Towards Minimax Online Learning with Unknown Time Horizon

Haipeng Luo HAIPENGL@CS.PRINCETON.EDU
Department of Computer Science, Princeton University, Princeton, NJ 08540

Robert E. Schapire SCHAPIRE@CS.PRINCETON.EDU
Department of Computer Science, Princeton University, Princeton, NJ 08540

Abstract

We consider online learning when the time horizon is unknown. We apply a minimax analysis, beginning with the fixed horizon case, and then moving on to two unknown-horizon settings, one that assumes the horizon is chosen randomly according to some distribution, and the other which allows the adversary full control over the horizon. For the random horizon setting with restricted losses, we derive a fully optimal minimax algorithm. And for the adversarial horizon setting, we prove a nontrivial lower bound which shows that the adversary obtains strictly more power than when the horizon is fixed and known. Based on the minimax solution of the random horizon setting, we then propose a new adaptive algorithm which "pretends" that the horizon is drawn from a distribution from a special family, but no matter how the actual horizon is chosen, the *worst-case* regret is of the optimal rate. Furthermore, our algorithm can be combined and applied in many ways, for instance, to online convex optimization, follow the perturbed leader, exponential weights algorithm and first order bounds. Experiments show that our algorithm outperforms many other existing algorithms in an online linear optimization setting.

1. Introduction

We study online learning problems with unknown time horizon with the aim of developing algorithms and approaches for the realistic case that the number of time steps is initially unknown.

We first adopt the standard Hedge setting (Freund & Schapire, 1997) where the learner chooses a distribution over N actions on each round, and the losses for each action are then selected by an adversary. The learner incurs loss equal to the expected loss of the actions in terms of the distribution it chose for this round, and its goal is to minimize the regret, the difference between its cumulative loss and that of the best action after T rounds.

Various algorithms are known to achieve the optimal (up to a constant) upper bound $O(\sqrt{T \ln N})$ on the regret. Most of them assume that the horizon T is known ahead of time, especially those which are minimax optimal (Cesa-Bianchi et al., 1997; Abernethy et al., 2008b). When the horizon is unknown, the so-called doubling trick (Cesa-Bianchi et al., 1997) is a general technique to make a learning algorithm adaptive and still achieve $O(\sqrt{T \ln N})$ regret uniformly for any T. The idea is to first guess a horizon, and once the actual horizon exceeds this guess, double it and restart the algorithm. Although, in theory, it is widely applicable, the doubling trick is aesthetically inelegant, and intuitively wasteful, since it repeatedly restarts itself, entirely forgetting all the preceding information. Other approaches have also been proposed, as we discuss shortly.

In this paper, we study the problem of learning with unknown horizon in a game-theoretic framework. We consider a number of variants of the problem, and make progress toward a minimax solution. Based on this approach, we give a new general technique which can also make other minimax or non-minimax algorithms adaptive and achieve low regret in a very general online learning setting. The resulting algorithm is still not exactly optimal, but it makes use of all the previous information on each round and achieves much lower regret in experiments.

We view the Hedge problem as a repeated game between the learner and the adversary. Abernethy et al. (2008b), and Abernethy & Warmuth (2010) proposed an exact minimax optimal solution for a slightly different game with binary losses, assuming that the loss of the best action is at most some fixed constant. They derived the solution under a very simple type of loss space; that is, on each round only one action suffers one unit loss. We call this the basis vector

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

loss space. As a preliminary of this paper, we also derive a similar minimax solution under this simple loss space for our setting where the horizon T is fixed and known to the learner ahead of time (see Theorem 1).

We then move on to the primary interest of this paper, that is, the case when the horizon is unknown to the learner. We study this unknown horizon setting in the minimax framework, with the aim of ultimately deriving game-theoretically optimal algorithms. Two types of models are studied. The first one assumes the horizon is chosen according to some known distribution, and the learner's goal is to minimize the expected regret. We show the exact minimax solution for the basis vector loss space in this case (see Theorem 2). It turns out that the distribution the learner should choose on each round is simply the conditional expectation of the distributions the learner would have chosen for the fixed horizon case.

The second model we study gives the adversary the power to decide the horizon on the fly, which is possibly the most adversarial case. In this case, we no longer use the regret as the performance measure. Otherwise the adversary would obviously choose an infinite horizon. Instead, we use a scaled regret to measure the performance. Specifically, we scale the regret at time t by the optimal regret under fixed horizon t. The exact optimal solution in this case is unfortunately not found and remains an open problem, even for the extremely simple case. However, we give a lower bound for this setting to show that the optimal regret is strictly greater than the one in the fixed horizon game. That is, the adversary does obtain strictly more power if allowed to pick the horizon (see Theorem 3).

We then propose our new adaptive algorithm based on the minimax solution in the random horizon setting. One might doubt how realistic a random horizon is in practice. Even if the true horizon is indeed drawn from a fixed distribution, how can we know this distribution? We address these problems at the same time. Specifically, we prove that no matter how the horizon is chosen, if we assume it is drawn from a distribution from a special family, and let the learner play in a way similar to the one in the random horizon setting, then the *worst-case* regret at any time T (not the expected regret) can still be of the optimal order. In other words, although the learner is behaving as if the horizon is random, its regret will be small even if the horizon is actually controlled by an adversary. Moreover, the results hold for not just the Hedge problem, but a general online learning setting—*online convex optimization*— that includes many interesting problems (see Theorem 5).

Our idea can be combined not only with the minimax algorithm, but also the "follow the perturbed leader" algorithm and the exponential weights algorithm (see Theorem 7 and 8). In addition, our technique can not only deal with unknown horizon, but also other unknown information such as the loss of the best action, thus leading to a first order regret bound that depends on the loss of the best action (see Theorem 9). Like the doubling trick, this seems to be a quite general way to make an algorithm adaptive. Furthermore, we conduct experiments showing that our algorithm outperforms many existing algorithms, including the doubling trick, in an online linear optimization setting within an ℓ_2 ball where our algorithm has an explicit closed form.

The rest of the paper is organized as follows. We define the Hedge setting formally in Section 2, and derive the minimax solution for the fixed horizon setting as the preliminary of this paper in Section 3. In Section 4, we study two unknown horizon settings in the minimax framework. We then turn to a general online learning setting and present our new adaptive algorithm in Section 5. Implementation issues, experiments, and applications are discussed in Section 6. We omit most of the proofs due to space limitations, but all details can be found in the supplementary material.

