.1, 5.2.3, **??**

[82] Cynthia Dwork and Adam Smith. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2):2, 2010. 5.4

[83] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT 2006*, pages 486–503. Springer, 2006. 2.27, 4.1, 4.5, **??**

[84] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography*, pages 265–284. Springer, 2006. 1, I, 2.2.2, 3.1, 5.1, 5.1, **??**

[85] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pages 2341–2349, 2015. 4.1, 4.4, 9, 4.4, 4.8, 10.1, 10.1, 10.4

[86] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349 (6248):636–638, 2015. 1, 2.1

[87] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *STOC'15*, pages 117–126. ACM, 2015. (document), I, 2.1, 2.9, 2.38, 3.3.3, 4.1, 4.5, 10.1, 10.1, 10.1, 10.4, 11.2

[88] Hamid Ebadi, David Sands, and Gerardo Schneider. Differential privacy: Now it's getting personal. In *ACM Symposium on Principles of Programming Languages*, pages 69–81. ACM, 2015. 4.1, 4.5, **??**, 5.2.3

[89] Ahmed El Alaoui, Xiang Cheng, Aaditya Ramdas, Martin J Wainwright, and Michael I Jordan. Asymptotic behavior of\ell_p-based laplacian regularization in semi-supervised learning. In *Conference on Learning Theory*, pages 879–906, 2016. 8

[90] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition (CVPR-05)*, volume 2, pages 524–531. IEEE, 2005. 3.1

[91] Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random*

Structures & Algorithms, 27(2):251–275, 2005. 6.9.3

[92] Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(98):298–305, 1973. 6.6.1

[93] Stephen E Fienberg, Alessandro Rinaldo, and Xiaolin Yang. Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *International Conference on Privacy in Statistical Databases*, pages 187–199. Springer, 2010. 5.1

[94] Stephen E Fienberg, Alessandro Rinaldo, and Xiaolin Yang. Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *Privacy in Statistical Databases*, pages 187–199. Springer, 2011. 4.1

[95] William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014. 10.1

[96] James Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. On the theory and practice of privacy-preserving bayesian data analysis. In *Uncertainty in Artificial Intelligence (UAI-16)*, pages 192–201. AUAI Press, 2016. 5.1, 5.4, 5.4, 5.4, 5.5, 5

[97] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001. 5.1, 5.3

[98] Andrew Gelman, John B Carlin, and Hal S Stern. *Bayesian data analysis*, volume 2. Taylor & Francis, 2014. 3.1

[99] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984. 3.1

[100] Subhashis Ghosal. *The Dirichlet process, related priors and posterior asymptotics*, volume 2. Chapter, 2010. 3.3.2, 3.3.2

[101] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011. 3.4.2

[102] S. Godunov and V. Ryabenkii. *Difference Schemes: An Introduction to the Underlying Theory*. Elsevier, Amsterdam, 1987. Number 19 in Studies in Mathematics and Its Applications. 6.9.7, 7.1

[103] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning*, 3(4):5, 2009. 9.5

[104] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012. 8.1, 8.5.2

[105] Adityanand Guntuboyina, Donovan Lieu, Sabyasachi Chatterjee, and Bodhisattva Sen. Spatial adaptation in trend filtering. *arXiv preprint arXiv:1702.05113*, 2017. 8

[106] Laszlo Gyorfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002. 7.3.1

[107] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998. 9.1, 9.3.2, 9.9.1, 9.11, 9.9.1, 9.16, 9.9.2

[108] Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Random differential privacy. *Journal of Privacy and Confidentiality*, 4(2):43–59, 2012. 4.1, 4.5

[109] Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 454–463. IEEE, 2014. 2.1, 4.1, 10.1, 10.1

[110] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. 3.2

[111] David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992. 2.2.1

[112] Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003. 9.1, 9.3.2, 9.12, 9.9.1

[113] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. 9.9

[114] Holger Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010. 6.2.1, 7.1

[115] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008. 3.3.1

[116] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47 (260):663–685, 1952. 9.1, 9.2

[117] Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul De Wolf. *Statistical Disclosure Control*. John Wiley & Sons, 2012. I

[118] Jan-Christian Hutter and Philippe Rigollet. Optimal rates for total variation denoising. In *Conference on Learning Theory (COLT-16)*, 2016. to appear. 7.1, 7.2, 7.8, 7.8, **??**, **??**, 8

