# On Fixed Convex Combinations of No-Regret Learners

Jan-P. Calliess*

September 2008
CMU-ML-08-112


Carnegie Mellon

# On Fixed Convex Combinations of No-Regret Learners

**Jan-P. Calliess***

September 2008
CMU-ML-08-112

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

* This research was completed while visiting Carnegie Mellon University, the author is currently a student at University of Karlsruhe, Germany.

**Abstract**

No-regret algorithms are powerful tools for learning in online convex problems that have received increased attention in recent years. Considering affine and external regret, we investigate what happens when a set of no-regret learners ($voters$) merge their respective strategies in each learning iteration to a single, common one in form of a convex combination. We show that an agent who executes this merged decision in each iteration of the online learning process and each time feeds back a reward function to the voters that is a correspondingly weighted version of its own reward, incurs sublinear regret itself. As a by-product, we obtain a simple method that allows us to construct new no-regret algorithms out of known ones.

**Keywords:** Machine Learning, No-Regret Algorithm, External Regret, Affine Regret, Online Convex Problem, Online Decision Problem, Computational Learning Theory, Online Learning.

# 1 Introduction

Regret minimizing algorithms are known since Hannan [12] presented the first one for repeated two-player games over 60 years ago. *Regret* is a measure of the quality of a sequence of actions that may be taken in the course of an online learning situation such as a repeated game or an *online convex problem (OCP)* [19]. Regret measures the difference of cumulative rewards between an actual sequence of actions taken and the best possible sequence one could have chosen from a predefined class. There are different types of regret that have been subject to investigations in the past years and vary with the definition of the before-mentioned class of reference action sequences (c.f. [6, 11]).
As the perhaps most prominent example, the *external regret* of a sequence of actions is defined as the cumulated reward incurred by their execution subtracted from the cumulative reward that would have been incurred had one chosen one single, optimal static solution instead, with the benefit of hindsight.
A *no-regret algorithm* is a procedure that learns (online) to generate a sequence of actions incurring regret that grows sublinearly with sequence length (i.e. with increasing learning experience).

No-regret algorithms have been proven to be powerful online learning tools that can distributively learn equilibrium points in multiagent game playing, planning scenarios and auctions (e.g. [6, 7, 13, 5, 3, 11]).
There are several recent works featuring the development of concrete no-regret algorithms such as *Greedy Projection* [19], *Lagrangian Hedging* [8] or *Follow the Perturbed Leader* [14].

Despite more recent advances (e.g. [2]) towards an understanding of general construction methods of no-regret algorithms for online convex problems, current knowledge is still limited. This is especially true when it comes to the nature of the underlying no-regret algorithm spaces.
This work aims at helping to close this gap to some extent. We show that a fixed convex combination of the output of an ensemble of no-external-regret learners results in a no-external-regret exhibiting learner again (provided each member of the ensemble is fed appropriate inputs). For restrictions to affine objective functions, an analogous statement is then derived for the class of what we named *no-affine-regret* learners. If we construe algorithms as points in a suitable space this insight spawns the intuition that the set of no-external-regret and no-affine-regret learning algorithms suitable for the same type of problems are each convex. Consequently, our findings will allow the construction of new no-regret algorithms as a combination of known ones.

Although the general idea of considering weighted sums of different learning entities is far from new the scope of the common multiplicative weights-based ensemble learning methods (e.g. [17, 7, 15]) is significantly different from ours. The latter strand of works is chiefly concerned with the problem of how to adapt the weights in order to combine different votes. For instance, Freund and Shapire provided a no-regret algorithm that, as a variation of *Weighted Majority* [15], adaptively learned weights of a weighted sum of pure strategies which corresponded to the voters (experts) [7]. In contrast, we consider settings where the adaptive behavior occurs only in the combined learning algorithms (solving online convex problems) while the weights are fixed. We do not focus

on finding a clever procedure to combine an arbitrary set of votes or class of learning algorithms but provide guarantees for a specific class (i.e. no-regret learners for OCPs) given constant weights.

