1 Perturbation Techniques in Online Learning and Optimization

Jacob Abernethy jabernet@umich.edu
University of Michigan
Ann Arbor, MI

Chansoo Lee chansool@umich.edu
University of Michigan
Ann Arbor, MI

Ambuj Tewari tewaria@umich.edu
University of Michigan
Ann Arbor, MI

In this chapter we give a new perspective on so-called perturbation methods *that have been applied in a number of different fields, but in particular for adversarial online learning problems. We show that the classical algorithm known as Follow The Perturbed Leader (FTPL) can be viewed through the lens of* stochastic smoothing*, a tool that has proven popular within convex optimization. We prove bounds on regret for several online learning settings, and provide generic tools for analyzing perturbation algorithms. We also consider the so-called* bandit setting*, where the feedback to the learner is significantly constrained, and we show that near-optimal bounds can be achieved as long as a simple condition on the perturbation distribution is met.*

1.1 Introduction

In this chapter we will study the problem of *online learning* with the goal of *minimizing regret*. A learner must iteratively play a sequence of actions, where each action is based on the data received up to the previous iteration.

We consider learning in a potentially adversarial environment, where we avoid making any stochastic assumptions about the sequence of data. The goal of the learner is to suffer as little regret as possible, where regret is defined as the difference between the learner's loss and the loss of the best fixed action in hindsight. The key to developing optimal algorithms is *regularization*, which may be interpreted either as *hedging* against bad future events, or similarly can be seen as avoiding *overfitting* to the observed data. In this paper, we focus on regularization techniques for online linear optimization problems where the learner's action is evaluated on a linear reward function.

In the present chapter, we will mostly focus on learning settings where our learner's decisions are chosen from a convex subset of $\mathbb{R}^N$, and where the "data" we observe arrives in the form of a (bounded) vector $g \in \mathbb{R}^N$, and the costs/gains will be linear in each. Specifically, the gain (equiv., reward) received on a given round, when the learner plays action w and Nature chooses vector g, is the inner product $\langle w, g \rangle$. Generally we will use the the symbol G to refer to the *cumulative gain vector* up to a particular time period.

The algorithm commonly known as Follow the Regularized Leader (FTRL) selects an action w on a given round by solving an explicit optimization problem, where the objective combines a "data fitness" term along with a regularization *via penalty function*. More precisely, FTRL selects an action by optimizing $\operatorname{argmax}_w \langle w, G \rangle - \mathcal{R}(w)$ where $\mathcal{R}$ is a strongly convex penalty function; a well-studied choice for $\mathcal{R}$ is the well-known ℓ_2-regularizer $\|\cdot\|_2^2$. The regret analysis of FTRL reduces to the analysis of the second-order behavior of the penalty function (Shalev-Shwartz, 2012), which is well-studied due to the powerful convex analysis tools. In fact, regularization via penalty methods for online learning in general are very well understood. Srebro et al. (2011) proved that Mirror Descent, a regularization via penalty method, achieves a nearly optimal regret guarantee for a general class of online learning problems, and McMahan (2011) showed that FTRL is equivalent to Mirror Descent under some assumptions.

Follow the Perturbed Leader (FTPL), on the other hand, uses implicit regularization *via perturbations*. At every iteration, FTPL selects an action by optimizing $\operatorname{argmax}_w \langle w, G + z \rangle$ where G is the observed data and z is some random noise vector, often referred to as a "perturbation" of the input. The early FTPL analysis tools lacked a generic framework and relied substantially on clever algebra tricks and heavy probabilistic analysis (Kalai and Vempala, 2005; Devroye et al., 2013; van Erven et al., 2014). This was in contrast to the elegant and simple convex analysis techniques that provided the basis for studying FTRL and proving tight bounds.

This book chapter focuses on giving a new perspective on perturbation methods and on providing a new set of analysis tools for controlling the

regret of FTPL. In particular, we show that the results hinge on certain *second-order properties* of stochastically-smoothed convex functions. Indeed, we show that both FTPL and FTRL naturally arise as *smoothing operations* of a non-smooth potential function and the regret analysis boils down to understanding the *smoothness* as defined in Section 1.3. This new unified analysis framework recovers known (near-)optimal regret bounds and provides tools for controlling regret.

An interesting feature of our analysis framework is that we can directly apply existing techniques from the optimization literature, and conversely, our new findings in online linear optimization may apply to optimization theory. In Section 1.4, a straightforward application of the results on Gaussian smoothing by Nesterov and Spokoiny (2011) and Duchi et al. (2012) gives a generic regret bound for an arbitrary online linear optimization problem. In Section 1.5 and 1.6, we improve this bound for the special cases that correspond to canonical online linear optimization problems; we analyze the so-called "experts setting" (Section 1.5) and we also look at the case where the decision set is the Euclidean ball (Section 1.6). Finally, in Section 1.7, we turn our attention to the *bandit setting* where the learner has limited feedback. For this case, we show that the perturbation distribution has to be chosen quite carefully, and indeed we show that near-optimal regret can be obtained as long as the perturbation distribution has a *bounded hazard rate* function.

1.2 Preliminaries

1.2.1 Convex Analysis

For this preliminary discussion, assume we are given an arbitrary norm $\|\cdot\|$. Throughout the chapter we will utilize various norms, such as the $\ell_1, \ell_2, \ell_\infty$, and the spectral norm of a matrix. In addition, we will often use $\|\cdot\|_*$ to refer to the *dual norm* of $\|\cdot\|$, defined as $\|\mathbf{z}\|_* = \max_{\mathbf{y}:\|\mathbf{y}\|\le 1}\langle \mathbf{y}, \mathbf{z}\rangle$.

Assume we are given f a differentiable, closed, and proper convex function whose domain is $\operatorname{dom} f \subseteq \mathbb{R}^N$. We say that f is *L-Lipschitz* with respect to a norm $\|\cdot\|$ when f satisfies $|f(x) - f(y)| \le L\|x - y\|$ for all $x, y \in \operatorname{dom}(f)$.

The *Bregman divergence* $D_f(y, x)$ is the gap between $f(y)$ and the linear approximation of $f(y)$ around x. Formally, $D_f(y, x) = f(y) - f(x) - \langle \nabla f(x), y - x\rangle$. We say that f is *β-strongly convex* with respect to a norm $\|\cdot\|$ if we have $D_f(y, x) \ge \frac{\beta}{2}\|y - x\|^2$ for all $x, y \in \operatorname{dom} f$. Similarly, f is said to be *β-strongly smooth* with respect to a norm $\|\cdot\|$ if we have $D_f(y, x) \le \frac{\beta}{2}\|y - x\|^2$ for all $x, y \in \operatorname{dom} f$.

The Bregman divergence measures how fast the gradient changes, or equivalently, how large the second derivative is. In fact, we can bound the Bregman divergence by analyzing the local behavior of Hessian, as the following adaptation of Abernethy et al. (2013, Lemma 4.6) shows.

Lemma 1.1. *Let f be a twice-differentiable convex function with* $\operatorname{dom} f \subseteq \mathbb{R}^N$. *Assume that the eigenvalues of* $\nabla^2 f(x)$ *all lie in the range* $[a, b]$ *for every* $x \in \operatorname{dom} f$. *Then,* $a\|v\|^2/2 \le D_f(x+v, x) \le b\|v\|^2/2$ *for any* $x, x+v \in \operatorname{dom} f$.

The *Fenchel conjugate* of f is defined as $f^\star(G) = \sup_{w \in \operatorname{dom}(f)}\{\langle w, G\rangle - f(w)\}$, and it is a dual mapping that satisfies $f = (f^\star)^\star$. If f is differentiable and strictly convex we also have $\nabla f^\star \in \operatorname{dom}(f)$. One can also show that the notions of strong convexity and strong smoothness are dual to each other. That is, f is β-strongly convex with respect to a norm $\|\cdot\|$ if and only if $f^\star$ is $\frac{1}{\beta}$-strongly smooth with respect to the dual norm $\|\cdot\|_\star$. For more details and proofs, readers are referred to an excellent survey by Shalev-Shwartz (2012).

1.2.2 Online Linear Optimization

Let $\mathcal{X}$ and $\mathcal{Y}$ be convex and closed subsets of $\mathbb{R}^N$. The online linear optimization (OLO) is defined to be the following repeated game between two entities that we call the *learner* and the *adversary*:

On round $t = 1, \dots, T$,

- the learner plays $w_t \in \mathcal{X}$;
- the adversary reveals $g_t \in \mathcal{Y}$;
- the learner receives a reward[1] $\langle w_t, g_t\rangle$.

We say $\mathcal{X}$ is the *decision set* and $\mathcal{Y}$ is the *reward set*. Let $G_t = \sum_{s=1}^t g_s$ be the cumulative reward. The learner's goal is to minimize the (external) regret, defined as:

$$\text{Regret} = \underbrace{\max_{w \in \mathcal{X}} \langle w, G_T\rangle}_{\text{baseline potential}} - \sum_{t=1}^T \langle w_t, g_t\rangle. \tag{1.1}$$

The *baseline potential function* $\Phi(G) := \max_{w \in \mathcal{X}} \langle w, G\rangle$ is the comparator term against which we define the regret, and it coincides with the *support function* of $\mathcal{X}$. For a bounded compact set $\mathcal{X}$, the support function of $\mathcal{X}$ is

1. Our somewhat less conventional choice of maximizing the reward instead of minimizing the loss was made so that we directly analyze the convex function $\max(\cdot)$ without cumbersome sign changes.

positively homogeneous, subadditive, and Lipschitz continuous with respect to any norm $\|\cdot\|$, where the Lipschitz constant is equal to $\sup_{x\in\mathcal{X}} \|x\|_*$. For more details and proofs, readers are referred to Rockafellar (1997, Section 13) or Molchanov (2005, Appendix F).

1.3 Gradient-Based Prediction Algorithm

Follow the Leader (FTL) style algorithms select the next action $w_t \in \mathcal{X}$ via an optimization problem: given the cumulative reward vector G_{t-1}, an FTL style algorithm selects $w_t = \operatorname{argmax}_{w\in\mathcal{X}} f(w, G_{t-1})$. The most simple algorithm, FTL, does not incorporate any perturbation or regularization into the optimization, and uses the objective $f(w, G) = \langle w, G\rangle$. Unfortunately FTL does not enjoy non-trivial regret guarantees in many scenarios, due to the inherent instability of vanilla linear optimization—that is, since the the optimal solution can fluctuate with small changes in the input. There are a couple of ways to induce stability in FTL. Follow the Regularized Leader (FTRL) sets $f(w, G) = \langle w, G\rangle - \mathcal{R}(w)$ where $\mathcal{R}$ is a strongly convex regularizer providing stability to the solution. Follow the Perturbed Leader (FTPL) sets $f(w, G) = \langle w, G + z\rangle$ where z is a random vector. The randomness in z imparts stability to the (expected) move of the FTPL algorithm.

We now proceed to show that a common property shared by all such algorithms is that the action w_t is exactly the *gradient* of some scalar-valued potential function $\widetilde{\Phi}_t$ evaluated at G_{t-1}. (For the remainder of the paper we will use the notation $\widetilde{\Phi}$ to refer to a modification of the baseline potential Φ). This perspective gives rise to what we call the Gradient-based Prediction Algorithm (GBPA), presented in Algorithm 1. In the following Section we give a full regret analysis of this algorithm. We note that Cesa-Bianchi and Lugosi (2006, Theorem 11.6) presented a similar algorithm, but our formulation eliminates all dual mappings.

