A Stochastic View of Optimal Regret through Minimax Duality

Jacob Abernethy
Computer Science Division
UC Berkeley

Alekh Agarwal
Computer Science Division
UC Berkeley

Peter L. Bartlett
Computer Science Division
Department of Statistics
UC Berkeley

Alexander Rakhlin
Department of Statistics
University of Pennsylvania

Abstract

We study the regret of optimal strategies for online convex optimization games. Using von Neumann's minimax theorem, we show that the optimal regret in this adversarial setting is closely related to the behavior of the empirical minimization algorithm in a stochastic process setting: it is equal to the maximum, over joint distributions of the adversary's action sequence, of the difference between a sum of minimal expected losses and the minimal empirical loss. We show that the optimal regret has a natural geometric interpretation, since it can be viewed as the gap in Jensen's inequality for a concave functional—the minimizer over the player's actions of expected loss—defined on a set of probability distributions. We use this expression to obtain upper and lower bounds on the regret of an optimal strategy for a variety of online learning problems. Our method provides upper bounds without the need to construct a learning algorithm; the lower bounds provide explicit optimal strategies for the adversary.

1 Introduction

Upon a review of the central results in adversarial online learning—most of which can be found in the recent book Cesa-Bianchi and Lugosi [7]—one cannot help but notice frequent similarities between the guarantees on performance of online algorithms and the analogous guarantees under stochastic assumptions. However, discerning an explicit link has remained elusive. Vovk [21] notices this phenomenon: "for some important problems, the adversarial bounds of online competitive learning theory are only a tiny amount worse than the average-case bounds for some stochastic strategies of Nature."

In this paper, we attempt to build a bridge between adversarial online learning and statistical learning. Using von Neumann's minimax theorem, we show that the optimal regret of an algorithm for online convex optimization is exactly the difference between a sum of minimal expected losses and the minimal empirical loss, under an adversarial choice of a stochastic process generating the data. This leads to upper and lower bounds for the optimal regret that exhibit several similarities to results from statistical learning.

The online convex optimization game proceeds in rounds. At each of these T rounds, the player (learner) predicts a vector in some convex set, and the adversary responds with a convex function which determines the player's loss at the chosen point. In order to emphasize the relationship with the stochastic setting, we denote the player's choice as $f \in \mathcal{F}$ and the adversary's choice as $z \in \mathcal{Z}$. Note that this differs, for instance, from the notation in [2].

Suppose $\mathcal{F}$ is a *convex compact* class of functions, which constitutes the set of Player's choices. The Adversary draws his choices from a *closed compact* set $\mathcal{Z}$. We also define a continuous bounded loss function $\ell : \mathcal{Z} \times \mathcal{F} \to \mathbb{R}$ and assume that ℓ is convex in the second argument. Denote by $\ell(\mathcal{F}) = \{\ell(\cdot, f) : f \in \mathcal{F}\}$ the associated loss class. Let $\mathscr{P}$ be the set of all probability distributions on $\mathcal{Z}$. Denote a sequence $(Z_1, \ldots, Z_T)$ by Z_1^T. We denote a joint distribution on $\mathcal{Z}^T$ by a bold-face $\mathbf{p}$ and its conditional and marginal distributions by $p_t\left(\cdot | Z_1^{t-1}\right)$ and p_t^m, respectively.

The online convex optimization interaction is described as follows.

Online Convex Optimization (OCO) Game

At each time step $t = 1$ to T,
- Player chooses $f_t \in \mathcal{F}$
- Adversary chooses $z_t \in \mathcal{Z}$
- Player observes z_t and suffers loss $\ell(z_t, f_t)$

The objective of the player is to minimize the regret

$$\sum_{t=1}^{T} \ell(z_t, f_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(z_t, f).$$

It turns out that many online learning scenarios can be realized as instances of OCO, including prediction with expert advice, data compression, sequential investment, and forecasting with side information (see, for example, [7]).

Let us briefly mention previous work and outline our contributions. Our starting point, Theorem 1, is similar to Sec. 2.10 of [7], but extends it to non-oblivious adversaries and arbitrary losses. Dealing with non-oblivious adversaries (viz., those whose actions depend on player's choices) entails considering a nested sequence of inf/sup pairs, and much of the generality and difficulty comes from this setup. The use of minimax duality has a long history in decision theory and Bayesian analysis. In the case of binary sequence prediction,

we refer to [9, 18] and references therein. For the log loss, the minimax strategy is known to have a closed form, and there is a large body of work on the subject in different communities (see [18, 7]). The use of Rademacher averages in the context of prediction first appeared in [6] for the absolute loss.

The aim of this paper is to provide a unifying framework, as well as tools for studying minimax regret in a general setting: a non-oblivious adversary and convex loss functions. One of the contributions is the upper bound in terms of Rademacher averages (Theorem 18), which extends that in [7] to non-linear losses. The lower bound of Theorem 19 generalizes the asymptotic bound for expert case (e.g. [7]). We provide an important geometric viewpoint and show that fast rates can be obtained by studying properties of the minimum expected risk functional. We show that strong convexity implies smoothness of this functional, thus recovering known results on logarithmic regret. When the functional is non-differentiable, an explicit optimal strategy of the adversary is to play the distribution that exhibits the non-differentiability.

2 Applying von Neumann's minimax theorem

Define the value of the OCO game—which we also call the *minimax regret*—as

$$\mathscr{R}_T := \inf_{f_1\in\mathcal{F}} \sup_{z_1\in\mathcal{Z}} \cdots \inf_{f_{T-1}\in\mathcal{F}} \sup_{z_{T-1}\in\mathcal{Z}} \inf_{f_T\in\mathcal{F}} \sup_{z_T\in\mathcal{Z}} \left(\sum_{t=1}^T \ell(z_t, f_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^T \ell(z_t, f) \right). \quad (1)$$

The OCO game has a purely "optimization" flavor. However, applying von Neumann's minimax theorem shows that its value is closely related to the behavior of the empirical minimization algorithm in a stochastic process setting.

Theorem 1 *Under the assumptions on $\mathcal{F}$, $\mathcal{Z}$, and ℓ given in the previous section,*

$$\mathscr{R}_T = \sup_{\mathbf{p}} \mathbb{E} \left[\sum_{t=1}^T \inf_{f_t\in\mathcal{F}} \mathbb{E}\left[\ell(Z_t, f_t) | Z_1^{t-1}\right] - \inf_{f\in\mathcal{F}} \sum_{t=1}^T \ell(Z_t, f) \right], \quad (2)$$

where the supremum is over all joint distributions $\mathbf{p}$ on $\mathcal{Z}^T$ and the expectations are over the sequence of random variables $\{Z_1, \ldots, Z_T\}$ drawn according to $\mathbf{p}$.

The proof relies on the following version of von Neumann's minimax theorem; it appears as Theorem 7.1 in [7].

Proposition 2 *Let $M(x, y)$ denote a bounded real-valued function on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are convex sets and $\mathcal{X}$ is compact. Suppose that $M(\cdot, y)$ is convex and continuous for each fixed $y \in \mathcal{Y}$ and $M(x, \cdot)$ is concave for each $x \in \mathcal{X}$. Then*

$$\inf_{x\in\mathcal{X}} \sup_{y\in\mathcal{Y}} M(x, y) = \sup_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} M(x, y).$$

Proof: [of Theorem 1]

Consider the last optimization choice z_T in Eq. (1). If we instead draw z_T according to a probability distribution, and compute the expected value of the quantity in the parentheses in Eq. (1), then maximizing this expected value over all distributions on $\mathcal{Z}$ is equivalent to maximizing over z_T. Hence,

$$\mathscr{R}_T = \inf_{f_1\in\mathcal{F}} \sup_{z_1\in\mathcal{Z}} \cdots \inf_{f_{T-1}\in\mathcal{F}} \sup_{z_{T-1}\in\mathcal{Z}} \inf_{f_T\in\mathcal{F}} \sup_{p_T\in\mathscr{P}} \mathbb{E}_{Z_T\sim p_T} \left[\sum_{t=1}^T \ell(z_t, f_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^T \ell(z_t, f) \right]. \quad (3)$$

In the last expression, it is understood that sums are over the sequence $\{z_1, \ldots, z_{T-1}, Z_T\}$, that is, the first $T-1$ elements are quantified in the suprema, while the last Z_T is a random variable.

