Perturbation Algorithms for Adversarial Online Learning

by

Zifan Li

Advisor: Ambuj Tewari

A senior thesis submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Science
(Honors Statistics)
in The University of Michigan
2017

ACKNOWLEDGEMENTS

During my time at the University of Michigan, I have had the fortune to meet and work with many extraordinary people who had made significant impact to my life. I would like to express my sincere gratitude to a few of them here.

First and foremost, I would like to thank my thesis advisor Prof. Ambuj Tewari, for his constant guidance and encouragement over the past two years. Throughout the course of multiple research projects, including ones that lead to this thesis, he was always willing to answer my questions and provide genuine advices, both in academics and in life. Given the numerous setbacks I have encountered writing this thesis, it cannot be more accurate to say that this thesis would not be possible without his dedication and support.

Second, I am very grateful to Prof. Ji Zhu, whose course on data mining inspired my interest in machine learning, for his supervision of some interesting statistical projects and the tremendous help he offered for my graduate school applications.

Third, I would like to recognize Prof. Honglak Lee, Prof. Andrew Snowden, Dr. Yuting Zhang, Kam Chung Wong, Yuan Chen, and all other people who have either supervised me for research or have worked collabratively with me. Working with them has greatly enriched my experience as a researcher and has truely been a pleasure for me.

Lastly, I would also like to thank my parents, my brother, and my friends for their unfailing support over the years, without whom my journey as an undergrduate student at the University of Michigan would not have been so wonderful.

ABSTRACT

Perturbation Algorithms for Adversarial Online Learning

by

Zifan Li

Online learning is an important paradigm of the machine learning, a field where people study ways to let machine achieve better performance by learning from data. One unique property of online learning is that data come in as a stream rather than in a batch. The goal of online learning is to make a sequence of accurate predictions given knowledge of the answers to previous prediction tasks. Online learning has been studied extensively during the past decades. It has also drew great interest to practitioners due to the recent emergence of large scale applications such as online advertisement placement and online web ranking. In this thesis, we approach the problem of online learning from a statistical and game-theoretic perspective. We aim to develop novel perturbation-based algorithms that have guaranteed worst-case performance for a variety of classical online learning problems, including the expert advice problem and the adversarial multi-armed bandit problem.

TABLE OF CONTENTS

CHAPTER I

Introduction

Regret has occupied a central place in online learning literature. In the setting of repeated games played in discrete time, the regret of a player, at any time point, is the difference between the payoffs she would have received had she played the best, in hindsight, constant strategy throughout, and the payoffs she did in fact receive. We approach the problem from an adversarial perspective, i.e, we do not make any stochastic assumption on the payoffs the player would receive, but rather try to design algorithms that have guaranteed performance against any payoff sequences. Hannan [1957] first showed the existence of algorithms with a "no-regret" property under the adversarial setting: algorithms for which the average regret per time goes to zero almost surely as the number of time points increases to infinity. Such algorithms are also called "Hannan Consistent". Unfortunately, the most naive algorithm one can think of, i.e, just choose among the strategies that have the best performance so far (called *Fictitious Play* or *Follow the Leader*), are not "Hannan Consistent". And it is well known that adding smoothness, either explicitly through regularization or implicitly through perturbations, to the cumulative payoffs before computing best response to other players' previous moves is crucial to achieve Hannan consistency. In this thesis, we study perturbation-based algorithms that enjoy the the no-regret

property for two classical online learning problem, i.e, the expert advice problem and the multi-armed bandit problem, through novel techniques such as Littlewood-Offord theory or "generalized hazard rate".

1.1 The Expert Advice Problem

If we think of other players collectively as the adversairial "environment" or "nature", then we arrive at the classical expert advice problem in online learning. Formally, consider the expert advice problem Cesa-Bianchi and Lugosi [2006] with N experts and T rounds. At each round t, the player choose one expert $i_t \in \{1, 2, \ldots, N\}$ and then the adversary reveals a payoff vector g_t where $g_{t,j} \in [-1, 1]$ is the payoff associated with the jth expert at round t and the player accumulates a gain of g_{t,k_t} where k_t (possibly randomized) is the expert the player choosed for round t. We define $G_t = \sum_{s=1}^{t} g_s$ the cumulative payoff vector at time t. The player's goal is to minimize the expected regret, which can be expressed as

$$\mathcal{R}_T = \mathbb{E}\left[\max_{i\in\{1,2,\ldots,N\}} \sum_{t=1}^{T} g_{t,i} - \sum_{t=1}^{T} g_{t,i_t}\right].$$

The fictitious play algorithm (which fails to be hannan consistent) has the simple rule that

$$i_t \in \arg\max_i G_{t,i}.$$

In the thesis, we will consider a variant of fictitious play, namely *sampled fictitious play*. Here, the player samples past time points using some (randomized) sampling scheme and plays best response to the plays of the other players restricted to the set of sampled time points. In other words, at time t, player randomly selects a subset $\mathbb{S}_t \subseteq \{1, \ldots, t-1\}$ of previous time points and plays best response to the other

players' moves only over $\mathbb{S}_t$. That is,

$$i_t \in \arg\max_i \sum_{\tau \in \mathbb{S}_t} g\tau, i$$

We consider *sampled fictitious play* with a natural sampling scheme, namely *Bernoulli sampling*, i.e., any particular round $\tau \in \{1, \ldots, t-1\}$ is included in $\mathbb{S}_t$ independently with probability 1/2. More specifically, if $\epsilon_1^{(t)}, \ldots, \epsilon_{t-1}^{(t)}$ are i.i.d. symmetric Bernoulli (or Rademacher) random variables taking values in $\{-1, +1\}$, then

$$\mathbb{S}_t = \{\tau \in \{1, \ldots, t-1\} \ : \ \epsilon_\tau^{(t)} = +1\}$$

As will become clear later, the sampling scheme can be viewed as a perturbation to the cumulative payoff vector before we compute the best response. In Chapter II, we will show that sampled fictitious play, using Bernoulli sampling, is Hannan consistent. Unlike several existing Hannan consistency proofs that rely on concentration of measure results, ours instead uses anti-concentration results from Littlewood-Offord theory.

1.2 The Multi-armed Bandit Problem

Let us now switch from the expert advice problem to the adversarial multi-armed bandit problem. In this problem, an agent must make a sequence of choices from a fixed set of options, just like in the expert advice probelm. However, what seperates the bandit problem from the expert advice problem is that after each decision is made, the agent receives some feedback associated with her choice, but no information is provided on the outcomes of alternative options. Mathematically, if we use N to denote the number of arms and T to denote the number of rounds. On each round $t = 1, \ldots, T$, a learner must choose a distribution p_t over the set of N available actions. The adversary (nature) chooses a payoff vector $g_t \in [-1, 0]^N$. The learner

plays action k_t sampled according to p_t and suffers the loss g_{t,k_t}. Note that here we use the term "loss" because of the fact that the payoff is restricted to be non-positive. The learner observes only a single coordinate g_{t,k_t} of the payoff vector and receives no information as to the values $g_{t,j}$ for $j \neq k_t$. This limited information feedback is what makes the bandit problem much more challenging than the full-information setting (expert advice problem) in which the entire g_t is observed.

As before, the learner's goal is to minimize the expected regret where the expected regret is defined the same way as before. We want to develop pertubation-based algorithms that have good guaranteed worst case regret. More specifically, we are interested in a class of algorithm called *Follow the Perturbed Leader (FTPL)*, defined as

$$p_t = \mathbb{E}_{Z_1,\ldots,Z_N \overset{\text{iid}}{\sim} \mathcal{D}} e_{i^*}, \text{ where } i^* = \arg\max_{i=1,\ldots,N} \{G_i + Z_i\},$$

where G_i is the cumulative loss vector and $\mathcal{D}$ is some probability distribution. Recent work by Abernethy et al. [2015] has highlighted the role of the hazard rate of the distribution generating the perturbations. Assuming that the hazard rate is bounded, it is possible to provide regret analyses for a variety of FTPL algorithms for the multi-armed bandit problem. In Chapter III, we push the inquiry into regret bounds for FTPL algorithms beyond the bounded hazard rate condition. There are good reasons to do so: natural distributions such as the uniform and Gaussian violate the condition. We give regret bounds for both bounded support and unbounded support distributions without assuming the hazard rate condition. We also disprove a conjecture that the Gaussian distribution cannot lead to a low-regret algorithm. In fact, it turns out that it leads to near optimal regret, up to logarithmic factors. A key ingredient in our approach is the introduction of a new notion called the generalized hazard rate.

1.3 Preprints

Both chapters in the thesis are preprints under review at the time of writing. We list the publication information:

- Prediction with Expert Advice and Sampled Fictitious Play - Under review for the *Games and Economic Behavior* journal (titled **Sampled Fictitous Play is Hannan Consistent**).
- Adversarial Multi-armed Bandit and Follow the Perturbed Leader - Under review for *Conference on Learning Theory 2017* (titled **Beyond the Hazard Rate: More Perturbation Algorithms for Adversarial Multi-armed Bandits**).

CHAPTER II

Prediction with Expert Advice and Sampled Fictitious Play

2.1 Introduction

In the setting of repeated games played in discrete time, the (unconditional) regret of a player, at any time point, is the difference between the payoffs she would have received had she played the best, in hindsight, constant strategy throughout, and the payoffs she did in fact receive. Hannan [1957] showed the existence of procedures with a "no-regret" property: procedures for which the average regret per time goes to zero for a large number of time points. His procedure was a simple modification of fictitious play: random perturbations are added to the cumulative payoffs of every strategy so far and the player picks the strategy with the largest perturbed cumulative payoff. No regret procedures are also called "universally consistent" [Fudenberg and Levine, 1998, Section 4.7] or "Hannan consistent" [Cesa-Bianchi and Lugosi, 2006, Section 4.2].

It is well known that smoothing the cumulative payoffs before computing the best response is crucial to achieve Hannan consistency. One way to achieve smoothness is through stochastic smoothing, or adding perturbations. Without perturbations, the procedure becomes identical to fictitious play, which fails to be Hannan consistent [Cesa-Bianchi and Lugosi, 2006, Exercise 3.8]. Besides Hannan's modification, other

variants of fictitious play are also known to be Hannan consistent, including (unconditional) regret matching, generalized (unconditional) regret matching and smooth fictitious play (for an overview, see Hart and Mas-Colell [2013, Section 10.9]).

In this note, we consider another variant of fictitious play, namely sampled fictitious play. Here, the player samples past time points using some (randomized) sampling scheme and plays the best response to the moves of the other players restricted to the set of sampled time points. Sampled fictitious play has been considered by other authors in the context of evolutionary games [Kaniovski and Young, 1995], the game of matching pennies [Gilliland and Jung, 2006], and games with identical payoffs [Lambert III et al., 2005]. To the best of our knowledge, it is not known whether sampled fictitious play is Hannan consistent. The purpose of this note is to show that it is indeed Hannan consistent when used with a natural sampling scheme, namely Bernoulli sampling.

2.2 Preliminaries

Consider a game in strategic form where M is the number of players, S_i is the set of strategies for player i, and $u_i : \prod_{j=1}^{M} S_i \to \mathbb{R}$ is the payoff function for player i. For simplicity assume that the payoff functions of all players are $[-1, 1]$ bounded. We also assume the number of pure strategies is the same for each player and that $S_i = \{1, \dots, N\}$. Let $S = \prod_{i=1}^{M} S_i$ be the set of M-tuples of player strategies. For $s = (s_i)_{i=1}^{M} \in S$, we denote the strategies of players other than i by $s_{-i} = (s_j)_{1 \le j \le M, j \ne i}$.

The game is played repeatedly over (discrete) time $t = 1, 2, \dots$. A learning procedure for player i is a procedure that maps the history $h_{t-1} = (s_\tau)_{\tau=1}^{t-1}$ of plays just prior to time t, to a strategy $s_{t,i} \in S_i$. The learning procedure is allowed to be randomized, i.e., player i has access to a stream of random variables $\epsilon_1, \epsilon_2, \dots$ and

she is allowed to use $\epsilon_1, \ldots, \epsilon_{t-1}$, in addition to h_{t-1}, to choose $s_{t,i}$. Player i's regret at time t is defined as

$$\mathcal{R}_{t,i} = \max_{k \in S_i} \sum_{\tau=1}^{t} u_i(k, s_{\tau,-i}) - \sum_{\tau=1}^{t} u_i(s_\tau).$$

This compares the player's cumulative payoff with the payoff she could have received had she selected the best constant (over time) strategy k with knowledge of the other players' moves.

A learning procedure for player i is said to be *Hannan consistent* if and only if

$$\limsup_{t \to \infty} \frac{\mathcal{R}_{t,i}}{t} \leq 0 \qquad \text{almost surely.}$$

Hannan consistency is also known as the "no-regret" property and as "universal consistency". The term "universal" refers to the fact that the regret per time goes to zero irrespective of what the other players do.