[119] Guido Imbens, Whitney Newey, and Geert Ridder. Mean-squared-error calculations for average treatment effects. Technical report, 2007. 9.1, 9.3.2

[120] Jeff Irion. *Multiscale Transformations for Signals on Graphs: Methods and Applications*. PhD thesis, Department of Mathematical Sciences, University of California at Davis, 2015. 6.8.1, 6.8.2

[121] Prateek Jain and Abhradeep Thakurta. Differentially private learning with kernels. In *International Conference on Machine Learning (ICML-13)*, pages 118–126, 2013. 2.1

[122] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4): 620, 1957. 4.3.3

[123] Shiva P Kasiviswanathan and Adam Smith. On the'semantics' of differential privacy: A

bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1):1, 2014. 3.6

[124] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011. 2.1, 2.1

[125] Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999. 2.2.3, 4.1

[126] Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 341–352. ACM, 1992. 2.2.1

[127] Jonathan Kelner, Lorenzo Orecchia, Aaron Sidford, and Zeyuan Allen Zhu. A simple, combinatorial algorithm for solving SDD systems in nearly-linear time. *ACM Annual Symposium on Theory of Computing (STOC-13)*, 45:911–920, 2013. 6.4.1

[128] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. *Journal of Machine Learning Research*, 1: 41, 2012. 2.1, 2.3, 2.4, 3.1, 3.3.1, 3.3.4, 3.5, 3.6, 5.5, 11.2

[129] Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM Review*, 51(2):339–360, 2009. 1, II, 6, 6.1, 6.4.2, 8.4, 11.1

[130] BJK Kleijn, AW van der Vaart, et al. The bernstein-von-mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012. 3.3.2, 3.3.2, 3

[131] Risi Kondor and John D. Lafferty. Diffusion kernels on graphs and other discrete structures. In *International Conference on Machine Learning (ICML-02)*, pages 315–322, San Francisco, CA, 2002. Morgan Kaufmann. 3

[132] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8), 2009. 1

[133] Ioannis Koutis, Gary Miller, and Richard Peng. A nearly-$m \log n$ time solver for SDD linear systems. *IEEE Annual Symposium on Foundations of Computer Science (FOCS-11)*, 52:590–598, 2011. 6.4.1

[134] Hans Kunsch. Robust priors for smoothing and image restoration. *Annals of the Institute of Statistical Mathematics*, 46(1):1–19, 1994. 7.1, 7.8

[135] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. 1

[136] John Lafferty, Han Liu, and Larry Wasserman. Minimax theory, 2008. URL `http://www.stat.cmu.edu/~larry/=sml/Minimax.pdf`. 9.7.1, 9.7

[137] Lucien Marie Le Cam. *On the Bernstein-von Mises theorem*. Department of Statistics, University of California, 1986. 3.1, 3.3.2

[138] Edward E Leamer. *Specification searches: Ad hoc inference with nonexperimental data*, volume 53. John Wiley & Sons Incorporated, 1978. 1, 10.1

[139] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[140] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015. 1

[141] Kai-Fu Lee. *Automatic speech recognition: the development of the SPHINX system*, volume 62. Springer Science & Business Media, 1988. 1

[142] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006. 2

[143] Erich Leo Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006. 10.7.1

[144] Jing Lei. Differentially private *m*-estimators. In *Advances in Neural Information Processing Systems (NIPS-11)*, pages 361–369, 2011. I

[145] Oleg V Lepski and VG Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, pages 2512–2546, 1997. II

[146] Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. In *AISTATS*, 2015. 9.1, 9.3.1, 9.3.2, 9.13

[147] Ziqi Liu, Yu-Xiang Wang, and Alexander Smola. Fast differentially private matrix factorization. In *ACM Conference on Recommender Systems (RecSys-15)*, pages 171–178. ACM, 2015. 4.1, 4.5, 5.1, **??**, 5.2.3, 5

[148] Ziqi Liu, Alex Smola, Kyle Soska, Yu-Xiang Wang, and Qinghua Zheng. Attributing hacks. *AISTATS'16*, 2016. 8

[149] Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014. 10.1

[150] Alexander Lubotzky, Ralph Phillips, and Peter Sarnak. Ramanujan graphs. *Combinatorica*, 8(3):261–277, 1988. 6.9.3

[151] Lester Mackey, Michael I Jordan, Richard Y Chen, Brendan Farrell, Joel A Tropp, et al. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42(3):906–945, 2014. 5.26

