Random Sampling Auctions for Limited Supply

Maria-Florina Balcan* **Nikhil Devanur†**
Jason D. Hartline‡ **Kunal Talwar‡**

September 2007
CMU-CS-07-154

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

* School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, ninamf@cs.cmu.edu.
† College of Computing, Georgia Institute of Technology, nikhil@gatech.edu
‡ Microsoft Research, Mountain View, CA, {hartline,kunal}@microsoft.com.

Abstract

Balcan et al. [3] show that the framework of the random sampling auction of Goldberg et al. [12] gives a generic reduction in the context of unlimited supply profit maximization from mechanism design (e.g., Goldberg et al. [12]) to algorithm design (e.g., Guruswami et al. [14]). One major open question left in [3] is to extend the results to limited supply settings. For the case that the seller has a large but limited supply of each commodity, we answer this question by giving a non-trivial adaptation of the random sampling auction framework to the limited supply setting, and prove bounds analogous to those of [3] for both the objectives of profit maximization and welfare maximization. These results generalize the prior work on random sampling limited supply auctions of Borgs et al. and Abrams [5, 1] which consider the special case where agents have linear valuations and budgets.

Keywords: Mechanism Design, Limited Supply, Random Sampling Auctions, Profit Maximization, Welfare Maximization.

1 Introduction

Random sampling is a fundamental technique in the design of economic mechanisms. Election polls are a vital part of a plurality voting system. Good pricing relies on good market analysis. In many contexts auctions based on random sampling give good approximations to the optimal mechanism (e.g., [12, 11, 3]). In this paper we extend the applicability of random sampling auctions to a very general multi-item, multi-unit settings with large but limited supply. The bidders may have arbitrary preferences over bundles of items and accompanying payments, though we assume that the demand by any one bidder for an item is small compared to the total supply of the item. In this context we show that the random sampling auction can optimize an arbitrary linear objective function such as the profit of the auctioneer or the social welfare of the economy.

This work generalizes theoretical mechanism design results related to web advertising and sponsored search when estimated supply is sold in batch. Borgs et al. [5] and Abrams [1] consider single-item, multi-unit auctions when bidders have linear valuations and budgets and the seller has a limited supply. Aggarwal and Hartline [2] consider single-item, multi-unit auctions for unit-demand advertisers with known click-through-rates when the seller has a limited supply. Balcan et al. [3] consider general advertiser preferences but in unlimited supply. In order for these models to be relevant for web advertising and sponsored search it is important to consider the general preferences (e.g., with budgets), multiple-items (i.e., different keywords, different web pages, etc.), and limited supply. This is setting is one of many covered by this paper.

The mechanism we discuss is based on the notion of an *offer*. An offer can be viewed as a menu describing possible outcomes of the mechanism and accompanying payments. We assume that a bidder's preferences maps any possible offer to their most desired outcome and payment. It is standard in mechanism design to assume that a bidder's preference is given by a *utility function* which maps outcomes and payments to a real number representing their happiness. Another standard assumption in mechanism design is that the utility functions are *quasi-linear*, i.e., there is a *valuation function* which maps outcomes to values and the utility function is just the difference between the valuation and the payment. In fact, even the celebrated Vickrey-Clarke-Groves (VCG) mechanism [16, 7, 13], requires quasi-linear preferences. Our general results make no such assumptions; however, it is natural to apply our results in these standard settings.

We consider *direct revelation* mechanisms, where each bidder is asked to submit as their "bid" a description of their preferences. A mechanism is *truthful* if, regardless of the actions of all other bidders, a bidder will obtain their most desired result by specifying their true preferences as their bid. We endow an offer with the following interpretation: given the preference of a bidder as their bid, select for the bidder the outcome and payment that their bid indicates is their most preferred. Thus, an offer is the simplest direct revelation, dominant strategy truthful mechanism.

The random sampling auction we consider can be stated simply as

1. Randomly partition the bidders into two parts.
2. Compute an "optimal offer" for each part.
3. "Use" each offer on the opposite part.

The terms "optimal offer" and "use" are deliberately vague, and we will specify their interpratations below.

This framework was first proposed for auctioning an unlimited supply of a single identical good to indistinguishable consumers with the objective of maximizing the profit[1] of the seller (a.k.a., the digital good auction problem) [12]. In unlimited supply settings there is a natural notion of an "optimal offer" for a

[1] Maximizing welfare in unlimited supply settings is trivial.

set of consumers. It is the one that maximizes the profit when made to each consumer independently. There is also a natural interpretation of "using" an offer on a set of consumers as independently assigning each consumer in the set the outcome that their reported preferences suggest that they most prefer. Since there are no supply constraints, both of these interpretations are well defined. Balcan et al. [3] give a simple analysis of this random sampling auction in any general unlimited supply setting. We believe that the simplicity and generality of this approach suggests that it is the "right way" to view the random sampling auction and its analysis in unlimited supply settings.

The challenge in adapting the random sampling auction framework to limited supply settings is in correctly specifying the notion of an "optimal offer" and "use" as it is no longer possible to making an offer to a set of consumers by independently making the offer to each consumer in the set. The main result of this paper provide a simple and general analysis that shows that if we

- interpret "use" as making an offer available "while supplies last" when the consumers are ordered in a random permutation (as proposed by Borgs et al. [5]), and
- interpret "optimal offer" of a set of consumers as the one that maximizes the expected payoff when made to consumers ordered by a random permutation,

then the performance of the random sampling auction is near optimal when specific, natural assumptions hold. Again, we believe that the simplicity and generality of this approach suggest that it is the "right way" to view the random sampling auction and its analysis in limited supply settings.

One of the main challenges addressed in this analysis is that for any given offer, the "bang per buck" of each need not be constant, as in [5]; and the demand of an agent need not be publicly known, as in [2]. This results in a knapsack like problem where each agent has a payoff and a demand and we wish to choose a set of agents such that the total demand is within our supply; however, incentive constraints require that we use neither the payoff nor demand when making our selection. Thus, our results are based on the study of this stochastic knapsack problem which may be of independent interest.

We follow an approach from [3]; a high level of generality is obtained by observing that when making an offer to a bidder the only relevant quantities are the *payoff* to our objective function and the *demand* imposed on our limited supply. Thus, we can remove the complicated details of a bidder's preference from the situation and assume that our payoff and demands are some function of the bidder and the offer. This allows us to express many objective functions with the only constraint being that the total payoff of our objective function is a simple sum of the payoffs from each bidder (i.e., it is linear). The two most natural objective functions to consider are *auctioneer profit* and *social welfare*. In the former, the payoff is simply the payment made by the bidder. In the latter, (assuming a utility function) the payoff is the utility of the bidder plus the payment made.[2] Consider the following examples:

Auctions for Web Advertising. The bidders are advertisers and the auctioneer is a webserver selling advertising space on a web page. The supply is given by statistics that estimate the number of times the web page will be viewed (in a given time interval). The advertisers' preferences may be described by a valuation per *impression* (i.e., views of their ad) and a *budget* for the maximum the advertiser is willing to pay (E.g., in the model proposed by Borgs et al. [5]). The utility of such a bidder is their valuation times the number of impressions obtained less their payment, as long as the payment is at most their budget. The bidder's utility for payments larger than their budget is negative infinity. Suppose the website made the offer "five cents per impression" to an advertiser with value 15 cents

[2] In the quasi-linear case, this sum is simply be the valuation of the bidder.

per impression and a budget of one dollar. The advertiser would buy 20 impressions for a total payment of one dollar. The utility of the advertiser is two dollars. The profit of the website is one dollar. The social welfare of the economy (the total valuation of the advertiser) would be three dollars. The supply consumed is 20 impressions.