Fictitious play is a (deterministic) learning procedure where player i plays the best response to the plays of the other players so far. That is,

$$s_{t,i} \in \arg \max_{k \in \{1,\ldots,N\}} \sum_{\tau=1}^{t-1} u_i(k, s_{\tau,-i}).$$

As mentioned earlier, fictitious play is not Hannan consistent. However, consider the following modification of fictitious play, called *sampled fictitious play*. At time t, player randomly selects a subset $\mathbb{S}_t \subseteq \{1, \ldots, t-1\}$ of previous time points and plays the best response to the other players' moves only over $\mathbb{S}_t$. That is,

$$s_{t,i} \in \arg \max_{k \in \{1,\ldots,N\}} \sum_{\tau \in \mathbb{S}_t} u_i(k, s_{\tau,-i}). \tag{2.1}$$

If multiple strategies achieve the maximum, then the tie is broken uniformly at random, and independently with respect to all previous randomness. Also, if $\mathbb{S}_t$

turns out to be empty (an event that happens with probability exactly $2^{-(t-1)}$ under the Bernoulli sampling described below), we adopt the convention that the argmax above includes all N strategies.

In this note, we consider *Bernoulli sampling*, i.e., any particular round $\tau \in \{1, \dots, t-1\}$ is included in $\mathbb{S}_t$ independently with probability $1/2$. More specifically, if $\epsilon_1^{(t)}, \dots, \epsilon_{t-1}^{(t)}$ are i.i.d. symmetric Bernoulli (or Rademacher) random variables taking values in $\{-1, +1\}$, then

$$\mathbb{S}_t = \{\tau \in \{1, \dots, t-1\} \ : \ \epsilon_\tau^{(t)} = +1\} \tag{2.2}$$

and therefore,

$$\sum_{\tau \in \mathbb{S}_t} u_i(k, s_{\tau,-i}) = \sum_{\tau=1}^{t-1} \frac{(1+\epsilon_\tau^{(t)})}{2} u_i(k, s_{\tau,-i}).$$

Note that the procedure defined by the combination of (2.1) and (2.2) is completely parameter free, i.e., there is no tuning parameter that has to be carefully tuned in order to obtain desired convergence properties.

2.3 Result and Discussion

Our main result is the following.

Theorem 1. *Sampled fictitious play* (2.1) *with Bernoulli sampling* (2.2) *is Hannan consistent.*

Before we move on to the proof, a few remarks are in order.

Rate of convergence Our proof gives the rate of convergence of (expected) average regret as $O(N^2\sqrt{\log\log t/t})$ where the constant hidden in $O(\cdot)$ notation is small and explicit. It is known that the optimal rate is $O(\sqrt{\log N/t})$ [Cesa-Bianchi and Lugosi, 2006, Section 2.10]. Therefore, our rate of convergence is almost optimal

in t but severely suboptimal in N. This raises several interesting questions. What is the best bound possible for Sampled Fictitious Play with Bernoulli sampling? Is there a sampling scheme for which Sampled Fictitious Play procedure achieves the optimal rate of convergence? The first question is partially answered by Theorem 32 in Appendix 1.2 which states that the dependency on N is likely to be polynomial instead of logarithmical, but there is still some gap between the lower bound and the upper bound we provide.

Asymmetric probabilities Instead of using symmetric Bernoulli probabilities, we can choose $\epsilon_\tau^{(t)}$ such that $P(\epsilon_\tau^{(t)} = +1) = \alpha$. As $\alpha \to 1$, the learning procedure becomes fictitious play and as $\alpha \to 0$, it selects strategies uniformly at random. Therefore, it is natural to expect that the regret bound will blow up near the two extremes of $\alpha = 1$ and $\alpha = 0$. We can make this intuition precise, but only for $\{-1, 0, 1\}$-valued payoffs (instead of $[-1, 1]$-valued). For details, see Appendix 1.3 in the supplementary material.

Follow the perturbed leader Note that

$$\arg\max_{k\in\{1,\dots,N\}} \sum_{\tau=1}^{t-1} \frac{(1+\epsilon_\tau^{(t)})}{2} u_i(k, s_{\tau,-i}) = \arg\max_{k\in\{1,\dots,N\}} \Big(\sum_{\tau=1}^{t-1} u_i(k, s_{\tau,-i}) + \sum_{\tau=1}^{t-1} \epsilon_\tau^{(t)} u_i(k, s_{\tau,-i}) \Big).$$

Therefore, we can think of sampled fictitious play as adding a random perturbation to the expression that fictitious play optimizes. Such algorithms are referred to as "follow the perturbed leader" (FPL) in the computer science literature ("fictitious play" is known as "follow the leader"). This family was originally proposed by Hannan [1957] and popularized by Kalai and Vempala [2005]. Closer to this paper are the FPL algorithms of Devroye et al. [2013] and van Erven et al. [2014]. However, none of these papers considered sampled fictitious play.

Extension to conditional (or internal) regret In this paper we focus on unconditional (or external) regret. Other notions of regret, especially conditional (or internal) regret can also be considered. Internal regret measures the worst regret, over $N(N-1)$ choices of $k \neq k'$, of the form "every time strategy k was picked, strategy k' should have been picked instead". There are generic conversions [Stoltz and Lugosi, 2005, Blum and Mansour, 2007] that will convert any learning procedure with small external regret to one with small internal regret. These conversion, however, require access to the probability distribution over strategies at each time point. This probability distribution can be approximated, to arbitrary accuracy, by making the choice of the strategy in (2.1) multiple times each time selecting the random subset $\mathbb{S}_t$ independently. However, doing so and using a generic conversion from external to internal regret will lead to a cumbersome overall algorithm. It will be nicer to design a simpler sampling based learning procedure with small internal regret.

2.4 Proof of the Main Result

We break the proof of our main result into several steps. The first and third steps involve fairly standard arguments in this area. Our main innovations are in step two.

2.4.1 Step 1: From Regret to Switching Probabilities

In this step, we assume that players other than player i (the "opponents") are *oblivious*, i.e., they do not adapt to what player i does. Mathematically, this means that the sequence $s_{t,-i}$ does not depend on the moves $s_{t,i}$ of player i. We will prove a uniform regret bound that holds for all deterministic payoff sequences $\{s_{t,-i}\}_{t=1}^{T}$, by which we can conclude that the same bound holds for oblivious but random payoff sequences as well. Since player i is fixed for the rest of the proof, we will not carry the index i in our notation further. Let the vector $g_t \in [-1,1]^N$ be defined as

$g_{t,k} = u_i(k, s_{t,-i})$ for $k \in \{1, \ldots, N\}$. Moreover, we denote player i's move $s_{t,i}$ as k_t. With this notation, regret at time T equals

$$\mathcal{R}_T = \max_{k \in \{1,\ldots,N\}} \sum_{t=1}^{T} g_{t,k} - \sum_{t=1}^{T} g_{t,k_t}.$$

In this step, we will look at the expected regret. Because the opponents are oblivious, this equals

$$\mathbb{E}\left[\mathcal{R}_T\right] = \max_{k \in \{1,\ldots,N\}} \sum_{t=1}^{T} g_{t,k} - \mathbb{E}\left[\sum_{t=1}^{T} g_{t,k_t}\right] = \max_{k \in \{1,\ldots,N\}} \sum_{t=1}^{T} g_{t,k} - \sum_{t=1}^{T} \mathbb{E}\left[g_{t,k_t}\right].$$

Recall that

$$k_t \in \arg \max_{k \in \{1,\ldots,N\}} \sum_{\tau=1}^{t-1} \frac{(1 + \epsilon_\tau^{(t)})}{2} g_{\tau,k}.$$

Since g_t's are fixed vectors, by independence we see that the distribution of k_t is exactly the same whether or not we share the Rademacher random variables across time points. Therefore, we do not have to draw a fresh sample $\epsilon_1^{(t)}, \ldots, \epsilon_{t-1}^{(t)}$ at time t. Instead, we fix a single stream $\epsilon_1, \epsilon_2, \ldots$ of i.i.d. Rademacher random variables and set $(\epsilon_1^{(t)}, \ldots, \epsilon_{t-1}^{(t)}) = (\epsilon_1, \ldots, \epsilon_{t-1})$ for all t. With this reduction in number of random variables used, we now have

$$k_t \in \arg \max_{k \in \{1,\ldots,N\}} \sum_{\tau=1}^{t-1} (1 + \epsilon_\tau) g_{\tau,k}. \tag{2.3}$$

We define $G_t = \sum_{\tau=1}^{t} g_\tau$, the cumulative payoff vector at time t. Define $\tilde{g}_t = (1 + \epsilon_t) g_t$ and $\tilde{G}_t = \sum_{\tau=1}^{t} \tilde{g}_\tau$. We also define

$$g_{t,i \ominus j} = g_{t,i} - g_{t,j} \ , \ \tilde{g}_{t,i \ominus j} = \tilde{g}_{t,i} - \tilde{g}_{t,j}.$$

With these definitions, we have

$$\tilde{G}_{t,i \ominus j} = \tilde{G}_{t,i} - \tilde{G}_{t,j} = \sum_{\tau=1}^{t} \tilde{g}_{\tau,i} - \sum_{\tau=1}^{t} \tilde{g}_{\tau,j}$$
$$= \sum_{\tau=1}^{t} (1 + \epsilon_\tau)(g_{\tau,i} - g_{\tau,j}) = \sum_{\tau=1}^{t} (1 + \epsilon_\tau) g_{\tau,i \ominus j}.$$

The following result upper bounds the regret in terms of downward zero-crossings of the process $\tilde{G}_{t,i \ominus j}$, i.e., the times t when it switches from being non-negative at time $t-1$ to non-positive at time t.

Theorem 2. *We have the following upper bound on the expected regret:*

$$\mathbb{E}\left[\mathcal{R}_T\right] \leq 2N^2 \max_{1\leq i,j\leq N} \sum_{t=1}^{T} |g_{t,i\ominus j}| P\left(\tilde{G}_{t-1,i\ominus j} \geq 0, \tilde{G}_{t,i\ominus j} \leq 0\right).$$

The proof of this theorem can be found in Appendix 1.1. We now focus on bounding the switching probabilities for a fixed pair i, j.

2.4.2 Step 2: Bounding Switching Probabilities Using Littlewood-Offord Theory

Our strategy is to do a "multi-scale" analysis and, within each scale, apply Littlewood-Offord theory to bound the switching probabilities. The need for a multi-scale argument arises from the requirement in Littlewood-Offord theorem (see Theorem 3 below) for a lower bound on the step sizes of random walks. We partition the set of T time points $[T] := \{1, \ldots, T\}$ into $K+1$ disjoint sets at different scales, denoted as $\{A_k\}_{k=0}^{K}$ where

$$A_k = \begin{cases} \{t \in [T] : |g_{t,i\ominus j}| \leq \frac{1}{\sqrt{T}}\} & k = 0 \\ \{t \in [T] : T^{-\frac{1}{2^k}} < |g_{t,i\ominus j}| \leq T^{-\frac{1}{2^{k+1}}}\} & k = 1, \ldots, K-1 \\ \{t \in [T] : T^{-\frac{1}{2^K}} < |g_{t,i\ominus j}| \leq 2\} & k = K \end{cases}.$$

Note that actually A_k depends on i, j as well but for the sake of clarity we drop this dependence in the notation. The cardinality of a finite set A will be denoted by $|A|$. The number $K+1$ of different scales is determined by

$$K = \arg\min\{k \in \mathbb{N} \,:\, T^{-\frac{1}{2^k}} \geq 1/2\}.$$

$\forall t, i,\ g_{t,i} \in [-1, 1]$ so $|g_{t,i\ominus j}| \in [0, 2]$. The scales here are chosen such that K is not very large (of order $O\Big(\log\log(T)\Big)$) and still covers the entire range of the payoffs.

It easily follows that,

$$\begin{aligned}
&\sum_{t=1}^{T} |g_{t,i\ominus j}| P\left(\tilde{G}_{t-1,i\ominus j} \geq 0, \tilde{G}_{t,i\ominus j} \leq 0\right) \\
=&\sum_{t=1}^{T} |g_{t,i\ominus j}| P\left(\sum_{\tau=1}^{t-1} \epsilon_\tau g_{\tau,i\ominus j} \geq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t} \epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j}\right) \\
=&\sum_{k=0}^{K} \sum_{t\in A_k} |g_{t,i\ominus j}| P\left(\sum_{\tau=1}^{t-1} \epsilon_\tau g_{\tau,i\ominus j} \geq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t} \epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j}\right).
\end{aligned}$$

We now want to argue that the probabilities involved above are small. The crucial observation is that, if a switch occurs, then the random sum $\sum_{\tau=1}^{t} \epsilon_\tau g_{\tau,i\ominus j}$ has to lie in a sufficiently small interval. Such "small ball" probabilities are exactly what the classic Littlewood-Offord theorem controls.

Theorem 3 (Littlewood-Offord Theorem of Erdös, Theorem 3 of Erdős [1945])**.** *Let $x_1, \ldots, x_n$ be n real numbers such that $|x_i| \geq 1$ for all i. For any given radius $\Delta > 0$, the small ball probability satisfies*

$$\sup_B P(\epsilon_1 x_1 + \cdots + \epsilon_n x_n \in B) \leq \frac{S(n)}{2^n}(\lfloor \Delta \rfloor + 1)$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables, B ranges over all closed balls (intervals) of radius Δ, and $\lfloor x \rfloor$ refers to the integral part of x, $S(n)$ is the largest binomial coefficient belonging to n.

Using elementary calculations to upper bound $\frac{S(n)}{2^n}$ gives us the following corollary.

