Fighting Bandits with a New Kind of Smoothness

Jacob Abernethy
University of Michigan
jabernet@umich.edu

Chansoo Lee
University of Michigan
chansool@umich.edu

Ambuj Tewari
University of Michigan
tewaria@umich.edu

Abstract

We provide a new analysis framework for the adversarial multi-armed bandit problem. Using the notion of convex smoothing, we define a novel family of algorithms with minimax optimal regret guarantees. First, we show that regularization via the *Tsallis entropy*, which includes EXP3 as a special case, matches the $O(\sqrt{NT})$ minimax regret with a smaller constant factor. Second, we show that a wide class of perturbation methods achieve a near-optimal regret as low as $O(\sqrt{NT \log N})$, as long as the perturbation distribution has a bounded hazard function. For example, the Gumbel, Weibull, Frechet, Pareto, and Gamma distributions all satisfy this key property and lead to near-optimal algorithms.

1 Introduction

The classic *multi-armed bandit* (MAB) problem, generally attributed to the early work of Robbins (1952), poses a generic online decision scenario in which an agent must make a sequence of choices from a fixed set of options. After each decision is made, the agent receives some feedback in the form of a loss (or gain) associated with her choice, but no information is provided on the outcomes of alternative options. The agent's goal is to minimize the total loss over time, and the agent is thus faced with the balancing act of both experimenting with the menu of choices while also utilizing the data gathered in the process to improve her decisions. The MAB framework is not only mathematically elegant, but useful for a wide range of applications including medical experiments design (Gittins, 1996), automated poker playing strategies (Van den Broeck et al., 2009), and hyperparameter tuning (Pacula et al., 2012).

Early MAB results relied on stochastic assumptions (e.g., IID) on the loss sequence (Auer et al., 2002; Gittins et al., 2011; Lai and Robbins, 1985). As researchers began to establish non-stochastic, *worst-case* guarantees for sequential decision problems such as *prediction with expert advice* (Littlestone and Warmuth, 1994), a natural question arose as to whether similar guarantees were possible for the bandit setting. The pioneering work of Auer, Cesa-Bianchi, Freund, and Schapire (2003) answered this in the affirmative by showing that their algorithm EXP3 possesses nearly-optimal regret bounds with matching lower bounds. Attention later turned to the bandit version of *online linear optimization*, and several associated guarantees were published the following decade (Abernethy et al., 2012; Dani and Hayes, 2006; Dani et al., 2008; Flaxman et al., 2005; McMahan and Blum, 2004).

Nearly all proposed methods have relied on a particular algorithmic blueprint; they reduce the bandit problem to the full-information setting, while using randomization to make decisions and to *estimate* the losses. A well-studied family of algorithms for the full-information setting is *Follow the Regularized Leader* (FTRL), which optimizes the objective function of the following form:

$$\arg\min_{x \in \mathcal{K}} L^\top x + \lambda R(x) \qquad (1)$$

where $\mathcal{K}$ is the decision set, L is (an estimate of) the cumulative loss vector, and R is a *regularizer*, a convex function with suitable curvature to stabilize the objective. The choice of regularizer R is

1

critical to the algorithm's performance. For example, the EXP3 algorithm (Auer, 2003) regularizes with the *entropy function* and achieves a nearly optimal regret bound when $\mathcal{K}$ is the probability simplex. For a general convex set, however, other regularizers such as *self-concordant barrier functions* (Abernethy et al., 2012) have tighter regret bounds.

Another class of algorithms for the full information setting is *Follow the Perturbed Leader* (FTPL) (Kalai and Vempala, 2005) whose foundations date back to the earliest work in adversarial online learning (Hannan, 1957). Here we choose a distribution $\mathcal{D}$ on $\mathbb{R}^N$, sample a random vector $Z \sim \mathcal{D}$, and solve the following *linear* optimization problem

$$\arg\min_{x \in \mathcal{K}} (L + Z)^\top x. \tag{2}$$

FTPL is computationally simpler than FTRL due to the linearity of the objective, but it is analytically much more complex due to the randomness. For every different choice of $\mathcal{D}$, an entirely new set of techniques had to be developed (Devroye et al., 2013; Van Erven et al., 2014). Rakhlin et al. (2012) and Abernethy et al. (2014) made some progress towards unifying the analysis framework. Their techniques, however, are limited to the full-information setting.

In this paper, we propose a new analysis framework for the multi-armed bandit problem that unifies the regularization and perturbation algorithms. The key element is a new kind of smoothness property, which we call *differential consistency*. It allows us to generate a wide class of both optimal and near-optimal algorithms for the adversarial multi-armed bandit problem. We summarize our main results:

1. We show that regularization via the *Tsallis entropy* leads to the state-of-the-art adversarial MAB algorithm, matching the minimax regret rate of Audibert and Bubeck (2009) with a tighter constant. Interestingly, our algorithm fully generalizes EXP3.

2. We show that a wide array of well-studied noise distributions lead to near-optimal regret bounds (matching those of EXP3). Furthermore, our analysis reveals a strikingly simple and appealing sufficient condition for achieving $O(\sqrt{T})$ regret: the *hazard rate* function of the noise distribution must be bounded by a constant. We conjecture that this requirement is in fact both necessary and sufficient.

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

Let us now introduce the adversarial multi-armed bandit problem. On each round $t = 1, \ldots, T$, a learner must choose a distribution $p_t \in \Delta_N$ over the set of N available actions. The adversary (Nature) chooses a vector $g_t \in [-1, 0]^N$ of losses, the learner samples $i_t \sim p_t$, and plays action i_t. After selecting this action, the learner observes only the value g_{t,i_t}, and receives no information as to the values $g_{t,j}$ for $j \neq i_t$. This limited information feedback is what makes the bandit problem much more challenging than the full-information setting in which the entire g_t is observed.

The learner's goal is to minimize the *regret*. Regret is defined to be the difference in the realized loss and the loss of the best fixed action in hindsight:

$$\text{Regret}_T := \max_{i \in [N]} \sum_{t=1}^{T} (g_{t,i} - g_{t,i_t}). \tag{3}$$

To be precise, we consider the *expected* regret, where the expectation is taken with respect to the learner's randomization.

Loss vs. Gain Note: We use the term "loss" to refer to g, although the maximization in (3) would imply that g should be thought of as a "gain" instead. We use the former term, however, as we impose the assumption that $g_t \in [-1, 0]^N$ throughout the paper.