Related work Besides the doubling trick, other adaptive algorithms have been studied (Auer et al., 2002; Gentile, 2003; Yaroshinsky et al., 2004; Chaudhuri et al., 2009; de Rooij et al., 2013). Auer et al. (2002) showed that for algorithms such as the exponential weights algorithm (Littlestone & Warmuth, 1994; Freund & Schapire, 1997; 1999), where a learning rate η should be set as a function of the horizon, typically in the form $\sqrt{(b \ln N)/T}$ for some constant b, one can simply set η adaptively as $\sqrt{(b \ln N)/t}$, where t is the current number of rounds. In other words, this algorithm always pretends the current round is the last round. Although this idea works with the exponential weights algorithm, we remark that assuming the current round is the last round does not always work. Specifically, one can show that it will fail if applied to the minimax algorithm (see Section 6.4). In another approach to online learning with unknown horizon, Chaudhuri et al. (2009) proposed an adaptive algorithm based on a novel potential function reminiscent of the half-normal distribution.

Other performance measures different from the usual regret were studied before. Foster & Vohra (1998) introduced internal regret comparing the loss of an online algorithm to the loss of a modified algorithm which consistently replaces one action by another; Herbster & Warmuth (1995), and Bousquet & Warmuth (2003) compared the learner's loss with the best k-shifting expert; Hazan & Seshadhri (2007) studied the usual regret within any time interval; Chernov & Zhdanov (2010) considered discounted losses. To the best of our knowledge, the form of scaled regret that we study is new. Lower bounds on anytime regret in terms of the quadratic variations for any loss sequence (instead of the worst case sequence this paper considers) were studied by Gofer & Mansour (2012).

2. Repeated Games

We first consider the following repeated game between a learner and an adversary. The learner has access to N actions. On each round $t = 1, \dots, T$, (1) the learner chooses a distribution $\mathbf{P}_t$ over the actions; (2) the adversary reveals the loss vector $\mathbf{Z}_t = (Z_{t,1}, \dots, Z_{t,N}) \in \mathbf{LS}$, where $Z_{t,i}$ is the loss for action i for this round, and the *loss space* $\mathbf{LS}$ is a subset of $[0,1]^N$; (3) the learner suffers loss $\ell_t = \mathbf{P}_t \cdot \mathbf{Z}_t$ for this round.

Notice that the adversary can choose the losses on round t with full knowledge of the history $\mathbf{P}_{1:t}$ and $\mathbf{Z}_{1:t-1}$, that is, all the previous choices of the learner and the adversary (we use notation $a_{1:t}$ to denote the multiset $\{a_1, \dots, a_t\}$). We also denote the cumulative loss up to round t for the learner and the actions by $L_t = \sum_{t'=1}^{t} \ell_{t'}$ and $\mathbf{M}_t = \sum_{t'=1}^{t} \mathbf{Z}_{t'}$ respectively. The goal for the learner is to minimize the difference between its total loss and that of the best action at the end of the game. In other words, the goal of the learner is to minimize $\mathbf{Reg}(L_T, \mathbf{M}_T)$, where we define the regret function $\mathbf{Reg}(L, \mathbf{M}) \triangleq L - \min_i M_i$, for $L \in \mathbb{R}$ and $\mathbf{M} \in \mathbb{R}^N$. The number of rounds T is called the *horizon*.

Regarding the loss space $\mathbf{LS}$, perhaps the simplest one is $\{\mathbf{e}_1, \dots, \mathbf{e}_N\}$, the N standard basis vectors in N dimensions. Playing with this loss space means that on each round, the adversary chooses one single action to incur one unit loss. In order to show the intuition of our main results, we mainly focus on this basis vector loss space in Sections 3 and 4, but we return to the most general case $[0,1]^N$ later.

3. Minimax Solution for Fixed Horizon

Although our primary interest in this paper is the case when the horizon is unknown to the learner, we first present some preliminary results on the setting where the horizon is known to both the learner and the adversary ahead of time. These will later be useful for the unknown horizon case.

If we treat the learner as an algorithm $\mathbf{Alg}$ that takes the information of previous rounds as inputs, and outputs a distribution $\mathbf{P}_t = \mathbf{Alg}(\mathbf{P}_{1:t-1}, \mathbf{Z}_{1:t-1})$ that the learner is going to play with, then finding the optimal solution in this fixed horizon setting can be viewed as solving the minimax expression

$$\inf_{\mathbf{Alg}} \sup_{\mathbf{Z}_{1:T}} \mathbf{Reg}(L_T, \mathbf{M}_T). \tag{1}$$

Alternatively, we can recursively define:

$$V(\mathbf{M}, 0) \triangleq -\min_i M_i \,;$$
$$V(\mathbf{M}, r) \triangleq \min_{\mathbf{P} \in \Delta(N)} \max_{\mathbf{Z} \in \mathbf{LS}} \left(\mathbf{P} \cdot \mathbf{Z} + V(\mathbf{M} + \mathbf{Z}, r-1)\right),$$

where $\mathbf{M} \in \mathbb{R}^N$ is a loss vector, r is a nonnegative integer, and $\Delta(N)$ is the N dimensional simplex. By a simple argument, one can show that the value of $V(\mathbf{M}, r)$ is the regret of a game with r rounds starting from the situation that each action has initial loss M_i, and assuming both the learner and the adversary will play optimally. In fact, the value of Eq. (1) is exactly $V(\mathbf{0}, T)$, and the optimal learner algorithm is the one that chooses the $\mathbf{P}^*$ which realizes the minimum in the definition of $V(\mathbf{M}, r)$ when the actions' cumulative loss vector is $\mathbf{M}$ and there are r rounds left. We call $V(\mathbf{0}, T)$ the *value* of the game.

As a concrete illustration of these ideas, we now consider the basis vector loss space[1], that is, $\mathbf{LS} = \{\mathbf{e}_1, \dots, \mathbf{e}_N\}$. It turns out that under this loss space, the value function V has a nice closed form. Similar to the results from Cesa-Bianchi et al. (1997) and Abernethy et al. (2008b), we show that V can be expressed in terms of a random walk. Suppose $R(\mathbf{M}, r)$ is the expectation of the loss of the best action if the adversary chooses each $\mathbf{e}_i$ uniformly randomly for the remaining r rounds, starting from loss vector $\mathbf{M}$. Formally, $R(\mathbf{M}, r)$ can be defined in a recursive way: $R(\mathbf{M}, 0) \triangleq \min_i M_i \,; R(\mathbf{M}, r) \triangleq \frac{1}{N} \sum_{i=1}^{N} R(\mathbf{M} + \mathbf{e}_i, r-1)$. The connection between V and R, and the optimal algorithm are then shown by the following theorem.