Corollary 4. *Under the same notation and conditions as Theorem 3, we have*

$$\sup_B P(\epsilon_1 x_1 + \cdots + \epsilon_n x_n \in B) \leq C_{LO}(\lfloor \Delta \rfloor + 1)\frac{1}{\sqrt{n}}$$

where $C_{LO} = \frac{2\sqrt{2}e}{\pi} < 3$.

The proof of this corollary can be found in Appendix 1.1.

The scale of payoffs for time periods in A_0 is so small that we do not need any Littlewood-Offord theory to control their contribution to the regret. Simply bounding the probabilities by 1 gives us the following.

Theorem 5. *The following upper bound holds for switching probabilities for time periods within A_0:*

$$\sum_{t\in A_0} |g_{t,i\ominus j}| P\left(\sum_{\tau=1}^{t-1}\epsilon_\tau g_{\tau,i\ominus j} \geq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t}\epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j}\right) \leq \sqrt{|A_0|} \leq 20C_{LO}\sqrt{|A_0|}.$$

where $C_{LO} > 1$.

The proof of this theorem can also be found in Appendix 1.1.

The real work lies in controlling the switching probabilities for payoffs at intermediate scales. The idea in the proof of the results is to condition on the ϵ_t's outside A_k. Then the probability of interest is written as a small ball event in terms of the ϵ_t's in A_k. Applying Littlewood-Offord theorem concludes the argument.

Theorem 6. *For any $k \in \{1, \ldots, K\}$, we have*

$$\sum_{t\in A_k} |g_{t,i\ominus j}| P\left(\sum_{\tau=1}^{t-1}\epsilon_\tau g_{\tau,i\ominus j} \geq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t}\epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j}\right) \leq 20C_{LO}\sqrt{|A_k|}.$$

Again, the proof of this theorem is deferred to Appendix 1.1.

We finally have all the ingredients in place to control the switching probabilities.

Corollary 7. *The following upper bound on the switching probabilities holds.*

$$\sum_{t=1}^{T} |g_{t,i\ominus j}| P\left(\sum_{\tau=1}^{t-1}\epsilon_\tau g_{\tau,i\ominus j} \geq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t}\epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j}\right) \leq 20C_{LO}\sqrt{T\log_2(4\log_2 T)}.$$

Proof. Using Theorem 5 and Theorem 6, we have

$$\begin{aligned}
&\sum_{t=1}^{T} |g_{t,i\ominus j}| P\left(\sum_{\tau=1}^{t-1} \epsilon_\tau g_{\tau,i\ominus j} \geq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t} \epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j}\right) \\
=&\sum_{k=0}^{K} \sum_{t\in A_k} |g_{t,i\ominus j}| P\left(\sum_{\tau=1}^{t-1} \epsilon_\tau g_{\tau,i\ominus j} \geq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t} \epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j}\right) \\
\leq&\sum_{k=0}^{K} 20 C_{LO} \sqrt{|A_k|}.
\end{aligned}$$

Since $\sum_{k=0}^{K} \sqrt{|A_k|} \leq \sqrt{K+1} \cdot \sqrt{\sum_{k=0}^{K} |A_k|}$ and $\sum_{k=0}^{K} |A_k| = T$, we have

$$\sum_{k=0}^{K} 20 C_{LO} \sqrt{|A_k|} \leq 20 C_{LO} \sqrt{(K+1)T}.$$

By definition of K, we have that $T^{-\frac{1}{2^{K-1}}} < \frac{1}{2}$, $K \leq \log_2(\log_2(T)) + 1$ which finishes the proof. □

Thus, $\forall i, j \in \{1, \ldots, N, i \neq j\}$, we have

$$\sum_{t=1}^{T} |g_{t,i\ominus j}| P\left(\tilde{G}_{t-1,i\ominus j} \geq 0, \tilde{G}_{t,i\ominus j} \leq 0\right) \leq 20 C_{LO} \sqrt{T \log_2(4 \log_2 T)},$$

which, when plugged into Theorem 2, immediately yields the following corollary.

Corollary 8. *Against an oblivious opponent, both versions — the single stream version* (2.3) *and the fresh-randomization-at-each-round version* (2.1) *— of sampled fictitious play enjoy the following bound on the expected regret.*

$$\mathbb{E}\left[\mathcal{R}_T\right] \leq 40 C_{LO} N^2 \sqrt{T \log_2(4 \log_2 T)}.$$

2.4.3 Step 3: From Oblivious to Adaptive Opponents

Now we consider adaptive opponents. In this setting, we can no longer assume that player i plays against a fixed sequence of payoff vectors $\{g_t\}_{t=1}^{T}$. Note that $g_{t,k}$ is just shorthand for $u_i(k, s_{t,-i})$ and opponents can react to player i's moves $k_1, \ldots, k_{t-1}$ in selecting their strategy tuple $s_{t,-i}$. Thus, g_t is a function $g_t(k_1, \ldots, k_{t-1})$. Faced

with general adaptive opponents, the single stream version (2.3) can incur terrible expected regret as stated below.

Theorem 9. *The single stream version of the sampled fictitious play procedure* (2.3) *can incur linear expected regret against adaptive opponents.*

The proof of this theorem can be found at the end of Appendix 1.1.

However, for the fresh randomness at each round procedure (2.1), we can apply Lemma 4.1 of Cesa-Bianchi and Lugosi [2006] along with Corollary 8 to derive our next result that holds for adaptive opponents too. There are two conditions that we must verify before we apply that lemma. First, the learning procedure should use independent randomization at different time points. Second, the probability distribution of $s_{t,i}$ over the N available strategies should be fully determined by $s_{1,-i}, \ldots, s_{t-1,-i}$ and should not depend explicitly on player i own previous moves $s_{1,i}, \ldots, s_{t-1,i}$. Both of these conditions are easily seen to hold for sampled fictitious play as defined in (2.1) and (2.2).

Theorem 10. *For any T, for any $\delta_T > 0$, with probability at least $1 - \delta_T$, the actual regret $\mathcal{R}_T$ of sampled fictitious play as defined in* (2.1) *and* (2.2) *satisfies, for any adaptive opponent,*

$$\mathcal{R}_T \leq 40 C_{LO} N^2 \sqrt{T \log_2(4 \log_2 T)} + \sqrt{\frac{T}{2} \log \frac{1}{\delta_T}}.$$

Now pick $\delta_T = \frac{1}{T^2}$. Consider the events $E_T = \{\mathcal{R}_T \geq 12 C_{LO} N^2 \sqrt{T \log_2(4 \log_2 T)} + \sqrt{T \log T}\}$ with $P(E_T) \leq \delta_T$. Since $\sum_{T=1}^{\infty} \delta_T < \infty$, we have $\sum_{T=1}^{\infty} P(E_T) < \infty$. Therefore, using Borel-Cantelli lemma, the event "infinitely many E_T's occur" has probability 0. That is, with probability 1, we have $\limsup_{T \to \infty} \frac{\mathcal{R}_T}{T \log T} \leq C$ for some constant C. In particular, with probability 1, $\limsup_{T \to \infty} \frac{\mathcal{R}_T}{T} = 0$, which proves Theorem 1.

2.5 Conclusion

We proved that a natural variant of fictitious play is Hannan consistent. In the variant we considered, the player plays the best response to moves of her opponents at sampled time points in the history so far. We considered one particular sampling scheme, namely Bernoulli sampling. It will be interesting to consider other sampling strategies including sampling with replacement. It will also be interesting to consider notions of regret, such as tracking regret [Cesa-Bianchi and Lugosi, 2006, Section 5.2], that are more suitable for non-stationary environments by biasing the sampling to give more importance to recent time points.

CHAPTER III

Adversarial Multi-armed Bandit and Follow the Perturbed Leader

3.1 Introduction

Starting from the seminal work of Hannan [1957] and later developments due to Kalai and Vempala [2005], perturbation based algorithms (called "Follow the Perturbed Leader (FTPL)") have occupied a central place in online learning. Another major family of online learning algorithms, called "Follow the Regularized Leader (FTRL)", is based on the idea of regularization. In special cases, such as the exponential weights algorithm for the experts problem, it has been folk knowledge that regularization and perturbation ideas are connected. That is, the exponential weights algorithm can be understood as either using negative entropy regularization or Gumbel distributed perturbation (for example, see the discussion in Abernethy et al. [2014]).

Recent work have begun to further uncover the connections between perturbation and regularization. For example, in online linear optimization, one can understand regularization and perturbation as simply two different ways to smooth a non-smooth potential function. The former corresponds to infimal convolution smoothing and the latter corresponds to stochastic (or integral convolution) smoothing [Abernethy et al., 2014]. Having a generic framework for understanding perturbations allows one to

study a wide variety of online linear optimization games and a number of interesting perturbations.

FTRL and FTPL algorithms have also been used beyond "full information" settings. "Full information" refers to the fact that the learner observes the entire move of the adversary. The multi-armed bandit problem is one of the most fundamental examples of "partial information" settings. Regret analysis of the multi-armed bandit problem goes back to the work of Robbins [1952] who formulated the stochastic version of the problem. The non-stochastic, or adversarial, version was formulated by Auer et al. [2002], who provided the EXP3 algorithm achieving $O(\sqrt{NT \log N})$ regret in T rounds with N arms. They also showed a lower bound of $\Omega(\sqrt{NT})$, which was later matched by the Poly-INF algorithm [Audibert and Bubeck, 2009, Audibert et al., 2011]. The Poly-INF algorithm can be interpreted as an FTRL algorithm with negative Tsallis entropy regularization [Audibert et al., 2011, Abernethy et al., 2015]. For a recent survey of both stochastic and non-stochastic bandit problems, see Bubeck and Cesa-Bianchi [2012].

For the non-stochastic multi-armed bandit problem, Kujala and Elomaa [2005] and Poland [2005] both showed that using the exponential (actually double exponential/Laplace) distribution in an FTPL algorithm coupled with standard unbiased estimation technique yields near-optimal $O(\sqrt{NT \log N})$ regret. Unbiased estimation needs access to arm probabilities that are not explicitly available when using an FTPL algorithm. Neu and Bartók [2013] introduced the geometric resampling scheme to approximate these probabilities while still guaranteeing low regret. Recently, Abernethy et al. [2015] analyzed FTPL for adversarial multi-armed bandits and provided regret bounds under the condition that the hazard rate of the perturbation distribution is bounded. This condition allowed them to consider a vari-

ety of perturbation distributions beyond the exponential, such as Gamma, Gumbel, Frechet, Pareto, and Weibull.

Unfortunately, the bounded hazard rate condition is violated by two of the most widely known distributions: namely the uniform[1] and the Gaussian distributions. As a result, the results of Abernethy et al. [2015] say nothing about the regret incurred in an adversarial multi-armed bandit problem when we use these distributions to generate perturbations. Contrast this to the full information experts setting where using these distributions as perturbations yields optimal $\sqrt{T}$ regret and even yields the optimal $\sqrt{\log N}$ dependence on the dimension in the Gaussian case [Abernethy et al., 2014].

The Gaussian distribution has lighter tails than the exponential. The hazard rate of a Gaussian increases linearly on the real line (and is hence unbounded) whereas the exponential has a constant hazard rate. Does having too light a tail makes a perturbation inherently bad? The uniform is even worse from a light tail point of view: it has bounded support! In fact, Kujala and Elomaa [2005] had trouble dealing with the uniform distribution and remarked, "we failed to analyze the expert setting when the perturbation distribution was uniform." Does having a bounded support make a perturbation even worse? Or is it that the hazard rate condition is just a sufficient condition without being anywhere close to necessary for a good regret bound to exist. The analysis of Abernethy et al. [2015] suggests that perhaps a bounded hazard rate is critical. They even made the following conjecture.

Conjecture 1. *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 incur at least a linear regret.*

[1]The uniform distribution is also historically significant as it was used in the original FTPL algorithm of **?**.

The main contribution of this paper is to provide answers to the questions raised above. First, we show that boundedness of the hazard rate is certainly not a requirement for achieving sublinear (in T) regret. Bounded support distributions, like the uniform, violate the boundedness condition on the hazard rate in the most extreme way. Their hazard rate blows up not just asymptotically at infinity, as in the Gaussian case, but as one approaches the right edge of the support. Yet, we can show (Corollary 15) that using the uniform distribution results in a regret bound of $O((NT)^{2/3})$. This bound is clearly not optimal. But optimality is not the point here. What is surprising, especially if one regards Conjecture 1 as plausible, is that a non-trivial sublinear bound holds at all. In fact, we show (Corollary 16) that using *any* continuous distribution with bounded support and bounded density results in a sublinear regret bound.

Second, moving beyond bounded support distributions to ones with unbounded support, we settle Conjecture 1 in the negative. In Theorem 22 we show that, instead of suffering linear regret as predicted by Conjecture 1, a perturbation algorithm using the Gaussian distribution enjoys a near optimal regret bound of $O(\sqrt{NT \log N} \log T)$. A key ingredient in our approach is a new quantity that we call the *generalized hazard rate* of a distribution. We show that bounded generalized hazard rate is enough to guarantee sublinear regret in T (Theorem 18).