Algorithm 1: Gradient-Based Prediction Algorithm (GBPA)

Input: $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^N$
Require: convex potentials $\widetilde{\Phi}_1, \ldots, \widetilde{\Phi}_T : \mathbb{R}^N \to \mathbb{R}$, with $\nabla\widetilde{\Phi}_t(G) \in \mathcal{X}$, $\forall G$
Initialize: $G_0 = 0$
for *t = 1 to T* **do**
  The learner plays $w_t = \nabla\widetilde{\Phi}_t(G_{t-1})$
  The adversary reveals $g_t \in \mathcal{Y}$
  The learner receives a reward of $\langle w_t, g_t\rangle$
  Update the cumulative gain vector: $G_t = G_{t-1} + g_t$

1.3.1 GBPA Analysis

We begin with a generic result on the regret of GBPA in the full-information setting.

Lemma 1.2 (GBPA Regret)**.** *Let Φ be the baseline potential function for an online linear optimization problem. The regret of the GBPA can be decomposed as follows:*

$$\begin{aligned}\text{Regret} = \sum_{t=1}^{T} \Bigg(\underbrace{\left(\widetilde{\Phi}_t(G_{t-1}) - \widetilde{\Phi}_{t-1}(G_{t-1})\right)}_{\textit{overestimation penalty}} + \underbrace{D_{\widetilde{\Phi}_t}(G_t, G_{t-1})}_{\textit{divergence penalty}} \Bigg) \\ + \underbrace{\Phi(G_T) - \widetilde{\Phi}_T(G_T)}_{\textit{underestimation penalty}}, \end{aligned} \tag{1.2}$$

where $\widetilde{\Phi}_0 \equiv \Phi$.

Proof. We note that since $\widetilde{\Phi}_0(0) = 0$,

$$\begin{aligned}\widetilde{\Phi}_T(G_T) &= \textstyle\sum_{t=1}^{T} \widetilde{\Phi}_t(G_t) - \widetilde{\Phi}_{t-1}(G_{t-1}) \\ &= \textstyle\sum_{t=1}^{T} \Big(\left(\widetilde{\Phi}_t(G_t) - \widetilde{\Phi}_t(G_{t-1})\right) + \left(\widetilde{\Phi}_t(G_{t-1}) - \widetilde{\Phi}_{t-1}(G_{t-1})\right) \Big) \\ &= \textstyle\sum_{t=1}^{T} \Big(\left(\langle \nabla \widetilde{\Phi}_t(G_{t-1}), g_t \rangle + D_{\widetilde{\Phi}_t}(G_t, G_{t-1})\right) \\ &\qquad + \left(\widetilde{\Phi}_t(G_{t-1}) - \widetilde{\Phi}_{t-1}(G_{t-1})\right) \Big), \end{aligned}$$

where the last equality holds because:

$$\widetilde{\Phi}_t(G_t) - \widetilde{\Phi}_t(G_{t-1}) = \langle \nabla \widetilde{\Phi}_t(G_{t-1}), g_t \rangle + D_{\widetilde{\Phi}_t}(G_t, G_{t-1}).$$

We now have

$$\begin{aligned}\text{Regret} &:= \Phi(G_T) - \sum_{t=1}^{T} \langle w_t, g_t \rangle \\ &= \Phi(G_T) - \textstyle\sum_{t=1}^{T} \langle \nabla \widetilde{\Phi}_t(G_{t-1}), g_t \rangle \\ &= \Phi(G_T) - \widetilde{\Phi}_T(G_T) + \textstyle\sum_{t=1}^{T} D_{\widetilde{\Phi}_t}(G_t, G_{t-1}) + \widetilde{\Phi}_t(G_{t-1}) - \widetilde{\Phi}_{t-1}(G_{t-1}), \end{aligned}$$

which completes the proof. □

We point out a couple of important facts about Lemma 1.2:

1. If $\widetilde{\Phi}_1 \equiv \cdots \equiv \widetilde{\Phi}_T$, then the overestimation penalty sums up to $\widetilde{\Phi}_1(0) - \widetilde{\Phi}(0) = \widetilde{\Phi}_T(0) - \widetilde{\Phi}(0)$.

2. If $\widetilde{\Phi}_t$ is β-strongly smooth with respect to $\|\cdot\|$, the divergence penalty at t is at most $\frac{\beta}{2}\|g_t\|^2$.

One source of regret is the Bregman divergence of $\widetilde{\Phi}_t$; since g_t is not known until playing w_t, the GBPA always ascends along the gradient that is one step behind. The adversary can exploit this and play g_t to induce a large *gap* between $\widetilde{\Phi}_t(G_t)$ and the linear approximation of $\widetilde{\Phi}_t(G_t)$ around G_{t-1}. The learner can reduce this gap by choosing a *smooth* $\widetilde{\Phi}_t$ whose gradient changes slowly.

The learner, however, cannot achieve low regret by choosing an arbitrarily smooth $\widetilde{\Phi}_t$, because the other source of regret is the difference between $\widetilde{\Phi}_t$ and Φ. In short, the GBPA achieves low regret if the potential function $\widetilde{\Phi}_t$ gives a favorable tradeoff between the two sources of regret. This tradeoff is captured by the following definition of *smoothing parameters*, adapted from Beck and Teboulle (2012, Definition 2.1).

Definition 1.1. *Let f be a closed proper convex function. A collection of functions $\{\tilde{f}_\eta : \eta \in \mathbb{R}_+\}$ is said to be an* η-smoothing *of f with* smoothing parameters $(\alpha, \beta, \|\cdot\|)$*, if for every $\eta > 0$:*

1. *There exists α_1 (underestimation bound) and α_2 (overestimation bound) such that*

$$\sup_{G \in \mathrm{dom}(f)} f(G) - \tilde{f}_\eta(G) \le \alpha_1 \eta \quad \text{and} \quad \sup_{G \in \mathrm{dom}(f)} \tilde{f}_\eta(G) - f(G) \le \alpha_2 \eta$$

with $\alpha_1 + \alpha_2 = \alpha$.

2. *$\tilde{f}_\eta$ is $\frac{\beta}{\eta}$-strongly smooth with respect to $\|\cdot\|$.*

We say α is the deviation parameter*, and β is the* smoothness parameter*.*

A straightforward application of Lemma 1.2 gives the following statement:

Corollary 1.3. *Let Φ be the baseline potential for an online linear optimization problem. Suppose $\{\widetilde{\Phi}_\eta\}$ is an η-smoothing of Φ with parameters $(\alpha, \beta, \|\cdot\|)$. Then, the GBPA run with $\widetilde{\Phi}_1 \equiv \cdots \equiv \widetilde{\Phi}_T \equiv \widetilde{\Phi}_\eta$ enjoys the following regret bound,*

$$\text{Regret} \le \alpha\eta + \frac{\beta}{2\eta} \sum_{t=1}^{T} \|g_t\|^2.$$

Choosing η to optimize the bound gives $\text{Regret} \le \sqrt{2\alpha\beta \sum_{t=1}^{T} \|g_t\|^2}$.

In OLO, we often consider the settings where the reward vectors $g_1, \ldots, g_t$ are constrained in norm, i.e., $\|g_t\| \le r$ for all t. In such settings, the regret grows in $O(r\sqrt{\alpha\beta T})$ for the optimal choice of η. The product $\alpha\beta$ of the devation and smoothness parameters is, therefore, at the core of the GBPA regret analysis.

An important smoothing technique for this chapter is *stochasting smoothing*, which is the convolution of a function with a probability density function.

Definition 1.2 (Stochastic Smoothing)**.** *Let $f : \mathbb{R}^N \to \mathbb{R}$ be a function. We define $\tilde{f}(\cdot; \mathcal{D}_\eta)$ to be the* stochastic smoothing *of f with distribution $\mathcal{D}$ and scaling parameter $\eta > 0$. The function value at G is obtained as:*

$$\tilde{f}(G; \mathcal{D}_\eta) := \mathbb{E}_{z' \sim \mathcal{D}_\eta}[f(G + z')] = \mathbb{E}_{z \sim \mathcal{D}}[f(G + \eta z)],$$

where we adopt the convention that if z has distribution $\mathcal{D}$ then the distribution of ηz is denoted by $\mathcal{D}_\eta$.

Notes on estimation penalty If the perturbation used has mean zero, it follows from Jensen's inequality that the stochastic smoothing will overestimate the convex function Φ. Hence, for mean zero perturbations, the underestimation penalty is always non-positive. When the scaling parameter η_t changes every iteration, the overestimation penalty becomes a sum of T terms. The following lemma shows that we can collapse them into one since the baseline potential Φ in OLO problems is sub-additive: $\Phi(G + H) \leq \Phi(G) + \Phi(H)$.

Lemma 1.4. *Let $\Phi : \mathbb{R}^N \to \mathbb{R}$ be a baseline potential function of an OLO problem. Let $\mathcal{D}$ be a continuous distribution with zero mean and support $\mathbb{R}^N$. Consider the GBPA with $\widetilde{\Phi}_t(G) = \widetilde{\Phi}(G; \mathcal{D}_{\eta_t})$ for $t = 0, \ldots, T$ where $(\eta_1, \ldots, \eta_T)$ is a non-decreasing sequence of non-negative numbers. Then the overestimation penalty has the following upper bound,*

$$\sum_{t=1}^{T} \widetilde{\Phi}_t(G_{t-1}) - \widetilde{\Phi}_{t-1}(G_{t-1}) \leq \eta_T \mathbb{E}_{u \sim \mathcal{D}}[\Phi(u)],$$

and the underestimation penalty is non-positive which gives gives a regret bound of

$$\text{Regret} \leq \eta_T \mathbb{E}_{u \sim \mathcal{D}}[\Phi(u)] + \sum_{t=1}^{T} D_{\widetilde{\Phi}_t}(G_t, G_{t-1}). \tag{1.3}$$

Proof. By virtue of the fact that Φ is a support function, it is also subadditive and satisfies the triangle inequality. Hence we can see that, for any $0 < \eta' \leq \eta$,

$$\begin{aligned}\widetilde{\Phi}(G; \mathcal{D}_\eta) - \widetilde{\Phi}(G; \mathcal{D}_{\eta'}) &= \mathbb{E}_{u \sim \mathcal{D}}[\Phi(G + \eta u) - \Phi(G + \eta' u)] \\ &\leq \mathbb{E}_{u \sim \mathcal{D}}[\Phi((\eta - \eta')u)] = (\eta - \eta')\mathbb{E}_{u \sim \mathcal{D}}[\Phi(u)],\end{aligned}$$

where the final line follows from the positive homogeneity of Φ. Since we

implicitly assume that $\widetilde{\Phi}_0 \equiv \Phi$ we can set $\eta_0 = 0$. We can then conclude that

$$\sum_{t=1}^{T} \widetilde{\Phi}_t(G_{t-1}) - \widetilde{\Phi}_{t-1}(G_{t-1}) \leq \left(\sum_{t=1}^{T} \eta_t - \eta_{t-1} \right) \mathbb{E}_{u \sim \mathcal{D}}[\Phi(u)] = \eta_T \mathbb{E}_{u \sim \mathcal{D}}[\Phi(u)],$$

which completes the proof. □

1.3.2 Understanding Follow the Perturbed Leader via Stochastic Smoothing

The technique of *stochastic smoothing* has been well-studied in the optimization literature for gradient-free optimization algorithms (Glasserman, 1991; Yousefian et al., 2010) and accelerated gradient methods for non-smooth optimizations (Duchi et al., 2012).