Theorem 1. *If $\mathbf{LS} = \{\mathbf{e}_1, \dots, \mathbf{e}_N\}$, then for any vector $\mathbf{M}$ and integer $r \geq 0$, we have $V(\mathbf{M}, r) = r/N - R(\mathbf{M}, r)$. Let $c_N = \frac{1}{N}\sqrt{2(N-1)\ln N}$. Then the value of the game satisfies*

$$V(\mathbf{0}, T) \leq c_N \sqrt{T}. \tag{2}$$

Moreover, on round t, the optimal learner algorithm is the one that chooses weight $P_{t,i} = V(\mathbf{M}_{t-1}, r) - V(\mathbf{M}_{t-1} + \mathbf{e}_i, r-1)$ for each action i, where $\mathbf{M}_{t-1}$ is the current cumulative loss vector and r is the number of remaining rounds, that is, $r = T - t + 1$.

Theorem 1 tells us that under the basis vector loss space, the best way to play is to assume that the adversary is playing uniformly randomly, since r/N and $R(\mathbf{M}, r)$ are exactly the expected losses for the learner and for the best action respectively. Note that c_N is decreasing when $N \geq 4$ (with maximum value about 0.72). So contrary to the $O(\sqrt{T \ln N})$ regret bound for the general loss space $[0,1]^N$ which is increasing in N, here $V(\mathbf{0}, T)$ is of order $O(\sqrt{T})$.

4. Playing without Knowing the Horizon

We turn now to the case in which the horizon T is unknown to the learner, which is often more realistic in practice. There are several ways of modeling this setting. For example, the horizon can be chosen ahead of time according to some fixed distribution, or it can even be chosen by the adversary. We will discuss these two variants separately.

[1] For other loss spaces, finding minimax solutions seems difficult. However, we show the relation of the values of the game for different loss spaces in the supplementary file, see Theorem 10.

4.1. Random Horizon

Suppose the horizon T is chosen according to some fixed distribution Q which is known to both the learner and the adversary. Before the game starts, a random T is drawn, and neither the learner nor the adversary knows the actual value of T. The game stops after T rounds, and the learner aims to minimize the expectation of the regret. Using our earlier notation, the problem can be formally defined as

$$\inf_{\mathbf{Alg}} \sup_{\mathbf{Z}_{1:\infty}} \mathbb{E}_{T\sim Q}[\mathbf{Reg}(L_T, \mathbf{M}_T)],$$

where we assume the expectation is always finite. We sometimes omit the subscript $T \sim Q$ for simplicity.

Continuing the example in Section 3 of the basis vector loss space, we can again show the exact minimax solution, which has a strong connection with the one for the fixed horizon setting.

Theorem 2. *If* $\mathbf{LS} = \{\mathbf{e}_1, \dots, \mathbf{e}_N\}$*, then*

$$\begin{aligned} &\inf_{\mathbf{Alg}} \sup_{\mathbf{Z}_{1:\infty}} \mathbb{E}_{T\sim Q}[\mathbf{Reg}(L_T, \mathbf{M}_T)] \\ &= \mathbb{E}_{T\sim Q}[\inf_{\mathbf{Alg}} \sup_{\mathbf{Z}_{1:T}} \mathbf{Reg}(L_T, \mathbf{M}_T)]. \end{aligned} \tag{3}$$

Moreover, on round t, the optimal learner plays with the distribution $\mathbf{P}_t = \mathbb{E}_{T\sim Q}[\mathbf{P}_t^T | T \geq t]$, where $\mathbf{P}_t^T$ is the optimal distribution the learner would play if the horizon is T, that is, $P_{t,i}^T = V(\mathbf{M}_{t-1}, T-t+1) - V(\mathbf{M}_{t-1} + \mathbf{e}_i, T-t)$.

Eq. (3) tells us that if the horizon is drawn from some distribution, then even though the learner does not know the actual horizon before playing the game, as long as the adversary does not know this information either, it can still do as well as the case when they are both aware of the horizon.

However, so far this model does not seem to be quite useful in practice for several reasons. First of all, the horizon might not be chosen according to a distribution. Even if it is, this distribution is probably unknown. Secondly, what we really care about is the performance which holds uniformly for any horizon, instead of the expected regret. Last but not least, one might conjecture that the similar result stated in Theorem 2 should hold for other more general loss spaces, which is in fact not true (see Example 1 in the supplementary file), making the result seem even less useful.

Fortunately, we address all these problems and develop new adaptive algorithms based on the result in this section. We discuss these in Section 5 after first introducing the fully adversarial model.

4.2. Adversarial Horizon

The most adversarial setting is the one where the horizon is completely controlled by the adversary. That is, we let the adversary decide whether to continue or stop the game on each round according to the current situation. However, notice that the value of the game is increasing in the horizon. So if the adversary can determine the horizon and its goal is still to maximize the regret, then the problem would not make sense because the adversary would clearly choose to play the game forever and never stop leading to infinite regret. One reasonable way to address this issue is to scale the regret by the value of the fixed horizon game $V(\mathbf{0}, T)$, so that the scaled regret $\mathbf{Reg}(L_T, \mathbf{M}_T)/V(\mathbf{0}, T)$ indicates how many times worse is the regret compared to the one that is optimal given the horizon. Under this setting, the corresponding minimax expression is

$$\tilde{V} = \inf_{\mathbf{Alg}} \sup_{T} \sup_{\mathbf{Z}_{1:T}} \frac{\mathbf{Reg}(L_T, \mathbf{M}_T)}{V(\mathbf{0}, T)}. \tag{4}$$

Unfortunately, finding the minimax solution to this setting seems to be quite challenging, even for the simplest case $N = 2$. It is clear, however, that $\tilde{V}$ is at most some constant due to the existence of adaptive algorithms such as the doubling trick, which can achieve the optimal regret bound up to a constant without knowing T. Another clear fact is $\tilde{V} \geq 1$, since it is impossible for the learner to do better than the case when it is aware of the horizon. Below, we derive a nontrivial lower bound that is greater than 1, thus proving that the adversary does gain strictly more power when it can stop the game whenever it wants.