Finally, we investigate the relationship between tail behavior of random perturbations and the regret they induce. We show that heavy tails, along with some fairly mild assumptions, guarantee a bounded hazard rate (Theorem 25) and hence previous results can yield regret bounds for these perturbations. However, light tails can fail to have a bounded hazard rate. Nevertheless, we show that under reasonable conditions, light tailed distributions do have a bounded *generalized* hazard rate

(Theorem 26). This result allows us to show that reasonably behaved light-tailed distributions lead to near optimal regret (Corollary 27). In particular, the exponential power (or generalized normal) family of distributions yields near optimal regret (Theorem 29)

All proofs in the chapter are deferred to the appendix.

3.2 Follow the Perturbed Leader Algorithm for Bandits

Recall the setting of the adversarial multi-armed bandit problem [Auer et al., 2002]. An adversary (or Nature) chooses loss vectors $g_t \in [-1, 0]^N$ for $1 \leq t \leq T$ ahead of the game. Such an adversary is called *oblivious*. At round $t = 1, \ldots, T$ in a repeated game, the learner must choose a distribution $p_t \in \Delta_N$ over the set of N available arms (or actions). The learner plays action i_t sampled according to p_t and incurs the loss $g_{t,i_t} \in [-1, 0]$. The learner observes only g_{t,i_t} and receives no information about the values $g_{t,j}$ for $j \neq i_t$.

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:

$$\text{Regret}_T := \max_{i \in [N]} \sum_{t=1}^{T} (g_{t,i} - g_{t,i_t}). \tag{3.1}$$

To be precise, we consider the *expected* regret, where the expectation is taken with respect to the learner's randomization. Note that, under an oblivious adversary, the only random variables in the above expression are the actions i_t of the learner.

The maximization in (3.1) implies that g is strictly speaking a negative *gain* vector, not a loss vector. Nevertheless, we use the term *loss*, as we impose the assumption that $g_t \in [-1, 0]^N$ throughout the paper. The decision to consider the loss setting is important: our proof will not work for gains. It is known that the adversarial multi-armed bandit problem does not exhibit symmetry with respect to

gains versus losses. Often losses are easier to handle than gains [Bubeck and Cesa-Bianchi, 2012]. Finally, our decision to treat losses as negative gains stems from the desire to work with convex, not concave, potential functions.

3.2.1 The Gradient-Based Algorithmic Template

We will consider the algorithmic template described in Framework 1, which is the Gradient Based Prediction Algorithm (GBPA) (see, for example, Abernethy et al. [2015]). Let Δ^N be the $(N-1)$-dimensional probability simplex in $\mathbb{R}^N$. Denote the standard basis vector along the ith dimension by $\mathbf{e}_i$. At any round t, the action choice i_t is made by sampling from the distribution p_t which is obtained by applying the gradient of a convex function $\tilde{\Phi}$ to the estimate $\hat{G}_{t-1}$ of the cumulative gain vector so far. The choice of $\tilde{\Phi}$ is flexible but it must be a differentiable convex function such that its gradient is always in Δ^N.

Note that we do not require that the range of $\nabla\tilde{\Phi}$ be contained in the *interior* of the probability simplex. If we required the gradient to lie in the interior, we would not be able to deal with bounded support distributions such as the uniform distribution. Even though some entries of the probability vector p_t might be 0, the estimation step is always well defined since $p_{t,i_t} > 0$. But allowing $p_{t,i}$ to be zero means that $\hat{g}_t$ is not exactly an unbiased estimator of g_t. Instead, it is an unbiased estimator on the support of p_t. That is, $\mathbb{E}[\hat{g}_{t,i}|i_{1:t-1}] = g_{t,i}$ for any i such that $p_{t,i} > 0$. Here, $i_{1:t-1}$ is shorthand for $i_1, \ldots, i_{t-1}$. Therefore, irrespective of whether $p_{t,i} = 0$ or not, we always have

$$\mathbb{E}[p_{t,i}\hat{g}_{t,i}|i_{1:t-1}] = p_{t,i}g_{t,i}. \tag{3.2}$$

When $p_{t,i} = 0$, we have $\hat{g}_{t,i} = 0$ but $g_{t,i} \leq 0$, which means that $\hat{g}_t$ overestimates g_t

outside the support of p_t. Hence, we also have

$$\mathbb{E}[\hat{g}_t | i_{1:t-1}] \succeq g_t, \tag{3.3}$$

where $\succeq$ means element-wise greater than.

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

GBPA($\tilde{\Phi}$): $\tilde{\Phi}$ is a differentiable convex function such that $\nabla\tilde{\Phi} \in \Delta^N$
Nature: Adversary chooses "gain" vectors $g_t \in [-1, 0]^N$ for $t = 1, \ldots, T$
Learner initializes $\hat{G}_0 = 0$
for $t = 1$ to T **do**
  Sampling: Learner chooses i_t according to the distribution $p_t = \nabla\tilde{\Phi}(\hat{G}_{t-1})$
  Cost: Learner incurs (and observes) "gain" $g_{t,i_t} \in [-1, 0]$
  Estimation: Learner creates estimate of gain vector $\hat{g}_t := \frac{g_{t,i_t}}{p_{t,i_t}} \mathbf{e}_{i_t}$
  Update: Cumulative gain estimate so far $\hat{G}_t = \hat{G}_{t-1} + \hat{g}_t$
end for

We now present a basic result bounding the expected regret of GBPA in the multi-armed bandit setting. It is basically just a simple modification of the arguments in Abernethy et al. [2015] to deal with the possibility that $p_{t,i} = 0$. We state and prove this result here for completeness without making any claim of novelty.

Lemma 11. (Decomposition of the Expected Regret) *Define the non-smooth potential* $\Phi(G) = \max_i G_i$. *The expected regret of GBPA*($\tilde{\Phi}$) *can be written as*

$$\mathbb{E}\text{Regret}_T = \Phi(G_T) - \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t, g_t \rangle\right]. \tag{3.4}$$

Furthermore, the expected regret of GBPA($\tilde{\Phi}$) *can be bounded by the sum of an overestimation, an underestimation, and a divergence penalty:*

(3.5)

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

where the expectations are over the sampling of i_t *and* $D_{\tilde{\Phi}}$ *is the Bregman divergence induced by* $\tilde{\Phi}$.

3.2.2 Stochastic Smoothing of Potential Function

Let $\mathcal{D}$ be a continuous distribution with finite expectation, probability density function f, and cumulative distribution function F. Consider GBPA with potential function of the form:

$$\tilde{\Phi}(G;\mathcal{D}) = \mathbb{E}_{Z_1,\ldots,Z_N \overset{\text{i.i.d}}{\sim} \mathcal{D}} \Phi(G_i + Z_i), \tag{3.6}$$

which is a *stochastic smoothing* of the non-smooth function $\Phi(G) = \max_i G_i$. We will often hide the dependence on the distribution $\mathcal{D}$ if the distribution is obvious from the context or when the dependence on $\mathcal{D}$ is not of importance in the argument. Since Φ is convex, $\tilde{\Phi}$ is also convex. For stochastic smoothing, we have the following result to control the underestimation and overestimation penalty.

Lemma 12. *For any G, we have*

$$\Phi(G) + \mathbb{E}[Z_1] \leq \tilde{\Phi}(G) \leq \Phi(G) + EMAX(N) \tag{3.7}$$

where $EMAX(N)$ is any function such that

$$\mathbb{E}_{Z_1,\ldots,Z_N}[\max_i Z_i] \leq EMAX(N).$$

In particular, this implies that the overestimation penalty $\tilde{\Phi}(0)$ is upper bounded by $\Phi(0) + EMAX(N) = EMAX(N)$ and the underestimation penalty $\Phi(\hat{G}_T) - \tilde{\Phi}(\hat{G}_T)$ is upper bounded by $-\mathbb{E}[Z_1]$.

Note that Φ is differentiable with probability 1 (under the randomness of the Z_i's) due to the fact that Z_i's are random variables with a density. By Proposition 2.3 of Bertsekas [1973], we can swap the order of differentiation and expectation:

$$\nabla\tilde{\Phi}(G;\mathcal{D}) = \mathbb{E}_{Z_1,\ldots,Z_N \overset{\text{i.i.d}}{\sim} \mathcal{D}} e_{i^*}, \text{ where } i^* = \arg\max_{i=1,\ldots,N}\{G_i + Z_i\}. \tag{3.8}$$

Note that, for any G, the random index i^* is unique with probability 1. Hence, ties between arms can be resolved arbitrarily. It is clear from above that $\nabla\tilde{\Phi}$, being an expectation of vectors in the probability simplex, is in the probability simplex. Thus, it is a valid potential to be used in Framework 1. Note that

$$\begin{aligned}\nabla_i\tilde{\Phi}(G) = \frac{\partial\tilde{\Phi}}{\partial G_i} &= \mathbb{E}_{Z_1,\dots,Z_N}\mathbf{1}\{G_i + Z_i > G_j + Z_j, \forall j \neq i\} \\ &= \mathbb{E}_{\tilde{G}_{-i}}[\mathbb{P}_{Z_i}[Z_i > \tilde{G}_{-i} - G_i]] = \mathbb{E}_{\tilde{G}_{-i}}[1 - F(\tilde{G}_{-i} - G_i)].\end{aligned} \tag{3.9}$$

where $\tilde{G}_{-i} = \max_{j\neq i} G_j + Z_j$. If $\mathcal{D}$ has unbounded support then this partial derivative is non-zero for all i given any G. However, it can be zero if $\mathcal{D}$ has bounded support. Moreover, we have the following useful identity that writes the Hessian of the smoothed potential function in terms of the expectation of the probability density function.

$$\begin{aligned}\nabla^2_{ii}\tilde{\Phi}(G) = \frac{\partial}{\partial G_i}\nabla_i\tilde{\Phi}(G) &= \frac{\partial}{\partial G_i}\mathbb{E}_{\tilde{G}_{-i}}[1 - F(\tilde{G}_{-i} - G_i)] \\ &= \mathbb{E}_{\tilde{G}_{-i}}\left[\frac{\partial}{\partial G_i}(1 - F(\tilde{G}_{-i} - G_i))\right] = \mathbb{E}_{\tilde{G}_{-i}} f(\tilde{G}_{-i} - G_i).\end{aligned} \tag{3.10}$$

3.2.3 Connection to Follow the Perturbed Leader

The sampling step of Framework 1 with a stochastically smoothed Φ as the potential $\tilde{\Phi}$ (Equation 3.6) can be done efficiently. Instead of evaluating the expectation (Equation 3.8), we just take a random sample. Doing so gives us an equivalent of Follow the Perturbed Leader Algorithm (FTPL) [**?**] applied to the bandit setting. On the other hand, 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), an iterative resampling process to estimate $\nabla\tilde{\Phi}$ (with bias). They showed that the extra regret after stopping at M iterations of GR introduces an estimation bias that is at most $\frac{NT}{eM}$ as an additive term. That is, all GBPA regret bounds that

we prove will hold for the corresponding FTPL algorithm that does M iterations of GR at every time step, with an extra additive $\frac{NT}{eM}$ term. This extra term does not affect the regret rate as long as $M = \sqrt{NT}$, because the lower bound for any adversarial multi-armed bandit algorithm is of the order $\sqrt{NT}$.

3.2.4 The Role of the Hazard Rate and Its Limitation

In previous work, Abernethy et al. [2015] proved that for a continuous random variable Z with finite and nonnegative expectation and support on the whole real line $\mathbb{R}$, if the hazard rate of the random variable is bounded, i.e,

$$\sup_z \frac{f(z)}{1 - F(z)} < \infty,$$

then the expected regret of GBPA can be upper bounded as

$$\mathbb{E}\text{Regret}_T = O\Big(\sqrt{NT \times EMAX(N)}\Big).$$

Common families of distributions whose regret can be controlled in this way include Gumbel, Frechet, Weibull, Pareto, and gamma (see Abernethy et al. [2015] for details). However, there are many other families of distributions where the hazard rate condition fails. For example, if the random variable has a bounded support, then the hazard rate would certainly explode at the end of the support. This is, in some sense, an extreme case of violation because the random variable does not even have a tail. There are also some random variables that do have support on $\mathbb{R}$ but have unbounded hazard rate, e.g. Gaussian, where the hazard rate monotonically increases to infinity. How can we perform analyses of the expected regret of GBPA using those random variables as perturbations? To address these issues, we need to go beyond the hazard rate.

3.3 Perturbations with Bounded Support

In this section, we prove that GBPA with any continuous distribution that has bounded support, bounded density enjoys sublinear expected regret. From Lemma 11 we see that the expected regret can be upper bounded by the sum of three terms. The overestimation penalty can be bounded very easily via Lemma 12 for a distribution with bounded support. The underestimation penalty is non-positive as long as the distribution has non-negative expectation. The only term that needs to be controlled with some effort is the divergence penalty.

We first present a general lemma that allows us to write the divergence penalty under a stochastic smoothing potential $\tilde{\Phi}$ as a sum involving certain double integrals.

Lemma 13. *When using a stochastically smoothed potential as in* (3.6)*, the divergence penalty can be written as*

(3.11)

$$\mathbb{E}\left[D_{\tilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1})|i_{1:t-1}\right] = \sum_{i \in supp(p_t)} p_{t,i} \int_0^{\left|\frac{g_{t,i}}{p_{t,i}}\right|} \mathbb{E}_{\hat{G}_{-i}}\left[\int_0^s f(\hat{G}_{-i} - \hat{G}_{t-1,i} + r)dr\right] ds$$

where $p_t = \nabla\tilde{\Phi}(\hat{G}_{t-1})$, $\hat{G}_{-i} = \max_{j \neq i} \hat{G}_{t-1,j} + Z_j$ *and* $supp(p_t) = \{i \,:\, p_{t,i} > 0\}$.