2.1   The Gradient-Based Algorithmic Template

Our results focus on a particular algorithmic template described in Framework 1, which is a slight variation of the Gradient Based Prediction Algorithm (GBPA) of Abernethy et al. (2014). Note that

the algorithm (i) maintains an *unbiased estimate* of the cumulative losses $\hat{G}_t$, (ii) updates $\hat{G}_t$ by adding a single round estimate $\hat{g}_t$ that has only one non-zero coordinate, and (iii) uses the gradient of a convex function $\tilde{\Phi}$ as sampling distribution p_t. The choice of $\tilde{\Phi}$ is flexible but $\tilde{\Phi}$ must be a differentiable convex function and its derivatives must always be a probability distribution.

Framework 1 may appear restrictive but it has served as the basis for much of the published work on adversarial MAB algorithms (Auer et al., 2003; Kujala and Elomaa, 2005; Neu and Bartók, 2013). First, the GBPA framework essentially encompasses all FTRL and FTPL algorithms (Abernethy et al., 2014), which are the core techniques not only for the full information settings, but also for the bandit settings. Second, the estimation scheme ensures that $\hat{G}_t$ remains an unbiased estimate of G_t. Although there is some flexibility, any unbiased estimation scheme would require some kind of inverse-probability scaling—information theory tells us that the unbiased estimates of a quantity that is observed with only probabilty p must necessarily involve fluctuations that scale as $O(1/p)$.

Framework 1: Gradient-Based Prediction Alg. (GBPA) Template for Multi-Armed Bandit

---

GBPA($\tilde{\Phi}$): $\tilde{\Phi}$ is a differentiable convex function such that $\nabla\tilde{\Phi} \in \Delta^N$ and $\nabla_i\tilde{\Phi} > 0$ for all i.

Initialize $\hat{G}_0 = 0$

for *t = 1 to T* **do**

> **Nature:** A loss vector $g_t \in [-1, 0]^N$ is chosen by the Adversary
> **Sampling:** Learner chooses i_t according to the distribution $p(\hat{G}_{t-1}) = \nabla\Phi_t(\hat{G}_{t-1})$
> **Cost:** Learner "gains" loss g_{t,i_t}
> **Estimation:** Learner "guesses" $\hat{g}_t := \frac{g_{t,i_t}}{p_{i_t}(\hat{G}_{t-1})}\mathbf{e}_{i_t}$
>
> **Update:** $\hat{G}_t = \hat{G}_{t-1} + \hat{g}_t$

Lemma 2.1. *Define $\Phi(G) \equiv \max_i G_i$ so that we can write the expected regret of GBPA($\tilde{\Phi}$) as*

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

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

$$\mathbb{E}\text{Regret}_T \leq \underbrace{\tilde{\Phi}(0) - \Phi(0)}_{\text{overestimation penalty}} + \mathbb{E}_{i_1,\dots,i_{t-1}}\left[\underbrace{\Phi(\hat{G}_T) - \tilde{\Phi}(\hat{G}_T)}_{\text{underestimation penalty}} + \sum_{t=1}^T \underbrace{\mathbb{E}_{i_t}[D_{\tilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1})|\hat{G}_{t-1}]}_{\text{divergence penalty}}\right],$$

(4)

where the expectations are over the sampling of i_t.

Proof. Let $\tilde{\Phi}$ be a valid convex function for the GBPA. Consider GBPA($\tilde{\Phi}$) being 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-NE($\tilde{\Phi}$), which is GBPA($\tilde{\Phi}$) 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

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

and $\Phi(G_T) \leq \Phi(\hat{G}_T)$ by the convexity of Φ. Hence, it suffices to show that the GBPA-NE($\tilde{\Phi}$) has regret at most the righthand side of Equation 4, which is a fairly well-known result in online learning literature; see, for example, (Cesa-Bianchi and Lugosi, 2006, Theorem 11.6) or (Abernethy et al., 2014, Section 2). For completeness, we included the full proof in Appendix A. $\square$

2.2 A New Kind of Smoothness

What has emerged as a guiding principle throughout machine learning is that enforcing *stability* of an algorithm can often lead immediately to performance guarantees—that is, small modifications of the input data should not dramatically alter the output. In the context of GBPA, algorithmic stability is guaranteed as long as the dervative $\nabla\tilde{\Phi}(\cdot)$ is Lipschitz. Abernethy et al. (2014) explored a set of conditions on $\nabla^2\tilde{\Phi}(\cdot)$ that lead to optimal regret guarantees for the full-information setting. Indeed,

this work discussed different settings where the regret depends on an upper bound on either the nuclear norm or the operator norm of this hessian.

In short, regret in the full information setting relies on the *smoothness* of the choice of $\tilde{\Phi}$. In the bandit setting, however, merely a uniform bound on the magnitude of $\nabla^2\tilde{\Phi}$ is insufficient to guarantee low regret; the regret (Lemma 2.1) involves terms of the form $D_{\tilde{\Phi}}(\hat{G}_{t-1} + \hat{g}_t, \hat{G}_{t-1})$, where the incremental quantity $\hat{g}_t$ can scale as large as *the inverse of the smallest probability* of $p(\hat{G}_{t-1})$. What is needed is a stronger notion of the smoothness that bounds $\nabla^2\tilde{\Phi}$ in correspondence with $\nabla\tilde{\Phi}$, and we propose the following definition:

Definition 2.2 (Differential Consistency). *For constants $\gamma, C > 0$, we say that a convex function $\tilde{\Phi}(\cdot)$ is (γ, C)-differentially-consistent if for all $G \in (-\infty, 0]^N$,*

$$\nabla_{ii}^2 \tilde{\Phi}(G) \leq C(\nabla_i\tilde{\Phi}(G))^\gamma.$$

We now prove a useful bound that emerges from differential consistency, and in the following two sections we shall show how this leads to regret guarantees.