Theorem 3. *If* $N = 2$ *and* $\mathbf{LS} = [0, 1]^2$*, then* $\tilde{V} \geq \sqrt{2}$*. That is, for every algorithm, there exists an adversary and a horizon T such that the regret of the learner after T rounds is at least* $\sqrt{2}V(\mathbf{0}, T)$*.*

5. A New General Adaptive Algorithm

We study next how the random-horizon algorithm of Section 4.1 can be used when the horizon is entirely unknown and furthermore, for a much more general class of online learning problems. In Theorem 2, we proposed an algorithm that simply takes the conditional expectation of the distributions we would have played if the horizon were given. Notice that even though it is derived from the random horizon setting, it can still be used in any setting as an adaptive algorithm in the sense that it does not require the horizon as a parameter. However, to use this algorithm, we should ask two questions: What distribution should we use? And what can we say about the algorithm's performance for an arbitrary horizon instead of in expectation?

As a first attempt, suppose we use a uniform distribution over $1, \dots, T_0$, where T_0 is a huge integer. From what we observe in some numerical calculations, $\mathbb{E}[\mathbf{P}_t^T | T \geq t]$ tends to be a uniform distribution in this case. Clearly it cannot be a good algorithm if for each round, it just

places equal weights for each action regardless of the actions' behaviors. In fact, one can verify that the exponential distribution (that is, $\Pr[T=t] \propto \alpha^t$ for some constant $0<\alpha<1$) also does not work. These examples show that even though this algorithm gives us the optimal expected regret, it can still suffer a big regret for a particular trial of the game, which we definitely want to avoid.

Nevertheless, it turns out that there does exist a family of distributions that can guarantee the regret to be of order $O(\sqrt{T})$ for any T. Moreover, this is true for a very general online learning problem that includes the Hedge setting we have been discussing. Before stating our results, we first formally describe this general setting, which is sometimes called the *online convex optimization* problem (Zinkevich, 2003; Shalev-Shwartz, 2011). Let S be a compact convex set, and $\mathcal{F}$ be a set of convex functions defined on S. On each round $t=1,\dots,T$: (1) the learner chooses a point $\mathbf{x}_t \in S$; (2) the adversary chooses a loss function $f_t \in \mathcal{F}$; (3) the learner suffers loss $f_t(\mathbf{x}_t)$ for this round. The regret after T rounds is defined by

$$\mathbf{Reg}(\mathbf{x}_{1:T}, f_{1:T}) = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in S} \sum_{t=1}^{T} f_t(\mathbf{x}).$$

It is clear that the Hedge problem is a special case of the above setting with S being the probability simplex, and $\mathcal{F}$ being a set of linear functions defined by a point in the loss space, that is, $\mathcal{F} = \{f(\mathbf{x}) = \mathbf{x}\cdot\mathbf{w} : \mathbf{w} \in \mathbf{LS}\}$. Similarly, to study the minimax algorithm we define the following $V_{S,\mathcal{F}}$ function of the multiset $\mathcal{M}$ of loss functions we have encountered and the number of remaining rounds r:

$$V_{S,\mathcal{F}}(\mathcal{M}, 0) \triangleq -\min_{\mathbf{x}\in S} \sum_{f\in\mathcal{M}} f(\mathbf{x});$$

$$V_{S,\mathcal{F}}(\mathcal{M}, r) \triangleq \min_{\mathbf{x}\in S} \max_{f\in\mathcal{F}} \left(f(\mathbf{x}) + V_{S,\mathcal{F}}(\mathcal{M}\uplus\{f\}, r-1)\right),$$

where $\uplus$ denotes multiset union. We omit the subscript of $V_{S,\mathcal{F}}$ whenever there is no confusion. Let $\mathbf{x}_t^T$ be the output of the minimax algorithm on round t. In other words, $\mathbf{x}_t^T$ realizes the minimum in the definition of $V(f_{1:t-1}, T-t+1)$. We will adapt the idea in Section 4.1 and study the adaptive algorithm that outputs $\mathbb{E}_{T\sim Q}[\mathbf{x}_t^T | T \ge t] \in S$ on round t for a distribution Q on the horizon. One mild assumption needed is

Assumption 1. *$\forall \mathcal{M}$ and $r>0$, $V(\mathcal{M}, r) \ge V(\mathcal{M}, 0)$.*

Roughly speaking, this assumption implies that the game is in the adversary's favor: playing more rounds leads to greater regret. It holds for the Hedge setting with basis vector loss space (see Property 7 in the supplementary file). In fact, it also holds as long as $\mathcal{F}$ contains the zero function $f_0(\mathbf{x}) \equiv 0$. To see this, simply observe that

$$\begin{aligned} V(\mathcal{M}, r) &= \min_{\mathbf{x}\in S} \max_{f\in\mathcal{F}} \left(f(\mathbf{x}) + V(\mathcal{M}\uplus\{f\}, r-1)\right) \\ &\ge V(\mathcal{M}\uplus\{f_0\}, r-1) \\ &\ge \dots \ge V(\mathcal{M}\uplus\{f_0,\dots,f_0\}, 0) = V(\mathcal{M}, 0). \end{aligned}$$

So the assumption is mild and will hold for all the examples we consider.

Below, we first give a general upper bound on the regret that holds for any distribution and has no dependence on the choices of the adversary. After that we will show what the appropriate distributions are to make this bound $O(\sqrt{T})$.

Theorem 4. *Let $\bar{V}_t(\mathcal{M}) = \mathbb{E}_{T\sim Q}[V(\mathcal{M}, T-t+1) | T \ge t]$ and $q_t = \Pr_{T\sim Q}[T=t | T\ge t]$. Suppose Assumption 1 holds, and on round t the learner chooses $\mathbf{x}_t = \mathbb{E}_{T\sim Q}[\mathbf{x}_t^T | T \ge t]$ where $\mathbf{x}_t^T$ is the output of the minimax algorithm for horizon T as described above. Then for any T_s, the regret after T_s rounds is at most $\bar{V}_1(\emptyset) + \sum_{t=1}^{T_s} q_t \bar{V}_{t+1}(\emptyset)$.*

To prove Theorem 4, we first show the following lemma.