# 2 Preliminaries

Before proceeding, we will briefly review the notions of *online convex programming problems (OCPs)* and *no-regret* assuming an underlying maximization problem. The corresponding statements for minimizations are analogous.
Online convex programming (OCP) [19, 10] is a useful paradigm for designing and analyzing learning algorithms. While first approaches addressing this setting date back to Hannan [12], the name *online convex problem* was later coined by Zinkevich [19]. He also contributed a gradient-ascent based no-regret algorithm solving a general OCP that is similar to another one introduced in [9]. Learning algorithms solving online convex problems are tools applicable in problem domains not amenable to other machine learning methods and have become subject to increasingly active research over the past years.

## 2.1 Online Convex Problems

A *convex programming problem* can be stated as follows [1]: Given a convex feasible set $F \subseteq \mathbb{R}^d$ and a convex mapping $\Gamma : F \to \mathbb{R}$ find the optimal solution given by the optimization problem $\inf_{\mathbf{x}\in F} \Gamma(\mathbf{x})$. If objective function $\Gamma$ determines a cost, the optimization task translates to finding a cost-optimal feasible strategy. Acknowledging that $\gamma := -\Gamma$ is concave, we can restate the problem as a maximization problem of a concave function $\gamma$ over a convex set. That is, the problem becomes to solve $\sup_{\mathbf{x}\in F} \gamma(\mathbf{x})$. In this context, $\gamma$ is interpreted as a *reward* or *revenue* function. Since both problems are completely analogous, we will limit our descriptions to the case where our problem is stated in terms of reward maximization. Notice, this choice also affects the definitions of regret given below but the emerging results are equivalent.

In an *online convex program* [19, 9], a (possibly adversarial) sequence $(\gamma_{(t)})_{t\in\mathbb{N}}$ of concave reward functions is revealed step by step. (Equivalently, one could substitute convex cost functions.) At each time step $t$, the convex programming algorithm must choose $\mathbf{x}_{(t)} \in F$ while only knowing the *past* reward functions $\gamma_{(\tau)}$ and choices $\mathbf{x}_{(\tau)}$ $(\tau \in \{1, ..., t-1\})$. After the choice is made, the current reward function $\gamma_{(t)}$ is revealed, and the algorithm receives a revenue amounting to $\gamma_{(t)}(\mathbf{x}_{(t)})$.

Note, there is a close connection between learning in an online convex problem and learning to play in repeated games. For instance, consider an individual agent playing a repeated matrix-game. In each round it picks a mixed strategy as a distribution over actions and receives a reward according to its choice of strategy in return. Then the process starts over. We can model this setting as an OCP: if the local convex set $F$ is a polytope and we interpret its corners as pure strategies then we can construe the choice $\mathbf{x}_{(t)}$ of an interior feasible point as a mixed strategy. We then let $\gamma_{(t)}$ be the

[1] For detailed background regarding convex optimization cf. e.g. [4].

resulting payoff function of the game such that $\gamma_{(t)}(\mathbf{x}_{(t)})$ reflects the current expected payoff of the player in round $t$.

## 2.2 No-Regret

To measure the performance of an OCP algorithm, we can compare its accumulated cost until step $T$ to an estimate of the best cost attainable against the sequence $(\gamma_{(t)})_{t=1...T}$. The notion *best* can be situation dependent. It could be expressed in rules such as *whenever action* $\mathbf{a} \in F$ *was chosen one should have chosen* $\phi(\mathbf{a}) \in F$ *instead* where $\phi : F \to F$ originates from a predefined class $\Phi$ of mappings on feasible set $F$. This idea leads to a measure called $\Phi$-*regret* $R_\Phi(T) := \sup_{\phi\in\Phi} \sum_{t=1}^{T} \gamma_{(t)}(\phi(\mathbf{x}_{(t)})) - \sum_{t=1}^{T} \gamma_{(t)}(\mathbf{x}_{(t)})$ [18, 11]. An algorithm is *no-*$\Phi$*-regret* with regret bound $\Delta$ iff $\forall T \in \mathbb{N} : R_\Phi(T) \leq \Delta(T) \in o(T)$.