Natural objectives to consider in this setting are profit maximization (as in [5]) and welfare maximization. Our work allows for results along the same lines as Borgs et al. but in much more general contexts: e.g., if there are multiple web pages, multiple ads on each web page, or user search queries such as in *auctions for sponsored search*. It also allows for more general advertiser preferences, e.g., with complicated valuations and budgeting conditions. Our bounds are similar to those of [5], though we require stronger assumptions to obtain a wider applicability.

Multi-unit, Multi-item Combinatorial Auctions. An auctioneer has multiple units of multiple items for sale (but limited supply). Bidders have quasi-linear preferences, i.e., a bidder has a function which specifies their value for every bundle that they might receive. Their utility is the difference between their value for a bundle and the price charged for it. Without computational constraints this mechanism design problem is solved by the VCG mechanism. With computational constraints, Bartal et al. [4] give mechanism that approximates the optimal social welfare.

Our techniques give near optimal results for the problem of maximizing profit or welfare. Our mechanism can be viewed as reducing the mechanism design problem to an algorithmic pricing problem (computing the optimal offer for a set of consumers). The incentive properties and performance bounds are preserved if an approximation algorithm is used in place of an optimal pricing algorithm.

For the objective of welfare maximization our results are interesting because the outcome we achieve is based on making a single offer to all bidders. This restriction is closely related to the requirement that the allocation and payments produced in the core, a mechanism design challenge recently considered by Milgrom [15] and Day and Raghavan [8]. To our knowledge, there was no truthful mechanism for this problem prior to our work.

There are a number of immediate extensions of our work that we do not explicitly consider here. First, the offers made to each bidder may use distinguishing features of the bidder to price discriminate. This ability is captured by the treatment of *attribute auctions* by [3]. Our results solve the limited supply attribute auction problem. Second, computational constraints may prevent exactly computing the optimal offers (as required in Step 2, above). A β-approximation algorithm can be used instead and results in a β factor loss in our objective function and slightly worse convergence rates (i.e., stronger assumptions on the size of the input and the available supply). We do not give the proof details. Finally, approaches for reducing the complexity of the class of offers by discretizing and running the auction restricted to a smaller class of offers or by doing data-dependent analyses to show that the effective complexity of the class of offers is small can be applied in our setting as well. We refer readers to [3] for details.

This paper is organized as follows. We start with our terminology and formal definitions in Section 2. We analyze the simpler single commodity case in Section 3, and the multiple commodities case in Section 4. We discuss computational issues and give conclusions in Section 5.

2 Preliminaries

The notions described in this section apply abstractly to mechanism design in limited supply settings. See Appendix A to see the constructs below instantiated in in the context of a simple web advertising example.

Suppose a seller is selling a limited supply C of some commodity. The seller could post an offer for units of this commodity as a pricing schedule which indicates how much any given quantity of the commodity costs. A buyer, endowed with a utility function which maps a quantity of the good and a price to a utility value representing their happiness, given such an offer would choose the quantity and price pair that maximizes their utility.

Let $\mathcal{G}$ be a class of allowable offers. As we have just discussed, given a particular offer $g \in \mathcal{G}$, a consumer will choose a quantity and price. Let $p(i, g)$ be the *payoff* to the mechanism designer's objective function when offering g to consumer i. For example, if our objective is *maximizing profit*, then $p(i, g)$ is the price paid by the consumer i when given offer g. Similarly, let $x(i, g)$ be the number of units they obtain.

Let $S = \{1, \ldots, n\}$ be a set of consumers. We extend the definitions of $p(\cdot, \cdot)$ and $x(\cdot, \cdot)$ to be the sum of the payoffs and quantities desired if each consumer was able to take their most desired quantity. I.e., $p(S, g) = \sum_{i \in S} p(i, g)$ and $x(S, g) = \sum_{i \in S} x(i, g)$.

We are going to be interested in the payoff we obtain from offering g to a set of consumers S. If we are lucky, the total quantity demanded by the consumers will be at most the supply, i.e, $x(S, g) \leq C$, and the payoff obtained by making such an offer will be $p(S, g)$. On the other hand, if the demanded quantity exceeds the supply, then the payoff obtained by such an offer is ill defined. To define the payoff in this case we need to be more explicit about how such an offer is made to the set of consumers.

Definition 1 *For a given offer g:*

1. *Let π be a random permutation.*
2. *Let k be the size of the largest prefix of π of consumers who can take what they want without exceeding the supply constraint. I.e., k satisfies*

$$0 \leq C - \sum_{i=1}^{k} x(\pi(i), g) < x(\pi(k+1), g).$$

3. *Define our payoff as $P(S, g, C) = \sum_{i=1}^{k} p(\pi(i), g)$.*

The value $P(S, g, C)$ is a random variable and is a lower bound on the true payoff we can obtain by an offer. In fact, we define our benchmark based on the expected value of this quantity. Formally:

Definition 2 *We define $\mathrm{OPT}_{\mathcal{G}}(S, C)$ to be $\max_{g \in \mathcal{G}} [\mathbf{E}_{\pi}[P(S, g, C)]]$.*

Since in general it is challenging to calculate the expected value of $P(S, g, C)$ we instead estimate its value with the following which, by design, is a simple function of $x(S, g)$ and $p(S, g)$.

Definition 3 *The* estimated payoff *for g on S is*

$$\tilde{P}(S, g, C) = \frac{C \cdot p(S, g)}{\max[C, x(S, g)]}.$$

Definition 4 *We define $\widetilde{\mathrm{OPT}}_{\mathcal{G}}(S, C) = \max_{g \in \mathcal{G}} \tilde{P}(S, g, C)$ and $\widetilde{\mathrm{opt}}_{\mathcal{G}}(S, C) = \mathrm{argmax}_{g \in \mathcal{G}} \tilde{P}(S, g, C)$.*

Now we formalize the auction we will be discussing throughout this paper. It is a generalization of the random sampling auctions of [12, 3, 5].