Lemma 1. *For any $r \ge 0$ and multiset $\mathcal{M}_1$ and $\mathcal{M}_2$,*

$$V(\mathcal{M}_1\uplus\mathcal{M}_2, r) - V(\mathcal{M}_1, 0) \le V(\mathcal{M}_2, r). \tag{5}$$

Proof. If $r=0$, then Eq. (5) holds since

$$\min_{\mathbf{x}\in S} \sum_{f\in\mathcal{M}_1} f(\mathbf{x}) + \min_{\mathbf{x}\in S} \sum_{f\in\mathcal{M}_2} f(\mathbf{x}) \le \min_{\mathbf{x}\in S} \sum_{f\in\mathcal{M}_1\uplus\mathcal{M}2} f(\mathbf{x}).$$

Now assume Eq. (5) holds for $r-1$. By induction one has

$$\begin{aligned} &V(\mathcal{M}_1\uplus\mathcal{M}_2, r) - V(\mathcal{M}_1, 0) \\ =&\min_{\mathbf{x}\in S} \max_{f\in\mathcal{F}} \left(f(\mathbf{x}) + V(\mathcal{M}_1\uplus\mathcal{M}_2\uplus\{f\}, r-1)\right) \\ &- V(\mathcal{M}_1, 0) \\ \le&\min_{\mathbf{x}\in S} \max_{f\in\mathcal{F}} \left(f(\mathbf{x}) + V(\mathcal{M}_2\uplus\{f\}, r-1)\right) = V(\mathcal{M}_2, r), \end{aligned}$$

concluding the proof. □

Proof of Theorem 4. By definition of $\mathbf{x}_t^T$, we have

$$\begin{aligned} &V(f_{1:t-1}, T-t+1) \\ =&\max_{f\in\mathcal{F}} (f(\mathbf{x}_t^T) + V(f_{1:t-1}\uplus\{f\}), T-t) \\ \ge& f_t(\mathbf{x}_t^T) + V(f_{1:t}, T-t). \end{aligned}$$

Therefore, by the convexity of f_t and the fact that $\Pr[T=t' | T\ge t] = (1-q_t)\Pr[T=t' | T\ge t+1]$ for any $t'>t$, the loss of the algorithm on round t is

$$\begin{aligned} &f_t(\mathbf{x}_t) = f_t(\mathbb{E}[\mathbf{x}_t^T | T\ge t]) \le \mathbb{E}[f_t(\mathbf{x}_t^T) | T\ge t] \\ \le&\mathbb{E}[V(f_{1:t-1}, T-t+1) - V(f_{1:t}, T-t) | T\ge t] \\ =&\bar{V}_t(f_{1:t-1}) - q_t V(f_{1:t}, 0) - (1-q_t)\bar{V}_{t+1}(f_{1:t}) \\ \le&\bar{V}_t(f_{1:t-1}) - \bar{V}_{t+1}(f_{1:t}) + q_t \bar{V}_{t+1}(\emptyset), \end{aligned}$$

where the last equality holds because $\bar{V}_{t+1}(f_{1:t}) - V(f_{1:t}, 0) = \mathbb{E}[V(f_{1:t}, T-t) - V(f_{1:t}, 0) | T \geq t+1] \leq \mathbb{E}[V(\emptyset, T-t) | T \geq t+1] = \bar{V}_{t+1}(\emptyset)$ by Lemma 1. We conclude the proof by summing up $f_t(\mathbf{x}_t)$ over $t = 1, \ldots, T_s$ and pointing out that $\bar{V}_{T_s+1}(f_{1:T_s}) = \mathbb{E}[V(f_{1:T_s}, T - T_s) | T \geq T_s + 1] \geq \mathbb{E}[V(f_{1:T_s}, 0) | T \geq T_s + 1] = -\min_{\mathbf{x} \in S} \sum_{t=1}^{T_s} f_t(\mathbf{x}_t)$ by Assumption 1. □

As a direct corollary, we now show an appropriate choice of Q. We assume that the optimal regret under the fixed horizon setting is of order $O(\sqrt{T})$. That is:

Assumption 2. *For any T, $V(\emptyset, T) \leq c_N \sqrt{T}$ for some constant c_N that might depend on N.*

This is proven to be true in the literature for all examples we consider, especially when $\mathcal{F}$ contains linear functions.

Theorem 5. *Under Assumption 2 and the same conditions of Theorem 4, if $\Pr[T = t] \propto 1/t^d$ with constant $d > \frac{3}{2}$, then for any T_s, the regret after T_s rounds is at most*

$$\frac{\Gamma(d - \frac{3}{2})}{\Gamma(d)} (d-1)^2 c_N \sqrt{\pi T_s} + o(\sqrt{T_s}),$$

where Γ is the gamma function. Choosing $d \approx 2.35$ approximately minimizes the main term in the bound, leading to regret approximately $3 c_N \sqrt{T_s} + o(\sqrt{T_s})$.

Theorem 5 tells us that pretending that the horizon is drawn from the distribution $\Pr[T = t] \propto 1/t^d$ ($d > 3/2$) can always achieve low regret, even if the actual horizon T_s is chosen adversarially. Also notice that the constant 3 in the bound for the term $c_N \sqrt{T_s}$ is less than the one for the doubling trick with the fixed horizon optimal algorithm, which is $2 + \sqrt{2}$ (Cesa-Bianchi & Lugosi, 2006). We will see in Section 6.1 an experiment showing that our algorithm performs much better than the doubling trick.

It is straightforward to apply our new algorithm to different instances of the online convex optimization framework. Examples include Hedge with basis vector loss space, predicting with expert advice (Cesa-Bianchi et al., 1997), online linear optimization within an ℓ_2 ball (Abernethy et al., 2008a) or an ℓ_∞ ball (McMahan & Abernethy, 2013). These are examples where minimax algorithms for fixed horizon are already known. In theory, however, our algorithm is still applicable when the minimax algorithm is unknown, such as Hedge with the general loss space $[0, 1]^N$.

6. Implementation and Applications

In this section, we discuss the implementation issue of our new algorithm, and also show that the idea of using a "pretend prior distribution" is much more applicable in online learning than we have discussed so far.