The choice of the transformation class $\Phi$ leads to different types of no-regret algorithms. For instance, if $\Phi$ is chosen to be the set of all endomorphisms on $F$ we obtain the class of the so-called *no-linear-regret* algorithms [11].
Perhaps the most prominent case arises if $\Phi$ is restricted to all constant transformations on $F$. Then, the *best* attainable reward corresponds to the reward gained by the best constant choice $\mathbf{s}_{(T)} \in F$, chosen with knowledge of $\gamma_{(1)} \dots \gamma_{(T)}$, i.e. $\mathbf{s}_{(T)} \in \arg\sup_{\mathbf{x}\in F} \sum_{t=1}^{T} \gamma_{(t)}(\mathbf{x})$. This choice leads to a measure called *external regret* $R(T) := \sum_{t=1}^{T} \gamma_{(t)}(\mathbf{s}_{(T)}) - \sum_{t=1}^{T} \gamma_{(t)}(\mathbf{x}_{(t)})$. Consequently, a no-external-regret algorithm for a maximizing OCP is defined as an algorithm that generates a sequence of feasible vectors $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \mathbf{x}_{(3)}, \dots$ such that

$$\exists \Delta \in o(T) \forall T \in \mathbb{N} : \Delta(T) + \sum_{t=1}^{T} \gamma_{(t)}(\mathbf{x}_{(t)}) \geq \sup_{\mathbf{x}\in F} \sum_{t=1}^{T} \gamma_{(t)}(\mathbf{x}). \tag{1}$$

If $\Phi$ is composed of all affine functions we could speak of *no-affine-regret* properties. Obviously, the set of all no-affine-regret algorithms comprises both the set of no-external-regret and the set of no-linear-regret algorithms and may therefore be an important class to consider.

In order to ensure that a no-regret algorithm can even exist in principle it is common to introduce further restrictions to the OCP such as requiring a compact feasible set and continuous reward functions. Doing so implies that $\sup_{\mathbf{x}\in F} \gamma(\mathbf{x})$ exists and equals $\max_{\mathbf{x}\in F} \gamma(\mathbf{x})$. We will assume this condition to hold throughout the most part of this paper.

# 3 Convex Combinations of No-Regret Learners

Consider a society of $q \in \mathbb{N}$ agents $A_1, ..., A_q$. Each agent $A_v$ is capable of no-external-regret learning in an online convex problem and shares the same feasible set $F$ with its peers $A_j$ $(j \neq v)$. That is: If in every time step $t$, each $A_v$ chooses a vector $\mathbf{a}_{v(t)} \in F$ and then observes a reward function $\Omega_{v(t)}$ which is both additional learning experience and used to calculate the magnitude