One very useful property of stochastic smoothing is that as long as $\mathcal{D}$ has a support over $\mathbb{R}^N$ and has a differentiable probability density function μ, $\tilde{f}$ is always differentiable. To see this, we use the change of variable technique:

$$\tilde{f}(G; \mathcal{D}) = \int f(G + z)\mu(z)\, dz = \int f(\tilde{G})\mu(\tilde{G} - G)\, d\tilde{G},$$

and it follows that

$$\begin{aligned} \nabla_G \tilde{f}(G; \mathcal{D}) &= -\int f(\tilde{G}) \nabla_G \mu(\tilde{G} - G)\, d\tilde{G}, \\ \nabla_G^2 \tilde{f}(G; \mathcal{D}) &= \int f(\tilde{G}) \nabla_G^2 \mu(\tilde{G} - G)\, d\tilde{G}. \end{aligned} \tag{1.4}$$

This change of variable trick leads to the following useful expressions for the first and second derivatives of $\tilde{f}$ in case the density $\mu(G)$ is proportional to $\exp(-\nu(G))$ for a sufficiently smooth ν.

Lemma 1.5 (Exponential Family Smoothing)**.** *Suppose $\mathcal{D}$ is a distribution over $\mathbb{R}^N$ with a probability density function μ of the form $\mu(G) = \exp(-\nu(G))/Z$ for some normalization constant Z. Then, for any twice-differentiable ν, we have*

$$\begin{aligned} \nabla \tilde{f}(G) &= \mathbb{E}[f(G + z)\nabla_z \nu(z)], \\ \nabla^2 \tilde{f}(G) &= \mathbb{E}[f(G + z)\left(\nabla_z \nu(z) \nabla_z \nu(z)^T - \nabla_z^2 \nu(z)\right)]. \end{aligned} \tag{1.5}$$

Furthermore, if f is convex, we have

$$\nabla^2 \tilde{f}(G) = \mathbb{E}[\nabla f(G + z) \nabla_z \nu(z)^T].$$

Proof. If ν is twice-differentiable, $\nabla \mu = -\mu \cdot \nabla \nu$ and $\nabla^2 \mu = \left(\nabla \nu \nabla \nu^T - \nabla^2 \nu\right) \mu$. Plugging these in (1.4) and using the substitution $z = \tilde{G} - G$ immediately

gives the first two claims of the lemma. For the last claim, we first directly differentiate the expression for $\nabla \tilde{f}$ in (1.5) by swapping the expectation and gradient. This is justified because f is convex (and is hence differentiable almost everywhere) and μ is absolutely continuous w.r.t. Lebesgue measure everywhere (Bertsekas, 1973, Proposition 2.3). □

Let $\mathcal{D}$ be a probability distribution over $\mathbb{R}^N$ with a well-defined density everywhere. Consider the GBPA run with a stochastic smoothing of the baseline potential:

$$\forall t, \widetilde{\Phi}_t(G) = \widetilde{\Phi}(G; \mathcal{D}_{\eta_t}) = \mathbb{E}_{z\sim\mathcal{D}}\Big[\max_{w\in\mathcal{X}} \langle w, G + \eta_t z\rangle\Big]. \tag{1.6}$$

Then, from the convexity of $G \mapsto \max_{w\in\mathcal{X}}\langle w, G + \eta_t z\rangle$ (for any fixed z), we can swap the expectation and gradient (Bertsekas, 1973, Proposition 2.2) and evaluate the gradient at $G = G_{t-1}$ to obtain

$$\nabla\widetilde{\Phi}_t(G_{t-1}) = \mathbb{E}_{z\sim\mathcal{D}}\Big[\operatorname*{argmax}_{w\in\mathcal{X}} \langle w, G_{t-1} + \eta_t z\rangle\Big]. \tag{1.7}$$

Taking a single random sample of argmax inside expectation is equivalent to the decision rule of FTPL (Hannan, 1957; Kalai and Vempala, 2005); the GBPA on a stochastically smoothed potential can thus be seen as playing the *expected action* of FTPL. Since the learner gets a linear reward in online linear optimization, the regret of the GBPA on a stochastically smoothed potential is equal to the *expected regret* of FTPL. For this reason, we will use the terms FTPL and GBPA with stochastic smoothing interchangably.

1.3.3 Connection between FTPL and FTRL via Duality

We have been discussing a method of smoothing out an objective (potential) function by taking the average value of the objective over a set of nearby "perturbed" points. Another more direct method of smoothing the objective function is via a regularization penalty. We can define the *regularized potential* as follows:

$$\widetilde{\Phi}(G) = \mathcal{R}^\star(G) = \max_{w\in\mathcal{X}}\{\langle w, G\rangle - \mathcal{R}(w)\} \tag{1.8}$$

where $\mathcal{R} : \mathcal{X} \to \mathbb{R}$ is some strictly convex function. This technique has been referred to as "inf-conv" smoothing of Φ with $\mathcal{R}^*$. The connection between regularization and smoothing is further developed by Abernethy et al. (2014), and the terminology draws from the work of Beck and Teboulle (2012) among others. The class of FTRL algorithms can be viewed precisely as an instance of GBPA where the potential is chosen according to Eqn. (1.8). This follows

because of the following fact, which is a standard result of Fenchel duality:

$$\nabla f^\star(\theta) = \arg\max_x \langle x, \theta\rangle - f(x),$$

under the condition that f is differentiable and strictly convex (Rockafellar, 1997). In other words, if we consider $f(\cdot)$ to be the regularizer for an FTRL function, then solution to the FTRL objective corresponds directly with the gradient of the potential function $f^\star(\cdot)$.

Now that we have see that FTRL and FTPL can be viewed as a certain type of smoothing operation, a natural question one might ask is: to what extent are stochastic smoothing and inf-conv smoothing related? That is, can we view FTRL and FTPL as really two sides of the same coin? The answer here is "partially yes" and "partially no":

1. When $\mathcal{X}$ is 1-dimensional then (nearly) every instance of FTRL can be seen as a special case of FTPL, and vice versa. In other words, stochastic smoothing and inf-conv smoothing are effectively one and the same, and we describe this equivalence in detail below.

2. For problems of dimension larger than 1, every instance of FTPL can be described as an instance of FTRL. More precisely, if we have a distribution $\mathcal{D}_\eta$ which leads to a stochastically smoothed potential $\widetilde{\Phi}(\cdot) = \widetilde{\Phi}(\cdot; \mathcal{D}_\eta)$, then we can always write the gradient of $\widetilde{\Phi}(\cdot)$ as the solution of an FTRL optimization. That is,

$$\nabla\widetilde{\Phi}(G, \mathcal{D}_\eta) = \arg\max_{x\in\mathcal{X}} \langle x, \theta\rangle - \mathcal{R}(x) \quad \text{where} \quad \mathcal{R}(x) := \widetilde{\Phi}^\star(x),$$

and we recall that $\widetilde{\Phi}^\star$ denotes the Fenchel Conjugate. In other words, the perturbation $\mathcal{D}$ induces an implicit regularizer defined as the cojugate of $\mathbb{E}_{z\sim\mathcal{D}}[\max_{g\in\mathcal{X}}\langle g, G\rangle]$

3. In general, however, stochastic smoothing is not as general as inf-conv smoothing. FTPL is in some sense less general than FTRL, as there are examples of regularizers that can not be "induced" via a specific perturbation. One particular case is given by Hofbauer and Sandholm (2002).

We now given a brief description of the equivalence between stochastic smoothing and inf-conv smoothing for the 1-dimensional case.

On the near-equivalence between FTRL and FTPL in one dimension. Consider a one-dimensional online linear optimization prediction problem where the player chooses an action w_t from $\mathcal{X} = [0, 1]$ and the adversary chooses a reward g_t from $\mathcal{Y} = [0, 1]$. This can be interpreted as a two-expert setting; the player's action $w_t \in \mathcal{X}$ is the probability of following the first

expert and g_t is the net excess reward of the first expert over the second. The baseline potential for this setting is $\widetilde{\Phi}(G) = \max_{w\in[0,1]} wG$.

Let us consider an instance of FTPL with a continuous distribution $\mathcal{D}$ whose cumulative density function (cdf) is $F_{\mathcal{D}}$. Let $\widetilde{\Phi}$ be the smoothed potential function (Equation 1.6) with distribution $\mathcal{D}$. Its derivative is

$$\widetilde{\Phi}'(G) = \mathbb{E}[\operatorname*{argmax}_{w\in\mathcal{Y}} w(G+u)] = \mathbb{P}[u > -G] \tag{1.9}$$

because the maximizer is unique with probability 1. Notice, crucially, that the derivative $\widetilde{\Phi}'(G)$ is exactly the expected solution of our FTPL instance. Moreover, by differentiating it again, we see that the second derivative of $\widetilde{\Phi}$ at G is exactly the pdf of $\mathcal{D}$ evaluated at $(-G)$.

We can now precisely define the mapping from FTPL to FTRL. Our goal is to find a convex regularization function $\mathcal{R}$ such that $\mathbb{P}(u > -G) = \arg\max_{w\in\mathcal{X}} (wG - \mathcal{R}(w))$. Since this is a one-dimensional convex optimization problem, we can differentiate for the solution. The characterization of $\mathcal{R}$ is:

$$\mathcal{R}(w) - \mathcal{R}(0) = -\int_0^w F_{\mathcal{D}}^{-1}(1-z)\mathrm{d}z. \tag{1.10}$$

Note that the cdf $F_{\mathcal{D}}(\cdot)$ is indeed invertible since it is a strictly increasing function.

The inverse mapping is just as straightforward. Given a regularization function $\mathcal{R}$ well-defined over $[0,1]$, we can always construct its Fenchel conjugate $\mathcal{R}^\star(G) = \sup_{w\in\mathcal{X}}\langle w, G\rangle - \mathcal{R}(w)$. The derivative of $\mathcal{R}^\star$ is an increasing convex function, whose infimum is 0 at $G = -\infty$ and supremum is 1 at $G = +\infty$. Hence, $\mathcal{R}^\star$ defines a cdf, and an easy calculation shows that this perturbation distribution exactly reproduces FTRL corresponding to $\mathcal{R}$.

1.4 Generic Bounds

In this section, we show how the general result in Corollary 1.3, combined with stochastic smoothing results from the existing literature, painlessly yield regret bounds for two generic settings: one in which the learner/adversary sets are bounded in ℓ_1/ℓ_∞ norms and another in which they are bounded in the standard Euclidean (i.e., ℓ_2) norm.

1.4.1 ℓ_1/ℓ_∞ Geometry

With slight abuse of notation, we will use $\|\mathcal{X}\|$ to denote $\sup_{x\in\mathcal{X}}\|x\|$ where $\|\cdot\|$ is a norm and $\mathcal{X}$ is a set of vectors.