6.1. Closed Form of the Algorithm

Among the examples listed at the end of Section 5, we are especially interested in online linear optimization within an ℓ_2 ball since our algorithm enjoys an explicit closed form in this case. Specifically, we consider the following problem (all the norms are ℓ_2 norms): take $S = \{\mathbf{x} \in \mathbb{R}^N : \|\mathbf{x}\| \leq 1\}$, and $\mathcal{F} = \{f(\mathbf{x}) = \mathbf{x} \cdot \mathbf{w} : \mathbf{w} \in S\}$. In other words, the adversary also chooses a point in S on each round, which we denote by $\mathbf{w}_t$. Abernethy et al. (2008a) showed a simple but exact minimax optimal algorithm for the fixed horizon setting (for $N > 2$): on each round t, choose

$$\mathbf{x}_t^T = -\mathbf{W}_{t-1} \Big/ \sqrt{\|\mathbf{W}_{t-1}\|^2 + (T - t + 1)}\,, \quad (6)$$

where $\mathbf{W}_t = \sum_{t'=1}^{t} \mathbf{w}_{t'}$. This strategy guarantees the regret to be at most $\sqrt{T}$. To make it adaptive, we again assign a distribution over the horizon. However, in order to get an explicit form, a continuous distribution on T is necessary. It does not seem to make sense at first glance since the horizon is always an integer, but keep in mind that the random variable T is merely an artifact of our algorithm, and Eq. (6) is well defined with $T \geq t$ being a real number. As long as the output of the learner is in the set S, our algorithm is valid. The analysis for our algorithm also holds with minor changes. Specifically, we show the following:

Theorem 6. *Let $T \geq 1$ be a continuous random variable with probability density $f(T) \propto 1/T^2$. If the learner chooses $\mathbf{x}_t = \mathbb{E}[\mathbf{x}_t^T | T \geq t]$ on round t, where $\mathbf{x}_t^T$ is defined by Eq. (6), then the regret after T_s rounds is at most $\pi \sqrt{T_s} + o(\sqrt{T_s})$ for any T_s. Moreover, with $c = 1 + \|\mathbf{W}_{t-1}\|^2$, $\mathbf{x}_t$ has the following explicit form*

$$\mathbf{x}_t = \begin{cases} \left(\frac{t \cdot \tanh^{-1}\left(\sqrt{1 - t/c}\right)}{(c-t)^{3/2}} - \frac{\sqrt{c}}{c-t} \right) \mathbf{W}_{t-1} & \text{if } c \neq t, \\ -\frac{2t}{3c^{3/2}} \mathbf{W}_{t-1} & \text{else.} \end{cases} \quad (7)$$

The algorithm we are proposing in Eq. (7) looks quite inexplicable if one does not realize that it comes from the expression $\mathbb{E}[\mathbf{x}_t^T | T \geq t]$ with an appropriate distribution. Yet the algorithm not only enjoys a low theoretic regret bound as shown in Theorem 6, but also achieves very good performance in simulated experiments.

To show this, we conduct an experiment that compares the regrets of four algorithms at any time step within 1000 rounds against an adversary that chooses points in S uniformly at random ($N = 10$). The results are shown in Figure 1, where each data point is the *maximum* regret over 1000 randomly generated adversaries for the corresponding algorithm and horizon. The four algorithms are: the minimax algorithm (OPT) in Eq. (6) with T fixed to 1000; the one we proposed in Theorem 6 (DIST); online gradient descent (OGD, with parameter η_t being $\sqrt{2/t}$), a general algorithm for online optimization (Zinkevich, 2003);

and the doubling trick (DOUBLE) with the minimax algorithm. Note that OPT is not really an adaptive algorithm: it "cheats" by knowing the horizon $T = 1000$ in advance, and thus performs best at the end of the game. We include this algorithm merely as a baseline. Figure 1 shows that our algorithm DIST achieves consistently much lower regret than any other adaptive algorithm, including OGD which seems to enjoy a better constant in the regret bound ($2\sqrt{2T_s}$, see Zinkevich, 2003). Moreover, for the first 450 rounds or so, our algorithm performs even better than OPT, implying that using the optimal algorithm with a large guess on the horizon is inferior to our algorithm. Finally, we remark that although the doubling trick is widely applicable in theory, in experiments it is beaten by most of the other algorithms.

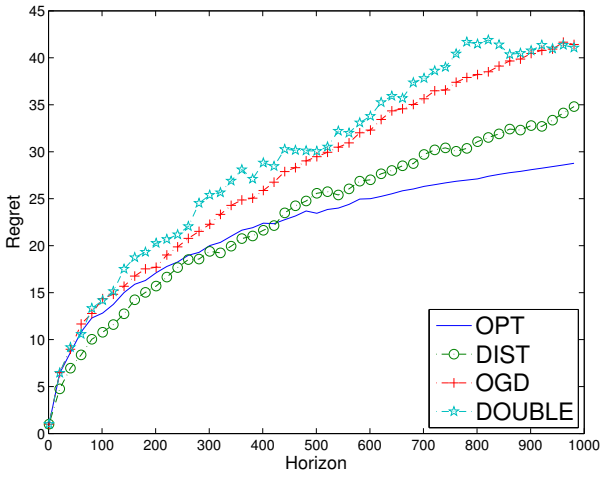

Figure 1. Comparison of four algorithms.

6.2. Randomized Play and Efficient Implementation

Implementation is an issue for our algorithm if $\mathbb{E}[\mathbf{x}_t^T | T \geq t]$ has no closed form, which is usually the case. One way to address this problem is to compute the sum of the first sufficient number of terms in the series to serve as a good estimate, since the weight for each term decreases rapidly.

However, there is another more natural way to deal with the implementation issue when we are in a similar setting but allowed to play randomly. Specifically, consider a modified Hedge setting where on each round t, the learner can bet on one and only one action I_t, and then the loss vector $\mathbf{Z}_t \in [0,1]^N$ is revealed with the learner suffering loss Z_{t,I_t} for this round. It is well known that in this kind of problem, randomization is necessary for the learner to achieve sub-linear regret (see for example Cover, 1967). That is, I_t is a random variable and $\mathbf{Z}_t$ is decided without knowing the actual draw of I_t. In addition, suppose $\mathbf{P}_t$, the conditional distribution of I_t given the past, only depends on $\mathbf{Z}_{1:t-1}$, and the learner achieves sub-linear regret in the usual Hedge setting (sometimes called *pseudo-regret*):

$$\sum_{t=1}^{T} \mathbf{P}_t \cdot \mathbf{Z}_t - \min_i M_{T,i} \leq c_N \sqrt{T} \tag{8}$$

(recall $\mathbf{M}_t = \sum_{t'=1}^{t} \mathbf{Z}_{t'}$) for any $\mathbf{Z}_{1:T}$ and a constant c_N. Then the learner also achieves sub-linear regret with high probability in the randomized setting. That is, with probability at least $1 - \delta$, the actual regret satisfies:

$$\sum_{t=1}^{T} Z_{t,I_t} - \min_i M_{T,i} \leq c_N \sqrt{T} + \sqrt{\frac{T}{2} \ln \frac{1}{\delta}}. \tag{9}$$

We refer the interested reader to Lemma 4.1 of Cesa-Bianchi & Lugosi (2006) for more details.