Note that each summand in the divergence penalty expression above involves an integral of the density function of the distribution $\mathcal{D}$ over an interval. The main idea to control the divergence penalty for a bounded support distribution is to truncate the interval at the end of the support. For points that are close to the end of the support, we bound the integral by the product of the bound on the density and the interval length. For points that are far from the end of the support, we bound the integral through the hazard rate as was done by Abernethy et al. [2015].

For a general continuous random variable Z with bounded density, bounded support, we first shift it (which obviously does not change the action choice i_t and

hence the expected regret) and scale it so that the support is a subset of $[0, 1]$ with $\inf\{z : F(z) = 0\} = 0$ and $\inf\{z : F(z) = 1\} = 1$ where F denotes the CDF of Z. A benefit of this normalization is that the expectation of the random variable becomes non-negative so the underestimation penalty is guaranteed to be non-positive. After scaling, we assume that the bound on the density is L. We consider the perturbation ηZ where $\eta > 0$ is a tuning parameter. Write $F_\eta(x)$ and $f_\eta(x)$ to denote the CDF and PDF of the scaled random variable ηZ respectively . If F is strictly increasing, we know that F^{-1} exists. If not, define $F^{-1}(y) = \inf\{z : f(z) = y\}$. Elementary calculation gives the following useful facts:

$$F_\eta(z) = F(\frac{z}{\eta}), f_\eta(z) = \frac{f(\frac{z}{\eta})}{\eta}, F_\eta^{-1}(y) = \eta F^{-1}(y).$$

Theorem 14. (Divergence Penalty Control, Bounded Support) *The divergence penalty in the GBPA regret bound using the perturbation ηZ, where Z is drawn from a bounded support distribution satisfying the conditions above, can be upper bounded, for any $\epsilon > 0$, by*

$$NL\Big(\frac{1}{2\eta\epsilon} + 1 - F^{-1}(1-\epsilon)\Big).$$

The regret bound for the uniform distribution is now an easy corollary.

Corollary 15. (Regret Bound for Uniform) *For GBPA run with a stochastic smoothing using an appropriately scaled* $[0, 1]$ *uniform perturbation, the expected regret can be upper bounded by* $3(NT)^{2/3}$.

For a general perturbation with bounded support and bounded density, the rate at which $1 - F^{-1}(1-\epsilon)$ goes to 0 as $\epsilon \to 0$ can vary but we can always guarantee sublinear expected regret.

Corollary 16. (Asymptotic Regret Bound for Bounded Support) *For stochastically smoothed GBPA using general continuous random variable Z with bounded density and bounded support contained in $[0, 1]$, the expected regret grows sublinearly, i.e.,*

$$\lim_{T\to\infty} \frac{\mathbb{E}\text{Regret}_T}{T} = 0.$$

3.4 Perturbations with Unbounded Support

Unlike perturbations with bounded support, perturbations with unbounded support (on the right) do have non-zero right tail probabilities, ensuring that $p_{t,i} > 0$ always. However, the tail behavior may be such that the hazard rate is unbounded. Still, under mild assumptions, perturbations with unbounded support (on the right) can also be shown to have near optimal expected regret in T, using the notion of *generalized hazard rate* that we now introduce.

3.4.1 Generalized Hazard Rate

We already know how to control the underestimation and overestimation penalties via Lemma 12. So our main focus will be to control the divergence penalty. Towards this end, we define the generalized hazard rate for a continuous random variable Z with support unbounded on the right, parameterized by $\alpha \in [0, 1)$, as

$$h_\alpha(z) := \frac{f(z)|z|^\alpha}{(1 - F(z))^{1-\alpha}}, \tag{3.12}$$

where $f(z)$ and $F(z)$ denotes the PDF and CDF of Z respectively. Note that by setting $\alpha = 0$ we recover the standard hazard rate.

One of the main results of this paper is the following. Note that it includes the result (Lemma 4.3) of Abernethy et al. [2015] as a special case.

Theorem 17. (Divergence Penalty Control via Generalized Hazard Rate) *Let $\alpha \in [0, 1)$. Suppose we have $\forall z \in \mathbb{R}, h_\alpha(z) \leq C$. Then,*

$$\mathbb{E}[D_{\tilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1})|i_{1:t-1}] \leq \frac{2C}{1-\alpha} \times N.$$

A regret bound now easily follows.

Theorem 18. (Regret Bound via Generalized Hazard Rate) *Suppose we use a stochastic smoothing with a perturbation distribution whose generalized hazard rate is bounded: $h_\alpha(x) \leq C, \forall x \in \mathbb{R}$ for some $\alpha \in [0, 1)$, and*

$$\mathbb{E}_{Z_1,\ldots,Z_N}[\max_i Z_i] - \mathbb{E}[Z_1] \leq Q(N),$$

where $Q(N)$ is some function of N. Then, the expected regret of GBPA is no greater than

$$2 \times (\frac{2C}{1-\alpha})^{1/(2-\alpha)} \times (NT)^{1/(2-\alpha)} \times Q(N)^{(1-\alpha)/(2-\alpha)}.$$

In particular, this implies that the algorithm has sublinear expected regret.

3.4.2 Gaussian Perturbation

In this section we prove that GBPA with the standard Gaussian perturbation incurs a near optimal expected regret in both N and T. Let $F(z)$ and $f(z)$ denote the CDF and PDF of standard Gaussian distribution.

Lemma 19 (Baricz [2008])**.** *For standard Gaussian random variable, we have*

$$z < \frac{f(z)}{1-F(z)} < \frac{z}{2} + \frac{\sqrt{z^2+4}}{2}.$$

This lemma together with example 2.6 in Thomas [1971] show that the hazard rate of a standard Gaussian random variable increases monotonically to infinity. However, we can still bound the generalized hazard rate for strictly positive α.

Lemma 20. (Generalized Hazard Bound for Gaussian) *For any $\alpha \in (0,1)$, we have*

$$\frac{f(z)|z|^{\alpha}}{(1-F(z))^{1-\alpha}} \leq C_1$$

where $C_1 = \frac{2}{\alpha}$.

The bounded generalized hazard rate shown in the above lemma can be used to control the divergence penalty. Combined with other knowledge of the standard Gaussian random variable we are able to give a bound on the expected regret.

Corollary 21. *The expected regret of GBPA with standard Gaussian random variable as perturbation has an expected regret at most*

$$2(C_1 C_2 NT)^{1/(2-\alpha)}(\sqrt{2\log N})^{(1-\alpha)/(2-\alpha)}$$

where $C_1 = \frac{2}{\alpha}$, $C_2 = \frac{2}{1-\alpha}$, and $\alpha \in (0,1)$.

It remains to optimally tune α in the above bound. Note the tuning parameter α appears only in the analysis, not in the algorithm.

Theorem 22. (Regret Bound for Gaussian) *The expected regret of GBPA with standard Gaussian random variable as perturbation has an expected regret at most*

$$96\sqrt{NT} \times N^{1/\log T}\sqrt{\log N}\log T$$

for $T > 4$. If we assume that $T > N$, the expected regret can be upper bounded by

$$278\sqrt{NT} \times \sqrt{\log N}\log T.$$

3.4.3 Sufficient Condition for Near Optimal Regret

In Section 3.4.1 we showed that if the generalized hazard rate of a distribution is bounded, the expected regret of the GBPA can be controlled. In this section,

we are going to prove that under reasonable assumptions on the distribution of the perturbation, the FTPL enjoys near optimal expected regret.

Assumptions (a)-(c). Before we proceed, let us formally state our assumptions on the distributions we will consider. The distribution needs to (a) be continuous and has bounded density (b) has finite expectation (c) has support unbounded in the $+\infty$ direction.

Note that if the expectation of the random perturbation is negative, we shift it so that the expectation is zero. Hence the underestimation penalty is non-positive. In addition to the assumptions we have made above, we make another assumption on the eventual monotonicity of the hazard rate.

$$\textbf{Assumption (d)} \quad h_0(z) = \frac{f(z)}{1 - F(z)} \text{ is eventually monotone.}$$

"Eventually monotone" means that $\exists z_0 \geq 0$ such that if $z > z_0$, $\frac{f(z)}{1-F(z)}$ is non-decreasing or non-increasing. This assumption might appear hard to check, but numerous theorems are available to establish the monotonicity of hazard rate, which is much stronger than what we are assuming here. For example, see Theorem 2.4 in Thomas [1971], Theorem 2 and Theorem 4 in Chechile [2003], Chechile [2009]. In fact, most natural distributions do satisfy this assumption [Bagnoli and Bergstrom, 2005].

Before we proceed, we mention a standard classification of random variables into two classes based on their tail property.

Definition 23 (see, for example, Foss et al. [2009])**.** A function $f(z) \geq 0$ is said to be heavy-tailed if and only if

$$\lim_{z \to \infty} \sup f(z) e^{\lambda z} = \infty \quad \text{for all } \lambda > 0.$$

A distribution with CDF $F(z)$ and $\overline{F}(z) = 1 - F(z)$ is said to be heavy-tailed if and only if $\overline{F}(z)$ is heavy-tailed. If the distribution is not heavy-tailed, we say that it is light-tailed.

It turns out that under assumptions (a)-(d), if the distribution is also heavy-tailed, then the hazard rate itself is bounded. If the distribution is light-tailed, we need an additional assumption on the eventual monotonicity of a function similar to generalized hazard rate to ensure the boundedness of the generalized hazard rate. But before we state and prove the main results, we introduce some functions and prove an intermediate lemma that will be useful to prove the main results.

Define $R(z) = -\log \overline{F}(z)$ so that we have $\overline{F}(z) = e^{-R(x)}$ and $R'(z) = \frac{f(z)}{\overline{F}(z)} = h_0(z)$.

Lemma 24. *Under assumptions (a)-(d), we have*

$$\overline{F}(z)e^{\lambda z} \text{ is eventually monotone } \forall \lambda > 0.$$

We are finally ready to present the main results in this section.

Theorem 25. (Heavy Tail Implies Bounded Hazard) *Under assumptions (a) - (d), if the distribution is also heavy-tailed, then the hazard rate is bounded, i.e,*

$$\sup_z \frac{f(z)}{\overline{F}(z)} < \infty.$$

Unlike heavy-tailed distributions, the hazard rate of light-tailed distributions might be unbounded. However, it turns out that if we make an additional assumption on the eventual monotonicity of a function similar to the generalized hazard rate, we can still guarantee the boundedness of the generalized hazard rate.

Assumption (e) $\exists \delta \in (0, 1]$ such that $\dfrac{f(z)}{(1 - F(z))^{1-\delta}}$ is eventually monotone.

Theorem 26. (Light Tail Implies Bounded Generalized Hazard) *Under assumptions (a) - (e), if the distribution is also light-tailed, then for any $\alpha \in (\delta, 1)$, the generalized hazard rate $h_\alpha(z)$ is bounded, i.e,*

$$\sup_z \frac{f(z)|z|^\alpha}{(\overline{F}(z))^{1-\alpha}} < \infty.$$

Combining the above result with control of the divergence penalty gives us the following corollary.

Corollary 27. *Under assumptions (a)-(e), if the distribution is also light-tailed, the expected regret of GBPA with perturbations drawn from that distribution is, for any $\alpha \in (\delta, 1)$ and $\xi > 0$,*

$$O\Big((TN)^{1/(2-\alpha)} N^\xi\Big).$$

In particular, if assumption (e) holds for any $\delta \in (0, 1)$, then the expected regret of GBPA is $O\Big((TN)^{1/2+\epsilon}\Big)$ for any $\epsilon > 0$, i.e, it is near optimal in both N and T.

Next we consider a family of light-tailed distributions that do not have a bounded hazard rate.

Definition 28. The exponential power (or generalized normal) family of distributions, denoted as $\mathcal{D}_\beta$ where $\beta > 1$, is defined via the cdf

$$f_\beta(z) = C_\beta e^{-z^\beta}, \quad z \geq 0.$$

The next theorem shows that GBPA with perturbations from this family of distributions enjoys near optimal expected regret in both N and T.

Theorem 29. (Regret Bound for Power Exponential Family) *$\forall \beta > 1$, the expected regret of GBPA with perturbations drawn from $\mathcal{D}_\beta$ is, for any $\epsilon > 0$, $O\Big((TN)^{1/2+\epsilon}\Big)$.*

3.5 Conclusion and Future Work

Previous work on providing regret guarantees for FTPL algorithms in the adversarial multi-armed bandit setting required a bounded hazard rate condition. We have shown how to go beyond the hazard rate condition but a number of questions remain open. For example, what if we use FTPL with perturbations from discrete distributions such as Bernoulli distribution? In the full information setting Devroye et al. [2013] and van Erven et al. [2014] have considered random walk perturbation and dropout perturbation, both leading to minimax optimal regret. But to the best of our knowledge those distributions have not been analyzed in the adversarial multi-armed bandit problem.