Definition 5 (RSLS) *The* random sampling limited supply (RSLS) *auction is parameterized by a class of allowable offers $\mathcal{G}$ and an available supply C. For a set of bidders S:*

1. *Partition S into S_1 and S_2 by assigning each bidder to S_1 or S_2 independently with probability $1/2$.*
2. *Compute $g_1 = \widetilde{\mathrm{opt}}_{\mathcal{G}}(S_1, C/2)$ and $g_2 = \widetilde{\mathrm{opt}}_{\mathcal{G}}(S_2, C/2)$.*
3. *Offer g_1 to S_2 and g_2 to S_1.*

The estimated payoff *of* RSLS *is* $\tilde{P}(S_2, g_1, C/2) + \tilde{P}(S_1, g_2, C/2)$. *The* actual payoff *of* RSLS *is* $P(S_2, g_1, C/2) + P(S_1, g_2, C/2)$.

We will show in the following section, that as long as both the maximum estimated payoff $\widetilde{\mathrm{OPT}}_{\mathcal{G}}(S, C)$ and the supply C are large compared to the complexity of our class of functions, the estimated payoff of the RSLS mechanism is nearly as large as the maximum estimated payoff. Note that the following lemma is immediate.

Lemma 1 *Mechanism* RSLS *is truthful.*

We state now a lemma from [3] which is essential in our analysis. We use the term *tally function* and notation $f(\cdot,\cdot)$ to refer generically to both $p(\cdot,\cdot)$ and $x(\cdot,\cdot)$.

Lemma 2 *Given S, a tally function $f(\cdot,\cdot)$, an offer g satisfying $f(i,g) \leq f_{\max}$ for all $i \in S$, and a tally level T; if we randomly partition S into S_1 and S_2, then the probability that $|f(S_1,g) - f(S_2,g)| \geq \epsilon \max\left[f(S,g), T\right]$ is at most $2e^{\left[-\frac{\epsilon^2 T}{2 f_{\max}}\right]}$.*

The following theorem follows from Lemma 2 by taking the union bound over all $g \in \mathcal{G}$ and all $f \in \mathcal{T}$.

Theorem 1 *Given*

- *a class of offers, $\mathcal{G}$,*
- *a set $\mathcal{T}$ of m tally functions,*
- *Upper bound $f_{\max}$ on $f(i,g)$ for all $i \in S$, and*
- *corresponding tally levels $T_f(\epsilon) = \frac{2 f_{\max}}{\epsilon^2} \ln\left(\frac{2m|\mathcal{G}|}{\delta}\right)$ for all $f \in \mathcal{T}$,*

With probability $1-\delta$ a random partitioning of S into two sets S_1 and S_2 satisfies

$$|f(S_1,g) - f(S_2,g)| \leq \epsilon \max\left[f(S,g), T_f(\epsilon)\right],$$

simultaneously for all $f \in \mathcal{T}$ and $g \in \mathcal{G}$.

Corollary 2 *For a random partition of S into S_1 and S_2, with probability at least $1-\delta$, every offer $g \in \mathcal{G}$ satisfies both*

$$|x(S_1,g) - x(S_2,g)| \leq \epsilon \max\left[x(S,g), T_x(\epsilon)\right]$$

and

$$|p(S_1,g) - p(S_2,g)| \leq \epsilon \max\left[p(S,g), T_p(\epsilon)\right].$$

Definition 6 (ϵ-good) *A partitioning of S into S_1 and S_2 is* ϵ-good *for g and tally function f if it satisfies the conditions of Lemma 2. The partitioning is* ϵ-good *for $\mathcal{G}$ and tally set $\mathcal{T}$ if it satisfies the conditions of Theorem 1.*

For convenience, we also state here a simple lemma that we will be useful in our analysis.

Lemma 3 *Let a, b, C be arbitrary non-negative real numbers. Then*

1. $\max(a, \frac{C}{2}) - \max(b, \frac{C}{2}) \le a - b$
2. $2\max(a, \frac{C}{2}) \le \max(a + b, C) + |a - b|$
3. $2\max(a, \frac{C}{2}) \ge \max(a + b, C) - |a - b|$

Proof: The first part follows immediately from the fact that function $f(x) = \max(x, \frac{C}{2})$ has slope at most 1 everywhere and is therefore 1-lipschitz. For the second part, note that $2a \le 2\max(a, b) = (a+b)+|a-b|$. Thus $\max(2a, C) \le \max(a + b + |a - b|, C) \le \max(a + b, C) + |a - b|$. The proof of the third part is analogous. $\square$

3 The Single Commodity Case

The structure of this section is as follows. We start by showing, in Section 3.1, that if both the maximum estimated payoff $\widetilde{\mathrm{OPT}}_{\mathcal{G}}(S, C)$ and the supply C are large compared to the complexity of our class of offers, then the estimated payoff of the `RSLS` mechanism is nearly as large as the maximum estimated payoff. We describe a natural randomized knapsack algorithm in Section 3.2 and analyze its performance. In Section 3.3 we combine the bounds of the preceding sections to relate the actual payoff of `RSLS` its estimated payoff.

For clarity, we will assume in the following that the offer class $\mathcal{G}$ is finite. We can easily extend our results to the case when the offer class is not finite, by using covering arguments as described in [3].

3.1 Analyzing the Estimated Payoff of the `RSLS` Mechanism

Using Corollary 2 we now prove that:

Theorem 3 *Given the offer class $\mathcal{G}$ and a limited supply C, with probability at least $1 - \delta$ the estimated payoff of* `RSLS` *mechanism is at least $(1 - \epsilon)\widetilde{\mathrm{OPT}}_{\mathcal{G}}(S, C)$ as long as*

$$\frac{\widetilde{\mathrm{OPT}}_{\mathcal{G}}(S,C)}{p_{\max}} \ge \frac{50}{\epsilon^2} \ln\left(\frac{4|\mathcal{G}|}{\delta}\right), \qquad \textit{and} \qquad \frac{C}{x_{\max}} \ge \frac{100}{\epsilon^2} \ln\left(\frac{4|\mathcal{G}|}{\delta}\right).$$

Proof: Let g_1 be the offer in $\mathcal{G}$ selected by our mechanism over S_1 and g_2 be the offer in $\mathcal{G}$ selected over S_2. Let $g^* = \widetilde{\mathrm{opt}}_{\mathcal{G}}(S, C)$. Since the optimal offer over S_1 is at least as good as g^* on S_1 (and likewise for S_2), we have $\tilde{P}(S_1, g_1, C/2) \ge \tilde{P}(S_1, g^*, C/2)$ and $\tilde{P}(S_2, g_2, C/2) \ge \tilde{P}(S_2, g^*, C/2)$.