$\Omega_{v(t)}(\mathbf{a}_{v(t)})$ of $A_v$'s reward for round $t$, then we can guarantee that its external regret $R_v(T)$ is always sublinear, i.e. $R_v(T) = \max_{\mathbf{x}\in F} \sum_{t=1}^{T} \Omega_{v(t)}(\mathbf{x}) - \sum_{t=1}^{T} \Omega_{v(t)}(\mathbf{a}_{v(t)}) \in o(T)$.
The interpretation of the generated vectors $\mathbf{a}_{v(t)}$ is application dependent. They may constitute strategies in a repeated game (e.g. [7]) or even represent plans. For instance, $\mathbf{a}_{v(t)}$ could be a routing plan in a network with each vector component entry representing the magnitude of traffic the agent intents to send through a corresponding link (e.g. [1, 5]). Alternatively, it may be a price for an item the agent sells or, it could conceivably be a representation of a tactical decision in a game of robotic soccer [16]. Regardless of the concrete interpretation, we will refer to $\mathbf{a}_{v(t)}$ as a *vote* and to agent $A_v$ as the corresponding *voter*.
Let $A$ be a *proxy* agent faced with the problem of having to solve an online convex problem: In each time step $t$ it has to choose an action $\mathbf{a}_{(t)} \in F$ and receives a concave reward function $\Omega_{A(t)}$ in return. If $A$ is able to consult the voters, i.e. to feed them learning experience in form of reward functions and to receive their votes in return, is it in the position to benefit from the voters' no-regret learning capabilities?
One trivial way to accomplish this is for $A$ to choose one single $A_v$ and let her solve his own OCP: In time step $t$, $A$ executes vote $\mathbf{a}_{v(t)}$ he was recommended by selected voter $A_v$ and after perceiving reward function $\Omega_{A(t)}$ this is sent back as further learning experience to $A_v$ (i.e. she perceives $\Omega_{A_v(t)} = \Omega_{A(t)}$ as her reward feedback) so she can generate a new recommendation $\mathbf{a}_{v(t+1)}$ in the next time step,... and so on. In the robotic soccer example, this could translate to a coach who selects a single agent (e.g. player) and leaves the tactical decision making to her from then on.
However, this approach may be less than optimal. Assume, the decision of which voter to select was made according to some distribution. Let $p_v$ denote the probability that $A$ chooses voter $A_v$. Then $A$'s expected reward in the first time step equals $\sum_{v=1}^{q} p_v \Omega_{A(1)}(\mathbf{a}_{v(1)})$. As an alternative option, $A$ could have consulted all voters and executed a compromise $\sum_{v=1}^{q} p_v \mathbf{a}_{v(1)}$ of their votes. Due to concavity we have by Jensen's inequality (e.g. [4])

$$\sum_{v=1}^{q} p_v \Omega_{A(1)}(\mathbf{a}_{v(1)}) \leq \Omega_{A(1)}(\sum_{v=1}^{q} p_v \mathbf{a}_{v(1)}). \tag{2}$$

Thus, consulting all voters and executing the convex combination of their votes $\sum_{v=1}^{q} p_v \mathbf{a}_{v(1)}$ would have gained him a higher reward[2] than the expected reward $A$ received in the first round otherwise.
Of course, depending on the nature of the OCP, future reward functions may depend on past choices of feasible vectors. Therefore, without further assumptions it would become more involved to generally assess whether relying on a convex combination of the individual votes would necessarily be a superior approach in the long run. Still, this simple example should already motivate that the execution of convex combinations of votes may be an approach worth examining - which we will start doing next.

[2]Or, at least not a lower reward.

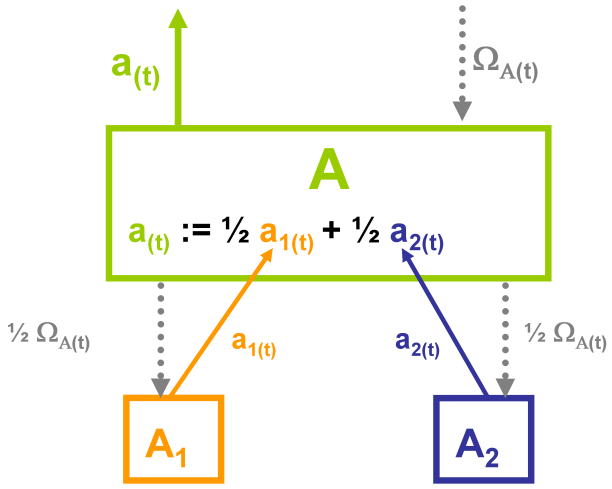


Figure 1: Illustration of a strategy voting situation with proxy agent $A$ and two voters $A_1$ and $A_2$.

## 3.1 Setup and Theorems

Let $z_1, ..., z_q$ be nonnegative constants such that $\sum_{v=1}^{q} z_v = 1$. If each voter $A_v$ submits vote $\mathbf{a}_{v(t)}$ then we will refer to $\mathbf{a}_{(t)} := z_1\mathbf{a}_{1(t)} + ... + z_q\mathbf{a}_{q(t)}$ as their *compromise strategy*. How well would $A$ perform in an OCP if it would always execute such a compromise strategy after providing each voter with learning experience depending on its own reward function $\Omega_{A(t)}$ in each round $t$? Of course this does not only depend on the individual learning algorithms each of the voters employs but also on the learning experience they are exposed to in the course of the online process.
We consider the following setup: In every round $t$, $A$ calls each voter $A_v$ and provides him a reward function which is a weighted split of her own, i.e. she sets $\Omega_{A_v(t)} = z_v\Omega_{A(t)}$. Upon receiving the outputs $\mathbf{a}_{v(t)}$ $(t = 1, ..., q)$ of the voters $A$ executes the compromise strategy of these votes. (The setup is depicted in Fig. 1 for $q = 2, z_1 = z_2 = \frac{1}{2}$.)