[152] Enno Mammen and Sara van de Geer. Locally apadtive regression splines. *Annals of Statistics*, 25(1):387–413, 1997. II, 6.6.1, 6.6.3, 6.9.9, 7.1, 7.1, 8.4.1, 8.7.5, 8.8.1

[153] Adam W Marcus, Daniel A Spielman, and Nikhil Srivastava. Ramanujan graphs and the solution of the Kadison-Singer problem. arXiv: 1408.4421, 2014. 6.9.3

[154] Julian McAuley and Jure Leskovec. Learning to discover social circles in ego networks. *Advances in Neural Information Processing Systems*, 25, 2012. 6.5.1

[155] Frank McSherry and Ilya Mironov. Differentially private recommender systems: building privacy into the netflix price contenders. In *ACM SIGKDD international conference on Knowledge discovery and data mining (KDD-09)*, pages 627–636. ACM, 2009. 5.1

[156] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS'07*, pages 94–103. IEEE, 2007. 2.1, 2.2.2, 2.4, 2.4, 2.8.5, 2.8.5, 3.1, 3.3.1, 4.3.1, 4.3, 5.4, 11.1

[157] Kentaro Minami, HItomi Arai, Issei Sato, and Hiroshi Nakagawa. Differential privacy without sensitivity. In *Advances in Neural Information Processing Systems*, pages 956–964,

2016. 5

[158] Darakhshan J. Mir. *Differential privacy: an exploration of the privacy-utility landscape*. PhD thesis, Rutgers University, 2013. 3.1, 1, 3.6

[159] Darakhshan J Mir. Information-theoretic foundations of differential privacy. In *Foundations and Practice of Security*, pages 374–381. Springer, 2013. 4.3.3

[160] Ilya Mironov. Rényi differential privacy. *arXiv preprint arXiv:1702.07476*, 2017. 5.2.3, **??**

[161] Frederick Mosteller and John W Tukey. *Data analysis, including statistics*. 1968. 4.1

[162] Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006. 2.2.3, 4.1

[163] R Neal. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, pages 113–162, 2011. 3.2, 3.4.2

[164] Deanna Needell and Rachel Ward. Stable image reconstruction using total variation minimization. *SIAM Journal on Imaging Sciences*, 6(2):1035–1058, 2013. 7.2

[165] Michael Ng, Raymond Chan, and Wun-Cheung Tang. A fast algorithm for deblurring models with Neumann boundary conditions. *SIAM Journal on Scientific Computing*, 21(3): 851–866, 1999. 7.1, 7.8

[166] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *ACM symposium on Theory of computing*, pages 75–84. ACM, 2007. 5.1, 5.2

[167] Michael Nussbaum. Spline smoothing in regression models and asymptotic efficiency in l_2. *Annals of Statistics*, 13(3):984–997, 1985. 8.4.1

[168] Peter Orbanz. Lecture notes on bayesian nonparametrics. *Journal of Mathematical Psychology*, 56:1–12, 2012. 3.3.2

[169] Oscar Hernan Madrid Padilla, James G Scott, James Sharpnack, and Ryan J Tibshirani. The dfs fused lasso: Linear-time denoising over general graphs. *arXiv preprint arXiv:1608.03384*, 2016. 8

[170] Mijung Park, James Foulds, Kamalika Chaudhuri, and Max Welling. Variational bayes in private settings (vips). *arXiv preprint arXiv:1611.00340*, 2016. 5

[171] William D Penny, Karl J Friston, John T Ashburner, Stefan J Kiebel, and Thomas E Nichols. *Statistical parametric mapping: the analysis of functional brain images: the analysis of functional brain images*. Academic press, 2011. 3.1

[172] James Propp and David Wilson. Coupling from the past: a userąŕs guide. *Microsurveys in Discrete Probability*, 41:181–192, 1998. 3.3.3

[173] Lawrence Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. 3.1

[174] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS-07)*, pages 1177–1184, 2007.