Theorem 2.3. *Suppose $\tilde{\Phi}$ is (γ, C)-differentially-consistent for constants $C, \gamma > 0$. Then divergence penalty at time t in Lemma 2.1 can be upper bounded as:*

$$\mathbb{E}_{i_t}[D_{\tilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1})|\hat{G}_{t-1}] \leq C \sum_{i=1}^{N} \left(\nabla_i\tilde{\Phi}(\hat{G}_{t-1})\right)^{\gamma-1}.$$

Proof. For the sake of clarity, we drop the subscripts; 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 1, $\hat{g}$ is a sparse vector with one non-zero and non-positive coordinate $\hat{g}_{i_t} = g_{t,i}/\nabla_{i_t}\tilde{\Phi}(\hat{G})$. Plus, i_t is conditionally independent given $\hat{G}$. For a fixed i_t, Let

$$h(r) := D_{\tilde{\Phi}}(\hat{G} + r\hat{g}/\|\hat{g}\|, \hat{G}) = D_{\tilde{\Phi}}(\hat{G} + r\mathbf{e}_{i_t}, \hat{G}),$$

so that $h''(r) = (\hat{g}/\|\hat{g}\|)^\top \nabla^2\tilde{\Phi}\left(\hat{G} + t\hat{g}/\|\hat{g}\|\right)(\hat{g}/\|\hat{g}\|) = \mathbf{e}_{i_t}^\top \nabla^2\tilde{\Phi}\left(\hat{G} - t\mathbf{e}_{i_t}\right)\mathbf{e}_{i_t}$. Now we can write

$$
\begin{aligned}
\mathbb{E}_{i_t}[D_{\tilde{\Phi}}(\hat{G} + \hat{g}, \hat{G})|\hat{G}] &= \sum_{i=1}^N \mathbb{P}[i_t = i] \int_0^{\|\hat{g}\|} \int_0^s h''(r)\, dr\, ds \\
&= \sum_{i=1}^N \nabla_i\tilde{\Phi}(\hat{G}) \int_0^{\|\hat{g}\|} \int_0^s \mathbf{e}_i^\top \nabla^2\tilde{\Phi}\left(\hat{G} - r\mathbf{e}_i\right)\mathbf{e}_i\, dr\, ds \\
&\leq \sum_{i=1}^N \nabla_i\tilde{\Phi}(\hat{G}) \int_0^{\|\hat{g}\|} \int_0^s C\left(\nabla_i\tilde{\Phi}(\hat{G} - r\mathbf{e}_i)\right)^\gamma\, dr\, ds \\
&\leq \sum_{i=1}^N \nabla_i\tilde{\Phi}(\hat{G}) \int_0^{\|\hat{g}\|} \int_0^s C\left(\nabla_i\tilde{\Phi}(\hat{G})\right)^\gamma\, dr\, ds \\
&= C \sum_{i=1}^N \left(\nabla_i\tilde{\Phi}(\hat{G})\right)^{1+\gamma} \int_0^{\|\hat{g}\|} \int_0^s\, dr\, ds \\
&= \frac{C}{2} \sum_{i=1}^N \left(\nabla_i\tilde{\Phi}(\hat{G})\right)^{\gamma-1} g_i^2 \leq C \sum_{i=1}^N \left(\nabla_i\tilde{\Phi}(\hat{G})\right)^{\gamma-1}.
\end{aligned}
$$

The first inequality is by the supposition and the second inequality is due to the convexity of $\tilde{\Phi}$ which guarantees that ∇_i is an increasing function in the i-th coordinate. Interestingly, this part of the proof critically depends on the fact that the we are in the "loss" setting where g is always non-positive. $\qquad\square$

3 A Minimax Bandit Algorithm via Tsallis Smoothing

The design of a multi-armed bandit algorithm in the adversarial setting proved to be a challenging task. Ignoring the dependence on N for the moment, we note that the initial published work on EXP3 provided only an $O(T^{2/3})$ guarantee (Auer et al., 1995), and it was not until the final version of this work (Auer et al., 2003) that the authors obtained the optimal $O(\sqrt{T})$ rate. For the more

general setting of online linear optimization, several sub-optimal rates were achieved (Dani and Hayes, 2006; Flaxman et al., 2005; McMahan and Blum, 2004) before the desired $\sqrt{T}$ was obtained (Abernethy et al., 2012; Dani et al., 2008).

We can view EXP3 as an instance of GBPA where the potential function $\tilde{\Phi}(\cdot)$ is the Fenchel conjugate of the *Shannon entropy*. For any $p \in \Delta_N$, the (negative) Shannon entropy is defined as $H(p) := \sum_i p_i \log p_i$, and its Fenchel conjugate is $H^\star(G) = \sup_{p \in \Delta_N} \{\langle p, G \rangle - \eta H(p)\}$. In fact, we have a closed-form expression for the supremum: $H^\star(G) = \frac{1}{\eta} \log \left(\sum_i \exp(\eta G_i) \right)$. By inspecting the gradient of the above expression, it is easy to see that EXP3 chooses the distribution $p_t = \nabla H^\star(G)$ every round.

The tighter EXP3 bound given by Auer et al. (2003) scaled according to $O(\sqrt{TN \log N})$ and the authors provided a matching lower bound of the form $\Omega(\sqrt{TN})$. It remained an open question for some time whether there exists a minimax optimal algorithm that does not contain the log term until Audibert and Bubeck (2009) proposed the Implicitly Normalized Forecaster (INF). The INF is implicitly defined via a specially-designed potential function with certain properties. It was not immediately clear from this result how to define a minimax-optimal algorithm using the now-standard tools of regularization and Bregman divergence.

More recently, Audibert et al. (2011) improved upon Audibert and Bubeck (2009), extending the results to the combinatorial setting, and they also discovered that INF can be interpreted in terms of Bregman divergences. We give here a reformulation of INF that leads to a very simple analysis in terms of our notion of *differential consistency*. Our reformulation can be viewed as a variation of EXP3, where the key modification is to replace the Shannon entropy function with the *Tsallis entropy*[1] for parameter $0 < \alpha < 1$:

$$S_\alpha(p) = \frac{1}{1-\alpha} \left(1 - \sum p_i^\alpha \right).$$

This particular function, proposed by Tsallis (1988), possesses a number of natural properties. The Tsallis entropy is in fact a generalization of the Shannon entropy, as one obtains the latter as a special case of the former asymptotically. That is, it is easy to prove the following uniform convergence:

$$S_\alpha(\cdot) \to H(\cdot) \qquad \text{as } \alpha \to 1.$$

We emphasize again that one can easily show that Tsallis-smoothing bandit algorithm is indeed *identical* to INF using the appropriate parameter mapping, although our analysis is simpler due to the notion of differential consistency (Definition 2.2).