### 3.1.1 External Regret

The next theorem tells us that as a result of this setup, $A$ will incur sublinear external regret if $A_1, ..., A_n$ do.

**Theorem 3.1** (Preservation of No-External-Regret)**.** *Let $F$ be a convex set, $A$ be a proxy agent always executing $\mathbf{a} = z_1\mathbf{a}_1 + ... + z_q\mathbf{a}_q$ where $\mathbf{a}_v \in F$ denotes the strategy of agent $A_v$ $(v = 1, ..., q)$ and $z_1, ..., z_q$ are nonnegative weights such that $\sum_{v=1}^{q} z_v = 1$. Let $\Omega_{A(t)}(\cdot)$ be $A$'s concave - and for all $v \in \{1, ..., q\}$ let $\Omega_{A_v(t)}(\cdot)$ be $A_v$'s individual revenue function for iteration $t$ where $\forall t \in \mathbb{N} \forall v \in \{1, ..., q\} : z_v\Omega_{A(t)}(\cdot) = \Omega_{A_v(t)}(\cdot)$. Let $A_1, ..., A_q$ each solve a maximizing online convex problem employing no-external-regret algorithms with regret bounds $\Delta_1, ..., \Delta_q \in o(T)$, respectively. Furthermore, let each element in the sequence of $A$'s revenue functions $(\Omega_{A(1)}, \Omega_{A(2)}, ...)$ be concave and $A$ solve a maximizing online convex problem observing the elements of this sequence one by one.*

*Then we have: $A$ is guaranteed to incur sublinear external regret, i.e. it effectively employs no-external-regret learning. A regret bound is $\sum_{v=1}^{q} \Delta_v$.*

*Proof.* First, we note $\Delta_1, ..., \Delta_q \in o(T) \Rightarrow \Delta := \sum_{v=1}^{q} \Delta_v \in o(T)$. Let $(\mathbf{a}_{v(t)})_{t\in\mathbb{N}}$ denote $A_v$'s sequence of strategies generated by a no-regret algorithm. Due to the no-regret property (cf. Eq. 1), we know:
$\forall v \in \{1, ..., q\}\, \forall T \in \mathbb{N} : \sum_{t=1}^{T} \Omega_{A_v(t)}(\mathbf{a}_{v(t)}) \geq \max_{\mathbf{a}_v} \sum_{t=1}^{T} \Omega_{A_v(t)}(\mathbf{a}_v) - \Delta_v(T)$.
For all $T \in \mathbb{N}$ we have:
$\sum_{t=1}^{T} \Omega_{A(t)}(\mathbf{a}_{(t)})$
$= \sum_{t=1}^{T} \Omega_{A(t)}(z_1\mathbf{a}_{1(t)} + ... + z_q\mathbf{a}_{q(t)})$
$\geq^{3} \sum_{t=1}^{T} \sum_{v=1}^{q} z_v \Omega_{A(t)}(\mathbf{a}_{v(t)})$
$= \sum_{t=1}^{T} \sum_{v=1}^{q} \Omega_{A_v(t)}(\mathbf{a}_{v(t)})$
$= \sum_{v=1}^{q} \sum_{t=1}^{T} \Omega_{A_v(t)}(\mathbf{a}_{v(t)})$
$\geq^{4} \sum_{v=1}^{q} (\max_{\mathbf{a}_v} \sum_{t=1}^{T} \Omega_{A_v(t)}(\mathbf{a}_v) - \Delta_v(T))$
$= \sum_{v=1}^{q} \max_{\mathbf{a}_v} \sum_{t=1}^{T} z_v \Omega_{A(t)}(\mathbf{a}_v) - \sum_{v=1}^{q} \Delta_v(T)$
$= (\sum_{v=1}^{q} z_v \max_{\mathbf{a}_v} \sum_{t=1}^{T} \Omega_{A(t)}(\mathbf{a}_v)) - \Delta(T)$
$= (\sum_{v=1}^{q} z_v \max_{\mathbf{a}} \sum_{t=1}^{T} \Omega_{A(t)}(\mathbf{a})) - \Delta(T)$
$= (\max_{\mathbf{a}} \sum_{t=1}^{T} \Omega_{A(t)}(\mathbf{a})) - \Delta(T)$.