Theorem 1.6. *Consider GBPA run with a potential* $\widetilde{\Phi}_t(G) = \widetilde{\Phi}(G; \mathcal{D}_\eta)$ *where* $\mathcal{D}$ *is the uniform distribution on the unit* ℓ_∞ *ball. Then we have,*

$$\text{Regret} \leq \frac{1}{2\eta} T \|\mathcal{X}\|_\infty \|\mathcal{Y}\|_1^2 + \eta \frac{\|\mathcal{X}\|_\infty N}{2}.$$

Choosing η *to optimize the bound gives* $\text{Regret} \leq \|\mathcal{X}\|_\infty \|\mathcal{Y}\|_1 \sqrt{NT}$.

Proof. The baseline potential function Φ is $\|\mathcal{X}\|_\infty$-Lipschitz with respect to $\|\cdot\|_1$. Also note that $\|g_t\|_1 \leq \|\mathcal{Y}\|_1$. Now, by Corollary 1.3, it suffices to prove that the stochastic smoothing of Φ with the uniform distribution on the unit ℓ_∞ ball is an η-smoothing with parameters

$$\left(\frac{\|\mathcal{X}\|_\infty N}{2}, \|\mathcal{X}\|_\infty, \|\cdot\|_1 \right).$$

These smoothing parameters have been shown to hold by Duchi et al. (2012, Lemma E.1). □

FTPL with perturbations drawn from the uniform distribution over the hypercube was considered by Kalai and Vempala (2005). The above theorem gives essentially the same result as their Theorem 1.1(a). The proof above not only uses our general smoothing based analysis but also yields better constants.

1.4.2 Euclidean Geometry

In this section, we will use a generic property of Gaussian smoothing to derive a regret bound that holds for any arbitrary online linear optimization problem.

Theorem 1.7. *Consider GBPA run with a potential* $\widetilde{\Phi}_t(G) = \widetilde{\Phi}(G; \mathcal{D}_\eta)$ *where* $\mathcal{D}$ *is the uniform distribution on the unit* ℓ_2 *ball. Then we have,*

$$\text{Regret} \leq \frac{1}{2\eta} T \sqrt{N} \|\mathcal{X}\|_2 \|\mathcal{Y}\|_2^2 + \eta \|\mathcal{X}\|_2.$$

If we choose $\mathcal{D}$ *to be the standard multivariate Gaussian distribution, then we have,*

$$\text{Regret} \leq \frac{1}{2\eta} T \|\mathcal{X}\|_2 \|\mathcal{Y}\|_2^2 + \eta \sqrt{N} \|\mathcal{X}\|_2.$$

In either case, optimizing over η *we get* $\text{Regret} \leq \|\mathcal{X}\|_2 \|\mathcal{Y}\|_2 N^{1/4} \sqrt{2T}$.

Proof. The baseline potential function Φ is $\|\mathcal{X}\|_2$-Lipschitz with respect to $\|\cdot\|_2$. Also note that $\|g_t\|_2 \leq \|\mathcal{Y}\|_2$. Duchi et al. (2012, Lemma E.2) show that the stochastic smoothing of Φ with the uniform distribution on the Euclidean

unit ball is an η-smoothing with parameters

$$\left(\|\mathcal{X}\|_2, \|\mathcal{X}\|_2\sqrt{N}, \|\cdot\|_1\right).$$

Further, Duchi et al. (2012, Lemma E.3) shows that the stochastic smoothing of Φ with the standard Gaussian distribution is an η-smoothing with parameters

$$\left(\|\mathcal{X}\|_2\sqrt{N}, \|\mathcal{X}\|_2, \|\cdot\|_1\right).$$

The result now follows from Corollary 1.3. □

We are not aware of a previous result for FTPL of generality comparable to Theorem 1.7 above. However, Rakhlin et al. (2012) prove a regret bound for $4\sqrt{2}\sqrt{T}$ when $\mathcal{X}, \mathcal{Y}$ are unit balls of the ℓ_2 norm. Their FTPL algorithm, however, draws $T - t$ samples from the uniform distribution over the unit *sphere*. In contrast, we will show that, for this special case, a dimension independent $O(\sqrt{T})$ bound can be obtained via an FTPL algorithm using a single Gaussian perturbation per time step (see Theorem 1.10 below).

1.5 Experts setting

Now we apply the GBPA analysis framework to the classical online learning problem of the *hedge setting*, or often referred to as *prediction with expert advice*[2]. Here we assume a learner is presented with a set of fixed actions, and on each round must (randomly) select one such action. Upon commiting to her choice, the learner then receives a vector of gains (or losses), one for each action, where the ith gain (loss) value is the reward (cost) for selecting action i. The learner's objective is to continually update the sampling distribution over actions in order to accumulate an expected gain (loss) that is not much worse than the gain (loss) of the optimal fixed action.

The important piece to note about this setting is that it may be cast as an instance of an OLO problem. To see this, we set $\mathcal{X} = \Delta^N \stackrel{\text{def}}{=} \{w \in \mathbb{R}^N : \sum_i w_i = 1, w_i \geq 0 \ \forall i\}$, the N-dimensional probability simplex, and we set $\mathcal{Y} = \{g \in \mathbb{R}^N : \|g\|_\infty \leq 1\}$, a set of bounded gain vectors. We may define the

2. The use of the term "expert" is historical and derives from an early version of the problem where one was given advice (a prediction) from a set of experts (Littlestone and Warmuth, 1994), and the learner's goal is to aggregate this advice. In the version we discuss here, proposed by Freund (1997), a more appropriate intuition is to imagine the task of choosing among a set of "actions" that each receive a "gain" or "loss" on every round.

baseline potential function therefore as

$$\Phi(G) = \max_{w \in \mathcal{X}} \langle w, G \rangle = \max_{i=1,\ldots,N} G_i = G_{i^*(G)}$$

where $i^*(G) := \min\{i : G_i = \max_j G_j\}$ (We need the outer $\min\{\cdot\}$ to define i^* in order to handle possible ties; in such cases we select the lowest index). In our framework we have used language of maximizing gain, in contrast to the more common theme of minimizing loss. However, the loss-only setting can be easily obtained by simply changing the domain $\mathcal{Y}$ to contain only vectors with negative-valued coordinates.

1.5.1 The Exponential Weights Algorithm, and the Equivalence of Entropy Regularization and Gumbel Perturbation

The most well-known and widely used algorithm in the experts setting is the *Exponential Weights Algorithm* (EWA), often referred to as the *Multiplicative Weights Algorithm* and strongly related to the classical *Weighted Majority Algorithm* (Littlestone and Warmuth, 1994). On round t, EWA specifies a set of unnormalized weights based on the cumulative gains thus far,

$$\tilde{w}_{t,i} := \exp(\eta G_{t-1,i}) \qquad i = 1, \ldots, N,$$

where $\eta > 0$ is a parameter. The learner's distribution on this round is then obtained by normalizing $\tilde{w}_t$

$$w_{t,i} := \frac{\tilde{w}_{t,i}}{\sum_{j=1}^N \tilde{w}_{t,j}} \qquad i = 1, \ldots, N. \tag{1.11}$$

More recent perspectives of EWA have relied on an alternative interpretation via an optimization problem. Indeed the weights obtained in Eqn. 1.11 can be equivalently obtained as follows,

$$w_t = \operatorname*{argmax}_{w \in \Delta^N} \left\{ \langle \eta G_{t-1}, w \rangle - \sum_{i=1}^N w_i \log w_i \right\}.$$

We have cast the exponential weights algorithm as an instance of FTRL where the regularization function $\mathcal{R}$ corresponds to the *negative entropy function*, $\mathcal{R}(w) := \sum_i w_i \log w_i$. Applying Lemma 1.2 one can show that EWA obtains a regret of order $\sqrt{T \log N}$.

A third interpretation of EWA is obtained via the notion of stochastic smoothing (perturbations) using the *Gumbel distribution*:

$$\begin{aligned} \mu(z) &:= e^{-(z+e^{-z})} && \text{is the PDF of the standard Gumbel; and} \\ \Pr(Z \le z) &= e^{-e^{-z}} && \text{is the CDF of the standard Gumbel.} \end{aligned}$$

The Gumbel distribution has several natural properties, including for example that it is *max-stable*: the maximum value of several Gumbel-distributed random variables is itself distributed according to a Gumbel distribution[3]. But another nice fact is that the distribution of the maximizer of N fixed values perturbed with Gumbel noise leads to an exponentially-weighted distribution. Precisely, if we have a values $v_1, \ldots, v_N$, and we draw n IID samples $Z_1, \ldots, Z_N$ from the standard Gumbel, then a straightforward calculus exercise gives that

$$\Pr\left[v_i + Z_i = \max_{j=1,\ldots,N}\{v_j + Z_j\}\right] = \frac{\exp(v_i)}{\sum_{j=1,\ldots,N}\exp(v_j)} \qquad i = 1, \ldots, N.$$

What we have just arrived at is that EWA is indeed an instance of FTPL with Gumbel-distributed noise. This was described by Adam Kalai in personal communication, and later Warmuth (2009) expanded it into a short note available online. However, the result appears to be folklore in the area of probabilistic choice models, and it is mentioned briefly by Hofbauer and Sandholm (2002).

1.5.2 Experts Bounds via Laplacian, Gaussian, and Gumbel Smoothing

We will now apply our stochastic smoothing analysis to derive bounds on a class of algorithms for the Experts Setting using three different perturbations: the *Exponential*, *Gaussian*, and *Gumbel*. The latter noise distribution generates an algorithm which is equivalent to EWA, as discussed above, but we prove the same bound using new tools. Note, however that we use a mean-zero Gumbel whereas the standard Gumbel has mean 1.

The key lemma for the GBPA analysis is Lemma 1.2, which decomposes the regret into overestimation, underestimation, and divergence penalty. By Lemma 1.4, the underestimation is less than or equal to 0 and the overestimation penalty is upper-bounded by $\mathbb{E}_{z\sim\mathcal{D}}\left[\max_{i=1,\ldots,N} z_i\right]$. This expectation for commonly used distributions $\mathcal{D}$ is well-studied in extreme value theory.

In order to upper bound the divergence penalty, it is convenient to analyze the Hessian matrix, which has a nice structure in the experts setting. We will be especially interested in bounding the trace of this Hessian.

Lemma 1.8. *Let Φ be the baseline potential for the N-experts setting, and $\mathcal{D}$ be a continuous distribution with a differentiable probability density function. We will consider the potential $\widetilde{\Phi}(G) = \widetilde{\Phi}(G; \mathcal{D}_\eta)$. If for some constant β we*

3. Above we only defined the standard Gumbel, but in general the Gumbel has both a scaling and shift parameter.

have a bound $\mathrm{Tr}(\nabla^2\widetilde{\Phi}(G)) \le \beta/\eta$ *for every* G*, then it follows that*

$$D_{\widetilde{\Phi}}(G+g, G) \le \beta\|g\|_\infty^2/\eta. \tag{1.12}$$

Proof. The Hessian exists because μ is differentiable (Equation 1.4). Let H denote the Hessian matrix of the stochastic smoothing of Φ, i.e., $H(\cdot) = \nabla^2\widetilde{\Phi}(\cdot; \mathcal{D}_\eta)$. First we claim two properties on H:

1. Diagonal entries are non-negative and off diagonal entries are non-positive.
2. Each row or column sums up to 0.

All diagonal entries of H are non-negative because $\widetilde{\Phi}$ is convex. Note that $\nabla_i\widetilde{\Phi}$ is the probability that the i-th coordinate of $G+z$ is the maximum coordinate, and an increase in the j-th of G where $j \neq i$ cannot increase that probability; hence, the off-diagonal entries of H are non-positive. To prove the second claim, note that the gradient $\nabla\widetilde{\Phi}$ is a probability vector, whose coordinates always sum up to 1. Thus, each row (or each column) must sum up to 0.