From Corollary 2 we have that with probability $1 - \delta$, $\mathcal{G}$ is $\frac{\epsilon}{10}$-good, i.e., every offer g satisfies both $|x(S_1, g)-x(S_2, g)| \le \frac{\epsilon}{10} \max\left[x(S, g), T_x(\epsilon/10)\right]$ and $|p(S_1, g)-p(S_2, g)| \le \frac{\epsilon}{10} \max\left[p(S, g), T_p(\epsilon/10)\right]$. We will use this as well as the assumptions of the theorem statement that $\widetilde{\mathrm{OPT}}_{\mathcal{G}}(S) \ge T_p(\epsilon/10)$ and $C \ge 2T_x(\epsilon/10)$, to show that:

$$\tilde{P}(S_1, g^*, C/2) + \tilde{P}(S_2, g^*, C/2) \ge \left(1 - \frac{\epsilon}{5}\right) \tilde{P}(S, g^*, C) \tag{1}$$

and

$$\tilde{P}(S_j, g_i, C/2) \geq \left(1 - \frac{2\epsilon}{5}\right) \tilde{P}(S_i, g^*, C/2) - \frac{\epsilon}{10}\widetilde{\mathrm{OPT}}_{\mathcal{G}}(S), \tag{2}$$

for $i, j = 1, 2$, $j \neq i$. These then imply that the estimated payoff of the RSLS mechanism, namely the sum $\tilde{P}(S_2, g_1, C/2) + \tilde{P}(S_1, g_2, C/2)$, is at least $(1-\epsilon)\widetilde{\mathrm{OPT}}_{\mathcal{G}}(S)$.

We start by proving (1). Let $X \max(C, x(S, g^*))$, $X_1 = \max(C/2, x(S_1, g^*))$, $X_2 = \max(C/2, x(S_2, g^*))$. From the fact that $|x(S_1, g^*) - x(S_2, g^*)| \leq \frac{\epsilon}{5}X$ and Lemma 3,[3] we conclude that $2X_1 \leq (1+\frac{\epsilon}{5})X$ and $2X_2 \leq (1+\frac{\epsilon}{5})X$. Thus,

$$\tilde{P}(S_1, g^*, C/2) + \tilde{P}(S_2, g^*, C/2) = \left(\frac{Cp(S_1, g^*)}{2X_1} + \frac{Cp(S_2, g^*)}{2X_2}\right) \geq \left(\frac{Cp(S_1, g^*)}{(1+\frac{\epsilon}{5})X} + \frac{Cp(S_2, g^*)}{(1+\frac{\epsilon}{5})X}\right)$$
$$\geq (1-\tfrac{\epsilon}{5})\tilde{P}(S, g^*, C).$$

It remains to prove (2). We show this for $i = 1$ and $j = 2$, i.e., $\tilde{P}(S_2, g_1, C/2) \geq \left(1 - \frac{2\epsilon}{5}\right) \tilde{P}(S_1, g^*, C/2) - \frac{\epsilon}{10}\widetilde{\mathrm{OPT}}_{\mathcal{G}}(S)$; the other case is identical. Now let $X = \max(C, x(S, g_1))$, $X_1 = \max(C/2, x(S_1, g_1))$, $X_2 = \max(C/2, x(S_2, g_1))$. From the fact that $|x(S_1, g_1) - x(S_2, g_1)| \leq \frac{\epsilon}{10}X$ and Lemma 3, we conclude that $2X_2 \leq (1+\frac{\epsilon}{10})X$ and $2X_1 \geq (1-\frac{\epsilon}{10})X$. Thus $X_2 \leq \frac{1+\frac{\epsilon}{10}}{1-\frac{\epsilon}{10}}X_1 \leq \frac{1}{1-\frac{2\epsilon}{10}}X_1$. If $p(S, g_1) \geq T_p\,(\epsilon/10)$, then $p(S_2, g_1) \geq (1-\frac{2\epsilon}{10})p(S_1, g_1)$. In this case

$$\tilde{P}(S_2, g_1, C/2)\frac{C}{2}\frac{p(S_2, g_1)}{X_2} \geq \frac{C}{2}\frac{(1-\frac{2\epsilon}{10})p(S_1, g_1)}{X_1/(1-\frac{2\epsilon}{10})} \geq (1-\frac{2\epsilon}{5})\frac{C}{2}\frac{p(S_1, g_1)}{X_1}$$
$$= (1-\frac{2\epsilon}{5})\tilde{P}(S_1, g_1, C/2).$$

If $p(S, g_1) < T_p\,(\epsilon/10)$, then

$$\tilde{P}(S_2, g_1, C/2)\frac{C}{2}\frac{p(S_2, g_1)}{X_2} \geq \frac{C}{2}\frac{p(S_1, g_1) - \frac{\epsilon}{10}T_p\,(\epsilon/10)}{X_1/(1-\frac{2\epsilon}{10})} \geq \left(1-\frac{2\epsilon}{5}\right)\tilde{P}(S_1, g_1, C/2) - \frac{\epsilon}{10}\widetilde{\mathrm{OPT}}_{\mathcal{G}}(S)$$

This completes the proof. □

Notice that this bound holds for all ϵ and δ simultaneously as these are not parameters of the mechanism.

3.2 The Uniform Knapsack Algorithm

In this section we present bounds on the performance of the following *uniform knapsack* algorithm for both *weighted* and *non-weighted* instances. While this is a natural algorithm to consider, we know of no such analysis already in the literature. We need these bounds in our main theorem to show that the actual payoff of RSLS is close to its estimated payoff.

Definition 7 (Uniform Knapsack Algorithm, UKA) *Given a knapsack instance, the* uniform knapsack algorithm *(UKA) works as follows:*

1. *Pick an ordering of the objects uniformly at random.*

[3] Note that we are using here the assumption that $C \geq 2T_x(\epsilon/10)$.

2. *Insert the objects into the knapsack in order until the first time an object does not fit.*

First some notation. A knapsack instance I is defined by the capacity of the knapsack, C, the number of pieces, n, their sizes $x_1, x_2, \ldots, x_n$ and weights (a.k.a., payoffs) $p_1, p_2, \ldots, p_n$. For any permutation $\pi \in \mathcal{S}_n$, the set of all permutation, let $k(\pi)$ be the number of pieces that fit in the knapsack, if we fill the knapsack in the order given by π, i.e., $k(\pi)$ is such that $0 \leq C - \sum_{i=1}^{k(\pi)} x_{\pi(i)} < x_{\pi(k(\pi)+1)}$. Then the payoff of UKA on the instance I with permutation π is defined to be

$$P(I, \pi) = \sum\nolimits_{i=1}^{k(\pi)} p_{\pi(i)}.$$

Consider the following definitions. The maximum size of an object is $x_{\max} = \max_i\{x_i\}$. The total demand is $x_{\text{sum}} = \sum_i x_i$. The total payoff if we could fit all objects into the knapsack would be $p_{\text{sum}} = \sum_i p_i$. Assume without loss of generality that $x_{\max} \leq C \leq x_{\text{sum}}$.

3.2.1 Bounds on Expectation

The main theorem of this section is the following. The proof of this theorem can be viewed as an approximate reduction from the weighted case to the non-weighted case.