An unsatisfactory aspect of even the tightest bounds for FTPL algorithms from existing work, including ours, is that they never reach the minimax optimal $O(\sqrt{NT})$ bound. They come very close to it: up to logarithmic factors. It is known that FTRL algorithms, using the negative Tsallis entropy as the regularizer, can achieve the optimal bound [Audibert and Bubeck, 2009, Audibert et al., 2011, Abernethy et al., 2015]. Is there a perturbation that can achieve the optimal bound?

We only considered multi-armed bandits in this work. There has been some interest in using FTPL algorithms for combinatorial bandit problems (see, for example, Neu and Bartók [2013]). In future work, it will be interesting to extend our analysis to combinatorial bandit problems.

Appendix A

Proof(s) of Chapter II

1.1 Proofs

We first present a lemma that helps us in proving Theorem 2.

Lemma 30. *Let k_t and $\tilde{g}_t$ be defined as in* (2.3) *and the text following that equation. We have,*

$$\sum_{t=1}^{T} \tilde{g}_{t,k_{t+1}} \geq \sum_{t=1}^{T} \tilde{g}_{t,k_{T+1}} = \max_{k\in\{1,\ldots,N\}} \sum_{t=1}^{T} \tilde{g}_{t,k}.$$

Proof. This is a classical lemma, for example, see Lemma 3.1 in [Cesa-Bianchi and Lugosi, 2006]. We follow the same idea, i.e, proving through induction but adapt it to handle gains instead of losses. The statement is obvious for $T = 1$. Assume now that

$$\sum_{t=1}^{T-1} \tilde{g}_{t,k_{t+1}} \geq \sum_{t=1}^{T-1} \tilde{g}_{t,k_T}.$$

Since, by definition, $\sum_{t=1}^{T-1} \tilde{g}_{t,k_T} \geq \sum_{t=1}^{T-1} \tilde{g}_{t,k_{T+1}}$, the inductive assumption implies

$$\sum_{t=1}^{T-1} \tilde{g}_{t,k_{t+1}} \geq \sum_{t=1}^{T-1} \tilde{g}_{t,k_{T+1}}.$$

Add $\tilde{g}_{T,k_{T+1}}$ to both sides to obtain the result. □

Proof of Theorem 2. We will prove a result for Bernoulli sampling with general probabilities, i.e., when $P(\epsilon_t = +1) = \alpha$ where α is not necessarily $1/2$. We will show

that

$$\mathbb{E}\left[\mathcal{R}_T\right] \leq \frac{N^2}{\alpha} \max_{1\leq i,j\leq N} \sum_{t=1}^{T} |g_{t,i\ominus j}| P\left(\tilde{G}_{t-1,i\ominus j} \geq 0, \tilde{G}_{t,i\ominus j} \leq 0\right)$$

from which the theorem follows as a special case when $\alpha = 1/2$.

Obviously we have $\mathbb{E}(\tilde{g}_{t,i}) = 2\alpha g_{t,i}$ because of the fact that $\mathbb{E}(\epsilon_t) = 2\alpha - 1$. Furthermore, $\mathbb{E}[\tilde{g}_{t,k_t}|\epsilon_1, \ldots, \epsilon_{t-1}] = 2\alpha g_{t,k_t}$ because k_t is fully determined by past randomness $\epsilon_1, \ldots, \epsilon_{t-1}$ and past payoffs $g_1, \ldots, g_{t-1}$ that are given. This implies that $\mathbb{E}[\tilde{g}_{t,k_t}] = \mathbb{E}\left[\mathbb{E}[\tilde{g}_{t,k_t}|\epsilon_1, \ldots, \epsilon_{t-1}]\right] = 2\alpha\mathbb{E}[g_{t,k_t}]$. We now have,

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}_T\right] &= \max_{k\in\{1,\ldots,N\}} \sum_{t=1}^{T} g_{t,k} - \mathbb{E}\left[\sum_{t=1}^{T} g_{t,k_t}\right] \\
&= \frac{1}{2\alpha} \max_{k\in\{1,\ldots,N\}} \mathbb{E}\left[\sum_{t=1}^{T} \tilde{g}_{t,k}\right] - \frac{1}{2\alpha}\mathbb{E}\left[\sum_{t=1}^{T} \tilde{g}_{t,k_t}\right] \\
&\leq \frac{1}{2\alpha}\mathbb{E}\left[\max_{k\in\{1,\ldots,N\}} \sum_{t=1}^{T} \tilde{g}_{t,k} - \sum_{t=1}^{T} \tilde{g}_{t,k_t}\right].
\end{aligned}$$

Using Lemma 30, we can further upper bound the last expression as follows,

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}_T\right] &\leq \frac{1}{2\alpha}\mathbb{E}\left[\sum_{t=1}^{T}\tilde{g}_{t,k_{t+1}} - \sum_{t=1}^{T}\tilde{g}_{t,k_t}\right] \\
&= \frac{1}{2\alpha}\sum_{t=1}^{T}\mathbb{E}\left[(1+\epsilon_t)(g_{t,k_{t+1}} - g_{t,k_t})\right] \\
&\leq \frac{1}{2\alpha}\sum_{t=1}^{T}\mathbb{E}\left[(1+\epsilon_t)|g_{t,k_{t+1}} - g_{t,k_t}|\right] \\
&\leq \frac{1}{\alpha}\sum_{t=1}^{T}\mathbb{E}\left[|g_{t,k_t} - g_{t,k_{t+1}}|\right] \\
&= \frac{1}{\alpha}\sum_{t=1}^{T}\sum_{1\leq i,j\leq N}\mathbb{E}\left[|g_{t,i} - g_{t,j}|1_{(k_t=i,k_{t+1}=j)}\right] \\
&= \frac{1}{\alpha}\sum_{1\leq i,j\leq N}\sum_{t=1}^{T}\mathbb{E}\left[|g_{t,i} - g_{t,j}|1_{(k_t=i,k_{t+1}=j)}\right] \\
&\leq \frac{N^2}{\alpha}\max_{1\leq i,j\leq N}\sum_{t=1}^{T}|g_{t,i} - g_{t,j}|P(k_t = i, k_{t+1} = j) \\
&\leq \frac{N^2}{\alpha}\max_{1\leq i,j\leq N}\sum_{t=1}^{T}|g_{t,i} - g_{t,j}|P\left(\tilde{G}_{t-1,i} \geq \tilde{G}_{t-1,j}, \tilde{G}_{t,i} \leq \tilde{G}_{t,j}\right) \\
&= \frac{N^2}{\alpha}\max_{1\leq i,j\leq N}\sum_{t=1}^{T}|g_{t,i\ominus j}|P\left(\tilde{G}_{t-1,i\ominus j} \geq 0, \tilde{G}_{t,i\ominus j} \leq 0\right).
\end{aligned}$$

□

The next lemma is useful to determine the appropriate constant in the Littlewood-Offord Theorem.

Lemma 31. *Suppose $X_1, \ldots, X_t$ are i.i.d. Bernoulli random variables that take value of 1 with probability α and 0 with probability $1-\alpha$. If $t > \max(\frac{2}{1-\alpha}, \frac{2}{\alpha}) \geq \max(\frac{2\alpha}{1-\alpha}, \frac{2}{\alpha})$, then for all k,*

$$P\left(\sum_{i=1}^{t}X_i = k\right) \leq \frac{e}{2\pi} \times \sqrt{\frac{2}{\alpha(1-\alpha)}} \times t^{-\frac{1}{2}}.$$

Proof. Note that for $0 \leq k < t$,

$$\frac{P(x=k+1)}{P(x=k)} = \frac{\binom{t}{k+1}\alpha^{k+1}(1-\alpha)^{t-k-1}}{\binom{t}{k}\alpha^k(1-\alpha)^{t-k}} = \frac{\alpha(t-k)}{(1-\alpha)(k+1)}.$$

Therefore, the maximum probability of Bernoulli distribution $P(X = k)$ is achieved when $k = \hat{k} = \lfloor (t+1)\alpha \rfloor$ where $\lfloor x \rfloor$ denotes the integral part of x. Clearly $\hat{k} \in [t\alpha - 1, (t+1)\alpha]$. Thus,

$$\begin{aligned}
\sqrt{\hat{k}(t-\hat{k})} &\geq \min\left(\sqrt{(t\alpha-1)(t-t\alpha+1)}, \sqrt{(t+1)\alpha(t-t\alpha-\alpha)}\right) \\
&= t \times \min\left(\sqrt{(\alpha - \frac{1}{t})(1-\alpha+\frac{1}{t})}, \sqrt{(1+\frac{1}{t})\alpha(1-\alpha-\frac{\alpha}{t})}\right) \\
&\geq t \times \min\left(\sqrt{(\alpha - \frac{\alpha}{2})(1-\alpha)}, \sqrt{\alpha(1-\alpha-\frac{1-\alpha}{2})}\right) \\
&= \sqrt{\frac{\alpha(1-\alpha)}{2}}t.
\end{aligned}$$

With this preliminary inequality, we are ready to prove the lemma.

$$\begin{aligned}
P(\sum_{i=1}^{t} X_i = k) &\leq P(\sum_{i=1}^{t} X_i = \hat{k}) \\
&= \binom{t}{\hat{k}} \times \alpha^{\hat{k}}(1-\alpha)^{t-\hat{k}} \\
&= \frac{t!}{(\hat{k})!(t-\hat{k})!} \times \alpha^{\hat{k}}(1-\alpha)^{t-\hat{k}} \\
&\leq \frac{t^{t+\frac{1}{2}}e^{1-t}}{\left(\sqrt{2\pi}\hat{k}^{\hat{k}+\frac{1}{2}}e^{-\hat{k}}\right)\left(\sqrt{2\pi}(t-\hat{k})^{t-\hat{k}+\frac{1}{2}}e^{-(t-\hat{k})}\right)} \times \alpha^{\hat{k}}(1-\alpha)^{t-\hat{k}} \\
&= \frac{e}{2\pi} \times \frac{1}{\sqrt{\hat{k}(t-\hat{k})}} \times \frac{t^{t+\frac{1}{2}}}{\hat{k}^{\hat{k}}(t-\hat{k})^{t-\hat{k}}} \times \alpha^{\hat{k}}(1-\alpha)^{t-\hat{k}} \\
&\leq \frac{e}{2\pi} \times \sqrt{\frac{2}{\alpha(1-\alpha)}} \times t^{t-\frac{1}{2}} \times \frac{\alpha^{\hat{k}}(1-\alpha)^{t-\hat{k}}}{\hat{k}^{\hat{k}}(t-\hat{k})^{t-\hat{k}}}.
\end{aligned}$$

Let $f(x) = \frac{\alpha^x(1-\alpha)^{t-x}}{x^x(t-x)^{t-x}}, f'(x) = \left(\log(\frac{\alpha}{1-\alpha}) - \log(\frac{x}{t-x})\right) \times f(x)$. Obviously $f'(x)$ is 0 when $x = \alpha t$, positive when $x < \alpha t$, and negative when $x > \alpha t$. Thus,

$$f(x) \leq \frac{\alpha^{\alpha t}(1-\alpha)^{t-\alpha t}}{(\alpha t)^{\alpha t}(t-\alpha t)^{t-\alpha t}} = t^{-t}.$$

Hence,

$$P(\sum_{i=1}^{t} X_i = a) \leq \frac{e}{2\pi} \times \sqrt{\frac{2}{\alpha(1-\alpha)}} \times t^{t-\frac{1}{2}} \times f(\hat{k})$$
$$\leq \frac{e}{2\pi} \times \sqrt{\frac{2}{\alpha(1-\alpha)}} \times t^{-\frac{1}{2}}.$$

□

Proof of Corollary 4. Note that when $\alpha = \frac{1}{2}$, Lemma 31 provides a bound on $\frac{S(n)}{2^n}$. Plug in $\alpha = \frac{1}{2}$ to Lemma 31 and combine with Theorem 3, we know that if $n > 4$, $C_{LO} = \frac{\sqrt{2}e}{\pi}$ will suffice. If $n \leq 4$, $\frac{2\sqrt{2}e}{\pi} \times n^{-\frac{1}{2}} > 1$ and Lemma 31 still holds. □

Proof of Theorem 5. We write $|A|$ to denote the cardinality of a finite set A.

$$\sum_{t\in A_0} |g_{t,i\ominus j}| P(\sum_{\tau=1}^{t-1} \epsilon_\tau g_{\tau,i\ominus j} \geq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t} \epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j})$$
$$\leq \sum_{t\in A_0} \frac{1}{\sqrt{T}} \times 1 = \frac{|A_0|}{\sqrt{T}} \leq \sqrt{|A_0|}.$$ □

Proof of Theorem 6. We write ϵ with a subset of $[T]$ as subscript to denote ϵ_t's at times that are within the subset. For example, $\epsilon_{[T]} = \{\epsilon_1, \ldots, \epsilon_T\}$. We also write ϵ_{-A} to denote the set of ϵ_t's that are within the complement of A with respect to $[T]$.