We now apply Proposition 2 to the last inf/sup pair in (3), with

$$M(f_T, p_T) = \mathbb{E}_{Z_T\sim p_T} \left[\sum_{t=1}^T \ell(z_t, f_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^T \ell(z_t, f) \right],$$

which is convex in f_T (by assumption) and linear in p_T. Moreover, the set $\mathcal{F}$ is compact, and both $\mathcal{F}$ and $\mathscr{P}$ are convex. We conclude that

$$\begin{aligned}
\mathscr{R}_T &= \inf_{f_1\in\mathcal{F}} \sup_{z_1\in\mathcal{Z}} \cdots \inf_{f_{T-1}\in\mathcal{F}} \sup_{z_{T-1}\in\mathcal{Z}} \sup_{p_T\in\mathscr{P}} \inf_{f_T\in\mathcal{F}} \mathbb{E}\left[\sum_{t=1}^T \ell(z_t, f_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^T \ell(z_t, f) \right] \\
&= \inf_{f_1\in\mathcal{F}} \sup_{z_1\in\mathcal{Z}} \cdots \inf_{f_{T-1}\in\mathcal{F}} \sup_{z_{T-1}\in\mathcal{Z}} \sup_{p_T\in\mathscr{P}} \left(\sum_{t=1}^{T-1} \ell(z_t, f_t) \right. \\
&\quad \left. + \inf_{f_T\in\mathcal{F}} \mathbb{E}\left[\ell(Z_T, f_T)\right] - \mathbb{E} \inf_{f\in\mathcal{F}} \sum_{t=1}^T \ell(z_t, f) \right).
\end{aligned}$$

Note that, in the last line, the maximizing distribution p_T depends on the previous choices z_1^{T-1}, but not on any of the f_t's. As we swap $\inf$ / $\sup$ from inside out, z_t's are taken to be random variables and denoted by Z_t. Pulling the expectation on the third term outside and repeating the process T times, we arrive at the statement of the Theorem. We refer to [1] for more details. ∎

We can think of Eq. (2) as a game where the adversary goes first. At every round he "plays" a distribution and the player responds with a function that minimizes the conditional expectation.

We remark that we can allow the player to choose f_t's non-deterministically in the original OCO game. In that case, the original infimum should be over distributions on $\mathcal{F}$. We then do not need convexity of ℓ in $f \in \mathcal{F}$'s in order to apply von Neumann's theorem, and the resulting expression for the value of the game is the same.

3 First Steps

The present work focuses on analyzing the expression in Equation (2) for a range of different choices of $\mathcal{Z}$ and $\mathcal{F}$, as

well as for various assumptions made about the loss function ℓ. We are not only interested in upper- and lower-bounding the value of the game $\mathscr{R}_T$, but also in determining the types of distributions $\mathbf{p}$ that maximize or almost maximize the expression in (2). To that end, define $\mathbf{p}$-regret as

$$\mathscr{R}_T(\mathbf{p}) = \mathbb{E}\left[\sum_{t=1}^{T} \min_{f_t \in \mathcal{F}} \mathbb{E}\left[\ell(Z_t, f_t) | Z_1^{t-1}\right] - \min_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(Z_t, f)\right] \tag{4}$$

for any joint distribution $\mathbf{p}$ of $(z_1, \ldots, z_T) \in \mathcal{Z}^T$. In this section we will provide an array of analytical tools for working with $\mathscr{R}_T(\mathbf{p})$.

3.1 Regret for IID and Product Distributions

It is natural to consider i.i.d. processes and product distributions $\mathbf{p}$ as candidates for maximizing $\mathscr{R}_T(\mathbf{p})$.

Lemma 3 *For any i.i.d. distribution* $\mathbf{p}$*,* $\mathscr{R}_T(\mathbf{p}) \geq 0$*. Hence,* $\mathscr{R}_T \geq 0$*.*

Proof: For an i.i.d. distribution Eq. (4) becomes

$$\begin{aligned}\frac{1}{T}\mathscr{R}_T(\mathbf{p}) &= \min_{f \in \mathcal{F}} \mathbb{E}\left[\ell(Z, f)\right] - \mathbb{E} \min_{f \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^{T} \ell(Z_t, f) \\ &\geq \min_{f \in \mathcal{F}} \mathbb{E}\left[\ell(Z, f)\right] - \min_{f \in \mathcal{F}} \mathbb{E}\, \frac{1}{T} \sum_{t=1}^{T} \ell(Z_t, f) = 0\end{aligned}$$

where the inequality is due to the fact that $\mathbb{E} \min \leq \min \mathbb{E}$. ■

Observe that $\mathscr{R}_T(\mathbf{p})$ for an i.i.d. process is the difference between the minimum expected loss and the expectation of the empirical loss of an empirical minimizer.

With the goal of studying various types of distributions, we now define the following hierarchy:

$$\mathscr{R}_T^{\text{i.i.d.}} := \sup_{\mathbf{p}=p^T} \mathscr{R}_T(\mathbf{p}); \quad \mathscr{R}_T^{\text{indep.}} := \sup_{\mathbf{p}=p_1 \times \ldots \times p_T} \mathscr{R}_T(\mathbf{p}),$$

where $p, p_1, \ldots, p_T$ are arbitrary distributions on $\mathcal{Z}$. It is immediately clear that

$$0 \quad \leq \quad \mathscr{R}_T^{\text{i.i.d}} \quad \leq \quad \mathscr{R}_T^{\text{indep.}} \quad \leq \quad \mathscr{R}_T. \tag{5}$$

We will see that, given particular assumptions on $\mathcal{F}, \mathcal{Z}$ and ℓ, some of the gaps in the above hierarchy are significant, while others are not. Before continuing, however, we need to develop some tools for analyzing the minimax regret.

3.2 Tools for a General Analysis

We now introduce two new objects that help to simplify the expression in (2) as well as derive properties of $\mathscr{R}_T(\mathbf{p})$.

Definition 4 *Given sets* $\mathcal{F}, \mathcal{Z}$*, we can define the* minimum expected loss *functional* Φ *as*

$$\Phi(p) := \inf_{f \in \mathcal{F}} \mathbb{E}_{Z \sim p}\left[\ell(Z, f)\right],$$

where p *is some distribution on* $\mathcal{Z}$*.*

Defining an inner product $\langle h, p \rangle = \int_z h(z) dp(z)$ for a distribution p, we observe that $\Phi(p) = \inf_{f \in \mathcal{F}} \langle \ell(\cdot, f), p \rangle$.

Definition 5 *For any* $Z_1, \ldots, Z_T \in \mathcal{Z}^T$*, we denote* $\hat{P}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{Z_t}(\cdot)$*, the empirical distribution.*

With this additional notation, we can rewrite (4) as

$$\frac{1}{T}\mathscr{R}_T(\mathbf{p}) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\,\Phi(p_t\left(\cdot | Z_1^{t-1}\right)) - \mathbb{E}\,\Phi(\hat{P}_T). \tag{6}$$

Thus, the Adversary's task is to induce a large deviation between the average sequence of conditional distributions $\{p_t\left(\cdot | Z_1^{t-1}\right)\}$ and an empirical sample $\hat{P}_T$ from these conditionals, where the deviation is defined by way of the functional Φ.

Lemma 6 *The functional* $\Phi(\cdot)$ *is concave on the space of distributions over* $\mathcal{Z}$ *and* $\mathscr{R}_T(\cdot)$ *is concave with respect to joint distributions on* $\mathcal{Z}^T$*.*

The (easy) proof of this lemma is in the full version [1]. It is indeed concavity of Φ that is key to understanding the behavior of $\mathscr{R}_T$. A hint of this can already be seen in the proof of Lemma 3, where the only inequality is due to the concavity of the $\min$. In the next section, we show how this description of regret can be interpreted through a *Bregman divergence* in terms of Φ.

3.3 Divergences and the Gap in Jensen's Inequality

We now show how to interpret regret through the lens of Jensen's Inequality by providing yet another expression for it, now in terms of Bregman Divergences. We begin by revisiting the i.i.d. case $\mathbf{p} = p^T = p \times \ldots \times p$, for some distribution p on $\mathcal{Z}$. Equation (6) simplifies to a very natural quantity,

$$\frac{1}{T}\mathscr{R}_T(p^T) = \Phi(p) - \mathbb{E}\,\Phi(\hat{P}_T). \tag{7}$$

Notice that $\hat{P}_T$ is a random quantity, and in particular that $\mathbb{E}\,\hat{P}_T = p$. As $\Phi(\cdot)$ is concave, with an immediate application of Jensen's Inequality we obtain $\mathscr{R}_T(p^T) \geq 0$. For arbitrary joint distributions p, we can similarly interpret regret as a "gap" in Jensen's Inequality, albeit with some added complexity.

Definition 7 *If* F *is any convex differentiable*[1] *functional on the space of distributions on* $\mathcal{Z}$*, we define* Bregman divergence *with respect to* F *as*

$$\mathcal{D}_F(q, p) = F(q) - F(p) - \langle \nabla F(p), q - p \rangle.$$

If F is non-differentiable, we can take a particular subgradient $v_p \in \partial F(p)$ in place of $\nabla F(p)$. Note that the notion of subgradients is well-defined even for infinite-dimensional convex functions. Having chosen[2] a mapping $p \mapsto v_p \in$

[1] Here, we mean differentiable with respect to the Fréchet or Gâteaux derivative. We refer the reader to [10] for precise definitions of functional Bregman Divergences.