Theorem 4

$$\mathbf{E}_{\pi \in_r \mathcal{S}_n}[P(I, \pi)] \geq \frac{p_{\text{sum}}\,(C - 4x_{\max})}{x_{\text{sum}}}.$$

Proof: Let $P_i(I, \pi)$ be the payoff from piece i under permutation π, i.e., $P_i(I, \pi)$ is equal to p_i if $\pi^{-1}(i) \leq k(\pi)$ and 0 otherwise. So

$$\mathbf{E}_{\pi \in_r \mathcal{S}_n}\left[P_i(I, \pi)\right] = p_i \cdot \mathbf{Pr}_{\pi \in_r \mathcal{S}_n}\left[\pi^{-1}(i) \leq k(\pi)\right].$$

One way to pick a random permutation π on n elements is to pick a random permutation π_{-i} on all the elements but i, and then insert i randomly at one of n positions. $\pi^{-1}(i) \leq k(\pi)$ if i is inserted in one of the first $k(\pi_{-i})$ positions in a knapsack of capacity $C - x_i$. Lemma 4 (see below) implies,

$$\mathbf{Pr}_{\pi \in_r \mathcal{S}_n}\left[\pi^{-1}(i) \leq k(\pi)\right] = \mathbf{E}_{\pi_{-i} \in_r \mathcal{S}_{n-1}}\left[\tfrac{k(\pi_{-i})}{n}\right] \geq \frac{(n-1)\left((C - x_i) - 2x_{\max}\right)}{n x_{\text{sum}}},$$

Therefore,

$$\mathbf{E}_{\pi \in_r \mathcal{S}_n}[P(I, \pi)] = \sum\nolimits_{i=1}^{n} \mathbf{E}_{\pi \in_r \mathcal{S}_n}[P_i(I, \pi)] = \sum\nolimits_{i=1}^{n} p_i \cdot \mathbf{E}_{\pi_{-i} \in_r \mathcal{S}_{n-1}}\left[\tfrac{k(\pi_{-i})}{n}\right] \geq \frac{(n-1)p_{\text{sum}}\,(C - 3x_{\max})}{n x_{\text{sum}}}.$$

Since $x_{\text{sum}} \geq C$ implies that $1/n \leq x_{\max}/C$, we have: $\mathbf{E}_{\pi \in_r \mathcal{S}_n}[P(I, \pi)] \geq \frac{p_{\text{sum}}(C - 4x_{\max})}{x_{\text{sum}}}$. □

Notice that $k(\pi)$ is the payoff of UKA in the non-weighted case (i.e., when $p_i = 1$ for all i). Thus, the following main technical lemma in this section gives a bound on the expected performance of the UKA algorithm in the non-weighted case.

Lemma 4 $\mathbf{E}_{\pi \in_r \mathcal{S}_n}\left[k(\pi)\right] \geq \frac{n(C - 2x_{\max})}{x_{\text{sum}}}$.

Proof: Step 1. We modify the distribution on the permutations so that the expectation only gets smaller. Let $\pi \in \mathcal{D}$ be the probability distribution on the set of all permutations $\mathcal{S}_n$ obtained as follows: pick $\pi(1)$ to be i with probability proportional to x_i, and then pick the rest of the permutation uniformly at random. The distribution $\pi \in \mathcal{D}$ is such that bigger pieces have a higher probability of being picked as the first element. Hence, the expected number of pieces that fit in the knapsack is only smaller. Thus, we have:

$$\mathbf{E}_{\pi \in_r \mathcal{S}_n}[k(\pi)] \geq \mathbf{E}_{\pi \in \mathcal{D}}[k(\pi)]. \tag{3}$$

Step 2. We define a "fractional" version of $k(\pi)$ that considers the total fraction of the number of objects in the knapsack of size C when we place the objects in on the $[0, x_{\text{sum}}]$ line according to $\pi \in_r \mathcal{S}_n$ but start the knapsack at a uniformly random position $q \in [0, x_{\text{sum}}]$ and wrap around cyclically. For a permutation π, starting position q, and supply C we let $k(\pi, q, C)$ represent the total fractional number of objects placed in the knapsack by this procedure. Formally, define the following regions on the $[0, x_{\text{sum}}]$ line:

- The region of objects taken into the knapsack is R defined as:

$$R = \begin{cases} [q, q+C] & \text{if } q + C \leq x_{\text{sum}}, \text{ and} \\ [q, x_{\text{sum}}] \cup [0, C - x_{\text{sum}} + q] & \text{otherwise.} \end{cases}$$

- The region taken up by object i is R_i defined as (the sum is over objects before i in π):

$$R_i = [0, x_i] + \sum_{j=1}^{\pi^{-1}(i)-1} x_{\pi(j)}.$$

Let $|R|$ be the size of a region. The value of $k(\pi, q, C) = \sum_i \frac{|R \cap R_i|}{|R_i|}$. Clearly,

$$\mathbf{E}_{\pi \in_r \mathcal{S}_n, q \in_r [0, x_{\text{sum}})}[k(\pi, q, C)] = \frac{nC}{x_{\text{sum}}}. \tag{4}$$

Step 3. We now compare $\mathbf{E}_{\pi \in_r \mathcal{S}_n, q \in_r [0, x_{\text{sum}})}[k(\pi, q, C)]$ with $\mathbf{E}_{\pi \in \mathcal{D}}[k(\pi)]$. Notice that the probability that an object is the first object in the knapsack according to π when we start at uniform q is proportional to its size. The subsequent objects are identical to that of a random permutation from $\mathcal{S}_{n-1}$. Thus, the order of objects we attempt to place in the knapsack are the same under $\pi \in \mathcal{D}$ starting $q = 0$ and under $\pi \in_r \mathcal{S}_n$ with start position $q \in_r [0, x_{\text{sum}})$. In the former case, however, the first object is placed in the knapsack in its entirety. This may block out an object at the end. Additionally, in the former case, we do not get credit for any fractional objects at the end. We can remedy this by comparing the former case to the latter case with a reduced knapsack capacity: the objects in the knapsack in the former case with capacity C is a superset of the objects in the knapsack in the latter case with capacity $C' = C - 2x_{\max}$. Thus,

$$\mathbf{E}_{\pi \in \mathcal{D}}[k(\pi)] \geq \mathbf{E}_{\pi \in_r \mathcal{S}_n, q \in_r [0, x_{\text{sum}})}[k(\pi, q, C - 2x_{\max})]. \tag{5}$$

Combining (3), (4), and (5), we obtain the statement of the lemma. □

3.2.2 Concentration Bounds

In this section we show that the random variable $k(\pi)$ in Lemma 4 is tightly concentrated around its expected value. Our main theorem on this topic is stated below, it is actually a immediate consequence of Theorem 9 which is given in the Appendix B. Notice that the assumption required for this bound is likely to be weaker than the assumption required in Theorem 3.

Theorem 5 *With probability at least $1-\delta$ the the number of objects in the knapsack, $k(\pi)$, is at least $(1-\epsilon)\mathbf{E}_\pi[k(\pi)]$ as long as,*

$$C \geq \frac{4(1+\epsilon)}{\epsilon^2}\frac{x_{\text{max}}^2}{x_{\text{avg}}}\ln(\tfrac{1}{\delta}).$$

From this theorem we can derive a similar concentration result for the weighted version of the knapsack problem; however, we do not need this result for the paper so we omit the details.