2.5.3

[175] Arun Rajkumar and Shivani Agarwal. A differentially private stochastic gradient descent algorithm for multiparty classification. In *International Conference on Artificial Intelligence and Statistics*, pages 933–941, 2012. 3.6

[176] Aaditya Ramdas and Ryan J Tibshirani. Fast and flexible admm algorithms for trend filtering. *Journal of Computational and Graphical Statistics*, 2015. 6.4.1

[177] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 3.2, 3.4.1

[178] James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429): 122–129, 1995. 9.1, 9.2

[179] Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958. 1

[180] Jeffrey S Rosenthal. Minorization conditions and convergence rates for markov chain monte carlo. *Journal of the American Statistical Association*, 90(430):558–566, 1995. 3.3.3

[181] Christoph Rothe. The value of knowing the propensity score for estimating average treatment effects. *IZA Discussion Paper Series*, 2016. 9.1, 9.2, 9.3.2, 9.9.1, 9.9.2, 9.17, 9.9.2

[182] Leonid Rudin, Stanley Osher, and Emad Faterni. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259–268, 1992. II, 7.1, 11.1

[183] Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. *AISTATS*, 2016. 4.4, 4.4, 4.4, 10, 10.1, 10.1, 10.1, 10.3.1, 10.3.1, 10.5

[184] Veeranjaneyulu Sadhanala and Ryan J Tibshirani. Additive models with trend filtering. *arXiv preprint arXiv:1702.05037*, 2017. 8

[185] Veeranjaneyulu Sadhanala, Yu-Xiang Wang, James Sharpnack, and Ryan Tibshirani. Higher-order total variation classes on grids: Minimax theory and trend filtering methods. *Advances in Neural Information Processing Systems (NIPS-17)*, 2017. 8, 8.1

[186] Aliaksei Sandryhaila and José MF Moura. Discrete signal processing on graphs. *IEEE transactions on signal processing*, 61(7):1644–1656, 2013. 8

[187] Issei Sato and Hiroshi Nakagawa. Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process. In *International Conference on Machine Learning (ICML-14)*, pages 982–990, 2014. 3.4.1, 3.15, 2, 3.4.1

[188] Robert E Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990. 2.5.1

[189] FW Scholz and MA Stephens. K-sample anderson–darling tests. *Journal of the American Statistical Association*, 82(399):918–924, 1987. 8.5.2

[190] Simon Setzer, Gabriel Steidl, and Tanja Teuber. Infimal convolution regularizations with discrete l1-type functionals. *Communications in Mathematical Science*, 9(3):797–827,

2011. 6.2.4

[191] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010. 2.1, 2.1, 2.2.1, 2.2.1, 2.1, 2.2.3, 2.5, 2.6, 2.3, 2.3, 2.3.1, 2.3.2, 2.3.4, 2.3.5, 2.4, 2.4, 2.4.1, 2.4.3, 2.5.1, 2.35, 2.8.2, 2.8.2, 2.8.4, 4.1

[192] James Sharpnack and Aarti Singh. Identifying graph-structured activation patterns in networks. *Advances in Neural Information Processing Systems (NIPS-10)*, 13, 2010. 7.1, 7.5

[193] James Sharpnack, Alessandro Rinaldo, and Aarti Singh. Sparsistency via the edge lasso. *International Conference on Artificial Intelligence and Statistics (AISTATS-12)*, 15:1028–1036, 2012. 6.2.1

[194] James Sharpnack, Alessandro Rinaldo, and Aarti Singh. Sparsistency of the edge lasso over graphs. *International Conference on Artificial Intelligence and Statistics*, 15:1028–1036, 2012. 7.1

[195] James Sharpnack, Aarti Singh, and Akshay Krishnamurthy. Detecting activations over graphs using spanning tree wavelet bases. In *International Conference on Artificial Intelligence and Statistics (AISTATS-13)*, volume 16, pages 536–544, 2013. 6.1, 6.5.1, 6.8.1, 6.8.2

[196] James Sharpnack, Aarti Singh, and Alessandro Rinaldo. Changepoint detection over graphs with the spectral scan statistic. In *International Conference on Artificial Intelligence and Statistics (AISTATS-13)*, volume 16, pages 545–553, 2013. 6.6.2

[197] Or Sheffet. Differentially private ordinary least squares. In *International Conference on Machine Learning*, pages 3105–3114, 2017. 5.1, 5.5

[198] David Shuman, Sunil Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30 (3):83–98, 2013. 3, 8, 11.1

[199] Wacław Sierpiński. Sur les fonctions d'ensemble additives et continues. *Fundamenta Mathematicae*, 3:240–246, 1922. 9.7.1

[200] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, page 0956797611417632, 2011. 10.1