[2] The assumption of compactness of $\mathcal{F}$, together with the characterization of the subgradient set in Section 4.2, allow us, for instance, to define the mapping $p \mapsto v_p$ by putting a uniform measure on the subgradient set and defining v_p to be the expected subgradient with respect to it. In fact, the choice of the mapping is not important, as long as it does not depend on q.

$\partial F(p)$, we define *a generalized divergence with respect to F and v_p* as

$$\mathcal{D}_F(q,p) = F(q) - F(p) - \langle v_p, q-p\rangle.$$

Throughout the paper, we focus only on the divergence $\mathcal{D}_{-\Phi}$, and thus we omit $-\Phi$ from the notation for simplicity.

Given the definition of divergence, it immediately follows that, for a random distribution q,

$$\Phi(\mathbb{E}\, q) - \mathbb{E}\,\Phi(q) = \mathbb{E}\,\mathcal{D}(q, \mathbb{E}\, q)$$

since the linear term disappears under the expectation. This simple observation is quite useful; notice we now have an even simpler expression for i.i.d. regret (7):

$$\frac{1}{T}\mathscr{R}(p^T) = \mathbb{E}\,\mathcal{D}(\hat{P}_T, p).$$

In other words, the p^T-regret is *equal* to the expected divergence between the empirical distribution and its expectation. This will be a starting point for obtaining lower bounds for $\mathscr{R}_T$. For general joint distributions p, let us rewrite the expression in (6) as

$$\mathbb{E}_{\,t\sim U}\mathbb{E}\,\Phi(p_t\left(\cdot|Z_1^{t-1}\right)) - \mathbb{E}\,\Phi(\hat{P}_T),$$

where we replaced the average with a uniform distribution on the rounds. Roughly speaking, the next lemma says that one can obtain $\mathbb{E}\,\Phi(\hat{P}_T)$ from $\mathbb{E}_{\,t\sim U}\mathbb{E}\,\Phi(p_t\left(\cdot|Z_1^{t-1}\right))$ through three applications of Jensen's inequality, due to various expectations being "pulled" inside or outside of Φ.

Lemma 8 *Suppose* $\mathbf{p}$ *is an arbitrary joint distribution. Denote by* $p_t\left(\cdot|Z_1^{t-1}\right)$ *and* p_t^m *the conditional and marginal distributions, respectively. Then*

$$\frac{1}{T}\mathscr{R}(\mathbf{p}) = -\Delta_0 - \Delta_1 + \Delta_2, \tag{8}$$

where

$$\Delta_0 = \frac{1}{T}\sum_t \mathcal{D}\left(p_t^m, \tfrac{1}{T}\textstyle\sum_{t'} p_{t'}^m\right),$$

$$\Delta_1 = \frac{1}{T}\sum \mathbb{E}_{\,\mathbf{p}}\mathcal{D}(p_t\left(\cdot|Z_1^{t-1}\right), p_t^m),$$

$$\Delta_2 = \mathbb{E}_{\,\mathbf{p}}\mathcal{D}\left(\hat{P}_T, \frac{1}{T}\sum p_t^m\right).$$

Proof: The marginal distribution satisfies $\mathbb{E}\, p_t\left(\cdot|Z_1^{t-1}\right) = p_t^m$, and it is easy to see that $\mathbb{E}\,\hat{P}_T = \frac{1}{T}\sum_t p_t^m$. Given this, we see that

$$\begin{aligned}
\frac{1}{T}\mathscr{R}(\mathbf{p}) &= \mathbb{E}_{\,\mathbf{p}}\left[\frac{1}{T}\sum_{t=1}^T \Phi(p_t\left(\cdot|Z_1^{t-1}\right)) - \Phi(\hat{P}_T)\right] \\
&= \overbrace{\mathbb{E}_{\,\mathbf{p}}\left[\frac{1}{T}\sum_t \left\{\Phi(p_t\left(\cdot|Z_1^{t-1}\right)) - \Phi(p_t^m)\right\}\right]}^{-\Delta_1} \\
&\quad + \overbrace{\frac{1}{T}\sum_t \Phi(p_t^m) - \Phi\left(\tfrac{1}{T}\textstyle\sum_t p_t^m\right)}^{-\Delta_0} \\
&\quad - \overbrace{\mathbb{E}_{\,\mathbf{p}}\left[\Phi(\hat{P}_T) - \Phi\left(\tfrac{1}{T}\textstyle\sum_t p_t^m\right)\right]}^{-\Delta_2}.
\end{aligned}$$

■

This lemma sheds some light on the influence of an i.i.d. vs. product vs. arbitrary joint distribution on the regret. For product distributions, every conditional distribution is identical to its marginal distributions, thus implying $\Delta_1 = 0$. Furthermore, for any i.i.d. distribution, each marginal distribution is identical to the average marginal, thus implying that $\Delta_0 = 0$. With this in mind, it is tempting to assert that the largest regret is obtained at an i.i.d. distribution, since transitions from i.i.d to product, and from product to arbitrary distribution, only subtract from the regret value. While appealing, this is unfortunately not the case: in many instances the final term, Δ_2, can be made larger with a non-i.i.d. (and even non-product) distribution, even at the added cost of positive Δ_0 and Δ_1 terms, so that $\mathscr{R}_T^{\text{i.i.d.}} = o(\mathscr{R}_T)$ as a function of T. In some cases, however, we show that a lower bound on the regret can be obtained with an i.i.d. distribution at a cost of only a constant factor.

4 Properties of Φ

In statistical learning, the rate of decay of prediction error is known to depend on the curvature of the loss: more curvature leads to faster rates (see, for example, [15, 16, 4]), and slow (e.g. $\Omega(T^{-1/2})$) rates occur when the loss is not strictly convex, or when the minimizer of the expected loss is not unique [15, 17]. There is a striking parallel with the behavior of the regret in online convex optimization; again the curvature of the loss plays a central role. Roughly speaking, if ℓ is strongly convex or exp-concave, second-order gradient-descent methods ensure that the regret grows no faster than $\log T$ (e.g. [11]); if ℓ is linear, the regret can grow no faster than $\sqrt{T}$ (e.g. [22]); intermediate rates can be achieved as well if the curvature varies [3].

The previous section expresses regret as a sum of divergences under Φ, and that suggests that the curvature of Φ should be an important factor in determining the rates of regret. We shall see that this is the case: curvature of Φ leads to large regret, while flatness of Φ implies small regret.

We will now show how properties of the loss function class determine the curvature of Φ. In later sections we will show how such curvature properties lead directly to particular rates for $\mathscr{R}_T$. First, let us provide a fruitful geometric picture, rooted in convex analysis. It allows us to see the function Φ, roughly speaking, as a mirror image of the function class.

4.1 Geometric interpretation of Φ

In general, the set $\mathcal{Z}$ is uncountable, so care must be taken with regard to various notions we are about to introduce. We refer the reader to Chapter 10 of [5] for the discussion of finite vs infinite-dimensional spaces in convex analysis. Since $\mathcal{Z}$ is compact by assumption, we can discretize it to a fine enough level such that the upper and lower bounds of this paper hold, as long as the results are non-asymptotic. In the present Section, for simplicity of exposition, we will suppose that the set $\mathcal{Z}$ is finite with cardinality d. This assumption is required only for the geometric interpretation; our proofs are correct as long as $\mathcal{Z}$ is compact.

Hence, distributions over the set $\mathcal{Z}$ are associated with d-dimensional vectors. Furthermore, each $f \in \mathcal{F}$ is specified by its d values on the points. We write $\ell_f \in \mathbb{R}^d$ for the loss vector of f, $\ell(\cdot, f)$. Let us denote the set of all such vectors by $\ell(\mathcal{F})$. We then have

$$-\Phi(p) = -\inf_{f\in\mathcal{F}} \mathbb{E}_p \ell(Z, f) = \sup_{f\in\mathcal{F}} \langle -\ell_f, p\rangle = \sigma_{-\ell(\mathcal{F})}(p),$$

where $\sigma_S(x) = \sup_{s\in S}\langle s, x\rangle$ is the support function for the set S. This function is one of the most basic objects of convex analysis (see, for instance, [12]). It is well-known that $\sigma_S = \sigma_{\mathrm{co}S}$; in other words, the support function does not change with respect to taking convex hull (see Proposition 2.2.1, page 137, [12]). To this end, let us denote $S = \mathrm{co}[-\ell(\mathcal{F})] \subset \mathbb{R}^d$.

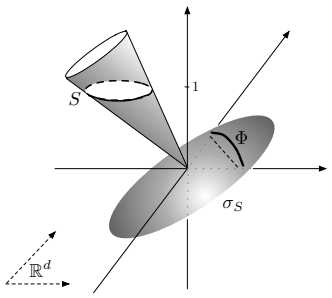

Figure 1: Dual cone as the epigraph of the support function. Φ is the restriction to the simplex.