3.3 Analyzing the Actual Payoff of the `RSLS` Mechanism

Putting together the results in Sections 3.1 and 3.2, we are now able to derive a bound on the expected actual payoff of the `RSLS` mechanism.

Theorem 6 *Given the offer class $\mathcal{G}$ and a limited supply C, with probability at least $1-\delta$ the expected actual payoff of* `RSLS` *is at least $(1-\epsilon)\,\text{OPT}_\mathcal{G}(S,C)$ as long as*

$$\frac{\widetilde{\text{OPT}}_\mathcal{G}(S)}{p_{\text{max}}} \geq O\left(\tfrac{1}{\epsilon^2}\ln\left(\tfrac{4|\mathcal{G}|}{\delta}\right)\right) \qquad \textit{and} \qquad \frac{C}{x_{\text{max}}} \geq O\left(\tfrac{1}{\epsilon^2}\ln\left(\tfrac{4|\mathcal{G}|}{\delta}\right)\right).$$

Proof: The expected *actual payoff* of `RSLS` mechanism which we denote by $\mathbf{E}[P(\texttt{RSLS})]$ is $\mathbf{E}_\pi[P(S_2,g_1,C/2)]+\mathbf{E}_\pi[P(S_1,g_2,C/2)]$; so using Theorem 4, we obtain:

$$\mathbf{E}[P(\texttt{RSLS})] \geq \frac{p(S_2,g_1)\cdot\left(\frac{C}{2}-4x_{\text{max}}\right)}{\max(x(S_2,g_1),\frac{C}{2})} + \frac{p(S_1,g_2)\cdot\left(\frac{C}{2}-4x_{\text{max}}\right)}{\max(x(S_1,g_2),\frac{C}{2})}.$$

Note that by the second assumption, $x_{\text{max}} < \epsilon^2 C$ so that $(\frac{C}{2}-4x_{\text{max}}) > (1-8\epsilon^2)(\frac{C}{2})$. Thus we get

$$\mathbf{E}[P(\texttt{RSLS})] \geq (1-8\epsilon^2)\left(\frac{p(S_2,g_1)\cdot(\frac{C}{2})}{\max(x(S_2,g_1),\frac{C}{2})} + \frac{p(S_1,g_2)\cdot(\frac{C}{2})}{\max(x(S_1,g_2),\frac{C}{2})}\right).$$

The expression in the parentheses is precisely the estimated payoff of mechanism. Thus Theorem 3 implies that with probability at least $1-\delta$ we have:

$$\mathbf{E}[P(\texttt{RSLS})] \geq (1-\tilde{\epsilon})\widetilde{\text{OPT}}_\mathcal{G}(S,C).$$

Finally, using Theorem 4 again as above implies that $\widetilde{\text{OPT}}_\mathcal{G}(S,C) \geq (1-4\epsilon^2)\,\text{OPT}_\mathcal{G}(S,C)$, the claim follows. □

4 Multiple Commodities

Consider the case that there are m commodities with supply constraint $\mathbf{C}=(C_1,\ldots,C_m)$ with C_j giving the available supply of commodity j. Now, each g is a menu giving a price for all possible bundles of the m commodities (these bundles are multisets). We define $p(i,g)$ as before and define $x_j(i,g)$ as the quantity of item j that consumer i desires when offered g. We now extend the definitions of $P(\cdot,\cdot,\cdot)$ and $\tilde{P}(\cdot,\cdot,\cdot)$ to the multi-commodity case.

Definition 8 *For a given offer g:*

1. *Let π be a random permutation.*
2. *Let k_j be the size of the largest prefix of π of consumers who can take what they want without exceeding the supply constraint of item j (ignoring the supply constraint of other bundles). I.e., k_j satisfies*

$$0 \le C_j - \sum_{i=1}^{k_j} x_j(\pi(i), g) < x_j(\pi(k_j+1), g).$$

3. *Define $P_j(S, g, C_j) = \sum_{i=1}^{k_j} p(\pi(i), g)$.*
4. *Define our payoff as $P(S, g, \mathbf{C}) = \min_j P_j(S, g, C_j)$.*

We estimate $P(S, g, \mathbf{C})$ value with the following which, by design, is a simple function of $\mathbf{x}(S, g)$ and $p(S, g)$.

Definition 9 *Let $\tilde{P}_j(S, g, C_j) = \frac{C_j \cdot p(S,g)}{\max[C_j, x_j(S,g)]}$ and define the* estimated payoff *for g on S as*

$$\tilde{P}(S, g, \mathbf{C}) = \min_j \tilde{P}_j(S, g, C_j).$$

Theorem 7 *Given the offer class $\mathcal{G}$ and a limited supply $\mathbf{C}$, with probability at least $1-\delta$ the estimated payoff of* RSLS *is at least $(1-\epsilon)\widetilde{\mathrm{OPT}}_{\mathcal{G}}(S, \mathbf{C})$ as long as*

$$\frac{\widetilde{\mathrm{OPT}}_{\mathcal{G}}(S,\mathbf{C})}{p_{\max}} \ge \frac{50}{\epsilon^2} \ln\left(\frac{2(m+1)|\mathcal{G}|}{\delta}\right) \qquad \textit{and} \qquad \frac{C_j x_{j,\mathrm{avg}}}{x_{j,\max}} \ge \frac{100}{\epsilon^2} \ln\left(\frac{2(m+1)|\mathcal{G}|}{\delta}\right)$$

for all items j.

Proof: Theorem 1 implies that with probability $1-\delta$ all the m demand tally functions $x_1(\cdot,\cdot), \ldots, x_m(\cdot,\cdot)$ and the payoff tally function $p(\cdot,\cdot)$ are $\frac{\epsilon}{5}$-good for sampling S_1 and S_2.

Let $g_1 = \widetilde{\mathrm{opt}}(S_1)$ and $g_2 = \widetilde{\mathrm{opt}}(S_2)$ and $g^* = \widetilde{\mathrm{opt}}(S)$. If follows from the proof of Theorem 3 and our assumptions on $\widetilde{\mathrm{OPT}}_{\mathcal{G}}(S, \mathbf{C})$ and $\mathbf{C}$ that for all items j,

$$\tilde{P}_j(S_2, g_1, C_j/2) + \tilde{P}_j(S_1, g_2, C_j/2) \ge (1-\epsilon)\tilde{P}_j(S, g^*, C_j).$$

Our estimated payoff is simply the minimum over these over items j of the estimated payoff for that item. Thus the profit of RSLS is at least $(1-\epsilon)\widetilde{\mathrm{OPT}}_{\mathcal{G}}(S, \mathbf{C})$. □

Theorem 8 *Given the offer class $\mathcal{G}$ and a limited supply $\mathbf{C}$, with probability at least $1-\delta$ the expected actual payoff of* RSLS *is at least $(1-\epsilon)\,\mathrm{OPT}_{\mathcal{G}}(S, \mathbf{C})$ as long as*

$$\frac{\widetilde{\mathrm{OPT}}_{\mathcal{G}}(S)}{p_{\max}} \ge O\left(\frac{1}{\epsilon^2} \ln\left(\frac{2(m+1)|\mathcal{G}|}{\delta}\right)\right) \qquad \textit{and} \qquad \frac{C_j}{x_{j,\max}^2} \ge O\left(\frac{1}{\epsilon^2} \ln\left(\frac{2(m+1)|\mathcal{G}|}{\delta}\right)\right),$$

for all items j.