Theorem 3.1. *Let $\tilde{\Phi}(G) = \max_{p \in \Delta_N} \{\langle p, G \rangle - \eta S_\alpha(p)\}$. Then the GBPA($\tilde{\Phi}$) has regret at most*

$$\mathbb{E}\text{Regret} \leq \eta \frac{N^{1-\alpha} - 1}{1-\alpha} + \frac{N^\alpha T}{\eta \alpha}. \tag{5}$$

Before proving the theorem, we note that it immediately recovers the EXP3 upper bound as a special case $\alpha \to 1$. An easy application of L'Hôpital's rule shows that as $\alpha \to 1$, $\frac{N^{1-\alpha}-1}{1-\alpha} \to \log N$ and $N^\alpha/\alpha \to N$. Choosing $\eta = \sqrt{(N \log N)/T}$, we see that the right-hand side of (5) tends to $2\sqrt{TN \log N}$. However the choice $\alpha \to 1$ is clearly not the optimal choice, as we show in the following statement, which directly follows from the theorem once we see that $N^{1-\alpha} - 1 < N^{1-\alpha}$.

Corollary 3.2. *For any $\alpha \in (0,1)$, if we choose $\eta = \sqrt{\frac{\alpha N^{1-2\alpha}}{(1-\alpha)T}}$ then we have*

$$\mathbb{E}\text{Regret} \leq 2\sqrt{\frac{NT}{\alpha(1-\alpha)}}.$$

In particular, the choice of $\alpha = \frac{1}{2}$ gives a regret of no more than $4\sqrt{NT}$.

Proof of Theorem 3.1. We will bound each penalty term in Lemma 2.1. Since S_α is non-positive, the **underestimation penalty** is upper bounded by 0 and the **overestimation penalty** is at most $(-\min S_\alpha)$. The minimum of S_α occurs at $(1/N, \ldots, 1/N)$. Hence,

$$(\text{overestimation penalty}) \leq -\frac{\eta}{1-\alpha} \left(1 - \sum_{i=1}^{N} \frac{1}{N^\alpha} \right) \leq \eta(N^{1-\alpha} - 1). \tag{6}$$

---

[1] More precisely, the function we give here is the *negative* Tsallis entropy according to its original definition.

Now it remains to upper bound the **divergence penalty** with $(\eta\alpha)^{-1}N^\alpha T$. We observe that straightforward calculus gives $\nabla^2 S_\alpha(p) = \eta\alpha \mathrm{diag}(p_1^{\alpha-2}, \ldots, p_N^{\alpha-2})$. Let $\mathbb{I}_{\Delta_N}(\cdot)$ be the indicator function of Δ_N; that is, $\mathbb{I}_{\Delta_N}(x) = 0$ for $x \in \Delta_N$ and $\mathbb{I}_{\Delta_N}(x) = \infty$ for $x \notin \Delta_N$. It is clear that $\tilde{\Phi}(\cdot)$ is the dual of the function $S_\alpha(\cdot) + \mathbb{I}_{\Delta_N}(\cdot)$, and moreover we observe that $\nabla^2 S_\alpha(p)$ is a *sub-hessian* of $S_\alpha(\cdot) + \mathbb{I}_{\Delta_N}(\cdot)$ at $p(G)$, following the setup of Penot (1994). Taking advantage of Proposition 3.2 in the latter reference, we conclude that $\nabla^{-2} S_\alpha(p(G))$ is a *super-hessian* of $\tilde{\Phi} = S_\alpha^*$ at G. Hence,

$$\nabla^2 \tilde{\Phi}(G) \preceq (\eta\alpha)^{-1}\mathrm{diag}(p_1^{2-\alpha}(G), \ldots, p_N^{2-\alpha}(G))$$

for any G. What we have stated, indeed, is that $\tilde{\Phi}$ is $(2 - \alpha, (\eta\alpha)^{-1})$-differentially-consistent, and thus applying Theorem 2.3 gives

$$D_{\tilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1}) \leq (\eta\alpha)^{-1}\sum_{i=1}^N \left(p_i(\hat{G}_{t-1})\right)^{1-\alpha}.$$

Noting that the $\frac{1}{\alpha}$-norm and the $\frac{1}{1-\alpha}$-norm are dual to each other, we can apply Hölder's inequality to any probability distribution $p_1, \ldots, p_N$ to obtain

$$\sum_{i=1}^N p_i^{1-\alpha} = \sum_{i=1}^N p_i^{1-\alpha} \cdot 1 \leq \left(\sum_{i=1}^N p_i^{\frac{1-\alpha}{1-\alpha}}\right)^{1-\alpha}\left(\sum_{i=1}^N 1^{\frac{1}{\alpha}}\right)^{\alpha} = (1)^{1-\alpha}N^\alpha = N^\alpha.$$

So, the divergence penalty is at most $(\eta\alpha)^{-1}N^\alpha$, which completes the proof. $\square$

4   Near-Optimal Bandit Algorithms via Stochastic Smoothing

Let $\mathcal{D}$ be a continuous distribution over an unbounded support with probability density function f and cumulative density function F. Consider the GBPA($\tilde{\Phi}(G; \mathcal{D})$) where

$$\tilde{\Phi}(G; \mathcal{D}) = \mathbb{E}_{Z_1, \ldots, Z_N \overset{\text{iid}}{\sim} \mathcal{D}} \max_i \{G_i + Z_i\}$$

which is a *stochastic smoothing* of $(\max_i G_i)$ function. Since the max function is convex, $\tilde{\Phi}$ is also convex. By Bertsekas (1973), we can swap the order of differentiation and expectation:

$$\tilde{\Phi}(G; \mathcal{D}) = \mathbb{E}_{Z_1, \ldots, Z_N \overset{\text{iid}}{\sim} \mathcal{D}} e_{i^*}, \text{ where } i^* = \underset{i=1,\ldots,N}{\arg\max}\{G_i + Z_i\}. \tag{7}$$

Even if the function is not differentiable everywhere, the swapping is still possible with any subgradient as long as they are bounded. Hence, the ties between coordinates (which happen with probability zero anyways) can be resolved in an arbitrary manner. It is clear that $\nabla\tilde{\Phi}$ is in the probability simplex, and note that

$$\frac{\partial \tilde{\Phi}}{\partial G_i} = \mathbb{E}_{Z_1, \ldots, Z_N} \mathbf{1}\{G_i + Z_i > G_j + Z_j, \forall j \neq i\}$$

$$= \mathbb{E}_{\tilde{G}_{j^*}}[\mathbb{P}_{Z_i}[Z_i > \tilde{G}_{j^*} - G_i]] = \mathbb{E}_{\tilde{G}_{j^*}}[1 - F(\tilde{G}_{j^*} - G_i)] \tag{8}$$

where $\tilde{G}_{j^*} = \max_{j\neq i} G_j + Z_j$. The unbounded support condition guarantees that this partial derivative is non-zero for all i given any G. So, $\tilde{\Phi}(G; \mathcal{D})$ satisfies the requirements of Algorithm 1.