Therefore, in this setting we can implement our algorithm in an efficient way: on round t, first draw a horizon $T \geq t$ according to distribution $\Pr[T = t'] \propto 1/t'^d$, then draw I_t according to $\mathbf{P}_t^T$. It is clear that the marginal distribution of I_t of this process is exactly $\mathbb{E}[\mathbf{P}_t^T | T \geq t]$. Hence, Eq. (8) is satisfied by Theorem 5 and as a result Eq. (9) holds.

6.3. Combining with the FPL algorithm

Even if we have an efficient randomized implementation, or sometimes even have a closed form of the output, it is still too constrained if we can only apply our technique to minimax algorithms since they are usually difficult to derive and sometimes even inefficient to implement. It turns out, however, that the "pretend prior distribution" idea is applicable for many other non-minimax algorithms, which we will discuss from this section on.

Continuing the randomized setting discussed in the previous section, we study the well-known "follow the perturbed leader (FPL)" algorithm (Kalai & Vempala, 2005), which chooses $I_t \in \arg\min_i (M_{t-1,i} + \xi_{t,i})$ where $\boldsymbol{\xi}_t \in R^N$ is a random variable drawn from some distribution. This distribution sometimes requires the horizon T as a parameter. If this is the case, applying our technique would have a simple *Bayesian interpretation*: put a prior distribution on an unknown parameter of another distribution. Working out the marginal distribution of $\boldsymbol{\xi}_t$ would then give an adaptive variant of FPL.

Kalai & Vempala (2005) and Devroye et al. (2013) showed different choices of $\boldsymbol{\xi}_t^T$ that lead to optimal regrets. Here, for simplicity, we only consider drawing $\boldsymbol{\xi}_t^T$ uniformly at random from the hypercube $[0, \Delta_T]^N$, which gives a sub-optimal pseudo-regret $2\sqrt{TN}$ for $\Delta_T = \sqrt{TN}$ (see Cesa-Bianchi & Lugosi, 2006, Chapter 4.3). Now again let $T \geq 1$ be a continuous random variable with probability density $f(T) \propto 1/T^d (d > 3/2)$, and $\boldsymbol{\xi}_t$ be obtained by first drawing T given $T \geq t$, and then drawing a point uniformly from $[0, \Delta_T]^N$. We show the following:

Lemma 2. *If $\Delta_t = \sqrt{btN}$ for some constant $b > 0$, the marginal density function of $\boldsymbol{\xi}_t$ is*

$$f_t(\boldsymbol{\xi}) \propto \begin{cases} 0 & \text{if } \min_i \xi_i < 0 \\ \min\left\{1, \left(\frac{\Delta_t}{\|\boldsymbol{\xi}\|_\infty}\right)^{2d-2+N}\right\} & \text{else.} \end{cases} \tag{10}$$

The normalization factor is $\frac{d-1}{d-1+N/2}\Delta_t^{-N}$.

Theorem 7. *Suppose on round t, the learner chooses*

$$I_t \in \arg\min_i (M_{t-1,i} + \xi_{t,i}),$$

where $\boldsymbol{\xi}_t$ is a random variable with density function (10). Then the pseudo-regret after T_s rounds is at most

$$\left(\frac{d-1}{\sqrt{b}(d-1/2)} + \frac{\sqrt{b}(d-1)^2}{d-3/2}\right) 2\sqrt{T_s N}.$$

Choosing $b = \frac{d-3/2}{(d-1/2)(d-1)}$ and $d = 1 + \frac{\sqrt{3}}{2}$ minimizes the main term in the bound, leading to about $4.6\sqrt{T_s N}$.

By the exact same argument, the actual regret is bounded by the same quantity plus $\sqrt{\frac{T}{2}\ln\frac{1}{\delta}}$ with probability $1-\delta$.

6.4. Generalizing the Exponential Weights Algorithm

Now we come back to the usual Hedge setting and consider another popular non-minimax algorithm (note that it is trivial to generalize the results to the randomized setting). When dealing with the most general loss space $[0,1]^N$, the minimax algorithm is unknown even for the fixed horizon setting. However, generalizing the weighted majority algorithm of Littlestone & Warmuth (1994), Freund & Schapire (1997; 1999) presented an algorithm using exponential weights that can deal with this general loss space and achieve the $O(\sqrt{T\ln N})$ bound on the regret. The algorithm takes the horizon T as a parameter, and on round t, it simply chooses $P_{t,i} \propto \exp(-\eta M_{t-1,i})$, where $\eta = \sqrt{(8\ln N)/T}$ is the learning rate. It is shown that the regret of this algorithm is at most $\sqrt{(T\ln N)/2}$. Auer et al. (2002) proposed a way to make this algorithm adaptive by simply setting a time-varying learning rate $\eta = \sqrt{(8\ln N)/t}$, where t is the current round, leading to a regret bound of $\sqrt{T\ln N}$ for any T (see Chapter 2.5 of Bubeck, 2011). In other words, the algorithm always treats the current round as the last round. Below, we show that our "pretend distribution" idea can also be used to make this exponential weights algorithm adaptive, and is in fact a generalization of the adaptive learning rate algorithm by Auer et al. (2002).

Theorem 8. *Let $\mathbf{LS} = [0,1]^N$, $\Pr[T=t] \propto 1/t^d$ $(d > 3/2)$ and $\eta_T = \sqrt{(b\ln N)/T}$, where b is a constant. If on round t, the learner assigns weight $\mathbb{E}_{T\sim Q}[P^T_{t,i}|T \geq t]$ to each action i, where $P^T_{t,i} \propto \exp(-\eta_T M_{t-1,i})$, then for any T_s, the regret after T_s rounds is at most*

$$\left(\frac{\sqrt{b}(d-1)}{4(d-1/2)} + \frac{d-1}{(d-3/2)\sqrt{b}}\right)\sqrt{T_s\ln N} + o(\sqrt{T_s\ln N}).$$

Setting $b = \frac{4d-2}{d-3/2}$ minimizes the main term, which approaches 1 as $d \to \infty$.