It is known that the support function is *sublinear* and its epigraph is a cone. To visualize the support function, consider the $\mathbb{R}^d \times \mathbb{R}$ space. Embed the set $S \subset \mathbb{R}^d$ in $\mathbb{R}^d \times \{1\}$. Then construct the conic hull of $S \times \{1\}$. It turns out that the cone which is dual to the constructed conic hull is the epigraph of the support function σ_S. The dual cone is the set of vectors which form obtuse or right angles with all the vectors in the original cone. Hence, one can visualize the surface σ_S as being at right angles to the conic hull of $S \times \{1\}$. Now, the function Φ is just the restriction of σ_S to the simplex (see Figure 1). We can now deduce properties of Φ from properties of the loss class.

4.2 Differentiability of Φ

Lemma 9 *The subdifferential set of Φ is the set of expected minimizers:*

$$\partial\Phi(p) = \{\ell_f : f \in \arg\min_{f\in\mathcal{F}} \mathbb{E}_p \ell(Z, f)\}.$$

Hence, the functional Φ is differentiable at a distribution p iff $\arg\min_{f\in\mathcal{F}} \mathbb{E}_p \ell(z, f)$ is unique.

Proof: The statement follows from Proposition 2.1.5 in [12]. ■

In particular, for Φ to be differentiable for all distributions, the loss function class should not have a "face" exposed to the origin. This geometrical picture and its implications will be studied further in Section 6.

It is easy to verify that *strict* convexity of $\ell(z, f)$ in f implies uniqueness of the minimizer for any p and, hence, differentiability of Φ.

4.3 Flatness of Φ through curvature of ℓ

In this section we show that curvature in the loss function leads to flatness of Φ. We would indeed expect such a result to hold since regret decaying faster than $O(T^{-1/2})$ is known to occur in the case of curved losses (e.g. [3]), and decomposition (6) suggests that this should imply flatness of Φ. More precisely, we show that if $\ell(f, z)$ is strongly convex in f with respect to some norm $\|\cdot\|$, then Φ is strongly flat with respect to the ℓ_1 norm on the space of distributions. Before stating the main result, we provide several definitions.

Definition 10 *A convex function F is α-*flat *(or α-*smooth*) with respect to a norm $\|\cdot\|$ when*

$$F(y) - F(x) \le \langle \nabla F(x), y - x\rangle + \alpha\|x - y\|^2 \qquad (9)$$

for all x, y. We will say that a concave function G is α-flat if $-G$ satisfies (9)*.*

Let us also recall the definition of ℓ_1 (or variational) norm on distributions.

Definition 11 *For two distributions p, q on $\mathcal{Z}$, we define*

$$\|p - q\|_1 = \int_{\mathcal{Z}} |dp(z) - dq(z)|.$$

Theorem 12 *Suppose $\ell(z, f)$ is σ-strongly convex in f, that is,*

$$\ell\left(z, \frac{f+g}{2}\right) \le \frac{\ell(z, f) + \ell(z, g)}{2} - \frac{\sigma}{8}\|f - g\|^2$$

for any $z \in Z$ and $f, g \in \mathcal{F}$. Suppose further that ℓ is $L-Lipschitz$, that is,

$$|\ell(z, f) - \ell(z, g)| \le L\|f - g\|.$$

Under these conditions, the Φ-functional is $\frac{2L^2}{\sigma}$-flat with respect to $\|\cdot\|_1$.

The proof uses the following lemma, which shows stability of the minimizers. Its proof is in the full version [1].

Lemma 13 *Fix two distributions p, q. Let f_p and f_q be the functions achieving the minimum in $\Phi(p)$ and $\Phi(q)$, respectively. Under the conditions of Theorem 12,*

$$\|f_p - f_q\| \le \frac{2L}{\sigma}\|p - q\|_1.$$

Proof:[of Theorem 12] We have

$$\begin{aligned}\Phi(p) - \Phi(q) &= \mathbb{E}_p \ell(z, f_p) - \mathbb{E}_q \ell(z, f_q)\\ &= (\mathbb{E}_p \ell(z, f_p) - \mathbb{E}_q \ell(z, f_p))\\ &\quad + (\mathbb{E}_q \ell(z, f_p) - \mathbb{E}_q \ell(z, f_q)). \qquad (10)\end{aligned}$$

Let us first study the second term in the expression above. As f_p is the minimizer of $\mathbb{E}_p \ell(z, f)$, we have:

$$\mathbb{E}_p [\ell(z, f_p) - \ell(z, f_q)] \le 0$$

So

$$\begin{aligned}\mathbb{E}_q [\ell(z, f_p) - \ell(z, f_q)] &\le \mathbb{E}_q [\ell(z, f_p) - \ell(z, f_q)]\\ &\quad - \mathbb{E}_p [\ell(z, f_p) - \ell(z, f_q)]\\ &= \int (\ell(z, f_p) - \ell(z, f_q))(q(z) - p(z))dz\\ &\le L \int \|f_p - f_q\||p(z) - q(z)|dz.\end{aligned}$$

Using Lemma 13, we get:

$$\mathbb{E}_q\left[\ell(z,f_p)-\ell(z,f_q)\right]\le\frac{2L^2}{\sigma}\|p-q\|_1^2. \qquad (11)$$

As for the first term in (10),

$$\begin{aligned}\mathbb{E}_p\ell(z,f_p)-\mathbb{E}_q\ell(z,f_p)&=\int_z\ell(z,f_p)(p(z)-q(z))dz\\&=\langle\ell(\cdot,f_p),(p-q)\rangle. \qquad (12)\end{aligned}$$

The fact that $\ell(\cdot,f_p)$ is a subdifferential of Φ at p is proved in [1]. We conclude that the terms in (10) are the first and the second order terms in the expansion of Φ.

■

We remark that we can arrive at above results by explicitly considering the dual function Φ^*, proving strong convexity of Φ^* with respect to $\|\cdot\|_\infty$ (which follows from our assumption on ℓ), and then concluding strong flatness of Φ with respect to $\|\cdot\|_1$. This is indeed the main intuition at the heart of our proof.

5 Upper Bounds on $\mathscr{R}_T$

In this section, we exhibit two general upper bounds on $\mathscr{R}_T$ that hold for a wide class of OCO games. The first bound, which holds when the functional Φ is differentiable and not too curved, is of the form $\mathscr{R}_T = O(\log T)$. The second, which holds for *arbitrary* Φ, e.g. where the functional may even have a non-differentiability, is stated in terms of the *Rademacher complexity* of the class $\mathcal{F}$. Such Rademacher complexity results imply a regret upper bound on the order of $\sqrt{T}$.

An intriguing observation is that these bounds are proved without actually exhibiting a strategy for the Player, as is typically done. This illustrates the power of the minimax duality approach: we can prove the existence of an optimal algorithm, and determine its performance, all without providing its construction.

5.1 Fast Rates: Exploiting the Curvature

For differentiable Φ with bounded second derivative, we can prove that the regret grows no faster than logarithmically in T. Of course, rates of $\log T$ have been given previously [11, 20, 21]. We build upon these results in the present work by showing that logarithmic regret must always arise when Φ satisfies a flatness condition.

Theorem 14 *Suppose the Φ functional is differentiable and α-flat with respect the norm $\|\cdot\|_1$ on $\mathscr{P}$. Then $\mathscr{R}_T\le 4\alpha\log T$.*

We immediately obtain the following corollary.

Corollary 15 *Suppose functions $\ell(z,f)$ are σ-strongly convex and $L-$ Lipschitz in f. Then $\mathscr{R}_T\le\frac{8L^2}{\sigma}\log T$.*

Furthermore, as we show in Section 7.3, the $\log T$ bound is tight for quadratic functions; there is an explicit joint distribution for the adversary which attains this value.

The proof of Theorem 14 involves the following lemma.

Lemma 16 *The* $\mathbf{p}$*-regret can be upper-bounded as*

$$\mathscr{R}_T(\mathbf{p})\le\mathbb{E}\left[\sum_{t=1}^T t\cdot\mathcal{D}\left(\hat{P}_t,\bar{P}_t\right)\right]$$

where $\bar{P}_t(\cdot)=\left(\frac{t-1}{t}\right)\hat{P}_{t-1}(\cdot)+\frac{1}{t}p_t\left(\cdot|Z_1^{t-1}\right)$.