By Taylor's theorem in the mean-value form, we have $D_{\widetilde{\Phi}}(G+g, G) = \frac{1}{2}g^T\nabla^2\widetilde{\Phi}(\tilde{G})g$ where $\tilde{G}$ is some convex combination on G and $G+g$. Now we have

$$D_{\widetilde{\Phi}}(G+g, G) \le \tfrac{1}{2}\|\nabla^2\widetilde{\Phi}(\tilde{G})\|_{\infty\to 1}\|g\|_\infty^2,$$

where $\|M\|_{\infty\to 1} := \sup_{u\neq 0}\|Mv\|_1/\|v\|_\infty$. Finally note that, for any M, $\|M\|_{\infty\to 1} \le \sum_{i,j}|M_{i,j}|$. We can now conclude the proof by noting that the sum of absolute values of the entries of $\nabla^2\widetilde{\Phi}(\tilde{G})$ is upper bounded by twice its trace given the two properties of the Hessian above. □

The above result will be very convenient in proving bounds on the divergence penalty associated with different noise distributions. In particular, assume we have a noise distribution with exponential form, then IID sample $z = (z_1, \ldots, z_n)$ has density $\mu(z) \propto \prod_i \exp(-\nu(z_i))$. Now applying Lemma 1.5 we have a nice expression for the diagonal Hessian values:

$$\begin{aligned}\nabla^2_{ii}\widetilde{\Phi}(G; \mathcal{D}_\eta) &= \frac{1}{\eta}\mathop{\mathbb{E}}_{(z_1,\ldots,z_n)\sim\mu}\left[\nabla_i\Phi(G+\eta z)\frac{d}{dz_i}\nu(z_i)\right] \\ &= \frac{1}{\eta}\mathop{\mathbb{E}}_{(z_1,\ldots,z_n)\sim\mu}\left[\mathbf{1}\{i = i^*(G+\eta z)\}\frac{d\nu(z_i)}{dz_i}\right]. \end{aligned}\tag{1.13}$$

The above formula now gives us a natural bound on the trace of the Hessian for the three distributions of interest.

▪ **Laplace:** For this distribution we have $\nu(z) = |z| \implies \frac{d\nu(z)}{dz} = \mathrm{sign}(z)$, where the sign function returns $+1$ if the argument is positive, -1 if the

argument is negative, and 0 otherwise. Then we have

$$\begin{aligned}\mathrm{Tr}(\nabla^2\widetilde{\Phi}(G)) &= \frac{1}{\eta}\mathop{\mathbb{E}}_{(z_1,\ldots,z_n)\sim\mu}\left[\textstyle\sum_{i=1}^N \mathbf{1}\{i=i^*(G+\eta z)\}\frac{d\nu(z_i)}{dz_i}\right]\\ &= \frac{1}{\eta}\mathop{\mathbb{E}}_{z}\left[\textstyle\sum_{i=1}^N \mathbf{1}\{i=i^*(G+\eta z)\}\mathrm{sign}(z_i)\right]\\ &\le \frac{1}{\eta}\mathop{\mathbb{E}}_{z}\left[\textstyle\sum_{i=1}^N \mathbf{1}\{i=i^*(G+\eta z)\}\right] = \frac{1}{\eta}.\end{aligned}$$

- **Gumbel:** Here, using zero-mean Gumbel, we have $\nu(z) = z+1+e^{-z-1} \implies \frac{d\nu(z)}{dz} = 1-e^{-z-1}$. Applying the same arguments we obtain

$$\begin{aligned}\mathrm{Tr}(\nabla^2\widetilde{\Phi}(G)) &= \frac{1}{\eta}\mathop{\mathbb{E}}_{z}\left[\textstyle\sum_{i=1}^N \mathbf{1}\{i=i^*(G+\eta z)\}(1-e^{-z_i-1})\right]\\ &\le \frac{1}{\eta}\mathop{\mathbb{E}}_{z}\left[\textstyle\sum_{i=1}^N \mathbf{1}\{i=i^*(G+\eta z)\}\right] = \frac{1}{\eta}.\end{aligned}$$

- **Gaussian:** Here we have $\nu(z) = \frac{z^2}{2} \implies \frac{d\nu(z)}{dz} = z$. Bounding the sum of diagonal Hessian terms requires a slightly different trick:

$$\begin{aligned}\mathrm{Tr}(\nabla^2\widetilde{\Phi}(G)) &= \frac{1}{\eta}\mathop{\mathbb{E}}_{z}\left[\textstyle\sum_{i=1}^N \mathbf{1}\{i=i^*(G+\eta z)\}z_i\right]\\ &= \frac{1}{\eta}\mathop{\mathbb{E}}_{z}\left[z_{i^*(G+\eta z)}\right] \le \frac{1}{\eta}\mathop{\mathbb{E}}_{z}[\max_i z_i] \le \frac{\sqrt{2\log N}}{\eta}.\end{aligned}$$

where the last inequality follows according to moment generating function arguments given below.

To obtain regret bounds, all that remains is a bound on the overestimation penalty. As we showed in Lemma 1.4, the overestimation penalty is upper bounded as $\eta\mathbb{E}_{z\sim\mathcal{D}}[\Phi(z)] = \eta\mathbb{E}[\max_i z_i]$. We can bound this quantity using moment generating functions. Let $s > 0$ be some parameter and notice

$$s\mathbb{E}[\max_i z_i] \le \log\mathbb{E}[\exp(s\max_i z_i)] \le \log\sum_i \mathbb{E}[\exp(sz_i)] \le \log N + \log m(s)$$

where $m(s)$ is the *moment generating function*[4] (mgf) of the distribution $\mathcal{D}$ (or an upper bound thereof). The statement holds for any positive choice of s in the domain of $m(\cdot)$, hence we have

$$\mathbb{E}_{z\sim\mathcal{D}}[\Phi(z)] \le \inf_{s>0}\frac{\log N + \log m(s)}{s}. \tag{1.14}$$

- **Laplace:** The mgf of the standard Laplace is $m(s) = \frac{1}{1-s}$. Choosing $s = \frac{1}{2}$

4. The mgf of a distribution $\mathcal{D}$ is the function $m(s) := \mathbb{E}_{X\sim\mathcal{D}}[\exp(sX)]$.

gives us that $\mathbb{E}[\max_i z_i] \leq 2\log 2N$.

- **Gumbel:** The mgf of the mean-zero Gumbel is $m(s) = \Gamma(1-s)e^{-s}$. Choosing $s = 1/2$ gives that $\mathbb{E}[\max_i z_i] \leq 2\log 2N$ since $m(0.5) < 2$.
- **Gaussian:** The mgf of the standard Gaussian is $m(s) = \exp(s^2/2)$. Choosing $s = \sqrt{2\log N}$ gives $\mathbb{E}[\max_i z_i] \leq \sqrt{2\log N}$.

Theorem 1.9. *Let Φ be the baseline potential for the experts setting. Suppose we GBPA run with $\widetilde{\Phi}_t(\cdot) = \widetilde{\Phi}(\cdot; \mathcal{D}_\eta)$ for all t where the mean-zero distribution $\mathcal{D}$ is such that $\mathbb{E}_{z\sim\mathcal{D}}[\Phi(z)] \leq \alpha$ and $\forall G, \mathrm{Tr}(\nabla^2\widetilde{\Phi}(G)) \leq \beta/\eta$. Then we have*

$$\text{Regret} \leq \eta\alpha + \frac{\beta T}{\eta}.$$

Choosing η to optimize the bound gives Regret $\leq 2\sqrt{\alpha\beta T}$. *In particular, for Laplace, (mean-zero) Gumbel and Gaussian perturbations, the regret bound becomes $2\sqrt{2T\log 2N}$, $2\sqrt{2T\log 2N}$ and $2\sqrt{2T\log N}$ respectively.*

Proof. Result follows by plugging in bounds into Lemma 1.2. Mean-zero perturbations imply that the underestimation penalty is zero. The overestimation penalty is bounded by $\eta\alpha$ and the divergence penalty is bounded by $\beta T/\eta$ because of Lemma 1.8 and the assumption that $\|g_t\|_\infty \leq 1$. Our calculations above showed that for the Laplace, (mean-zero) Gumbel and Gaussian perturbations, we have $\alpha = 2\log 2N$, $2\log 2N$ and $\sqrt{2\log N}$ respectively. Furthermore, we have $\beta = 1$, 1 and $\sqrt{2\log N}$ respectively. □

1.6 Euclidean Balls Setting

The Euclidean balls setting is where $\mathcal{X} = \mathcal{Y} = \{x \in \mathbb{R}^N : \|x\|_2 \leq 1\}$. The baseline potential function is $\Phi(G) = \max_{w\in\mathcal{X}}\langle w, G\rangle = \|G\|_2$. We show that the GBPA with Gaussian smoothing achieves a minimax optimal regret (Abernethy et al., 2008) up to a constant factor.

Theorem 1.10. *Let Φ be the baseline potential for the Euclidean balls setting. The GBPA run with $\widetilde{\Phi}_t(\cdot) = \widetilde{\Phi}(\cdot; \mathcal{N}(0, I)_{\eta_t})$ for all t has regret at most*

$$\text{Regret} \leq \eta_T\sqrt{N} + \tfrac{1}{2\sqrt{N}}\textstyle\sum_{t=1}^T \tfrac{1}{\eta_t}\|g_t\|_2^2. \tag{1.15}$$

If the algorithm selects $\eta_t = \sqrt{\sum_{s=1}^T \|g_s\|_2^2/(2N)}$ for all t, we have

$$\text{Regret} \leq \sqrt{2\textstyle\sum_{t=1}^T \|g_t\|_2^2}.$$

If the algorithm selects η_t adaptively according to $\eta_t = \sqrt{(1 + \sum_{s=1}^{t-1}\|g_s\|_2^2))/N}$,

we have

$$\text{Regret} \leq 2\sqrt{1 + \sum_{t=1}^{T} \|g_t\|_2^2}$$

Proof. The proof is mostly similar to that of Theorem 1.9. In order to apply Lemma 1.2, we need to upper bound (i) the overestimation and underestimation penalty, and (ii) the Bregman divergence.

The Gaussian smoothing always overestimates a convex function, so it suffices to bound the overestimation penalty. Furthermore, it suffices to consider the fixed η_t case due to Lemma 1.1. The overestimation penalty can be upper-bounded as follows:

$$\begin{aligned}\widetilde{\Phi}_T(0) - \widetilde{\Phi}(0) &= \mathbb{E}_{u\sim\mathcal{N}(0,I)}\|G + \eta_T u\|_2 - \|G\|_2 \\ &\leq \eta_T \mathbb{E}_{u\sim\mathcal{N}(0,I)}\|u\|_2 \leq \eta_T \sqrt{\mathbb{E}_{u\sim\mathcal{N}(0,I)}\|u\|_2^2} = \eta_T\sqrt{N}.\end{aligned}$$

The first inequality is from the triangle inequality, and the second inequality is from the concavity of the square root.