Case I: $k \in \{1, \ldots, K-1\}$

$$\sum_{t\in A_k} |g_{t,i\ominus j}| P\left(\sum_{\tau=1}^{t-1} \epsilon_\tau g_{\tau,i\ominus j} \geq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t} \epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j}\right)$$
$$= \sum_{t\in A_k} |g_{t,i\ominus j}| \mathbb{E}_{\epsilon_{[T]}}\left[\mathbb{1}_{(\sum_{\tau=1}^{t-1} \epsilon_\tau g_{\tau,i\ominus j} \geq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t} \epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j})}\right]$$
$$= \sum_{t\in A_k} |g_{t,i\ominus j}| \mathbb{E}_{\epsilon_{-A_k}}\left[\mathbb{E}_{\epsilon_{A_k}}\left[\mathbb{1}_{(\sum_{\tau=1}^{t-1} \epsilon_\tau g_{\tau,i\ominus j} \geq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t} \epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j})} | \epsilon_{-A_k}\right]\right]$$
$$= \mathbb{E}_{\epsilon_{-A_k}}\left[\sum_{t\in A_k} |g_{t,i\ominus j}| \mathbb{E}_{\epsilon_{A_k}}\left[\mathbb{1}_{(\sum_{\tau=1}^{t-1} \epsilon_\tau g_{\tau,i\ominus j} \geq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t} \epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j})} | \epsilon_{-A_k}\right]\right]$$
$$\leq \sup_{\epsilon_{-A_k}} \sum_{t\in A_k} |g_{t,i\ominus j}| \mathbb{E}_{\epsilon_{A_k}}\left[\mathbb{1}_{(\sum_{\tau=1}^{t-1} \epsilon_\tau g_{\tau,i\ominus j} \geq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t} \epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j})} | \epsilon_{-A_k}\right].$$

Let $A_k = \{t_{k,1}, \ldots, t_{k,|A_k|}\}$ with elements listed in increasing order of time index. Also, define

$$D_n = D_n(\epsilon_{-A_k}) = -\sum_{\tau=1, \tau \in -A_k}^{t_{k,n}-1} (1+\epsilon_\tau) g_{\tau, i \ominus j}.$$

Then, we have

$$\begin{aligned}
&\sum_{t \in A_k} |g_{t,i\ominus j}| \mathbb{E}_{\epsilon_{A_k}} \left[\mathbb{1}_{(\sum_{\tau=1}^{t-1} \epsilon_\tau g_{\tau,i\ominus j} \geq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t} \epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j})} \middle| \epsilon_{-A_k} \right] \\
=&\sum_{n=1}^{|A_k|} |g_{t_{k,n},i\ominus j}| \mathbb{E}_{\epsilon_{A_k}} \left[\mathbb{1}_{(\sum_{s=1}^{n-1} \epsilon_{t_{k,s}} g_{t_{k,s},i\ominus j} \geq -\sum_{s=1}^{n-1} g_{t_{k,s},i\ominus j} + D_n, \sum_{s=1}^{n} \epsilon_{t_{k,s}} g_{s,i\ominus j} \leq -\sum_{s=1}^{n} g_{t_{k,s},i\ominus j} + D_n)} \middle| \epsilon_{-A_k} \right] \\
=&\sum_{n=1}^{|A_k|} |g_{t_{k,n},i\ominus j}| P \left(\sum_{s=1}^{n-1} \epsilon_{t_{k,s}} g_{t_{k,s},i\ominus j} \geq -\sum_{s=1}^{n-1} g_{t_{k,s},i\ominus j} + D_n, \right. \\
&\qquad \left. \sum_{s=1}^{n} \epsilon_{t_{k,s}} g_{t_{k,s},i\ominus j} \leq -\sum_{s=1}^{n} g_{t_{k,s},i\ominus j} + D_n \middle| \epsilon_{-A_k} \right).
\end{aligned}$$

By definition of the set A_k, we have $|g_{t_{k,s},i\ominus j}| \geq T^{-\frac{1}{2^k}}$, so $T^{\frac{1}{2^k}} |g_{t_{k,s},i\ominus j}| \geq 1$. Let $M_k = T^{\frac{1}{2^k}}$. Then, we have

$$\begin{aligned}
&\sum_{n=1}^{|A_k|} |g_{t_{k,n},i\ominus j}| P \left(\sum_{s=1}^{n-1} \epsilon_{t_{k,s}} g_{t_{k,s},i\ominus j} \geq -\sum_{s=1}^{n-1} g_{t_{k,s},i\ominus j} + D_n, \right. \\
&\qquad \left. \sum_{s=1}^{n} \epsilon_{t_{k,s}} g_{t_{k,s},i\ominus j} \leq -\sum_{s=1}^{n} g_{t_{k,s},i\ominus j} + D_n \middle| \epsilon_{-A_k} \right) \\
=&\sum_{n=1}^{|A_k|} |g_{t_{k,n},i\ominus j}| P \left(\sum_{s=1}^{n-1} \epsilon_{t_{k,s}} g_{t_{k,s},i\ominus j} M_k \geq -\sum_{s=1}^{n-1} g_{t_{k,s},i\ominus j} M_k + D_n M_k, \right. \\
&\qquad \left. \sum_{s=1}^{n} \epsilon_{t_{k,s}} g_{t_{k,s},i\ominus j} M_k \leq -\sum_{s=1}^{n} g_{t_{k,s},i\ominus j} M_k + D_n M_k \middle| \epsilon_{-A_k} \right) \\
\leq&\sum_{n=1}^{|A_k|} |g_{t_{k,n},i\ominus j}| P \left(\sum_{s=1}^{n} \epsilon_{t_{k,s}} g_{t_{k,s},i\ominus j} M_k \in B_{k,n} \middle| \epsilon_{-A_k} \right)
\end{aligned}$$

where

$$B_{k,n} = \left[-\sum_{s=1}^{n} g_{t_{k,s},i\ominus j} M_k + D_n M_k - 2|g_{t_{k,n},i\ominus j}| M_k, -\sum_{s=1}^{n} g_{t_{k,s},i\ominus j} M_k + D_n M_k \right]$$

is a one-dimensional closed ball with radius $\Delta = |g_{t_{k,n},i\ominus j}|M_k$. Note that this ball is fixed given ϵ_{-A_k}. Since $|g_{t_{k,s},i\ominus j}|M_k \geq 1$, we can apply Corollary 4 to get

$$P\left(\sum_{s=1}^{n}\epsilon_{t_{k,s}}g_{t_{k,s},i\ominus j}M_k \in B_{k,n}\middle|\epsilon_{-A_k}\right) \leq \frac{C_{LO}(\Delta+1)}{\sqrt{n}} = \frac{C_{LO}(|g_{t_{k,n},i\ominus j}|M_k+1)}{\sqrt{n}}.$$

Now we continue the derivation,

$$\begin{aligned}
&\sum_{n=1}^{|A_k|}|g_{t_{k,n},i\ominus j}|P\left(\sum_{s=1}^{n}\epsilon_{t_{k,s}}g_{t_{k,s},i\ominus j}M_k \in B_{k,n}\middle|\epsilon_{-A_k}\right)\\
\leq&\sum_{n=1}^{|A_k|}|g_{t_{k,n},i\ominus j}|\frac{C_{LO}(|g_{t_{k,n},i\ominus j}|M_k+1)}{\sqrt{n}}\\
\leq&C_{LO}\left(\sum_{n=1}^{|A_k|}\frac{|g_{t_{k,n},i\ominus j}|^2M_k}{\sqrt{n}}+\sum_{n=1}^{|A_k|}\frac{2}{\sqrt{n}}\right).
\end{aligned}$$

Since we have $|g_{t_{k,n},i\ominus j}| < T^{-\frac{1}{2^{k+1}}}$, $|g_{t_{k,n},i\ominus j}|^2T^{\frac{1}{2^k}} = |g_{t_{k,n},i\ominus j}|^2M_k < 1$. Thus we have the bound,

$$C_{LO}\left(\sum_{n=1}^{|A_k|}\frac{|g_{t_{k,n},i\ominus j}|^2M_k}{\sqrt{n}}+\sum_{n=1}^{|A_k|}\frac{2}{\sqrt{n}}\right) \leq 3C_{LO}\sum_{n=1}^{|A_k|}\frac{1}{\sqrt{n}} \leq 6C_{LO}\sqrt{|A_k|}.$$

Case II: $k = K$. Similar to the previous case, we have

$$\begin{aligned}
&\sum_{t\in A_K}|g_{t,i\ominus j}|P\left(\sum_{\tau=1}^{t-1}\epsilon_\tau g_{\tau,i\ominus j} \geq -\sum_{\tau=1}^{t-1}g_{\tau,i\ominus j}, \sum_{\tau=1}^{t}\epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t}g_{\tau,i\ominus j}\right)\\
\leq&\sup_{\epsilon_{-A_K}}\sum_{t\in A_K}|g_{t,i\ominus j}|\mathbb{E}_{\epsilon_{A_k}}[\mathbb{1}_{(\sum_{\tau=1}^{t-1}\epsilon_\tau g_{\tau,i\ominus j}\geq-\sum_{\tau=1}^{t-1}g_{\tau,i\ominus j},\sum_{\tau=1}^{t}\epsilon_\tau g_{\tau,i\ominus j}\leq-\sum_{\tau=1}^{t}g_{\tau,i\ominus j})}|\epsilon_{-A_K}]
\end{aligned}$$

and writing the elements of A_K in increasing order as $\{t_{K,1},\ldots,t_{K,|A_K|}\}$, we get

$$\begin{aligned}
&\sum_{t\in A_K}|g_{t,i\ominus j}|\mathbb{E}_{\epsilon_{A_K}}[\mathbb{1}_{(\sum_{\tau=1}^{t-1}\epsilon_\tau g_{\tau,i\ominus j}\geq-\sum_{\tau=1}^{t-1}g_{\tau,i\ominus j},\sum_{\tau=1}^{t}\epsilon_\tau g_{\tau,i\ominus j}\leq-\sum_{\tau=1}^{t}g_{\tau,i\ominus j})}|\epsilon_{-A_K}]\\
\leq&\sum_{n=1}^{|A_K|}|g_{t_{K,n},i\ominus j}|P\left(\sum_{s=1}^{n}\epsilon_{t_{K,s}}g_{t_{K,s},i\ominus j}M_K \in B_{K,n}\right)
\end{aligned}$$

where

$$D_n = D_n(\epsilon_{-A_K}) = -\sum_{\tau=1,\tau\in\{-A_K\}}^{t_{K,n}-1}(1+\epsilon_\tau)g_{\tau,i\ominus j},$$

$M_K = T^{\frac{1}{2^K}} \leq 2$, and

$$B_{K,n} = \left[-\sum_{s=1}^{n} g_{t_{K,s},i\ominus j} M_K + D_n M_K - 2|g_{t_{K,n},i\ominus j}|M_K, -\sum_{s=1}^{n} g_{t_{K,s},i\ominus j} M_K + D_n M_K \right]$$

is a one-dimensional closed ball with radius $\Delta = |g_{t_{K,n},i\ominus j}|M_K$. Note that this ball is fixed given ϵ_{-A_K} and hence, we can apply Corollary 4 to get

$$\begin{aligned}
&\sum_{n=1}^{|A_K|} |g_{t_{K,n},i\ominus j}| P\left(\sum_{s=1}^{n} \epsilon_{t_{K,s}} g_{t_{K,s},i\ominus j} M_K \in B_{K,n} \right) \\
\leq &\sum_{n=1}^{|A_K|} |g_{t_{K,n},i\ominus j}| \frac{C_{LO}(|g_{t_{K,n},i\ominus j}|M_K + 1)}{\sqrt{n}} \\
\leq &C_{LO}(4M_K + 2) \sum_{n=1}^{|A_K|} \frac{1}{\sqrt{n}} \leq 20 C_{LO} \sqrt{|A_K|}.
\end{aligned}$$

Combining the two cases proves the theorem. □

Proof of Theorem 9. Consider a game with two strategies, i.e., $N = 2$. We refer to player i as the "player" and the other players collectively as the "environment". On odd rounds, the environment plays payoff vector $(0, 0)$. This ensures that after odd rounds, the environment will know exactly which strategy the player will choose as long as there is no tie in the player's sampled cumulative payoffs, because no matter whether the Rademacher random variable is -1 or $+1$, the next strategy played will be the same as the strategy the player just played. On even rounds t, the environment plays the payoff vector $(0, 1 - 0.1^t)$ if the player chose the first strategy in the previous round, and $(1 - 0.1^t, 0)$ if the player chose the second strategy in the previous round. Under this scenario, we make a critical observation that, as long as it is not empty, there cannot be a tie in the cumulative payoffs of the two strategies. Moreover, without a tie, the player will not be able to switch strategy on even rounds so will not accumulate any payoff. Therefore, the total payoff acquired by the player by following sampled fictitious play procedure will be at most $2 + \frac{T}{8}$ since after the

second round the probbility of a tie is at most $\frac{1}{8}$. because the player gains 0 on all rounds. However, as evident from the environment's procedure, the total payoff for two strategies is at least $0.45T$ and thus the best strategy has a payoff no less than $0.225T$ because of the pigeonhole principle. Hence, the expected regret for the player is at least $0.225T - \frac{T}{8} - 2$, which is linear in T. □

1.2 Counterexample of Polynomial Dependence on N

In this section we present a counterexample which shows that the sampled fictitious play algorithm (2.1) with Bernoulli sampling (2.2) has expected regret of $\Omega(N)$ when T is $2N$ and $N \to \infty$. This roughly corresponds to a lower bound of the expected regret of order $\Omega(\sqrt{NT})$. The idea of this counterexample is from [**?**] and private communication with Manfred Warmuth and Gergely Neu.