Proof: Consider the following difference:

$$\begin{aligned}\delta_T:&=\frac{1}{T}\mathbb{E}\,\Phi\left(p_T(\cdot|Z_1^{T-1})\right)-\mathbb{E}\,\Phi\left(\hat{P}_T\right)\\&=\frac{1}{T}\mathbb{E}\,\Phi\left(p_T(\cdot|Z_1^{T-1})\right)-\mathbb{E}\,\Phi\left(\bar{P}_T\right)\\&\quad+\mathbb{E}\,\Phi\left(\bar{P}_T\right)-\mathbb{E}\,\Phi\left(\hat{P}_T\right)\end{aligned}$$

For the first difference we use concavity of Φ. The second difference can be written as a divergence because the linear term vanishes in expectation. Indeed,

$$\mathbb{E}\left\langle\nabla\Phi\left(\bar{P}_T\right),\frac{1}{T}(\mathbf{1}_{Z_T}(\cdot)-p_T(\cdot|Z_1^{T-1}))\right\rangle=0$$

because the gradient does not depend on Z_T, while

$$\mathbb{E}_{Z_T}\left[\mathbf{1}_{Z_T}(\cdot)|Z_1^{T-1}\right]=p_T(\cdot|Z_1^{T-1}).$$

Hence, δ_T is no more than

$$-\left(\frac{T-1}{T}\right)\mathbb{E}\,\Phi(\hat{P}_{T-1})+\mathbb{E}\,\mathcal{D}\left(\hat{P}_T,\bar{P}_T\right)$$

and so

$$\begin{aligned}\mathscr{R}_T(\mathbf{p})&=\sum_{t=1}^T\mathbb{E}\,\Phi\left(p_t(\cdot|Z_1^{t-1})\right)-T\mathbb{E}\,\Phi\left(\hat{P}_T\right)\\&=\sum_{t=1}^{T-1}\mathbb{E}\,\Phi\left(p_t(\cdot|Z_1^{t-1})\right)+T\delta_T\\&\le\sum_{t=1}^{T-1}\mathbb{E}\,\Phi\left(p_t(\cdot|Z_1^{t-1})\right)-(T-1)\mathbb{E}\,\Phi(\hat{P}_{T-1})\\&\quad+T\mathbb{E}\,\mathcal{D}\left(\hat{P}_T,\bar{P}_T\right).\end{aligned}$$

■

Before proceeding, note that we may interpret $\bar{P}_t$ as the conditional expectation of the uniform distribution $\hat{P}_t$ given $Z_1,\ldots,Z_{t-1}$. The flatness of Φ will allow us to show that $\bar{P}_t$ deviates very slightly from $\hat{P}_t$ in expectation—indeed, by no more than $O(\frac{1}{t^2})$. This is crucial for obtaining fast rates: for general Φ (which may be non-differentiable), it is natural to expect $\mathcal{D}\left(\hat{P}_t,\bar{P}_t\right)=\Omega(1/t)$. In this case, the regret would be bounded by $O(\sum_t t\cdot 1/t)=O(T)$, rendering the above lemma useless.

Proof:(of Theorem 14) We have that the divergence terms in Lemma 16 are bounded as

$$t\cdot\mathcal{D}\left(\hat{P}_t,\bar{P}_t\right)\le t\alpha\left\|\frac{1}{t}\mathbf{1}_{Z_t}(\cdot)-\frac{1}{t}p_t(\cdot|Z_1^{t-1})\right\|_1^2\le\frac{4\alpha}{t}$$

because the norm between distributions is bounded by 4:

$$\left(\int_z\left|\mathbf{1}_{Z_t}(z)-p_t(z|Z_1^{T-1})\right|dz\right)^2\le 4.$$

■

5.2 General $\sqrt{T}$ Upper Bounds

We start with the definition of Rademacher averages, one of the central notions of complexity of a function class.

Definition 17 *Denote by*

$$\widehat{Rad}_T(\ell(\mathcal{F})) := \frac{1}{\sqrt{T}}\mathbb{E}_{\epsilon_1^T}\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^T \epsilon_t \ell(f, Z_t)\right|\right)$$

the data-dependent Rademacher averages *of the class $\ell(\mathcal{F})$. Here, $\epsilon_1 \dots \epsilon_T$ are independent Rademacher random variables (uniform on $\{\pm 1\}$).*

We will omit the subscript T and dependence on Z_1^T, for the sake of simplicity. In statistical learning theory, Rademacher averages often provide the tightest guarantees on the performance of empirical risk minimization and other methods. The next result shows that the Rademacher averages play a key role in online convex optimization as well, as the minimax regret is upper bounded by the worst-case (over the sample) Rademacher averages. In the next section, we will also show lower bounds in terms of Rademacher averages for certain linear games, showing that this notion of complexity is fundamental for OCO.

Theorem 18 $\mathscr{R}_T \le 2\sqrt{T}\sup_{Z_1^T\in\mathcal{Z}^T}\widehat{Rad}_T(\ell(\mathcal{F}))$.

Proof: Let $\mathbf{p}$ be an arbitrary joint distribution. Let $\hat{f}$ be an empirical minimizer over Z_1^T, a sequence-dependent function. Then

$$\frac{1}{T}\mathscr{R}_T(\mathbf{p}) = \mathbb{E}\,\frac{1}{T}\sum_{t=1}^T\left[\Phi(p_t\left(\cdot|Z_1^{t-1}\right)) - \Phi(\hat{P}_T)\right]$$

$$\le \mathbb{E}\,\frac{1}{T}\sum_{t=1}^T\left[\mathbb{E}_{p_t\left(\cdot|Z_1^{t-1}\right)}\ell(Z,\hat{f}) - \frac{1}{T}\sum_{s=1}^T\ell(Z_s,\hat{f})\right],$$

as the particular choice of $\hat{f}$ is (sub)optimal. Replacing the $\hat{f}$ by the supremum over $\mathcal{F}$,

$$\begin{aligned}\frac{1}{T}\mathscr{R}_T(\mathbf{p}) &\le \mathbb{E}\,\frac{1}{T}\sum_{t=1}^T\left[\mathbb{E}_{p_t\left(\cdot|Z_1^{t-1}\right)}\ell(Z,\hat{f}) - \ell(Z_t,\hat{f})\right]\\ &\le \mathbb{E}\,\sup_{f\in\mathcal{F}}\frac{1}{T}\sum_{t=1}^T\left[\mathbb{E}_{p_t\left(\cdot|Z_1^{t-1}\right)}\ell(Z,f) - \ell(Z_t,f)\right]\\ &= \mathbb{E}\,\sup_{f\in\mathcal{F}}\frac{1}{T}\sum_{t=1}^T\left[\mathbb{E}_{p_t\left(\cdot|Z_1^{t-1}\right)}\ell(Z'_t,f) - \ell(Z_t,f)\right]\\ &\le \mathbb{E}\,\sup_{f\in\mathcal{F}}\frac{1}{T}\sum_{t=1}^T\left[\ell(Z'_t,f) - \ell(Z_t,f)\right],\end{aligned}$$

where we renamed each dummy variable Z as Z'_t. Even though Z_t and Z'_t have the same conditional expectation, we cannot generally exchange them keeping the distribution of the whole quantity intact. Indeed, the conditional distributions for $\tau > t$ will depend on Z_t and not on Z'_t. The trick is to exchange them one by one[3], starting from $t = T$ and going backwards, introducing an additional supremum. (One can view the sequence $\{Z'_t\}$ as being *tangent* to $\{Z_t\}$ (see [8]).) To this end, for any fixed $\epsilon_T \in \{-1, +1\}$,

$$\begin{aligned}&\mathbb{E}\,\sup_{f\in\mathcal{F}}\frac{1}{T}\sum_{t=1}^T\left[\ell(Z'_t,f) - \ell(Z_t,f)\right]\\ &= \mathbb{E}\,\sup_{f\in\mathcal{F}}\left(\frac{1}{T}\sum_{t=1}^{T-1}\left[\ell(Z'_t,f) - \ell(Z_t,f)\right]\right.\\ &\qquad\left.+\frac{1}{T}\epsilon_T\left(\ell(Z'_T,f) - \ell(Z_T,f)\right)\right)\end{aligned}$$

because for the last step, indeed, Z_T and Z'_T can be exchanged. Since this holds for any ϵ_T, we can take it to be a Rademacher random variable. Thus,

$$\begin{aligned}&\mathbb{E}\,\sup_{f\in\mathcal{F}}\left(\frac{1}{T}\sum_{t=1}^{T-1}\left[\ell(Z'_t,f) - \ell(Z_t,f)\right]\right.\\ &\qquad\left.+\frac{1}{T}\epsilon_T\left(\ell(Z'_T,f) - \ell(Z_T,f)\right)\right)\\ &= \mathbb{E}_{\epsilon_T}\mathbb{E}\,\sup_{f\in\mathcal{F}}\left(\frac{1}{T}\sum_{t=1}^{T-1}\left[\ell(Z'_t,f) - \ell(Z_t,f)\right]\right.\\ &\qquad\left.+\frac{1}{T}\epsilon_T\left(\ell(Z'_T,f) - \ell(Z_T,f)\right)\right)\\ &\le \sup_{Z_T,Z'_T}\mathbb{E}_{Z_1^{T-1}}\mathbb{E}_{\epsilon_T}\sup_{f\in\mathcal{F}}\left(\frac{1}{T}\sum_{t=1}^{T-1}\left[\ell(Z'_t,f) - \ell(Z_t,f)\right]\right.\\ &\qquad\left.+\frac{1}{T}\epsilon_T\left(\ell(Z'_T,f) - \ell(Z_T,f)\right)\right).\end{aligned}$$