For the divergence penalty, note that the upper bound on $\max_{v:\|g\|_2=1} g^T(\nabla^2\widetilde{\Phi})g$ is exactly the maximum eigenvalue of the Hessian, which we bound in Lemma 1.11. The final step is to apply Lemma 1.1. □

Lemma 1.11. *Let Φ be the baseline potential for the Euclidean balls setting. Then, for all $G \in \mathbb{R}^N$ and $\eta > 0$, the Hessian matrix of the Gaussian smoothed potential satisfies*

$$\nabla^2\widetilde{\Phi}(G; \mathcal{N}(0, I)_\eta) \preceq \tfrac{1}{\eta\sqrt{N}} I.$$

Proof. The Hessian of the Euclidean norm $\nabla^2\Phi(G) = \|G\|_2^{-1} I - \|G\|_2^{-3} GG^T$ diverges near $G = 0$. Expectedly, the maximum curvature is at origin even after Gaussian smoothing (See Appendix 1.8.1). So, it suffices to prove

$$\nabla^2\widetilde{\Phi}(0) = \mathbb{E}_{u\sim\mathcal{N}(0,I)}[\|u\|_2(uu^T - I)] \preceq \sqrt{\tfrac{1}{N}} I,$$

where the Hessian expression is from Lemma 1.5.

By symmetry, all off-diagonal elements of the Hessian are 0. Let $Y = \|u\|^2$, which is Chi-squared with N degrees of freedom. So,

$$\begin{aligned}\text{Tr}(\mathbb{E}[\|u\|_2(uu^T - I)]) &= \mathbb{E}[\text{Tr}(\|u\|_2(uu^T - I))] = \mathbb{E}[\|u\|_2^3 - N\|u\|_2] \\ &= \mathbb{E}[Y^{\frac{3}{2}}] - N\mathbb{E}[Y^{\frac{1}{2}}]\end{aligned}$$

Using the Chi-squared moment formula (Simon, 2002, p. 13):

$$\mathbb{E}[Y^k] = \frac{2^k\Gamma(\frac{N}{2} + k)}{\Gamma(\frac{N}{2})},$$

the above becomes:

$$\frac{2^{\frac{3}{2}}\Gamma(\frac{3}{2}+\frac{N}{2})}{\Gamma(\frac{N}{2})} - \frac{N2^{\frac{1}{2}}\Gamma(\frac{1}{2}+\frac{N}{2})}{\Gamma(\frac{N}{2})} = \frac{\sqrt{2}\Gamma(\frac{1}{2}+\frac{N}{2})}{\Gamma(\frac{N}{2})}. \tag{1.16}$$

From the log-convexity of the Gamma function,

$$\log\Gamma\left(\tfrac{1}{2}+\tfrac{N}{2}\right) \leq \tfrac{1}{2}\left(\log\Gamma\left(\tfrac{N}{2}\right) + \log\Gamma\left(\tfrac{N}{2}+1\right)\right) = \log\Gamma\left(\tfrac{N}{2}\right)\sqrt{\tfrac{N}{2}}.$$

Exponentiating both sides, we obtain

$$\Gamma\left(\tfrac{1}{2}+\tfrac{N}{2}\right) \leq \Gamma\left(\tfrac{N}{2}\right)\sqrt{\tfrac{N}{2}},$$

which we apply to Equation 1.16 and get $\mathrm{Tr}(\nabla^2\widetilde{\Phi}(0)) \leq \sqrt{N}$. To complete the proof, note that by symmetry, each entry must have the same expected value, and hence it is bounded by $\sqrt{1/N}$. □

1.7 The Multi-Armed Bandit Setting

Let us introduce the adversarial multi-armed bandit (MAB) setting. The MAB problem is a variation of the loss-only experts setting (Section 1.5) with $\mathcal{X} = \Delta^N$ and $\mathcal{Y} = [-1, 0]^N$. The two main differences are that (a) that learner is required to playing randomly, sampling an action $i_t \in \{1, \ldots, N\}$ according to w_t and then suffering loss/gain g_{t,i_t}, and (b) the learner then observes *only the scalar value* g_{t,i_t}, she receives no information regarding the losses/gains for the unplayed actions, i.e. the values $g_{t,j}$ for $j \neq i_t$ remain unobserved. Note that, while g_t is assumed to take only negative values, we will continue to refer to these quantities as *gains*.

This limited-information feedback makes the bandit problem much more challenging than the full-information setting we studied in Section 1.5, where the learner was given the entire g_t on each round. In the adversarial MAB problem the learner is indeed required to play randomly; it can be shown that a deterministic strategy will lead to linear regret in the worst case. Hence our focus will be on the *expected* regret over the learner's randomization, and we will assume that the sequence of gains are fixed in advance and thus non-random. While the present book chapter will explore this area, other work has considered the problem of obtaining high-probability bounds (Auer et al., 2003), as well as bounds that are robust to adaptive adversaries (Abernethy and Rakhlin, 2009).

The MAB framework is not only mathematically elegant, but useful for a wide range of applications including medical experiment design (Gittins, 1996), automated poker playing strategies (Van den Broeck et al., 2009), and

hyperparameter tuning (Pacula et al., 2012). For the survey of work on MAB, see Bubeck and Cesa-Bianchi (2012).

1.7.1 Gradient-Based Prediction Algorithms for the Multi-Armed Bandit

We give a generic template for constructing MAB strategies in Algorithm 2, and we emphasize that this template can be viewed as a bandit reduction to the (full information) GBPA framework. Randomization is used for making decisions and for *estimating* the losses via importance sampling.

Algorithm 2: GBPA Template for Multi-Armed Bandits.

Require: fixed convex potential $\widetilde{\Phi} : \mathbb{R}^N \to \mathbb{R}$, with $\nabla\widetilde{\Phi} \subset \text{interior}(\Delta^N)$.
Require: Adversary selects (hidden) seq. of loss vectors $g_1, \ldots, g_T \in [-1, 0]^N$
Initialize: $\hat{G}_0 = 0$
for *t = 1 to T* **do**
  Sampling: Learner chooses i_t according to dist. $p(\hat{G}_{t-1}) = \nabla\widetilde{\Phi}(\hat{G}_{t-1})$
  Cost: Learner "gains" g_{t,i_t}, and observes this value
  Estimation: Learner produces estimate of gain vector, $\hat{g}_t := \frac{g_{t,i_t}}{p_{i_t}(\hat{G}_{t-1})}\mathbf{e}_{i_t}$
  Update: $\hat{G}_t = \hat{G}_{t-1} + \hat{g}_t$

Nearly all proposed methods have relied on this particular algorithmic blueprint. For example, the EXP3 algorithm of Auer et al. (2003) proposed a more advanced version of the Exponential Weights Algorithm (discussed in Section 1.5) to set the sampling distribution $p(\hat{G}_{t-1})$, where the only real modification is to include a small probability of uniformly sampling the arms.[5] But EXP3 more or less fits the template we propose in Algorithm 2 when we select $\widetilde{\Phi}(\cdot) = \mathbb{E}_{z\sim\text{Gumbel}}\Phi(G + \eta z)$. We elaborated on the connection between EWA and Gumbel perturbations in Section 1.5.

Lemma 1.12. *The baseline potential for this setting is* $\Phi(G) \equiv \max_i G_i$ *so that we can write the expected regret of GBPA*$(\widetilde{\Phi})$ *as*

$$\mathbb{E}\text{Regret}_T = \Phi(G_T) - \mathbb{E}[\textstyle\sum_{t=1}^{T}\langle\nabla\widetilde{\Phi}(\hat{G}_{t-1}), g_t\rangle].$$

5. One of the conclusions we may draw from this section is that the uniform sampling of EXP3 is not necessary when we are only interested in expected-regret bounds and we focus on negative gains (that is, where $\hat{g}_t \in [-1, 0]^N$). It has been suggested that the uniform sampling may be necessary in the case of positive gains, although this point has not been resolved to the authors' knowledge.

Then, the expected regret of GBPA($\widetilde{\Phi}$) can be written as:

$$\mathbb{E}\text{Regret}_T \leq \mathbb{E}_{i_1,\dots,i_T}\Bigg[\underbrace{\Phi(\hat{G}_T) - \widetilde{\Phi}(\hat{G}_T)}_{\textit{underestimation penalty}} + \sum_{t=1}^{T} \underbrace{\mathbb{E}_{i_t}[D_{\widetilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1})|\hat{G}_{t-1}]}_{\textit{divergence penalty}}\Bigg] + \underbrace{\widetilde{\Phi}(0) - \Phi(0)}_{\textit{overestimation penalty}} \tag{1.17}$$

where the expectations are over the sampling of $i_t, t = 1, \dots, T$.

Proof. Let $\widetilde{\Phi}$ be a valid convex function for GBPA. Consider GBPA($\widetilde{\Phi}$) run on the loss sequence $g_1, \dots, g_T$. The algorithm produces a sequence of estimated losses $\hat{g}_1, \dots, \hat{g}_T$. Now consider GBPA-FI($\widetilde{\Phi}$), which is GBPA($\widetilde{\Phi}$) run with the full information on the deterministic loss sequence $\hat{g}_1, \dots, \hat{g}_T$ (there is no estimation step, and the learner updates $\hat{G}_t$ directly). The regret of this run can be written as

$$\Phi(\hat{G}_T) - \textstyle\sum_{t=1}^{T} \langle \nabla\widetilde{\Phi}(\hat{G}_{t-1}), \hat{g}_t \rangle \tag{1.18}$$

and $\Phi(G_T) \leq \mathbb{E}[\Phi(\hat{G}_T)]$ by the convexity of Φ. □

1.7.2 Implementation of Perturbation Methods

It is clear that $\nabla\widetilde{\Phi}$ is in the probability simplex, and note that

$$\begin{aligned}\frac{\partial\widetilde{\Phi}}{\partial G_i} &= \mathbb{E}_{Z_1,\dots,Z_N}\mathbf{1}\{G_i + Z_i > G_j + Z_j, \forall j \neq i\} \\ &= \mathbb{E}_{\tilde{G}_{j^*}}[\mathbb{P}_{Z_i}[Z_i > \tilde{G}_{j^*} - G_i]] = \mathbb{E}_{\tilde{G}_{j^*}}[1 - F(\tilde{G}_{j^*} - G_i)]\end{aligned} \tag{1.19}$$

where $\tilde{G}_{j^*} = \max_{j\neq i} G_j + Z_j$ and F is the cdf of Z_i. The unbounded support condition guarantees that this partial derivative is non-zero for all i given any G. So, $\widetilde{\Phi}(G; \mathcal{D})$ satisfies the requirements of Algorithm 2.

The sampling step of the bandit GBPA (Framework 2) with a stochastically smoothed function (Equation 1.6) can be implemented efficiently: we need not evaluate the full expectation (Equation 1.7) and instead rely on but a single random sample. On the other hand, the estimation step is challenging since generally there is no closed-form expression[6] for $\nabla\widetilde{\Phi}$.