Proof: (sketch)

1. Consider g_1 on S_2 and the supply constraint C_j for item j (assume all other supplies are infinite). Use Theorem 5 to argue that the actual number of bidders able to satisfy their demand for item j with respect to item j is close to the estimated number (with probability $1-\delta/m$).

2. Use the union bound on all items j to argue that the actual number of bidders able to satisfy their demand, with respect to the most constrained item supply, is close to the estimated number (with probability $1 - \delta$).

3. Use the reduction approach (which relates the weighted knapsack bound to the non-weighted bound) implied by the proof of Theorem 4 to show that the expected actual payoff from g_1 on S_2 is close to the estimated payoff for g_1 on S_2.

4. Use Theorem 7 to conclude that the actual payoff is close to the optimal estimated payoff which in turn is close to the optimal actual payoff.

□

5 Conclusions, Discussion and Open Problems

In this paper we have described a very general random sampling mechanism for multi-item, multi-unit settings with large but limited supply. We have proved bounds analogous to those of [3] that can be applied to both the objectives of profit maximization and welfare maximization.

Computational Issues. While our mechanism and analysis are very general, they rely on the existence of an algorithm that computes $\widetilde{\mathrm{opt}}_{\mathcal{G}}(S, C)$ or an approximation to it. This raises the natural *algorithmic pricing* question: Can this be computed efficiently? It turns out that even in the unlimited supply setting [3] (which we generalize here) computing these optimal offers is intractable for most interesting problems [14, 9, 6]. None-the-less there may be interesting pricing problems in our setting that are tractable and it is interesting to consider the problem for special classes of consumer preferences and offers.

We now consider special cases of the above problem. Suppose the class of offers $\mathcal{G}$ is the set of all *item-pricings* (i.e., prices are a linear function of the quantity demanded). Suppose further that the consumers preferences are specified by a linear valuation function and a budget that is their maximum willingness to pay. For the case of a single commodity, this precisely the model considered in the *auctions for web advertising* example in the introduction of this paper. It is easy to see that there is a simple algorithm for the case. For the case of multiple commodities, the prices that maximize profit are exactly the market clearing prices. Hence one can use efficient market clearing algorithms to find the prices. Such algorithms are known only for divisible goods [10], and in fact market clearing prices may not exist for indivisible goods. Since the supply for each commodity is large, it is a reasonable approximation to treat the goods as divisible.

References

[1] Z. Abrams. Revenue Maximization When Bidders Have Budgets. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2006.

[2] G. Aggarwal and J. Hartline. Knapsack Auctions. In *Proceedings of the 17th ACM-SIAM Symposium on Discrete Algorithms*, 2006.

[3] M.-F. Balcan, A. Blum, J. Hartline, and Y. Mansour. Mechanism Design via Machine Learning. In *46th Annual IEEE Symposium on Foundations of Computer Science*, pages 605 – 614, 2005.

[4] Y. Bartal, R. Gonen, and N.Nisan. Incentive compatible multi unit combinatorial auctions. In *Proceedings of the 9th conference on Theoretical aspects of rationality and knowledge*, pages 72–87, 2003.

[5] C. Borgs, J. T. Chayes, N. Immorlica, M. Mahdian, and A. Saberi. Multi-unit auctions with budget-constrained bidders. In *Proceedings of the 6th ACM Conference on Electronic Commerce*, pages 44–51, 2005.

[6] P. Briest and P. Krysta. Single-minded unlimited-supply pricing on sparse instances. In *Symposium on Discrete Algorithms*, 2006.

[7] E. Clarke. Multipart Pricing of Public Goods. *Public Choice*, 11:17–33, 1971.

[8] R. Day and S. Raghavan. Fair Payments and Efficient Allocations in Public Sector Combinatorial Auctions. In *Dagstuhl Procedings*, 2005.

[9] E. Demaine, U. Feige, M. Hajiaghayi, and M. Salavatipour. Combination Can Be Hard: Approximability of the Unique Coverage Problem . In *Proceedings of the 17th ACM-SIAM Symposium on Discrete Algorithms*, 2006.

[10] N. Devanur, C. H. Papadimitriou, A. Saberi, and V. V. Vazirani. Market equilibrium via a primal-dual-type algorithm. In *Proceedings of IEEE Annual Symposium on Foundations of Computer Science*, 2002.

[11] S. Dobzinski, N. Nisan, and M. Schapira. Truthful Randomized Mechanisms for Combinatorial Auctions. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 644 – 652, 2006.

[12] A. Goldberg, J. Hartline, and A. Wright. Competitive Auctions and Digital Goods. In *Proceeding of the 12th ACM-SIAM Symposium on Discrete Algorithms*, pages 735–744, 2001.

[13] T. Groves. Incentives in Teams. *Econometrica*, 41:617–631, 1973.

[14] V. Guruswami, J. Hartline, A. Karlin, D. Kempe, C. Kenyon, and F. McSherry. On Profit-Maximizing Envy-Free Pricing. In *Proceedings of the 16th ACM-SIAM Symposium on Discrete Algorithms*, pages 1164 – 1173, 2005.

[15] P. Milgrom. Incentives in core-selecting auctions. Levine's Bibliography 321307000000000503, UCLA Department of Economics, Oct. 2006. available at http://ideas.repec.org/p/cla/levrem/321307000000000503.html.

[16] W. Vickrey. Counterspeculation, Auctions, and Competitive Sealed Tenders. *Journal of Finance*, 16:8–37, 1961.

A Web Advertising Example

Our notions of offers, payoffs, and demands are fairly abstract. To gain some appreciation for them we demonstrate how they can be applied to the objective of profit maximization in the *web advertising* example discussed briefly in the introduction. For the purpose of this example define notation v_i as the per-impression valuation of advertiser i and B_i as their total budget. A reasonable class of offers, $\mathcal{G}$, are those of the form

g_q = "buy as many impressions as you want at price q each." As our objective is profit maximization, $p(i, g_q)$ is the payment advertiser i makes. Now consider what happens when advertiser i faces offer g_q. If $v_i \geq q$ then the advertiser has positive utility for buying and will buy up to their budget. Thus, the payment made is $p(i, g_q) = B_i$. The supply demanded is the total number of impressions they can afford given the price-per-impression cost q and their budget, i.e., $x(i, g_q) = B_i/q$.