[201] Adam Smith. Efficient, differentially private point estimators. *arXiv preprint arXiv:0809.4794*, 2008. 3.6, 5.4, 5.4, 5.5

[202] Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *ACM symposium on Theory of Computing*, pages 813–822, 2011. I

[203] Alexander Smola and Risi Kondor. Kernels and regularization on graphs. In Bernhard Scholkopf and Manfred Warmuth, editors, *Learning Theory and Kernel Machines*, volume 2777 of *Lecture Notes in Computer Science*, pages 144–158. Springer, Berlin, 2003. 1, 6.2.3

[204] Alexander Smola and Risi Kondor. Kernels and regularization on graphs. *Annual Conference on Learning Theory*, 16, 2003. 7.1

[205] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing*, 2013. 3.3.1, 3.4, 3.6

[206] David Sontag and Dan Roy. Complexity of inference in latent dirichlet allocation. In *Advances in neural information processing systems (NIPS-11)*, pages 1008–1016, 2011. 3.3.3

[207] Daniel Spielman and Shang-Hua Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. *ACM Annual Symposium on Theory of Computing (STOC-04)*, 36:81–90, 2004. 6.4.1

[208] Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011. 5.3

[209] Gabriel Steidl, Stephan Didas, and Julia Neumann. Splines in higher order TV regularization. *International Journal of Computer Vision*, 70(3):214–255, 2006. II, 6.1, 8

[210] Charles Stein. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6):1135–1151, 1981. 11.1

[211] Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. *arXiv preprint arXiv:1410.1228*, 2014. 4.1, 10.1, 10.1, 10.2, 10.4

[212] Gilbert W Stewart. Perturbation theory for the singular value decomposition. Technical report, 1998. 5.23

[213] Partha Pratim Talukdar and Koby Crammer. New regularized algorithms for transductive learning. In *Machine Learning and Knowledge Discovery in Databases*, pages 442–457. Springer, 2009. 6.5.2, 6.5.2

[214] Partha Pratim Talukdar and Fernando Pereira. Experiments in graph-based semi-supervised learning methods for class-instance acquisition. In *Annual Meeting of the Association for Computational Linguistics*, pages 1473–1481. Association for Computational Linguistics, 2010. 6.5.2

[215] Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015. 10.1

[216] Jonathan Taylor, Richard Lockhart, Ryan J Tibshirani, and Robert Tibshirani. Exact post-selection inference for forward stepwise and least angle regression. *arXiv preprint arXiv:1401.3889*, 7, 2014. 10.1

[217] Abhradeep Guha Thakurta and Adam Smith. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, pages 819–850, 2013. I, 2.1

[218] Philip S Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning (ICML-16)*, 2016. 2, 9.4.2

[219] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996. II

[220] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67 (1):91–108, 2005. II, 7.1

[221] Ryan J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, 42(1):285–323, 2014. 1, II, 6, 6.1, 6.1, 6.2, 6.2.2, 6.2.3, 6.6.1, 6.6.2, 6.6.3, 6.7, 7.1, 7.1, 7.2, 8.1, 8.2, 8.2.2, 8.4, 8.4, 8.4.1, 8.7.5, 8.7.5, 8.7.7

[222] Ryan J. Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. *Annals of Statistics*, 39(3):1335–1371, 2011. 6.2.1, 6.2.1, 6.3.1, 7.1

[223] Ryan J. Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. *Annals of Statistics*, 40(2):1198–1232, 2012. 6.3.1

[224] Ryan J Tibshirani et al. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014. **??**, **??**

[225] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 4.3.3

[226] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1521–1528. IEEE, 2011. 1

[227] Alan M Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. 1

[228] Caroline Uhlerop, Aleksandra Slavković, and Stephen E Fienberg. Privacy-preserving data sharing for genome-wide association studies. *The Journal of privacy and confidentiality*, 5 (1):137, 2013. 4.1

[229] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984. 1, 2.2.1, 2.2.1

[230] Sara van de Geer. Estimating a regression function. *Annals of Statistics*, 18(2):907—924, 1990. 6.6.3, 6.6.3, 6.9.9, 6.9.10

[231] Sara van de Geer and Johannes Lederer. The lasso, correlated design, and improved oracle inequalities. *IMS Collecions*, 9:303–316, 2013. 6.6.3