Repeating the process, we have that $\frac{1}{T}\mathscr{R}_T(\mathbf{p})$ is bounded by

$$\begin{aligned}&\sup_{Z_1^T,Z_1'^T}\mathbb{E}_{\epsilon_1^T}\sup_{f\in\mathcal{F}}\left(\frac{1}{T}\sum_{t=1}^T\epsilon_t\left(\ell(Z'_t,f) - \ell(Z_t,f)\right)\right)\\ &\le 2\sup_{Z_1^T}\mathbb{E}_{\epsilon_1^T}\sup_{f\in\mathcal{F}}\frac{1}{T}\left|\sum_{t=1}^T\epsilon_t\ell(Z_t,f)\right| = 2\frac{1}{\sqrt{T}}\sup_{Z_1^T}\widehat{\mathrm{Rad}}(\ell(F)).\end{aligned}$$

■

Properties of Rademacher averages are well-known. For instance, the Rademacher averages of a function class coincide with those of its convex hull. Furthermore, if ℓ is Lipschitz, the complexity of $\ell(\mathcal{F})$ can be upper bounded by the complexity of $\mathcal{F}$, multiplied by the Lipschitz constant. For example, we can immediately conclude that if the loss function is Lipschitz and the function class is a convex hull of a finite number M of functions, the minimax value of the game is bounded by $\mathscr{R}_T \le C\sqrt{T\log M}$ for some constant C. Similarly, a class with VC-dimension d would have $\log M$ replaced by d. Theorem 18 is, therefore, giving us the flexibility to upper bound the minimax value of OCO for very general classes of functions.

Finally, we remark that most known upper bounds on Rademacher averages do not depend on the underlying distribution, as they hold for the worst-case empirical measure (see [16], p. 27). Thus, the supremum over the sequences might not be a hinderance to using known bounds for $\widehat{\mathrm{Rad}}(\ell(\mathcal{F}))$.

[3] We thank Ambuj Tewari for pointing out a mistake in our original proof. We refer to [19] for a similar analysis.

5.3 Linear Losses: Primal-Dual Ball Game

Let us examine the linear loss more closely. Of particular interest are linear games when $\mathcal{F} = B_{\|\cdot\|_*}$ is a ball in some norm $\|\cdot\|_*$ and $\mathcal{Z} = B_{\|\cdot\|}$, the two norms being dual. For this case, Theorem 18 gives an upper bound of

$$\begin{aligned}\frac{1}{T}\mathscr{R}_T &\le 2\sup_{Z_1^T}\mathbb{E}_{\epsilon_1^T}\sup_{f\in\mathcal{F}} f^\top\left(\frac{1}{T}\sum_{t=1}^T \epsilon_t Z_t\right)\\ &= 2\sup_{Z_1^T}\mathbb{E}_{\epsilon_1^T}\left\|\frac{1}{T}\sum_{t=1}^T \epsilon_t Z_t\right\|. \end{aligned} \tag{13}$$

Fix $Z_1 \dots Z_{T-1}$ and observe that the expected norm is a convex function of Z_T. Hence, the supremum over Z_T is achieved at the boundary of $\mathcal{Z}$. The same statement holds for all Z_t's. Let $z_1^*, \dots, z_T^*$ be the sequence achieving the supremum. Now take a distribution for round t to be $p_t^*(z) = \frac{1}{2}\left(\mathbf{1}_{z_t^*}(\cdot) + \mathbf{1}_{-z_t^*}(\cdot)\right)$ and let $\mathbf{p}^* = p_1^* \times \dots \times p_T^*$ be the product distribution. It is easy to see that

$$\begin{aligned}\frac{1}{T}\mathscr{R}_T &\le 2\sup_{Z_1^T}\mathbb{E}_{\epsilon_1^T}\left\|\frac{1}{T}\sum_{t=1}^T \epsilon_t Z_t\right\| = 2\mathbb{E}_{\epsilon_1^T}\left\|\frac{1}{T}\sum_{t=1}^T \epsilon_t z_t^*\right\|\\ &= 2\mathbb{E}_{\mathbf{p}^*}\mathbb{E}_{\epsilon_1^T}\left\|\frac{1}{T}\sum_{t=1}^T \epsilon_t Z_t\right\| = \frac{2}{\sqrt{T}}\mathbb{E}_{\mathbf{p}^*}\widehat{\mathrm{Rad}}(\mathcal{F}). \end{aligned} \tag{14}$$

Also note that p_t^* has zero mean. It will be shown in Section 7.1 that the lower bound arising from this distribution is

$$\mathbb{E}_{\mathbf{p}^*}\mathbb{E}_{\epsilon_1^T}\left\|\frac{1}{T}\sum_{t=1}^T \epsilon_t Z_t\right\| = \frac{1}{\sqrt{T}}\mathbb{E}_{\mathbf{p}^*}\widehat{\mathrm{Rad}}(\mathcal{F}),$$

which is only a factor of 2 away. Thus, the adversary can play a product distribution that arises from the maximization in (13) and achieve regret at most a factor 2 from the optimum.

6 $\Omega(\sqrt{T})$ bounds for non-differentiable Φ

In this section, we develop lower bounds on the minimax value $\mathscr{R}_T$ based on the geometric view-point described in Section 4.1.

Theorem 14 shows that the regret is upper bounded by $\log T$ for the case of strongly convex losses, and this upper bound is tight if the loss functions are quadratic, as we show later in the paper. Thus, flatness of Φ implies low regret. What about the converse? It turns out that if Φ is non-differentiable (has a point of infinite curvature), the regret is lower-bounded by $\sqrt{T}$, and this rate is achieved with $\mathbf{p} = p^T$, where p corresponds to a point of non-differentiability of Φ.

The geometric viewpoint is fruitful here: vertices (points of non-differentiability) of Φ correspond to *exposed faces* in the loss class $S = \mathrm{co}[-\ell(\mathcal{F})]$, suggesting that the lower bounds of $\Omega(\sqrt{T})$ arise from having two distinct minimizers of expected error—a striking parallel to the analogous results for stochastic settings [15, 17].

To be more precise, vertices of σ_S (and Φ) translate into flat parts (non-singleton *exposed faces*) of $S(\mathrm{co}[-\ell(\mathcal{F})])$ and the other way around. Corresponding to an exposed face is a supporting hyperplane. If $\ell(\mathcal{F})$ is non-negative, then any exposed face facing the origin is supported by a hyperplane with positive co-ordinates (which can be normalized to get a distribution). So a non-singleton face exposed to the origin is equivalent to having at least two distinct minimizers f and g of $\mathbb{E}_p\ell(Z,\cdot)$ for some p, as discussed in Section 4.2.

Define the set of expected minimizers under p as

$$\mathcal{F}^* := \{f \in \mathcal{F} : \mathbb{E}_p\ell(Z,f) = \inf_{f\in\mathcal{F}}\mathbb{E}_p\ell(Z,f)\}.$$

Thus, $-\ell(\mathcal{F}^*) \subseteq F_S(p) \subseteq \mathrm{co}[-\ell(\mathcal{F})]$. The lower bound we are about to state arises from fluctuations of the empirical process over the set $\mathcal{F}^*$. To ease the presentation, we will refer to the sample average $\frac{1}{T}\sum_{t=1}^T \ell(Z_t,f)$ as $\hat{\mathbb{E}}\,\ell(Z,f)$.

Theorem 19 *Suppose $F_S(p)$ is a non-singleton face of $\mathrm{co}[-\ell(\mathcal{F})]$, supported by p (i.e. $|\mathcal{F}^*| > 1$). Fix any $f^* \in \mathcal{F}^*$ and let $Q \subseteq \ell(\mathcal{F}^*)$ be any subset containing $\ell(\cdot, f^*)$. Define $\bar{Q} = \{g - \ell(\cdot,f^*) : g \in Q\}$, the shifted loss class. Then for $T > T_0(\mathcal{F})$,*

$$\begin{aligned}\frac{1}{T}\mathscr{R}_T &\ge \frac{1}{T}\mathscr{R}_T(p^T)\\ &= \mathbb{E}\sup_{f\in\mathcal{F}^*}\left[\mathbb{E}_p\ell(Z,f) - \hat{\mathbb{E}}\,\ell(Z,f)\right]\\ &\ge \frac{c}{\sqrt{T}}\sup_{Q\subseteq\ell(\mathcal{F}^*)}\mathbb{E}\sup_{q\in\bar{Q}} G_q,\end{aligned}$$

where G_q is the Gaussian process indexed by the (centered) functions in $\bar{Q}$, and c is some absolute constant.