4.1   Connection to Follow the Perturbed Leader

There is a straightforward way to efficiently implement the sampling step of the bandit GBPA (Algorithm 1) with a stochastically smoothed function. Instead of evaluating the expectation of Equation 7, we simply take a random sample. In fact, this is equivalent to **Follow the Perturbed Leader Algorithm (FTPL)** (Kalai and Vempala, 2005) for bandit settings. On the other hand, implementing the estimation step is hard because generally there is no closed-form expression for $\nabla\tilde{\Phi}$.

To address this issue, Neu and Bartók (2013) proposed Geometric Resampling (GR). GR uses an iterative resampling process to estimate $\nabla_i\tilde{\Phi}$. This process gives an unbiased estimate when allowed

to run for an unbounded number of iterations. Even when we truncate the resampling process after M iterations, the extra regret due to the estimation bias is at most $\frac{NT}{eM}$ (additive term). Since the lower bound for the multi-armed bandit problem is $O(\sqrt{NT})$, any choice of $M = O(\sqrt{NT})$ does not affect the asymptotic regret of the algorithm. In summary, all our GBPA regret bounds in this section hold for the corresponding FTPL algorithm with an extra additive $\frac{NT}{eM}$ term in the bound.

Despite the fact that perturbation-based 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: the EXP3 algorithm, for example, 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?

4.2 Hazard Rate analysis

In this section, we show that the performance of the GBPA($\tilde{\Phi}(G; \mathcal{D})$) 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 4.1 (Hazard rate function)**.** *Hazard rate function of a distribution $\mathcal{D}$ is*

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

For the rest of the section, we assume that $\mathcal{D}$ is unbounded in the direction of $+\infty$, so that the hazard function is well-defined everywhere. This assumption is for the clarity of presentation and can be easily removed (Appendix B).

Theorem 4.2. *The regret of the GBPA on $\tilde{\Phi}(L) = \mathbb{E}_{Z_1,\ldots,Z_n \sim D} \max_i \{G_i + \eta Z_i\}$ is at most:*

$$\frac{N(\sup h_{\mathcal{D}})}{\eta} T + \eta \mathbb{E}_{Z_1,\ldots,Z_n \sim D} \left[\max_i Z_i \right]$$

Proof. We analyze each penalty term in Lemma 2.1. Due to the convexity of Φ, the underestimation penalty is non-positive. The overestimation penalty is clearly at most $\mathbb{E}_{Z_1,\ldots,Z_n \sim D}[\max_i Z_i]$, and Lemma 4.3 proves the $N(\sup h_{\mathcal{D}})$ upper bound on the divergence penalty.

It remains to provide 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 $\mathbb{E}[\max_{i=1,\ldots,n} \eta X_i] = \eta \mathbb{E}[\max_{i=1,\ldots,n} X_i]$. 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. $\square$

Lemma 4.3. *The divergence penalty of the GBPA with $\tilde{\Phi}(G) = \mathbb{E}_{Z \sim \mathcal{D}} \max_i \{G_i + Z_i\}$ is at most $N(\sup h_{\mathcal{D}})$ each round.*

Proof. Recall the gradient expression in Equation 8. The i-th diagonal entry of the Hessian is:

$$\nabla_{ii}^2 \tilde{\Phi}(G) = \frac{\partial}{\partial G_i} \mathbb{E}_{\tilde{G}_{j^*}}[1 - F(\tilde{G}_{j^*} - G_i)] = \mathbb{E}_{\tilde{G}_{j^*}} \left[\frac{\partial}{\partial G_i}(1 - F(\tilde{G}_{j^*} - G_i)) \right] = \mathbb{E}_{\tilde{G}_{j^*}} f(\tilde{G}_{j^*} - G_i)$$

$$= \mathbb{E}_{\tilde{G}_{j^*}}[h(\tilde{G}_{j^*} - G_i)(1 - F(\tilde{G}_{j^*} - G_i))] \tag{9}$$

$$\leq (\sup h) \mathbb{E}_{\tilde{G}_{j^*}}[1 - F(\tilde{G}_{j^*} - G_i)]$$

$$= (\sup h) \nabla_i(G)$$

where $\tilde{G}_{j^*} = \max_{j \neq i}\{G_j + Z_j\}$ which is a random variable independent of Z_i. We now apply Theorem 2.3 with $\gamma = 1$ and $C = (\sup h)$ to complete the proof. $\square$

7

Distribution	$\sup_x h_{\mathcal{D}}(x)$	$\mathbb{E}[\max_{i=1}^N Z_i]$	$O(\sqrt{TN \log N})$ Param.
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 $x = 0$	$O((\frac{1}{k})!(\log N)^{\frac{1}{k}})$	$k = 1$ (Exponential)
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 $x \to \infty$	$\log N + (\alpha - 1)\log\log N - \log\Gamma(\alpha) + \beta^{-1}\gamma_0$	$\beta = \alpha = 1$ (Exponential)

Table 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. Distributions marked with (*) need to be slightly modified using the conditioning trick explained in Appendix B.2. The maximum of Frechet hazard function has to be computed numerically (Elsayed, 2012, p. 47) but elementary calculations show that it is bounded by 2α (Appendix D).

Corollary 4.4. *Follow the Perturbed Leader Algorithm with distributions in Table 1 (restricted to a certain range of parameters), combined with Geometric Resampling (Section 4.1) with $M = \sqrt{NT}$, has an expected regret of order $O(\sqrt{TN \log N})$.*

Table 1 provides the two terms we need to bound. We derive the third column of the table in Appendix C using Extreme Value Theory (Embrechts et al., 1997). Note that our analysis in the proof of Lemma 4.3 is quite tight; the only place we have an inequality is when we upper bound the hazard rate. It is thus reasonable to pose the following conjecture:

Conjecture 4.5. *If a distribution $\mathcal{D}$ has a monotonically increasing hazard rate $h_{\mathcal{D}}(x)$ that does not converge as $x \to +\infty$ (e.g., Gaussian), then there is a sequence of losses that will incur at least a linear regret.*

The intuition is that if adversary keeps incurring a high loss for the i-th arm, then with high probability $\tilde{G}_{j^*} - G_i$ will be large. So, the expectation in Equation 9 will be dominated by the hazard function evaluated at large values of $\tilde{G}_{j^*} - G_i$.