To address this issue, Neu and Bartók (2013) proposed Geometric Resampling (GR). GR uses an iterative resampling process to estimate $\nabla\widetilde{\Phi}$. They

6. A case where we find a natural closed form solution occurs when the perturbation is chosen to be Gumbel, as we know this corresponds to the EXP3 algorithm which relies on exponential weighting of $\tilde{G}$.

showed that if we stop after M iterations, the extra regret due to the estimation bias is at most $\frac{NT}{eM}$ (additive term). That is, all our GBPA regret bounds in this section hold for the corresponding FTPL algorithm with an extra additive $\frac{NT}{eM}$ term.. This term, however, does not affect the asymptotic regret rate as long as $M = \sqrt{NT}$, because the lower bound for any algorithm is of the order $\sqrt{NT}$.

1.7.3 Differential Consistency

Recall that for the full information experts setting, if we have a uniform bound on the trace of $\nabla^2\widetilde{\Phi}$, then we immediately have a finite regret bound. In the bandit setting, however, the regret (Lemma 1.12) involves terms of the form $D_{\widetilde{\Phi}}(\hat{G}_{t-1} + \hat{g}_t, \hat{G}_{t-1})$, where the incremental quantity $\hat{g}_t$ can scale as large as *the inverse of the smallest probability* of $p(\hat{G}_{t-1})$. These inverse probabilities are essentially unavoidable, because unbiased estimates of a quantity that is observed with only probability p must necessarily involve fluctuations that scale as $O(1/p)$.

Therefore, we need a stronger notion of smoothness that counters the $1/p$ factor in $\|\hat{g}_t\|$. We propose the following definition which bounds $\nabla^2\widetilde{\Phi}$ in correspondence with $\nabla\widetilde{\Phi}$.

Definition 1.3 (Differential Consistency)**.** *For constant $C > 0$, we say that a convex function $f(\cdot)$ is* C-differentially-consistent *if for all $G \in (-\infty, 0]^N$,*

$$\nabla^2_{ii} f(G) \leq C \nabla_i f(G).$$

In other words, the rate in which we decrease p_i should approach 0 as p_i approaches 0. This guarantees that the algorithm reduces the rate of exploration slowly enough. We later show that smoothings obtaining using perturbations with bounded hazard rate satisfy the differential consistency property introduced above (see Lemma 1.15).

We now prove a generic bound that we will use in the following two sections, in order to derive regret guarantees.

Theorem 1.13. *Suppose $\widetilde{\Phi}$ is C-differentially-consistent for constant $C > 0$. Then divergence penalty at time t in Lemma 1.12 can be upper bounded as:*

$$\mathbb{E}_{i_t}[D_{\widetilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1}) | \hat{G}_{t-1}] \leq \frac{NC}{2}.$$

Proof. For the sake of clarity, we drop the t subscripts on $\hat{G}$ and $\hat{g}$; we use $\hat{G}$ to denote the cumulative estimate $\hat{G}_{t-1}$, $\hat{g}$ to denote the marginal estimate $\hat{g}_t = \hat{G}_t - \hat{G}_{t-1}$, and g to denote the true loss g_t.

Note that by definition of Algorithm 2, $\hat{g}$ is a sparse vector with one non-

zero (and negative) coordinate with value $\hat{g}_{i_t} = g_{t,i_t}/\nabla_{i_t}\widetilde{\Phi}(\hat{G})$. Plus, i_t is conditionally independent given $\hat{G}$. Now we can expand the expectation as

$$\begin{aligned}\mathbb{E}_{i_t}[D_{\widetilde{\Phi}}(\hat{G}+\hat{g},\hat{G})|\hat{G}] &= \sum_i \mathbb{P}[i_t=i]\mathbb{E}[D_{\widetilde{\Phi}}(\hat{G}+\hat{g},\hat{G})|\hat{G},i_t=i] \\ &= \sum_i \nabla_i\widetilde{\Phi}(\hat{G})\mathbb{E}[D_{\widetilde{\Phi}}(\hat{G}+\hat{g},\hat{G})|\hat{G},i_t=i]. \qquad (1.20)\end{aligned}$$

For each term in the sum on the right hand side, the conditional expectation given $\hat{G}$ is now,

$$\mathbb{E}[D_{\widetilde{\Phi}}(\hat{G}+\hat{g},\hat{G})|\hat{G},i_t=i] = D_{\widetilde{\Phi}}\left(\hat{G}+\frac{g_i}{\nabla_i\widetilde{\Phi}(\hat{G})}\mathbf{e}_i,\hat{G}\right) = \frac{g_i^2}{2(\nabla_i\widetilde{\Phi}(\hat{G}))^2}\nabla_{ii}^2\widetilde{\Phi}(J_i)$$

where J_i is some vector on the line segment joining $\hat{G}$ and $\hat{G}+\frac{g_i}{\nabla_i\widetilde{\Phi}(\hat{G})}\mathbf{e}_i$. Using differential consistency, we have $\nabla_{ii}^2\widetilde{\Phi}(J_i) \le C\nabla_i\widetilde{\Phi}(J_i)$. Note that J_i agrees with $\hat{G}$ in all coordinates except coordinate i where it is at most $\hat{G}_i$. Note that this conclusion depends crucially on the *loss-only assumption* that $g_i \le 0$. Convexity of $\widetilde{\Phi}$ guarantees that ∇_i is a non-decreasing function of coordinate i. Therefore, $\nabla_i\widetilde{\Phi}(J_i) \le \nabla_i\widetilde{\Phi}(\hat{G})$. This means that

$$\mathbb{E}[D_{\widetilde{\Phi}}(\hat{G}+\hat{g},\hat{G})|\hat{G},i_t=i] \le C\frac{g_i^2}{2(\nabla_i\widetilde{\Phi}(\hat{G}))^2}\nabla_i\widetilde{\Phi}(\hat{G}) \le \frac{C}{2\nabla_i\widetilde{\Phi}(\hat{G})},$$

since $g_i^2 \le 1$. Plugging this into (1.20), we get

$$\mathbb{E}_{i_t}[D_{\widetilde{\Phi}}(\hat{G}+\hat{g},\hat{G})|\hat{G}] \le \sum_i \nabla_i\widetilde{\Phi}(\hat{G})\frac{C}{2\nabla_i\widetilde{\Phi}(\hat{G})} = \frac{NC}{2}. \qquad \square$$

1.7.4 Hazard Rate analysis

Despite the fact that perturbation-based multi-armed bandit algorithms provide a natural randomized decision strategy, they have seen little applications mostly because they are hard to analyze. But one should expect general results to be within reach: as we mentioned above, the EXP3 algorithm can be viewed through the lens of perturbations, where the noise is distributed according to the Gumbel distribution. Indeed, an early result of Kujala and Elomaa (2005) showed that a near-optimal MAB strategy comes about through the use of exponentially-distributed noise, and the same perturbation strategy has more recently been utilized in the work of Neu and Bartók (2013) and Kocák et al. (2014). However, a more general understanding of perturbation methods has remained elusive. For example, would Gaussian noise be sufficient for a guarantee? What about, say, the Weibull distribution?

In this section, we show that the performance of the GBPA($\widetilde{\Phi}(G;\mathcal{D})$) can

be characterized by the *hazard function* of the smoothing distribution $\mathcal{D}$. The hazard rate is a standard tool in survival analysis to describe failures due to aging; for example, an increasing hazard rate models units that deteriorate with age while a decreasing hazard rate models units that improve with age (a counter intuitive but not illogical possibility). To the best of our knowledge, the connection between hazard rates and design of adversarial bandit algorithms has not been made before.

Definition 1.4 (Hazard rate function)**.** *Assume we are given a distribution $\mathcal{D}$ whose PDF is given by f and whose CDF is given by F. The hazard rate function of $\mathcal{D}$ is*

$$h_{\mathcal{D}}(x) := \frac{f(x)}{1 - F(x)}.$$

We will write $\sup h_{\mathcal{D}}$ *to mean the supremal value obtained by $h_{\mathcal{D}}$ on its domain; we drop the subscript $\mathcal{D}$ when it is clear.*

For the rest of the section, we assume that $F(x) < 1$ for all finite x, so that $h_{\mathcal{D}}$ is well-defined everywhere. This assumption is for the clarity of presentation but is not strictly necessary.

Theorem 1.14. *The regret of the GBPA for multi-armed bandits (Algorithm 2) with $\widetilde{\Phi}(G) = \mathbb{E}_{Z_1,\dots,Z_n \sim D} \max_i \{G_i + \eta Z_i\}$ is at most:*

$$\underbrace{\eta \mathbb{E}_{Z_1,\dots,Z_n \sim D}\left[\max_i Z_i\right]}_{\text{overestimation penalty}} + \underbrace{\frac{N \sup h_{\mathcal{D}}}{\eta} T}_{\text{divergence penalty}}$$

Proof. Due to the convexity of Φ, the underestimation penalty is non-positive. The overestimation penalty is clearly at most $\mathbb{E}_{Z_1,\dots,Z_n \sim D}[\max_i Z_i]$, and Lemma 1.15 proves the $N(\sup h_{\mathcal{D}})$ upper bound on the divergence penalty.

It remains to prove the tuning parameter η. Suppose we scale the perturbation Z by $\eta > 0$, i.e., we add ηZ_i to each coordinate. It is easy to see that $\mathbb{E}[\max_{i=1,\dots,n} \eta X_i] = \eta \mathbb{E}[\max_{i=1,\dots,n} X_i]$. For the divergence penalty, let F_η be the CDF of the scaled random variable. Observe that $F_\eta(t) = F(t/\eta)$ and thus $f_\eta(t) = \frac{1}{\eta} f(t/\eta)$. Hence, the hazard rate scales by $1/\eta$, which completes the proof. □

Lemma 1.15. *Consider implementing GBPA with potential function*

$$\widetilde{\Phi}(G) = \mathbb{E}_{Z_1,\dots,Z_n \sim D} \max_i \{G_i + \eta Z_i\}.$$

The divergence penalty on each round is at most $N(\sup h_{\mathcal{D}})$.