Now suppose we have two advertisers, $S = \{1, 2\}$, with valuation and budgets, $v_1 = 10$, $B_1 = 20$, $v_2 = 5$, $B_2 = 30$; and supply, $C = 5$. The relevant offers are g_5 and g_{10}, the offers of \$5 per impression and \$10 per impression, respectively. It is easy to verify that

$$\begin{array}{lclclcr} p(1, g_{10}) & = & 20 & x(1, g_{10}) & = & 2 \\ p(2, g_{10}) & = & 0 & x(2, g_{10}) & = & 0 \\ p(\{1, 2\}, g_{10}) & = & 20 & x(\{1, 2\}, g_{10}) & = & 2 \\ \\ p(1, g_5) & = & 20 & x(1, g_5) & = & 4 \\ p(2, g_5) & = & 30 & x(2, g_5) & = & 6 \\ p(\{1, 2\}, g_5) & = & 50 & x(\{1, 2\}, g_5) & = & 10 \end{array}$$

Clearly, $\tilde{P}(S, g_{10}, C) = 20$ and $\tilde{P}(S, g_5, C) = 25$ so $\widetilde{\text{OPT}}_{\{g_5, g_{10}\}}(S, C) = 25$ and $\widetilde{\text{opt}}_{\{g_5, g_{10}\}}(S, C) = g_5$.

B Proofs of Concentration Bounds for the Uniform Knapsack Algorithm

In this section we given the main theorem necessary to show the concentration result for the uniform knapsack algorithm in the non-weighted case (Theorem 5). We obtain this result by studying a much simpler random object, the sum of k numbers selected randomly without replacement from a set of n numbers. The concentration results we obtain for this problem follow naturally from the method of average bounded differences; proofs are given for completeness.

Theorem 9 *Let X_k be a random variable representing the sum of k numbers drawn without replacement from a set of n numbers with maximum magnitude $x_{\max}$ and average x_{avg}, then*

$$\mathbf{Pr}[X_k \geq (1 + \epsilon)\mathbf{E}[X_k]] \leq \exp(-\epsilon^2 (\tfrac{x_{\text{avg}}}{x_{\max}})^2 \tfrac{k)}{2}.$$

Notice that if $k > \frac{4}{\epsilon^2} (\frac{x_{\max}}{x_{\text{avg}}})^2 \ln(\frac{1}{\delta})$, then with probability $(1 - \delta)$, $X_k < (1 + \epsilon)\mathbf{E}[X_k] = (1 + \epsilon)kx_{\text{avg}}$. Setting $k = \frac{C}{(1+\epsilon)x_{\text{avg}}}$ and flipping this around, we get Theorem 5 which is given and discussed in Section 3.2.2.

Our proof is based on the following theorem from the method of average bounded differences.

Theorem 10 (Method of Average Bounded Differences) *Let $Z_1, \ldots, Z_k$ be an arbitrary set of random variables and let X_k be some function of $Z_1, \ldots, Z_k$ satisfying, for each $i \in \{1, \ldots, k\}$, there is a c_i with*

$$|\mathbf{E}[X_k | Z_1, \ldots, Z_i] - \mathbf{E}[X_k | Z_1, \ldots, Z_{i-1}]| \leq c_i.$$

Then,

$$\mathbf{Pr}[X_k > \mathbf{E}[X_k] + t] \leq \exp(-t^2/2 \sum_i c_i^2).$$

Lemma 5 *Drawing k numbers without replacement from a set of n numbers with magnitude at most $x_{\max}$, let Z_i represent the ith number drawn and $X_i = \sum_{j \le i} Z_j$. Then,*

$$\mathbf{E}[X_k | Z_1, \ldots, Z_i] - \mathbf{E}[X_k | Z_1, \ldots Z_{i-1}] \le x_{\max}.$$

Proof: Let $x_1, \ldots, x_n$ represent the n numbers with sum $x_{\text{sum}} = \sum_i x_i$. Picking k of n numbers has expected sum $\mathbf{E}[X_k] = \frac{k}{n} x_{\text{sum}}$. For $j \le i$,

$$\mathbf{E}[X_k | Z_1, \ldots, Z_j] = \mathbf{E}[X_i | Z_1, \ldots, Z_j] + \mathbf{E}[X_k - X_i | Z_1, \ldots, Z_j]$$

To calculate the last term above, notice that $X_k - X_i$ given $Z_1, \ldots, Z_j$ is the sum of $k - i$ numbers selected from $n - i$ numbers with total sum $x_{\text{sum}} - \mathbf{E}[X_i | Z_1, \ldots, Z_j]$. Thus,

$$\begin{aligned} \mathbf{E}[X_k | Z_1, \ldots, Z_j] &= \mathbf{E}[X_i | Z_1, \ldots, Z_j] + \tfrac{k-i}{n-i} \left(x_{\text{sum}} - \mathbf{E}[X_i | Z_1, \ldots, Z_j] \right) \\ &= \tfrac{n-k}{n-i} \mathbf{E}[X_i | Z_1, \ldots, Z_j] + \tfrac{k-i}{n-i} x_{\text{sum}}. \end{aligned} \tag{6}$$

Renumber the objects $x_1, \ldots, x_k$ in the order they are drawn, thus, for $j \le i$, $\mathbf{E}[Z_j | Z_1, \ldots, Z_i] = x_j$. Note that $\mathbf{E}[X_i | Z_1, \ldots, Z_{i-1}] = \mathbf{E}[Z_i | Z_1, \ldots, Z_{i-1}] + \mathbf{E}[X_{i-1} | Z_1, \ldots, Z_{i-1}]$. Using this fact and equation (6) substituting $j = i - 1$ and $j = i$ and subtracting, we have,

$$\begin{aligned} \left| \mathbf{E}[X_k | Z_1, \ldots, Z_i] - \mathbf{E}[X_k | Z_1, \ldots, Z_{i-1}] \right| &= \tfrac{n-k}{n-i} \left| \mathbf{E}[X_i | Z_1, \ldots, Z_i] - \mathbf{E}[X_i | Z_1, \ldots, Z_{i-1}] \right| \\ &= \tfrac{n-k}{n-i} \left| x_i - \mathbf{E}[Z_i | Z_1, \ldots, Z_{i-1}] \right| \\ &\le x_{\max}. \end{aligned}$$

□

The proof of Theorem 9 follows from Theorem 10 and Lemma 5 by plugging in $c_i = x_{\max}$.

Proof: (of Theorem 9) Recall that $\mathbf{E}[X_k] = \frac{k}{n} x_{\text{sum}} = k x_{\text{avg}}$. Using Theorem 10 with $t = \epsilon \mathbf{E}[X_k]$, $c_i = x_{\max}$, we get that the failure probability

$$\begin{aligned} \mathbf{Pr}[X_k > (1 + \epsilon) \mathbf{E}[X_k]] &\le \exp(-\epsilon^2 k^2 x_{\text{avg}}^2 / 2 x_{\max}^2 k) \\ &\le \exp(-\epsilon^2 (\frac{x_{\text{avg}}}{x_{\max}})^2 \frac{k}{2}) \end{aligned}$$

□