□

We can easily derive the analogous statement for convex cost functions and minimizing OCPs but chose to omit such redundant considerations in order to keep the exposition concise.

### 3.1.2 Affine and Linear Regret

We will now assume that each individual voter $A_v$ incurs sublinear affine regret, i.e it incurs sublinear $\Phi - regret$ where $\Phi$ is the class of affine mappings on the feasible set $F$. Furthermore, we restrict our considerations to the case where the aggregate reward function $\Omega_{A(t)}$ is affine. An example for a situation where the latter assumption holds is the case of the adversarial revenue functions considered in [5].

**Theorem 3.2** (Preservation of No-Affine-Regret)**.** *Let $F$ be a convex set, $A$ be a proxy agent always playing $\mathbf{a} = z_1\mathbf{a}_1 + ... + z_q\mathbf{a}_q$ where $\mathbf{a}_v \in F$ denotes the strategy of agent $A_v$ $(v = 1, ..., q)$ and $z_1, ..., z_q$ are nonnegative weights such that $\sum_{v=1}^{q} z_v = 1$. Let $\Omega_{A(t)}(\cdot)$ be $A$'s affine - and for all $v \in \{1, ..., q\}$ let $\Omega_{A_v(t)}(\cdot)$ be $A_v$'s individual revenue function for iteration $t$ where $\forall t \in \mathbb{N} \forall v \in \{1, ..., q\} : z_v\Omega_{A(t)}(\cdot) = \Omega_{A_v(t)}(\cdot)$. Let $A_1, ..., A_q$ each solve a maximizing online convex problem employing no-regret algorithms with regret bounds $\Delta_1, ..., \Delta_q \in o(T)$, respectively. Furthermore, let each element in the sequence of $A$'s revenue functions $(\Omega_{A(1)}, \Omega_{A(2)}, ...)$ be affine and $A$ solve a maximizing online convex problem observing the elements of this sequence one by one.*
*Then we have: $A$ is guaranteed to experience sublinear affine regret. A regret bound is $\sum_{v=1}^{q} \Delta_v$.*

[3] Owing to concavity.
[4] Due to individual no-regret learning.