Proof. Recall the gradient expression in Equation 1.19. We upper bound

Distribution	$\sup_x h_{\mathcal{D}}(x)$	$\mathbb{E}[\max_{i=1}^N Z_i]$	Parameters
Gumbel($\mu = 1, \beta = 1$)	1 as $x \to 0$	$\log N + \gamma_0$	N/A
Frechet ($\alpha > 1$)	at most 2α	$N^{1/\alpha}\Gamma(1 - 1/\alpha)$	$\alpha = \log N$
Weibull($\lambda = 1, k \leq 1$)	k at $x = 0$	$O(\left(\frac{1}{k}\right)!(\log N)^{\frac{1}{k}})$	$k = 1$
Pareto($x_m = 1, \alpha$)	α at $x = 0$	$\alpha N^{1/\alpha}/(\alpha - 1)$	$\alpha = \log N$
Gamma($\alpha \geq 1, \beta$)	β as $x \to \infty$	$\log N + (\alpha - 1)\log\log N - \log\Gamma(\alpha) + \beta^{-1}\gamma_0$	$\beta = \alpha = 1$

Table 1.1: *Distributions that give $O(\sqrt{TN\log N})$ regret FTPL algorithm.* The parameterization follows Wikipedia pages for easy lookup. We denote the Euler constant (≈ 0.58) by γ_0. Please see Abernethy et al. (2015) for a full description.

the i-th diagonal entry of the Hessian, as follows. First, let where $\tilde{G}_{j^*} = \max_{j \neq i}\{G_j + Z_j\}$ which is a random variable independent of Z_i. Now,

$$\begin{aligned}
\nabla^2_{ii}\widetilde{\Phi}(G) &= \frac{\partial}{\partial G_i}\mathbb{E}_{\tilde{G}_{j^*}}[1 - F(\tilde{G}_{j^*} - G_i)] = \mathbb{E}_{\tilde{G}_{j^*}}\left[\frac{\partial}{\partial G_i}(1 - F(\tilde{G}_{j^*} - G_i))\right] \\
&= \mathbb{E}_{\tilde{G}_{j^*}} f(\tilde{G}_{j^*} - G_i) \\
&= \mathbb{E}_{\tilde{G}_{j^*}}[h(\tilde{G}_{j^*} - G_i)(1 - F(\tilde{G}_{j^*} - G_i))] \qquad (1.21) \\
&\leq (\sup h)\mathbb{E}_{\tilde{G}_{j^*}}[1 - F(\tilde{G}_{j^*} - G_i)] \\
&= (\sup h)\nabla_i\widetilde{\Phi}(G).
\end{aligned}$$

We have just established that $\widetilde{\Phi}$ is differentially consistent with parameter $C = \sup h$. We apply Theorem 1.13 and the proof is complete. □

Corollary 1.16. *Algorithm 2 run with $\widetilde{\Phi}$ that is obtained by smoothing Φ using any of the distributions in Table 1.1 (restricted to a certain range of parameters), combined with Geometric Resampling with $M = \sqrt{NT}$, has an expected regret of order $O(\sqrt{TN\log N})$.*

Table 1.1 provides the two terms we need to bound. More details on these distributions and their relation to stochastic smoothing can be found in Abernethy et al. (2015).

Acknowledgements

We would like to thank Elad Hazan and Gergely Neu for many helpful and insightful conversations on this work. The research was supported by NSF CAREER Awards IIS-1453304 and IIS-1452099, as well as NSF grants IIS-1421391 and IIS-1319810.

Appendix: Detailed Proofs

1.8.1 Proof that the origin is the worst case (Lemma 1.11)

Proof. Let $\Phi(G) = \|G\|_2$ and η be a positive number. By continuity of eigenvectors, it suffices to show that the maximum eigenvalue of the Hessian matrix of the Gaussian smoothed potential $\widetilde{\Phi}(G; \eta, \mathcal{N}(0, I))$ is decreasing in $\|G\|$ for $\|G\| > 0$.

By Lemma 1.5, the gradient can be written as follows:

$$\nabla\Phi(G; \eta, \mathcal{N}(0, I)) = \frac{1}{\eta}\mathbb{E}_{u\sim\mathcal{N}(0,I)}[u\|G + \eta u\|] \tag{1.22}$$

Let u_i be the i-th coordinate of the vector u. Since the standard normal distribution is spherically symmetric, we can rotate the random variable u such that its first coordinate u_1 is along the direction of G. After rotation, the gradient can be written as

$$\frac{1}{\eta}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[u\sqrt{(\|G\| + \eta u_1)^2 + \sum_{k=2}^{N}\eta^2 u_k^2}\right]$$

which is clearly independent of the coordinates of G. The pdf of standard Gaussian distribution has the same value at $(u_1, u_2, \ldots, u_n)$ and its sign-flipped pair $(u_1, -u_2, \ldots, -u_n)$. Hence, in expectation, the two vectors cancel out every coordinate but the first, which is along the direction of G. Therefore, there exists a function α such that $\mathbb{E}_{u\sim\mathcal{N}(0,I)}[u\|G + \eta u\|] = \alpha(\|G\|)G$.

Now, we will show that α is decreasing in $\|G\|$. Due to symmetry, it suffices to consider $G = te_1$ for $t \in \mathbb{R}^+$, without loss of generality. For any $t > 0$,

$$\begin{aligned}
\alpha(t) &= \mathbb{E}[u_1\sqrt{(t + \eta u_1)^2 + u_{\text{rest}}^2})]/t \\
&= \mathbb{E}_{u_{\text{rest}}}[\mathbb{E}_{u_1}[u_1\sqrt{(t + \eta u_1)^2 + b^2}|u_{\text{rest}} = b]]/t \\
&= \mathbb{E}_{u_{\text{rest}}}[\mathbb{E}_{a=\eta|u_1|}[a\Big(\sqrt{(t + a)^2 + b^2} - \sqrt{(t - a)^2 + B}\Big)|u_{\text{rest}} = b]]/t
\end{aligned}$$

Let $g(t) = \Big(\sqrt{(t + a)^2 + B} - \sqrt{(t - a)^2 + B}\Big)/t$. Take the first derivative with respect to t, and we have:

$$\begin{aligned}
g'(t) &= \frac{1}{t^2}\left(\sqrt{(t - a)^2 + b^2} - \frac{t(t - a)}{\sqrt{(t + a)^2 + b^2}} - \sqrt{(t + a)^2 + b^2} + \frac{t(t - a)}{\sqrt{(t + a)^2 + b^2}}\right) \\
&= \frac{1}{t^2}\left(\frac{a^2 + b^2 - at}{\sqrt{(t - a)^2 + b^2}} - \frac{a^2 + b^2 + at}{\sqrt{(t + a)^2 + b^2}}\right)
\end{aligned}$$

$$\Big((a^2+b^2)-at\Big)^2\Big((t+a)^2+b^2\Big)-\Big((a^2+b^2)+at\Big)^2\Big((t-a)^2+b^2\Big)=-4ab^2t^3<0$$

because t, η, u', B are all positive. So, $g(t) < 0$, which proves that α is decreasing in G.

The final step is to write the gradient as $\nabla(\widetilde{\Phi}; \eta, \mathcal{N}(0, I))(G) = \alpha(\|G\|)G$ and differentiate it:

$$\nabla^2 f_\eta(G) = \frac{\alpha'(\|G\|)}{\|G\|}GG^T + \alpha(\|G\|)I$$

The Hessian has two distinct eigenvalues $\alpha(\|G\|)$ and $\alpha(\|G\|) + \alpha'(\|G\|)\|G\|$, which correspond to the eigenspace orthogonal to G and parallel to G, respectively. Since α' is negative, α is always the maximum eigenvalue and it decreases in $\|G\|$. □

1.9 References

J. Abernethy and A. Rakhlin. Beating the Adaptive Bandit with High Probability. In *Proceedings of Conference on Learning Theory (COLT)*, 2009.

J. Abernethy, P. L. Bartlett, A. Rakhlin, and A. Tewari. Optimal Stragies and Minimax Lower Bounds for Online Convex Games. In *Proceedings of Conference on Learning Theory (COLT)*, 2008.

J. Abernethy, Y. Chen, and J. W. Vaughan. Efficient Market Making via Convex Optimization, and a Connection to Online Learning. *ACM Transactions on Economics and Computation*, 1(2):12, 2013.

J. Abernethy, C. Lee, A. Sinha, and A. Tewari. Online Linear Optimization via Smoothing. In *Proceedings of Conference on Learning Theory (COLT)*, 2014.

J. Abernethy, C. Lee, and A. Tewari. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems 28*, 2015. to appear.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The Nonstochastic Multiarmed Bandit Problem. *SIAM Journal of Computuataion*, 32(1):48–77, 2003. ISSN 0097-5397.

A. Beck and M. Teboulle. Smoothing and First Order Methods: A Unified Framework. *SIAM Journal on Optimization*, 22(2):557–580, 2012.

D. P. Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231, 1973. ISSN 0022-3239.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5 (1):1–122, 2012.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006. ISBN 978-0-521-84108-5.

L. Devroye, G. Lugosi, and G. Neu. Prediction by Random-Walk Perturbation. In *Proceedings of Conference on Learning Theory (COLT)*, 2013.

J. Duchi, P. L. Bartlett, and M. J. Wainwright. Randomized Smoothing for Stochastic Optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

Y. Freund. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139, 1997.

J. Gittins. Quantitative methods in the planning of pharmaceutical research. *Drug Information Journal*, 30(2):479–487, 1996.

P. Glasserman. *Gradient Estimation Via Perturbation Analysis*. Kluwer international series in engineering and computer science: Discrete event dynamic systems. Springer, 1991. ISBN 9780792390954.

J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

J. Hofbauer and W. H. Sandholm. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6):2265–2294, 2002.

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

T. Kocák, G. Neu, M. Valko, and R. Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 613–621. Curran Associates, Inc., 2014.

J. Kujala and T. Elomaa. On following the perturbed leader in the bandit setting. In *Algorithmic Learning Theory*, pages 371–385. Springer, 2005.

N. Littlestone and M. K. Warmuth. The Weighted Majority Algorithm. *Information and Computation*, 108(2):212–261, 1994.

H. B. McMahan. Follow-the-Regularized-Leader and Mirror Descent: Equivalence Theorems and L1 Regularization. In *AISTATS*, pages 525–533, 2011.

I. S. Molchanov. *Theory of random sets*. Probability and its applications. Springer, New York, 2005. ISBN 1-85233-892-X.

Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, pages 1–40, 2011.

G. Neu and G. Bartók. An efficient algorithm for learning with semi-bandit feedback. In *Algorithmic Learning Theory*, pages 234–248. Springer, 2013.

M. Pacula, J. Ansel, S. Amarasinghe, and U.-M. O'Reilly. Hyperparameter tuning in bandit-based adaptive operator selection. In *Applications of Evolutionary Computation*, pages 73–82. Springer, 2012.

S. Rakhlin, O. Shamir, and K. Sridharan. Relax and randomize: From value to algorithms. In *Advances in Neural Information Processing Systems*, pages 2141–2149, 2012.

R. Rockafellar. *Convex Analysis*. Convex Analysis. Princeton University Press, 1997. ISBN 9780691015866.

S. Shalev-Shwartz. Online Learning and Online Convex Optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, Feb 2012.

M. K. Simon. *Probability distributions involving Gaussian random variables: A handbook for engineers and scientists*. Springer Science & Business Media, 2002.

N. Srebro, K. Sridharan, and A. Tewari. On the Universality of Online Mirror Descent. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 2645–2653, 2011.

G. Van den Broeck, K. Driessens, and J. Ramon. Monte-Carlo tree search in poker using expected reward distributions. In *Advances in Machine Learning*, pages 367–381. Springer, 2009.

T. van Erven, W. Kotlowski, and M. K. Warmuth. Follow the Leader with Dropout Perturbations. In *Proceedings of The 27th Conference on Learning Theory*, 2014.

M. Warmuth. A perturbation that makes "Follow the leader" equivalent to "Randomized Weighted Majority". `http://classes.soe.ucsc.edu/cmps290c/Spring09/lect/10/wmkalai-rewrite.pdf`, 2009. Accessed: March 19, 2016.

F. Yousefian, A. Nedić, and U. V. Shanbhag. Convex nondifferentiable stochastic optimization: A local randomized smoothing technique. In *Proceedings of American Control Conference (ACC), 2010*, pages 4875–4880, June 2010.