References

J. Abernethy, E. Hazan, and A. Rakhlin. Interior-point methods for full-information and bandit online learning. *IEEE Transactions on Information Theory*, 58(7):4164–4175, 2012.

J. Abernethy, C. Lee, A. Sinha, and A. Tewari. Online linear optimization via smoothing. In *COLT*, pages 807–823, 2014.

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.

J.-Y. Audibert, S. Bubeck, and G. Lugosi. Minimax policies for combinatorial prediction games. In *COLT*, 2011.

P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *The Journal of Machine Learning Research*, 3:397–422, 2003.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-arm bandit problem. In *FOCS*, 1995.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

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.

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

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

V. Dani and T. P. Hayes. Robbing the bandit: less regret in online geometric optimization against an adaptive adversary. In *SODA*, pages 937–943, 2006.

V. Dani, T. Hayes, and S. Kakade. The price of bandit information for online optimization. In *NIPS*, 2008.

L. Devroye, G. Lugosi, and G. Neu. Prediction by random-walk perturbation. In *Conference on Learning Theory*, pages 460–473, 2013.

E. Elsayed. *Reliability Engineering*. Wiley Series in Systems Engineering and Management. Wiley, 2012. ISBN 9781118309544. URL https://books.google.com/books?id=NdjF5G6tfLQC.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events: For Insurance and Finance*. Applications of mathematics. Springer, 1997. ISBN 9783540609315. URL https://books.google.com/books?id=BXOI2pICfJUC.

A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *SODA*, pages 385–394, 2005. ISBN 0-89871-585-7.

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

J. Gittins, K. Glazebrook, and R. Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

J. Hannan. Approximation to bayes risk in repeated play. *In M. Dresher, A. W. Tucker, and P. Wolfe, editors, Contributions to the Theory of Games, volume III*, pages 97–139, 1957.

A. 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 *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.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994. ISSN 0890-5401.

H. B. McMahan and A. Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *COLT*, pages 109–123, 2004.

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. OReilly. Hyperparameter tuning in bandit-based adaptive operator selection. In *Applications of Evolutionary Computation*, pages 73–82. Springer, 2012.

J.-P. Penot. Sub-hessians, super-hessians and conjugation. *Nonlinear Analysis: Theory, Methods & Applications*, 23(6):689–702, 1994. URL http://www.sciencedirect.com/science/article/pii/0362546X94902127.

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.

H. Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58(5):527–535, 1952.

C. Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 52(1-2):479–487, 1988.

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 *COLT*, 2014.

Algorithm 2: Gradient-Based Prediction Algorithm (GBPA) for Full Information Setting

---

Input: $\tilde{\Phi}$, a differentiable convex function such that $\nabla\tilde{\Phi} \in \Delta^N$ and $\nabla_i\tilde{\Phi} > 0$ for all i.
Initialize $G_0 = 0$
for $t = 1$ to T **do**

    Sampling: The learner chooses arm i_t with probability $p_i(\hat{G}_{t-1}) = \nabla_i\Phi_t(\hat{G}_{t-1})$
    Adversary chooses a loss vector $g_t \in [-1,0]^N$ and learner pays $g_{t,i}$
    Update $G_t = G_{t-1} + g_t$

A Proof of the GBPA Regret Bound (Lemma 2.1)

Lemma A.1. *The expected regret of Algorithm 2 can be written as:*

$$\mathbb{E}\text{Regret} = \underbrace{\tilde{\Phi}(0) - \Phi(0)}_{\text{overestimation penalty}} + \underbrace{\Phi(G_T) - \tilde{\Phi}(G_T)}_{\text{underestimation penalty}} + \sum_{t=1}^{T}\underbrace{D_{\tilde{\Phi}}(G_t, G_{t-1})}_{\text{divergence penalty}}$$

Proof. Note that since $\Phi_0(0) = 0$,

$$\tilde{\Phi}(G_T) = \underbrace{\left(\tilde{\Phi}(0) - \Phi_0(0)\right)}_{\text{overestimation penalty}} + \sum_{t=1}^{T}\tilde{\Phi}(G_t) - \tilde{\Phi}(G_{t-1})$$

$$= \underbrace{\left(\tilde{\Phi}(0) - \Phi_0(0)\right)}_{\text{overestimation penalty}} + \sum_{t=1}^{T}\langle\nabla\tilde{\Phi}(G_{t-1}), \ell_t\rangle + D_{\tilde{\Phi}}(G_t, G_{t-1})$$

Therefore,

$$\mathbb{E}\text{Regret} \overset{\text{def}}{=} \mathbb{E}\left[\Phi(G_T) - \sum_{t=1}^{T}\langle\tilde{\Phi}(G_{t-1}), g_t\rangle\right]$$

$$= \mathbb{E}\left[\underbrace{\left(\Phi(G_T) - \tilde{\Phi}(G_T)\right)}_{\text{underestimation penalty}} + \tilde{\Phi}(G_T) - \sum_{t=1}^{T}\langle\tilde{\Phi}(G_{t-1}), g_t\rangle\right]$$

$$= \mathbb{E}\left[\underbrace{\left(\Phi(G_T) - \tilde{\Phi}(G_T)\right)}_{\text{underestimation penalty}} + \underbrace{\left(\tilde{\Phi}(0) - \Phi_0(0)\right)}_{\text{overestimation penalty}} + D_{\tilde{\Phi}}(G_t, G_{t-1})\right]$$

$\square$

B Relaxing Assumptions on the Distribution

B.1 Mirroring trick for extending the support

Let X have support on $x > 0$ with density f and CDF F. Let us define Y by mirroring the density of X around zero, i.e., Y has density $g(y) = \frac{1}{2}f(|y|)$ and CDF $G(y) = \frac{1}{2}\left(1 + \text{sign}(y)F(|y|)\right)$. Note that $|Y|$ is distributed as X and hence,

$$\mathbb{E}[\max_i Y_i] \leq \mathbb{E}[\max_i |Y_i|] = \mathbb{E}[\max_i X_i].$$

The hazard $h_Y(y)$ for $y \geq 0$ is $f(y)/(1-F(y))$ and for $y < 0$ is $f(-y)/(1+F(-y)) \leq F(-y)/(1-F(-y))$. Therefore,

$$\sup_y h_Y(y) = \sup_{x>0} h_X(x).$$

This proves the following lemma.