Note that if $d \to \infty$, our algorithm simply becomes the one of Auer et al. (2002), because $\Pr[T=\tau|T\geq t]$ is 1 if $\tau = t$ and 0 otherwise. Therefore, our algorithm can be viewed as a generalization of the idea of treating the current round as the last round. However, we emphasize that the way we deal with unknown horizon is more applicable in the sense that if we try to make a minimax algorithm adaptive by treating each round as the last round, one can construct an adversary that leads to linear— and therefore grossly suboptimal—regret, whereas our approach yields nearly optimal regret. (See Example 2 and 3 in the supplementary file for details.)

6.5. First Order Regret Bound

So far all the regret bounds we have discussed are in terms of the horizon, which are also called *zeroth order bounds*. More refined bounds have been studied in the literature (Cesa-Bianchi & Lugosi, 2006). For example, the *first order bound* for Hedge, that depends on the loss of the best action m^* at the end of the game, usually is of order $O(\sqrt{m^*\ln N})$. Again, using the exponential weights algorithm with a slightly different learning rate $\eta = \ln(1+\sqrt{(2\ln N)/m^*})$, one can show that the regret is at most $\sqrt{2m^*\ln N} + \ln N$. Here, m^* is prior information on the loss sequence similar to the horizon. To avoid exploiting this information that is unavailable in practice, one can again use techniques like the doubling trick or the time-varying learning rate. Alternatively, we show that the "pretend distribution" technique can also be used here. Again it makes more sense to assign a continuous distribution on the loss of the best action instead of a discrete one.

Theorem 9. *Let $\mathbf{LS} = [0,1]^N$, $m_t = \min_i M_{t,i} + 1$, $\eta_m = \sqrt{(\ln N)/m}$, and $m \geq 1$ be a continuous random variable with probability density $f(m) \propto 1/m^d$ $(d > 3/2)$. If on round t, the learner assigns weight $\mathbb{E}[P^m_{t,i}|m \geq m_{t-1}]$ to each action i, where $P^m_{t,i} \propto \exp(-\eta_m M_{t-1,i})$, then for any T_s, the regret after T_s rounds is at most*

$$\frac{3(d-7/6)(d-1)}{(d-3/2)(d-1/2)}\sqrt{m^*\ln N} + (1+(d-1)\ln(m^*+1))\ln N + o(\sqrt{m^*\ln N}),$$

where $m^* = \min_i M_{T_s,i}$ is the loss of the best action after T_s rounds. Setting $d = 5/2 + \sqrt{2}$ minimizes the main term, which becomes $(3/2+\sqrt{2})\sqrt{m^*\ln N}$.

References

Abernethy, Jacob and Warmuth, Manfred K. Repeated games against budgeted adversaries. In *Advances in Neural Information Processing Systems 24*, 2010.

Abernethy, Jacob, Bartlett, Peter L., Rakhlin, Alexander, and Tewari, Ambuj. Optimal strategies and minimax lower bounds for online convex games. In *Proceedings of the 21st Annual Conference on Learning Theory*, 2008a.

Abernethy, Jacob, Warmuth, Manfred K., and Yellin, Joel. Optimal strategies from random walks. In *Proceedings of the 21st Annual Conference on Learning Theory*, 2008b.

Auer, Peter, Cesa-Bianchi, Nicolò, and Gentile, Claudio. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.

Berend, Daniel and Kontorovich, Aryeh. On the concentration of the missing mass. *Electron. Commun. Probab.*, 18:no. 3, 1–7, 2013. ISSN 1083-589X. doi: 10.1214/ECP.v18-2359.

Bousquet, Olivier and Warmuth, Manfred K. Tracking a small set of experts by mixing past posteriors. *Journal of Machine Learning Research*, 3:363–396, 2003.

Bubeck, Sébastien. Introduction to online optimization. Lecture notes, available at `http://www.princeton.edu/~sbubeck/BubeckLectureNotes.pdf`, 2011.

Cesa-Bianchi, Nicolò and Lugosi, Gábor. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Cesa-Bianchi, Nicolò, Freund, Yoav, Haussler, David, Helmbold, David P., Schapire, Robert E., and Warmuth, Manfred K. How to use expert advice. *Journal of the ACM*, 44(3):427–485, May 1997.

Chaudhuri, Kamalika, Freund, Yoav, and Hsu, Daniel. A parameter-free hedging algorithm. *Advances in Neural Information Processing Systems 23*, 2009.

Chernov, Alexey and Zhdanov, Fedor. Prediction with expert advice under discounted loss. In *Algorithmic Learning Theory*, volume 6331, pp. 255–269. 2010.

Cover, Thomas M. Behavior of sequential predictors of binary sequences. In *Trans. Fourth Prague Conf. on Information Theory, Statistical Decision Functions, Random Processes (Prague, 1965)*, pp. 263–272. Academia, Prague, 1967.

de Rooij, Steven, van Erven, Tim, Grünwald, Peter D., and Koolen, Wouter M. Follow the leader if you can, hedge if you must. *CoRR*, abs/1301.0534, 2013.

Devroye, Luc, Lugosi, Gábor, and Neu, Gergely. Prediction by random-walk perturbation. In *Proceedings of the 26th Annual Conference on Learning Theory*, 2013.

Foster, Dean P. and Vohra, Rakesh V. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

Freund, Yoav and Schapire, Robert E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.

Freund, Yoav and Schapire, Robert E. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29:79–103, 1999.

Gentile, Claudio. The robustness of the p-norm algorithms. *Machine Learning*, 53(3):265–299, 2003.

Gofer, Eyal and Mansour, Yishay. Lower bounds on individual sequence regret. In *Algorithmic Learning Theory*, pp. 275–289. Springer, 2012.

Hazan, Elad and Seshadhri, C. Adaptive algorithms for online decision problems. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 14, 2007.

Herbster, Mark and Warmuth, Manfred. Tracking the best expert. In *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 286–294, 1995.

Kalai, Adam and Vempala, Santosh. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Lehmer, D. H. Interesting series involving the central binomial coefficient. *The American Mathematical Monthly*, 92(7):449–457, 1985.

Littlestone, Nick and Warmuth, Manfred K. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

McMahan, H. Brendan and Abernethy, Jacob. Minimax optimal algorithms for unconstrained linear optimization. In *Advances in Neural Information Processing Systems 27*, 2013.

Rockafellar, R. Tyrrell. *Convex Analysis*. Princeton University Press, 1970.

Shalev-Shwartz, Shai. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4 (2):107–194, 2011.

Yaroshinsky, Rani, El-Yaniv, Ran, and Seiden, Steven S. How to better use expert advice. *Machine Learning*, 55(3):271–309, 2004.

Zinkevich, Martin. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.