Proof: Recalling that $\mathbb{E}\,\ell(Z,f) = \inf_{g\in\mathcal{F}}\mathbb{E}\,\ell(Z,g) = \Phi(p)$ for all $f \in \mathcal{F}^*$, we have

$$\begin{aligned}\frac{1}{T}\mathscr{R}_T &\ge \frac{1}{T}\mathscr{R}_T(p^T) = \Phi(p) - \mathbb{E}_p\Phi(\hat{P}_T)\\ &= \Phi(p) - \mathbb{E}\inf_{f\in\mathcal{F}}\frac{1}{T}\sum_{t=1}^T \ell(Z_t,f)\\ &\ge \Phi(p) - \mathbb{E}\inf_{f\in\mathcal{F}^*}\hat{\mathbb{E}}\,\ell(Z,f)\\ &= \mathbb{E}\sup_{f\in\mathcal{F}^*}\left[\mathbb{E}_p\ell(Z,f) - \hat{\mathbb{E}}\,\ell(Z,f)\right]\\ &\ge \sup_{Q\subseteq\ell(\mathcal{F}^*)}\mathbb{E}\sup_{f:\ell_f\in Q}\left[\mathbb{E}_p\ell(Z,f) - \hat{\mathbb{E}}\,\ell(Z,f)\right]\end{aligned}$$

Now, fix any $f^* \in \mathcal{F}^*$. The proof of Theorem 2.2 in [14] reveals that empirical fluctuations are lower bounded by the supremum of the Gaussian process indexed by $\bar{Q}$. To be precise, there exists $T_0(\mathcal{F})$ such that for $T > T_0(\mathcal{F})$ with probability greater than c_1,

$$\inf_{f:\ell_f\in Q}\hat{\mathbb{E}}\left(\ell(Z,f) - \ell(Z,f^*)\right) \le -c_2\frac{\mathbb{E}\sup_{q\in\bar{Q}} G_q}{\sqrt{T}},$$

for some absolute constants c_1, c_2. Rearranging and using the fact that $\mathbb{E}\,\ell(Z,f) - \mathbb{E}\,\ell(Z,f^*) = 0$ for $f \in \mathcal{F}^*$,

$$\begin{aligned}&\sup_{f:\ell_f\in Q}\left[\mathbb{E}\,\ell(Z,f) - \hat{\mathbb{E}}\,\ell(Z,f) + \hat{\mathbb{E}}\,\ell(Z,f^*) - \mathbb{E}\,\ell(Z,f^*)\right]\\ &\qquad\ge c_2\frac{\mathbb{E}\sup_{q\in\bar{Q}} G_q}{\sqrt{T}}\end{aligned}$$

with probability at least c_1. The supremum is non-negative because $f^* \in Q$ and therefore

$$\mathbb{E} \sup_{f:\ell_f \in Q} \left[\mathbb{E}\, \ell(Z,f) - \hat{\mathbb{E}}\, \ell(Z,f) \right] \geq c_1 c_2 \frac{\mathbb{E} \sup_{q \in \bar{Q}} G_q}{\sqrt{T}}.$$

■

We remark that in the experts case, the lower bound on regret becomes $\sqrt{T \log N}$, as the Gaussian process reduces to N independent Gaussian random variables. We discuss this and other examples in the next section.

7 Lower Bounds for Special Cases

We now provide lower bounds for particular games. Some of the results of the section are known: we show how the proofs follow from the general lower bounds developed in the previous section.

7.1 Linear Loss: Primal-Dual Ball Game

Here, we develop lower bounds for the case considered in Section 5.3. As before, to prove a lower bound it is enough to take an i.i.d. or product distribution. In particular, the product distribution described after Eq. (13) is of particular interest. To this end, choose $\mathbf{p} = p_1 \times \ldots \times p_T$ to be a product of *symmetric* distributions on the surface of the primal ball $\mathcal{Z}$ with $\mathbb{E}_{p_t} Z = 0$. We conclude that $\Phi(p_t) = 0$ and $\frac{1}{T}\mathscr{R}_T$ is greater than

$$-\mathbb{E}\, \Phi(\hat{P}_T) = -\mathbb{E} \inf_{f \in \mathcal{F}} f \cdot \left(\frac{1}{T} \sum_{t=1}^T Z_t \right) = \mathbb{E} \left\| -\frac{1}{T} \sum_{t=1}^T Z_t \right\|$$

by the definition of dual norm. Now, because of symmetry,

$$\mathbb{E} \left\| -\frac{1}{T} \sum_{t=1}^T Z_t \right\| = \frac{1}{T} \mathbb{E}\, \mathbb{E}_\epsilon \left\| \sum_{t=1}^T \epsilon_t Z_t \right\|. \qquad (15)$$

We conclude that $\mathscr{R}_T \geq \sqrt{T}\mathbb{E}\, \widehat{\mathrm{Rad}}(\mathcal{F})$, the expected Rademacher averages of the dual ball acting on the primal ball. This is within a factor of 2 of the upper bound (14) of Section 5.3.

Now, consider the particular case of $\mathcal{F} = \mathcal{Z} = B_2$, the Euclidean ball. We will consider three distributions $\mathbf{p}$.

- Suppose $\mathbf{p}$ is such that $p_t\left(\cdot | Z_1^{t-1}\right)$ puts mass on the intersection of B_2 and the subspace perpendicular to $\sum_{s=1}^{t-1} Z_s$ and $\mathbb{E}\left[Z | Z_1^{t-1}\right] = 0$. Then $\mathbb{E} \left\| \sum_{t=1}^T Z_t \right\| = \sqrt{T}$ by unraveling the sum from the end. In fact, this is shown to be the optimal value for this problem in [2]. We conclude that a non-product distribution achieves the optimal regret for this problem.

- Consider any symmetric i.i.d. distribution on the surface of the ball $\mathcal{Z}$. Note that for this case we still have the lower bound of Eq. (15). Kinchine-Kahane inequality then implies $\mathscr{R}_T \geq \sqrt{\frac{T}{2}}$ and the constant $\sqrt{2}$ is optimal (see [13]) in the absence of further assumptions.

- Consider another example of an i.i.d. distribution that puts equal mass on two points $\pm z_0$ on each round, with $\|z_0\| = 1$. It then follows that this i.i.d. distribution achieves the regret equal to the length of the random walk $\mathbb{E} \left| \sum_{t=1}^T \epsilon_t \right|$, which is known to be *asymptotically* $\sqrt{2T/\pi}$.

We conclude that for the Euclidean game, the best strategy of the adversary is a sequence of dependent distributions, while product and i.i.d. distributions come within a multiplicative constant close to 1 from it.

7.2 Experts Setting

The experts setting provides some of the easiest examples for linear games. We start with a simplified game, where $\mathcal{F} = \mathcal{Z} = \Delta_N$, the N-simplex. The Φ function for this case is easy to visualize. We then present the usual game, where the set $\mathcal{Z} = [0,1]^N$. In both cases, we are interested in lower-bounding regret.

7.2.1 The simplified game

Let us look at the game when only one expert can suffer a loss of 1 per round, i.e. the space of actions $\mathcal{Z}$ contains N elements $e_1, \ldots, e_N$. The probability over these choices of the adversary is an N-dimensional simplex, just as the space of functions $\mathcal{F}$. For any $p \in \Delta_N$,

$$\Phi(p) = \min_{f \in \Delta_N} \mathbb{E}_p Z \cdot f = \min_f p \cdot f = \min_{i \in [N]} p_i$$

and therefore the Φ has the shape of a pyramid with its maximum at $p^* = \frac{1}{N}\mathbf{1}$ and $\Phi(p^*) = 1/N$. The regret is lower-bounded by an i.i.d game with this distribution p^* at each round, i.e.

$$\begin{aligned} \mathscr{R}_T \geq \Phi(p^*) - \mathbb{E}\, \Phi(U) &= \frac{1}{N} - \mathbb{E} \min_{f \in \Delta_N} \left(\frac{1}{T} \sum Z_t \right) f \\ &= \mathbb{E} \max_{i \in [N]} \left[\frac{1}{N} - \frac{n_i}{T} \right], \end{aligned}$$

where n_i is the number of times e_i has been chosen out of T rounds. This is the expected maximum deviation from the mean of a multinomial distribution, i.e. $1/N$ minus the smallest proportion of balls in any bin after T balls have been distributed uniformly at random.