Lemma B.1. *If a random variable X has support on the non-negative reals with density $f(x)$ and we define Y as the mirrored version with density $g(y) = \frac{1}{2}f(|y|)$. Then, we have*

$$\mathbb{E}[\max_i Y_i] \leq \mathbb{E}[\max_i X_i],$$

$$\sup_y h_Y(y) = \sup_{x>0} h_X(x)$$

where h_X, h_Y are hazard rates of X, Y respectively.

B.2 Conditioning trick for unbounded hazard rate near zero

Suppose $F(x)$ is the CDF of a random variable X whose hazard rate is bounded for $x \geq 1$ but blows up near zero. Then define Y as X conditioned on $X \geq 1$. That is, Y has CDF, for $y > 0$:

$$G(y) = P(X \geq 1 + y | X > 1) = \frac{F(1+y) - F(1)}{1 - F(1)}$$

and density $g(y) = f(1+y)/(1 - F(1)), y > 0$. So the hazard rate $h_Y(y)$ is $g(y)/(1 - G(y)) = f(1+y)/(1 - F(1+y)) = h_X(1+y)$. Therefore,

$$\sup_{y>0} h_Y(y) = \sup_{x>1} h_X(x)$$

which makes the hazard rate of Y now bounded. This we have proved the lemma below.

Lemma B.2. *If a hazard function of X is bounded for $x > 1$ and blows up only for small values of x then we can condition on $X > 1$ to define a new random variable whose hazard rate is now bounded.*

The same technique can be applied for any arbitrary constant other than 1, but for the family of random variables we considered, it suffices to condition on $X \geq 1$.

C Detailed derivation of extreme value behavior

C.1 Maximum of iid Gumbel

The CDF of the Gumbel distribution is $\exp(-\exp(-x))$ and the expected value is γ_0, the Euler (Euler-Mascheroni) constant. Thus, the CDF of the maximum of n iid Gumbel random variables is $(\exp(-\exp(-x)))^N = \exp(-\exp(-(x - \log N)))$ which is also Gumbel but with the mean increased by $\log N$.

C.2 Maximum of iid Frechet

The CDF of Frechet is $\exp(-x^{-\alpha})$ and it has mean $\Gamma(1 - \frac{1}{\alpha})$ as long as $\alpha > 1$ (otherwise it is infinite). Hence, the CDF of the maximum of N iid Frechet random variables is

$$(\exp(-x^{-\alpha}))^N = \exp(-Nx^{-\alpha}) = \exp\left(-\left(\frac{x}{N^{\frac{1}{\alpha}}}\right)^{-\alpha}\right)$$

which is also Frechet but with mean scaled by $N^{1/\alpha}$.

C.3 Maximum of iid Weibull

Let X_i have modified Weibull distribution with CDF $1 - \exp(-(x+1)^k + 1)$. Thus, $P(\max_i X_i > t) \leq NP(X_1 > t) = N\exp(-(t+1)^k + 1)$. For non-negative random variable X and any $u > 0$, we have,

$$\mathbb{E}[X] = \int_0^\infty P(X > x)dx \leq u + \int_u^\infty P(X > x)dx.$$

11

Assume $k = 1/m$ where $m \geq 1$ is a positive integer. Therefore,

$$\mathbb{E}[\max_i X_i] \leq u + \int_u^\infty N \exp(-(x+1)^k + 1)dx$$

$$\leq u + 3N \int_u^\infty \exp(-(x+1)^k)dx$$

$$= u + 3N \int_{u+1}^\infty \exp(-x^{1/m})dx$$

$$= u + 3Nm\Gamma(m, (1+u)^{1/m})dx$$

where $\Gamma(m, x)$ is the incomplete Gamma function that for a positive integer m and $x > 1$ simplifies to

$$\Gamma(m, x) = (m-1)!e^{-x} \sum_{k=0}^{m-1} \frac{x^k}{k!} \leq (m-1)!e^{-x} \sum_{k=0}^{m-1} \frac{x^m}{k!}$$

$$= (m-1)!e^{-x}x^m \sum_{k=0}^{m-1} \frac{1}{k!} \leq (m-1)!e^{-x}x^m \sum_{k=0}^{\infty} \frac{1}{k!}$$

$$\leq 3(m-1)!e^{-x}x^m.$$

Plugging this back above, we get, for any $u > 0$,

$$\mathbb{E}[\max_i X_i] \leq u + 9Nm!e^{-(1+u)^{1/m}}(1+u).$$

Now choose $u = \log^m N + 1$ to get

$$\mathbb{E}[\max_i X_i] \leq \log^m N + 9Nm!\frac{\log^m N}{N} \leq 10\,m!\log^m N.$$

C.4   Maximum of iid Gamma

Let Y be the maximum of N iid $\text{Gamma}(\alpha, \beta)$ ramdom variables. Then, $\frac{Y-d_N}{c_N}$ follows Gumbel distribution, where $c_N = \beta^{-1}$ and $d_N = \beta^{-1}(\log N + (\alpha - 1) \log \log N - \log \Gamma(\alpha))$. In the language of extreme value theory, Gamma distribution belongs to the maximum domain of attraction of Gumbel distribution with parameters (Embrechts et al., 1997). As mentioned in Section C.1, Gumbel distribution has mean γ_0.