[232] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000. 3.1, 3.3.2

[233] Tim Van Erven and Peter Harremoës. Rényi divergence and kullback-leibler divergence. *Information Theory, IEEE Transactions on*, 60(7):3797–3820, 2014. 4.3.2

[234] Vladimir Vapnik. *Statistical Learning Theory*. Wiley, 1998. 1, 2.4, 2.24, 2.4.3

[235] Vladimir N Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995. 2.1, 2.2.1

[236] Nisheeth Vishnoi. $Lx = b$: Laplacian solvers and their algorithmic applications. *Foundations and Trends in Theoretical Computer Science*, 8(1–2):1–141, 2012. 6.3.2

[237] Curtis R. Vogel and M. E. Oman. Iterative methods for total variation denoising. *SIAM Journal on Scientific Computing*, 17(1):227–238, 1996. 7.1

[238] Sebastian J Vollmer, Konstantinos C Zygalakis, et al. (non-) asymptotic properties of stochastic gradient langevin dynamics. *arXiv preprint arXiv:1501.00438*, 2015. 3.4.1, 3.4.1

[239] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395–416, 2007. 6

[240] Grace Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia, 1990. 8.1

[241] Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008. 3.3.1

[242] Yilun Wang, Junfeng Yang, Wotao Yin, and Yin Zhang. A new alternating minimization algorithm for total variation image reconstruction. *SIAM Journal on Imaging Sciences*, 1 (3):248–272, 2008. 7.1, 7.8

[243] Yining Wang, Yu-Xiang Wang, and Aarti Singh. Differentially private subspace clustering. *Advances in Neural Information Processing Systems (NIPS-15)*, 2015. 5

[244] Yu-Xiang Wang, Alex Smola, and Ryan J. Tibshirani. The falling factorial basis and its statistical applications. *International Conference on Machine Learning (ICML-14)*, 31, 2014. II, 6.1, 6.2, 6.2.2, 6.6.2, 6.9.3, 6.9.4

[245] Yu-Xiang Wang, Stephen E. Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient Monte Carlo. In *International Conference on Machine Learning (ICML-15)*, 2015. I, 4.3, 5.1, 5.4, 5.4, 5.1, 5.6

[246] Yu-Xiang Wang, Jing Lei, and Stephen E. Fienberg. Learning with differential privacy: Stability, learnability and the sufficiency and necessity of ERM principle. *Journal of Machine Learning Research*, 2016. I, 4.1

[247] Yu-Xiang Wang, Jing Lei, and Stephen E Fienberg. On-average kl-privacy and its equivalence to generalization for max-entropy mechanisms. *Privacy in Statistical Databases (PSD-2016)*, 2016. I, **??**

[248] Yu-Xiang Wang, James Sharpnack, Alex Smola, and Ryan J Tibshirani. Trend filtering on graphs. *Journal of Machine Learning Research*, 17(105):1–41, 2016. 7.1, 7.2, 7.8

[249] Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010. I, 2.2.2, 5.2

[250] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011. 3.1, 3.2, 3.4.1, 3.4.1, 5

[251] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6): 80–83, 1945. 8.5.2

[252] Oliver Williams and Frank McSherry. Probabilistic inference and differential privacy. In *Advances in Neural Information Processing Systems*, pages 2451–2459, 2010. 3.6

[253] Yonghui Xiao and Li Xiong. Bayesian inference under differential privacy. *arXiv preprint arXiv:1203.0617*, 2012. 3.6

[254] Fei Yu, Stephen E Fienberg, Aleksandra B Slavković, and Caroline Uhler. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of biomedical informatics*, 50:133–141, 2014. I, 4.1

[255] Fei Yu, Michal Rybar, Caroline Uhler, and Stephen E Fienberg. Differentially-private logistic regression for detecting multiple-snp association in gwas databases. In *International Conference on Privacy in Statistical Databases*, pages 170–184. Springer, 2014. 5.1

[256] Dengyong Zhou, Jiayuan Huang, and Bernhard Scholkopf. Learning from labeled and unlabeled data on a directed graph. *International Conference on Machine Learning (ICML-05)*, 22, 2005. 7.1

[257] Shuheng Zhou, John Lafferty, and Larry Wasserman. Compressed and privacy-sensitive sparse regression. *Information Theory, IEEE Transactions on*, 55(2):846–866, 2009. 4.3.3

[258] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. *International Conference on Machine Learning (ICML-03)*, 20, 2003. 7.1