To obtain the lower bound on the maximum deviation, let us turn to Section 6. The convex hull of the (negative) loss class $\mathrm{co}[-\ell(\mathcal{F})]$ is the simplex itself. This is also the face supported by the uniform distribution p^*. The lower bound of Theorem 19 involves the Gaussian process indexed by a set Q. Let us take $f^* = \frac{1}{N}1$ and $\mathcal{F}^* = \{e_1, \ldots, e_N\} \cup \{f^*\}$. We can verify that $\mathbb{E}\, e_i^\intercal Z = \Phi(p^*) = \frac{1}{N}$, the covariance of the process indexed by $Q = \ell(\mathcal{F}^*)$ is $\mathbb{E}\,(e_i^\intercal Z - \frac{1}{N})(e_j^\intercal Z - \frac{1}{N}) = -\frac{1}{N^2}$ for $i \neq j$ and the variance is $\mathbb{E}\,(e_i^\intercal Z - \frac{1}{N})^2 = \frac{N-1}{N^2}$. Let $\{Y_i\}_1^N$ be the Gaussian random variables with the aforementioned covariance structure. Then $\|Y_i - Y_j\|^2 = \mathbb{E}\,(Y_i - Y_j)^2 = \frac{2}{N}$. We can now construct independent Gaussian random variables $\{X_i\}_1^N$ with the same distance by putting $\frac{2}{N}$ on the diagonal of the covariance matrix. By

Slepian's Lemma, $\frac{1}{2}\mathbb{E}\sup_i X_i \leq \mathbb{E}\sup_i Y_i$, thus giving us the lower bound

$$\mathscr{R}_T \geq c\sqrt{\frac{T\log N}{N}}$$

for this problem, for some absolute constant c and T large enough.

7.2.2 The general case

In the more general game, any expert can suffer a $0/1$ loss. Thus, p is a distribution on 2^N losses Z. To lower bound the regret, choose a uniform distribution on 2^N binary vectors as the i.i.d. choice for the adversary. We have $\Phi\left(\frac{1}{2^N}\mathbf{1}\right) = \min_{f\in\Delta_N} f\cdot\mathbb{E}\,Z = 1/2$. As for the other term, $\mathbb{E}\,\Phi(\hat{P}_T) = \mathbb{E}\,\min_{f\in\Delta_N} f\cdot\left(\frac{1}{T}\sum Z_t\right)$. Thus, the regret is

$$\frac{1}{T}\mathscr{R}_T \geq \mathbb{E}\max_{i\in[N]}\left[\frac{1}{2} - \frac{\sum_{t=1}^T \epsilon_{i,t}}{T}\right],$$

where $\epsilon_{i,t}$ are Rademacher $\{\pm 1\}$-valued random variables. It is easy to show that the expected maximum is lower bounded by $c\sqrt{\log N/T}$. This coincides with a result in [7], which shows that the asymptotic behavior is $\sqrt{\log N/(2T)}$.

7.3 Quadratic Loss

We consider the quadratic loss, $\ell(z,f) = \|f-z\|^2$. This loss function is 1-strongly convex, and therefore we already have the $O(\log T)$ bound of Corollary 15. In this section, we present an *almost* matching lower bound using a particular adversarial strategy. The problem of quadratic loss was previously addressed in [20]; we reprove their lower bound in our framework, borrowing a number of tricks from that work.

Following Section 6, it is tempting to use an i.i.d. distribution and compute the regret explicitly. Unfortunately, this only leads to a constant lower bound, whereas we would hope to match the upper bound of $\log T$. We can show this easily: let $\mathbf{p} := p^T$ be some i.i.d. distribution, then

$$\begin{aligned} T\mathbb{E}\,\Phi(\hat{P}_T) &= (T-1)\mathbb{E}\,\|Z_1\|^2 - \frac{T(T-1)}{T}\mathbb{E}\,\langle Z_1, Z_2\rangle \\ &= (T-1)\left(\mathbb{E}\,\|Z\|^2 - (\mathbb{E}\,Z)^2\right) = (T-1)\mathrm{var}(Z) \\ &= (T-1)\Phi(p). \end{aligned}$$

Thus $\mathscr{R}(p^T) = T\Phi(p) - T\mathbb{E}\,\Phi(\hat{P}_T) = \Phi(p)$, where we see that the last term is independent of T.

Indeed, obtaining $\log T$ regret requires that we look further than i.i.d. To this end, define $c_T := \frac{1}{T}$ and $c_{t-1} := c_t + c_t^2$ for $t = T, T-1, \ldots, 2$. We construct our distribution $\mathbf{p}$ using this sequence as follows. Assume $\mathcal{Z} = \mathcal{F} = [-1,1]$ and for convenience let $Z_{1:s} := \sum_{t=1}^s Z_t$. Also, for this section, we use a shorthand for the conditional expectation, $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot|Z_1,\ldots,Z_{t-1}]$. Each conditional distribution is chosen as

$$p_t(Z_t = z|Z_1,\ldots,Z_{t-1}) := \begin{cases} \frac{1+c_t Z_{1:t-1}}{2}, & \text{for } z = 1 \\ \frac{1-c_t Z_{1:t-1}}{2}, & \text{for } z = -1 \end{cases}.$$

Notice that this choice ensures that $\mathbb{E}_t Z_t = c_t Z_{1:t-1}$, i.e. the conditional expectation is identical to the observed sample mean scaled by some *shrinkage factor* c_t. That $\frac{1+c_t Z_{1:t-1}}{2} \in [0,1]$ follows from the statement $c_t \leq \frac{1}{t}$ which is proven by an easy induction. We now recall a result shown in [20]:

$$\sum_{t=1}^T c_t = \log T - \log\log T + o(1).$$

This crucial lemma leads directly to the main result of this section.

Theorem 20 *With* $\mathbf{p}$ *defined above,* $\mathscr{R}_T(\mathbf{p}) = \sum_{t=1}^T c_t$ *and therefore*

$$\mathscr{R}_T(\mathbf{p}) = \log T - \log\log T + o(1).$$

The proof follows closely along the lines of [20] and can be found in the full version [1].

Acknowledgements We gratefully acknowledge the support of the NSF under awards DMS-0707060 and DMS-0830410 and of DARPA under award FA8750-05-2-0249. Jacob is supported in part by a Yahoo! PhD Fellowship.

References

[1] J. Abernethy, A. Agarwal, P. L. Bartlett, and A. Rakhlin. A stochastic view of optimal regret through minimax duality. *CoRR*, abs/0903.5328, 2009. http://arxiv.org/abs/0903.5328.

[2] J. Abernethy, P. L. Bartlett, A. Rakhlin, and A. Tewari. Optimal strategies and minimax lower bounds for online convex games. In *COLT*, 2008.

[3] P. L. Bartlett, E. Hazan, and A. Rakhlin. Adaptive online gradient descent. In *NIPS*, 2007.

[4] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[5] J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization*. Advanced Books in Mathematics. Canadian Mathematical Society, to appear.

[6] N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *Ann. Stat.*, 27(6):1865–1895, 1999.

[7] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambr. Univ. Press, 2006.

[8] V. de la Peña and E. Gine. *Decoupling: From Dependence to Independence*. Springer, 1998.

[9] Y. Freund. Predicting a binary sequence almost as well as the optimal biased coin. *Inf. Comput.*, 182(2):73–94, 2003.

[10] B. A. Frigyik, S. Srivastava, and M. R. Gupta. Functional bregman divergence and bayesian estimation of distributions. *CoRR*, abs/cs/0611123, 2006.

[11] E. Hazan, A. Kalai, S. Kale, and A. Agarwal. Logarithmic regret algorithms for online convex optimization. In *COLT*, pages 499–513, 2006.

[12] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2001.

[13] R. Latala and K. Oleszkiewicz. On the best constant in the Khinchin-Kahane inequality. *Studia Math.*, 109(1):101–104, 1994.

[14] G. Lecué and S. Mendelson. Sharper lower bounds on the performance of the empirical risk minimization algorithm. Available at http://www.cmi.univ-mrs.fr/ lecue/LM2.pdf.

[15] W. S. Lee, P. L. Bartlett, and R. C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974–1980, 1998.

[16] S. Mendelson. A few notes on statistical learning theory. In S. Mendelson and A. J. Smola, editors, *Advanced Lectures in Machine Learning, LNCS 2600, Machine Learning Summer School 2002, Canberra, Australia, February 11-22*, pages 1–40. Springer, 2003.

[17] S. Mendelson. Lower bounds for the empirical minimization algorithm. *IEEE Transactions on Information Theory*, 2008. To appear.

[18] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44:2124–2147, 1998.

[19] K. Sridharan and A. Tewari. Convex games in banach spaces (working title), 2009. Unpublished.

[20] E. Takimoto and M. Warmuth. The minimax strategy for gaussian density estimation. In *COLT*, pages 100–106. Morgan Kaufmann, San Francisco, 2000.

[21] V. Vovk. Competitive on-line linear regression. In *NIPS '97: Proceedings of the 1997 conference on Advances in neural information processing systems 10*, pages 364–370, Cambridge, MA, USA, 1998. MIT Press.

[22] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003.