*Proof.* First, we note $\Delta_1, ..., \Delta_q \in o(T) \Rightarrow \Delta := \sum_{v=1}^{q} \Delta_v \in o(T)$. Let $(\mathbf{a}_{v(t)})_{t\in\mathbb{N}}$ denote $A_v$'s sequence of strategies generated by a no-regret algorithm. Furthermore, let $\Phi$ be the set of affine mappings on feasible set $F$. Due to the assumption that each individual learner $A_1, ..., A_q$ incurs sublinear affine regret, we know:
$\forall v \in \{1, ..., q\}\, \forall T \in \mathbb{N} : \sum_{t=1}^{T} \Omega_{A_v(t)}(\mathbf{a}_{v(t)}) \geq \sup_{\phi\in\Phi} \sum_{t=1}^{T} \Omega_{A_v(t)}(\phi(\mathbf{a}_{v(t)})) - \Delta_v(T)$.
For all $T \in \mathbb{N}$ we have:
$\sum_{t=1}^{T} \Omega_{A(t)}(\mathbf{a}_{(t)})$
$= \sum_{t=1}^{T} \Omega_{A(t)}(z_1\mathbf{a}_{1(t)} + ... + z_q\mathbf{a}_{q(t)})$
$\geq^{5} \sum_{t=1}^{T} \sum_{v=1}^{q} z_v \Omega_{A(t)}(\mathbf{a}_{v(t)})$
$= \sum_{v=1}^{q} \sum_{t=1}^{T} \Omega_{A_v(t)}(\mathbf{a}_{v(t)})$
$\geq^{6} \sum_{v=1}^{q} (\sup_{\phi\in\Phi} \sum_{t=1}^{T} \Omega_{A_v(t)}(\phi(\mathbf{a}_{v(t)})) - \Delta_v(T))$
$= -\sum_{v=1}^{q} \Delta_v(T) + \sum_{v=1}^{q} \sup_{\phi\in\Phi} \sum_{t=1}^{T} z_v \Omega_{A(t)}(\phi(\mathbf{a}_{v(t)}))$
$\geq -\sum_{v=1}^{q} \Delta_v(T) + \sup_{\phi\in\Phi} \sum_{v=1}^{q} \sum_{t=1}^{T} z_v \Omega_{A(t)}(\phi(\mathbf{a}_{v(t)}))$
$= -\Delta(T) + \sup_{\phi\in\Phi} \sum_{t=1}^{T} \sum_{v=1}^{q} z_v \Omega_{A(t)}(\phi(\mathbf{a}_{v(t)}))$
$=^{7} -\Delta(T) + \sup_{\phi\in\Phi} \sum_{t=1}^{T} \Omega_{A(t)}(\phi(\sum_{v=1}^{q} z_v \mathbf{a}_{v(t)}))$

□

Note, since any linear function is also affine we can conclude that a convex combination of *no-linear-regret* learners [11] results in a learner that exhibits no-linear-regret again as well (in settings with affine objective functions). Why is it worthwhile to consider affine regret properties? Of course, affinity is generally a handy property since for affine mappings, Jensen's inequality is tight. In fact, this was explicitly leveraged in the last line of the proof of Theorem 3.2. On the other hand, no-affine-regret is still a quite general notion that, as mentioned above, comprises the important cases of both no-linear-regret and no-external-regret. Unfortunately, Theorem 3.1 could not be stated as a corollary building upon Theorem 3.2 since the latter requires each member of the sequence of objective functions to be affine[8], while the former merely assumes them to be concave.

## 3.2 Convexity of No-Regret Algorithm Spaces

Of course, the above result is not restricted to cases where $A_1, \ldots, A_q$ are agents.
If $A_1, \ldots, A_q$ are different algorithms on respective problem domains $D_1, \ldots, D_q$ then following above procedure is a prescription of how to construct a new learning algorithm $A$ for domain $D_1 \cap \ldots \cap D_q$ as a convex combination of these previously known ones. If each $A_v$ exhibits no-external-regret Theorem 3.1 implies that the resulting algorithm $A$ exhibits no-external-regret as well. In case each $A_v$ solves an OCP with affine rewards and is guaranteed to incur sublinear affine regret, then by Theorem 3.2, combined algorithm $A$ will constitute a no-affine-regret algorithm.
Note, we can construe no-regret algorithms as points in a common vector space where the Abelian group operation (+) is constituted by pointwise addition of the algorithms' outputs and the scalar

[5] Owing to concavity.
[6] Due to individual no-regret learning.
[7] Leveraging that $\phi, \Omega_{A_{(t)}}$ were assumed to be affine.
[8] That is, both concave and convex.

operation ($*$) is simply pointwise multiplication with elements of a field (typically $\mathbb{R}$) that comprises the range of the reward functions. In this light, our results state that the set of all no-regret algorithms of the same type[9] is convex.