C.5   Maximum of iid Pareto

Let X_i have modified Pareto distribution with CDF $1 - 1/(1 + x)^\alpha$. Thus, $P(\max_i X_i > t) \leq NP(X_1 > t) = N/(1 + x)^\alpha$. For non-negative random variable X and any $u > 0$, we have,

$$\mathbb{E}[X] = \int_0^\infty P(X > x)dx \leq u + \int_u^\infty P(X > x)dx.$$

Therefore, for $\alpha > 1$,

$$\mathbb{E}[\max_i X_i] \leq u + \int_u^\infty \frac{N}{(1+x)^\alpha}dx$$

$$= u + \frac{N}{(\alpha - 1)(1+u)^{\alpha-1}}.$$

Setting $u = N^{1/\alpha} - 1$ gives the bound

$$\mathbb{E}[\max_i X_i] \leq \frac{\alpha}{\alpha - 1}N^{1/\alpha}.$$

D Hazard Functions of Modified Distributions and the Frechet Case

D.1 Pareto distribution

Using the conditioning trick, we consider, for $\alpha > 1$ (otherwise mean is infinite), the modified Pareto distribution with pdf $f(x) = \frac{\alpha}{(x+1)^{\alpha+1}}$ supported on $(0, \infty)$. Its CDF is $1 - 1/(x+1)^\alpha$. Its hazard function is $h(x) = \frac{\alpha}{x+1}$ which decreases in x and is bounded by α. Expected maximum of N iid Pareto random variables is bounded by $\alpha N^{1/\alpha}/(\alpha - 1)$ (see Appendix C.5). This gives a regret bound of $\sqrt{NT}\sqrt{\alpha^2 N^{1/\alpha}/(\alpha - 1)}$.

D.2 Frechet distribution

The CDF of Frechet is $\exp(-x^{-\alpha}), x > 0$ where $\alpha > 0$ is a shape parameter. The hazard function of Frechet distribution is $h(x) = \alpha x^{-\alpha-1} \frac{\exp(-x^{-\alpha})}{1 - \exp(-x^{-\alpha})}$ which is hard to optimize analytically but can be upper bounded, for $\alpha > 1$, via elementary calculations given below, by 2α. The CDF of the maximum of N iid Frechet random variables is $\exp(-(x/N^{1/\alpha})^{-\alpha})$ which is also Frechet (but with mean scaled by $N^{1/\alpha}$) with expected value $N^{1/\alpha}\Gamma(1 - \frac{1}{\alpha})$ (as long as $\alpha > 1$, otherwise expectation is infinite). Thus, the regret bound we get is $O\left(\sqrt{NT}\sqrt{\alpha N^{1/\alpha}\Gamma(1 - \frac{1}{\alpha})}\right)$. Setting $\alpha = \log N$ makes the regret bound $O(\sqrt{TN \log N})$. Our choice of α is larger than 1 as soon as $N > 2$.

D.2.1 Elementary calculations for bounding Frechet distribution's hazard rate

For $\alpha > 1$, we want to show that $\sup_{x>0} h(x) \le 2\alpha$ where

$$h(x) = \alpha x^{-\alpha-1} \frac{\exp(-x^{-\alpha})}{1 - \exp(-x^{-\alpha})}.$$

First, consider the case $x \ge 1$. In this case, define $y = x^\alpha$ and note that $y \ge 1$. Then, we have

$$h(x) = \frac{\alpha}{xy} \frac{\exp(-1/y)}{1 - \exp(-1/y)} \le \frac{\alpha}{y} \frac{\exp(-1/y)}{1 - \exp(-1/y)}$$
$$\le \frac{\alpha}{y} \frac{1}{1 - (1 - 1/(2y))} = 2\alpha.$$

The first inequality holds because $x \ge 1$. The second holds because $\exp(-1/y) < 1$ and $\exp(-1/y) \le 1 - 1/(2y)$ for $y \ge 1$.

Next, consider the case $x < 1$. Define $y = 1/x$ and note that $y > 1$. Then, we have

$$h(x) = \frac{\alpha}{x^{\alpha+1}} \frac{\exp(-x^{-\alpha})}{1 - \exp(-x^{-\alpha})} \le \frac{\alpha}{x^{\alpha+1}} \frac{\exp(-x^{-\alpha})}{1 - \exp(-1)}$$
$$= \frac{\alpha}{1 - e^{-1}} y^{\alpha+1} \exp(-y^\alpha) \le 2\alpha y^{\alpha+1} \exp(-y^\alpha).$$

To show an upper bound of 2α, it therefore suffices to show that $\sup_{y>1} g(y) \le 1$ where $g(y) = y^{\alpha+1} \exp(-y^\alpha)$. We will show this now. Note that

$$g'(y) = (\alpha + 1)y^\alpha \exp(-y^\alpha) - y^{\alpha+1}\alpha y^{\alpha-1} \exp(-y^\alpha) = y^\alpha \exp(-y^\alpha) \left((\alpha + 1) - \alpha y^\alpha\right),$$

which means that $g(y)$ is monotonically increasing on the interval $(1, y_0)$ and monotonically decreasing on the interval $(y_0, +\infty)$ where $y_0 = \left(\frac{\alpha+1}{\alpha}\right)^{1/\alpha}$. We therefore have,

$$\sup_{y>1} g(y) = g(y_0) = \left(1 + \frac{1}{\alpha}\right)^{(1+1/\alpha)} \exp\left(-(1 + 1/\alpha)\right) \le 2^2 \exp(-2) = 4/e^2 \le 1,$$

where the first inequality above holds because $\alpha > 1$. Note that, for $\alpha > 1$, the function $\alpha \mapsto \left(1 + \frac{1}{\alpha}\right)^{(1+1/\alpha)} \exp\left(-(1 + 1/\alpha)\right)$ decreases monotonically.

D.3 Weibull distribution

The CDF of Weibull is $1 - \exp(-x^k)$ for $x > 0$ (and 0 otherwise) where $k > 0$ is a shape parameter. The density is $kx^{k-1}\exp(-x^k)$ and hazard rate is kx^{k-1}. For $k > 1$, hazard rate monotonically increases and is therefore unbounded for large x. When $k < 1$, the hazard rate is unbounded for small values of x. Note that Weibull includes exponential as a special case when $k = 1$.

Let $k = 1/m$ for some positive integer $m \geq 1$ and using the conditioning trick, consider a modified Weibull with CDF $1 - \exp(-(x+1)^k + 1)$. Density is $k(x+1)^{k-1}\exp(-(x+1)^k + 1)$ and hazard is $k(x+1)^{k-1}$ which is bounded by k. When $k < 1$ we get tails heavier than the exponential but not as heavy as a Pareto or a Frechet. The expected value of the maximum of N iid (modified) Weibull random variables with parameter $k = 1/m$ scales as $O(m!(\log N)^m)$ (see Appendix C.3). Thus, we get the regret bound $O(\sqrt{NT}\sqrt{m!(\log n)^m})$. Thus, the entire modified Weibull family yields $O(\sqrt{N\mathrm{polylog}(N)}\sqrt{T})$ regret bounds. The best bound is obtained when $m = 1$, i.e. when the Weibull becomes an exponential.