# 4 Discussion and Future Work

This paper developed a general no-regret property regarding convex combinations of learners. For the class of no-external- and no-affine -regret learners, we established how a convex combination of a finite number of such learners ($voters$) can be employed to commonly solve an online convex problem in a manner that is guaranteed to incur sublinear regret, provided each of the voters does.[10] More specifically, the proofs of Theorem 3.1 and 3.2 reveal that the convex combination of no-regret learners results in a new one whose regret bound is not growing faster than the sum of the regret bounds of the voters. It may be worthwhile to explore conditions under which this sum is an overly conservative bound.
For instance, one could investigate whether convergence can be sped up by iteratively adapting the combing weights with a meta-learning algorithm such as *Weighted Majority* [15].
As this work derived the insight that no-regret algorithms of the same class suitable for the same problems constitute a convex set, exploring additional of its properties may be an interesting direction of future efforts. For instance, does this set have border points? That is, are there no-regret algorithms that inherently cannot be (nontrivially) found by the construction method we presented?
In conclusion, we believe the insights gained in this work may not only be of theoretical interest but also hope that they have the potential to serve as the outset for fruitful future efforts.

# Acknowledgements

I am grateful to Geoff Gordon for his kind supportiveness regarding this work. I would also like to thank Fabian Meyer for helping me proofread the final draft and making sure it is comprehensible.

# References

[1] A. Blum, E. Even-Dar, and K. Ligett, *Routing without regret: on convergence to nash equilibria of regret-minimizing algorithms in routing games*, PODC '06: Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing, 2006.

[2] A. Blum and Y. Mansour, *From external to internal regret*, International Conference on Learning Theory, 2005.

[3] Avrim Blum, Vijay Kumar, Atri Rudra, and Felix Wu, *Online learning in online auctions*, Theor. Comput. Sci. **324** (2004), no. 2-3, 137–146.

[9] Either no-external regret or no-affine-regret.

[10] Remember, for the case of affine regret we further restricted the objective functions to be affine as well.

[4] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[5] Jan-P. Calliess and Geoffrey J. Gordon, *No-regret learning and a mechanism for distributed multiagent planning*, Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS'08), 2008.

[6] D. Foster and R. Vohra, *Calibrated learning and correlated equilibrium*, Games and Economic Behavior (1997).

[7] Yoav Freund and Robert E. Shapire, *Game theory, on-line prediction and boosting*, COLT, 1996.

[8] Geoff Gordon, *No-regret algorithms for online convex programs*, Advances in Neural Information Processing Systems 19, 2007.

[9] Geoffrey J. Gordon, *Approximate solutions to markov decision processes*, Ph.D. thesis, Carnegie Mellon University, 1999.

[10] _____, *Regret bounds for prediction problems*, COLT: Workshop on Computational Learning Theory, 1999.

[11] Geoffrey J. Gordon, Amy Greenwald, and Casey Marks., *No-regret learning in convex games*, 25th Int. Conf. on Machine Learning (ICML '08), 2008.

[12] J. Hannan, *Contributions to the theory of games*, Princeton University Press, 1957.

[13] Amir Jafari, Amy R. Greenwald, David Gondek, and Gunes Ercal, *On no-regret learning, fictitious play, and nash equilibrium*, ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pp. 226–233.

[14] A. Kalai and S. Vempala, *Efficient algorithms for online decision problems*, 16th Annual Conf. on Learning Theory (COLT), 2003.

[15] Nick Littlestone and Manfred K. Warmuth, *The weighted majority algorithm*, IEEE Symposium on Foundations of Computer Science, 1989, pp. 256–261.

[16] Michael K. Sahota, Alan K. Mackworth, Rod A. Barman, and Stewart J. Kingdon, *Real-time control of soccer-playing robots using off-board vision: the dynamite testbed.*, IEEE International Conference on Systems, Man, and Cybernetics, 1995, pp. 3690–3663.

[17] R.E. Shapire, *The strength of weak learnability.*, Machine Learning **5(2)** (1990), 197–227, First boosting method.

[18] G. Stoltz and G. Lugosi, *Learning correlated equilibria in games with compact sets of strategies*, Games and Economic Behavior **59** (2007), 187–208.

[19] M. Zinkevich, *Online convex programming and generalized infinitesimal gradient ascent*, Twentieth International Conference on Machine Learning, 2003.


Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

Carnegie Mellon