Theorem 32 (**?**). *The sampled fictitious play algorithm has expected regret of* $\Omega(N)$ *when* T *is* $2N$ *and* $N \to \infty$.

Proof. Consider the payoff matrix of

$$\begin{bmatrix} 0 & -1 & -1 & -1 & \dots \\ -1 & 0 & 0 & -1 & \dots \\ -1 & 0 & -1 & 0 & \dots \\ -1 & 0 & -1 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & 0 & -1 & 0 & \dots \end{bmatrix}.$$

Each row represents a strategy and each column represents payoffs of the strategies in a particular round. For $m \in \{1, \dots, N\}$, in the $2m-1$th round, the adversary assigns a payoff of -1 to all strategy except strategy m. In the $2m$th round, the adversary assigns a payoff of -1 to strategy m and a payoff of 0 to the others. In all

rounds after $2m$, strategy m will always be given a payoff of -1. Therefore, we will have N strategies and $2N$ rounds in total, with the best constant strategy being the last strategy which accumulates payoff of $-N$.

To analyze the expected regret, we note that as long as round $2m-1$ is not ignored, which happends with probability $\frac{1}{2}$, the algorithm will always choose from strategy 1 to strategy m for round $2m$, all of which acquire a gain of -1 and so the algorithm will acquire an expected payoff of at most $-\frac{1}{2}$ on even rounds. On round $2m-1$, we observe that the leader set from the previous round will consist of at least all strategies from m to N, and possibly more strategies in the set {1,. . . ,m-1}. Since all strategies except strategy m acquire a gain of -1 on round $2m-1$, we conclude that the algorithm will acquire an expected gain of at most $-\frac{N-m}{N-m+1}$ on odd rounds. Hence, the expected regret of Sampled Fictitious Play algorithm under this scenario with N strategies and $2N$ rounds is

$$\mathcal{R}_T = -N - \Big(-\sum_{m=1}^{N} (\frac{N-m}{N-m+1}) - \frac{N}{2} \Big) \approx \frac{N}{2} - \log(N) = \Omega(N).$$

□

1.3 Asymmetric Probabilities

In this section we prove that for binary payoff and arbitrary probability $\alpha \in (0,1)$ insead of just 1/2, the expected regret is $O(\sqrt{T})$ where the constant hidden in $O(\cdot)$ notation blows up in either of the two extreme case: $\alpha \to 0$ and $\alpha \to 1$. Note that we are still considering the single stream version (2.3) of the learning procedure.

Theorem 33. *For $\alpha \in (0,1)$ and $g_t \in \{-1,0,1\}^N$, assuming that $T > \max(\frac{2}{1-\alpha}, \frac{2}{\alpha})$, the expected regret satisfies*

$$\mathbb{E}\left[\mathcal{R}_T\right] \leq \frac{40N^2 Q_\alpha}{\alpha}\sqrt{T}$$

where $Q_\alpha = \frac{e}{2\pi} \times \sqrt{\frac{2}{\alpha(1-\alpha)}}$.

Proof. We begin with the inequality obtained in the proof of Theorem 2:

$$\mathbb{E}\left[\mathcal{R}_T\right] \leq \frac{N^2}{\alpha} \max_{1\leq i,j\leq N} \sum_{t=1}^{T} |g_{t,i\ominus j}| P\left(\tilde{G}_{t-1,i\ominus j} \geq 0, \tilde{G}_{t,i\ominus j} \leq 0\right). \tag{1.1}$$

As before, we fix i and j, and will bound the expression

$$\sum_{t=1}^{T} |g_{t,i\ominus j}| P(\sum_{\tau=1}^{t-1}(1+\epsilon_\tau)g_{\tau,i\ominus j} \geq 0, \sum_{\tau=1}^{t}(1+\epsilon_\tau)g_{\tau,i\ominus j} \leq 0).$$

The rest of the proof is similar to the proof of Theorem 6. Define the classes $A_k = \{t : g_{t,i\ominus j} = k, t = 1,\ldots,T\}$ for $k \in \{-2,-1,1,2\}$. We have,

$$\begin{aligned}
&\sum_{t=1}^{T} |g_{t,i\ominus j}| P\left(\sum_{\tau=1}^{t-1}(1+\epsilon_\tau)g_{\tau,i\ominus j} \geq 0, \sum_{\tau=1}^{t}(1+\epsilon_\tau)g_{\tau,i\ominus j} \leq 0\right) \\
\leq& 2\sum_{t=1}^{T} P\left(\sum_{\tau=1}^{t-1}\epsilon_\tau g_{\tau,i\ominus j} \geq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t}\epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j}\right) \\
=& 2\sum_{k\in\{-2,-1,1,2\}} \sum_{t\in A_k} P\left(\sum_{\tau=1}^{t-1}\epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t}\epsilon_s g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j}\right).
\end{aligned}$$

For any $k \in \{-2,-1,1,2\}$,

$$\begin{aligned}
&\sum_{t\in A_k} P\left(\sum_{\tau=1}^{t-1}\epsilon_\tau g_{\tau,i\ominus j} \geq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t}\epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j}\right) \\
&\leq \sup_{\epsilon_{-A_k}} \sum_{t\in A_k} \mathbb{E}_{\epsilon_{A_k}}[\mathbb{1}_{\sum_{\tau=1}^{t-1}\epsilon_\tau g_{\tau,i\ominus j} \geq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t}\epsilon_\tau g_{\tau,i\ominus j} \leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j}} | \epsilon_{-A_k}].
\end{aligned}$$

Let $A_k = \{t_{k,1},\ldots,t_{k,|A_k|}\}$ with elements listed in increasing order of time index. Also define, for $n \in \{1,\ldots,|A_k|\}$,

$$D_n = D_n(\epsilon_{-A_k}) = -\sum_{\tau=1,\tau\in\{-A_k\}}^{t_{k,n}-1} \epsilon_\tau g_{\tau,i\ominus j} - \sum_{\tau=1,\tau\in\{-A_k\}}^{t_{k,n}-1} g_{\tau,i\ominus j}.$$

We then proceed as follows.

$$
\begin{aligned}
&\sum_{t\in A_k} \mathbb{E}_{\epsilon_{A_k}}\left[\mathbb{1}_{\sum_{\tau=1}^{t-1}\epsilon_\tau g_{\tau,i\ominus j}\geq -\sum_{\tau=1}^{t-1} g_{\tau,i\ominus j}, \sum_{\tau=1}^{t}\epsilon_\tau g_{\tau,i\ominus j}\leq -\sum_{\tau=1}^{t} g_{\tau,i\ominus j}}\middle|\epsilon_{-A_k}\right]\\
=&\sum_{n=1}^{|A_k|} \mathbb{E}_{\epsilon_{A_k}}\left[\mathbb{1}_{(\sum_{s=1}^{n-1}\epsilon_{t_{k,s}} g_{t_{k,s},i\ominus j}\geq -\sum_{s=1}^{n-1} g_{t_{k,s},i\ominus j}+D_n, \sum_{s=1}^{n}\epsilon_{t_{k,s}} g_{s,i\ominus j}\leq -\sum_{s=1}^{n} g_{t_{k,s},i\ominus j}+D_n)}\middle|\epsilon_{-A_k}\right]\\
=&\sum_{n=1}^{|A_k|} P\left(\sum_{s=1}^{n-1}\epsilon_{t_{k,s}} g_{t_{k,s},i\ominus j}\geq -\sum_{s=1}^{n-1} g_{t_{k,s},i\ominus j}+D_n,\right.\\
&\qquad\left.\sum_{s=1}^{n}\epsilon_{t_{k,s}} g_{t_{k,s},i\ominus j}\leq -\sum_{s=1}^{n} g_{t_{k,s},i\ominus j}+D_n\middle|\epsilon_{-A_k}\right)\\
\leq&\sum_{n=1}^{|A_k|} P\left(\bigcup_{u=0}^{4}(\sum_{s=1}^{n}\epsilon_{t_{k,s}} g_{t_{k,s},i\ominus j}=-\sum_{s=1}^{n-1} g_{t_{k,s},i\ominus j}+D_n-u)\middle|\epsilon_{-A_k}\right)\\
\leq&\sum_{n=1}^{|A_k|}\left(\sum_{u=0}^{4} P\left(\sum_{s=1}^{n}\epsilon_{t_{k,s}} g_{t_{k,s},i\ominus j}=-\sum_{s=1}^{n-1} g_{t_{k,s},i\ominus j}+D_n-u\middle|\epsilon_{-A_k}\right)\right)\\
\leq&5\sum_{n=1}^{|A_k|}\frac{Q_\alpha}{\sqrt{n}}\leq 10Q_\alpha\sqrt{|A_k|}
\end{aligned}
$$

where $Q_\alpha = \frac{e}{2\pi}\times\sqrt{\frac{2}{\alpha(1-\alpha)}}$ from Lemma 31. Putthing things together, we have

$$\mathbb{E}\left[\mathcal{R}_T\right]\leq\frac{20N^2Q_\alpha}{\alpha}\sum_{k\in\{-2,-1,1,2\}}\sqrt{|A_k|}\leq\frac{40N^2Q_\alpha}{\alpha}\sqrt{T}.$$

□

Appendix B

Proof(s) of Chapter III

2.1 Proofs

Proof of Lemma 11. First, note that the regret, by definition, is

$$\text{Regret}_T = \Phi(G_T) - \sum_{t=1}^{T} \langle \mathbf{e}_{i_t}, g_t \rangle .$$

Under an oblivious adversary, only the summation on the right hand side is random. Moreover $\mathbb{E}[\langle \mathbf{e}_{i_t}, g_t \rangle \, | i_{1:t-1}] = \langle p_t, g_t \rangle$. This proves the claim in (3.4).

From (3.2), we know that $\mathbb{E}[\langle p_t, \hat{g}_t \rangle \, | i_{1:t-1}] = \langle p_t, g_t \rangle$ even if some entries in p_t might be zero. Therefore, we have

$$\mathbb{E}\text{Regret}_T = \Phi(G_T) - \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t, \hat{g}_t \rangle\right]. \tag{2.1}$$

From (3.3), we know that $G_T \leq \mathbb{E}[\hat{G}_T]$. This implies

$$\Phi(G_T) \leq \Phi(\mathbb{E}[\hat{G}_T]) \leq \mathbb{E}[\Phi(\hat{G}_T)], \tag{2.2}$$

where the first inequality is because $G \succeq G' \Rightarrow \Phi(G) \geq \Phi(G')$, and the second inequality is due to the convexity of Φ. Plugging (2.2) into (2.1) yields

$$\mathbb{E}\text{Regret}_T \leq \mathbb{E}\left[\Phi(\hat{G}_T) - \sum_{t=1}^{T} \langle p_t, \hat{g}_t \rangle\right]. \tag{2.3}$$

Now considering the quantity inside the expectation above and recalling the definition of Bregman divergence

$$D_{\tilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1}) = \tilde{\Phi}(\hat{G}_t) - \tilde{\Phi}(\hat{G}_{t-1}) - \left\langle \nabla\tilde{\Phi}(\hat{G}_{t-1}), \hat{G}_t - \hat{G}_{t-1} \right\rangle$$

we get,

$$\begin{aligned}
\Phi(\hat{G}_T) - \sum_{t=1}^{T} \langle p_t, \hat{g}_t \rangle &= \Phi(\hat{G}_T) - \sum_{t=1}^{T} \left\langle \nabla\tilde{\Phi}(\hat{G}_{t-1}), \hat{g}_t \right\rangle \\
&= \Phi(\hat{G}_T) - \sum_{t=1}^{T} \left\langle \nabla\tilde{\Phi}(\hat{G}_{t-1}), \hat{G}_t - \hat{G}_{t-1} \right\rangle \\
&= \Phi(\hat{G}_T) + \sum_{t=1}^{T} \left(D_{\tilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1}) + \tilde{\Phi}(\hat{G}_{t-1}) - \tilde{\Phi}(\hat{G}_t) \right) \\
&= \Phi(\hat{G}_T) + \tilde{\Phi}(\hat{G}_0) - \tilde{\Phi}(\hat{G}_T) + \sum_{t=1}^{T} D_{\tilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1}). \qquad (2.4)
\end{aligned}$$

The proof ends by plugging in (2.4) into (2.3) and noting that $\tilde{\Phi}(\hat{G}_0) = \tilde{\Phi}(0)$ is not random. □

Proof of Lemma 12. We have,

$$\begin{aligned}
\Phi(G) + \mathbb{E}[Z_1] &= \max_i G_i + \mathbb{E}[Z_i] = \max_i (G_i + \mathbb{E}[Z_i]) \\
&\leq \mathbb{E}[\max_i (G_i + Z_i)] = \tilde{\Phi}(G) \\
&\leq \mathbb{E}[\max_i G_i + \max_i Z_i] = \max_i G_i + \mathbb{E}[\max_i Z_i] = \Phi(G) + \mathbb{E}[\max_i Z_i].
\end{aligned}$$

Noting that $\mathbb{E}[\max_i Z_i] \leq EMAX(N)$ finishes the proof. □

Proof of Lemma 13. To reduce clutter, we drop the time subscripts: we use $\hat{G}$ to denote the cumulative estimate $\hat{G}_{t-1}$, $\hat{g}$ to denote the marginal estimate $\hat{g}_t = \hat{G}_t - \hat{G}_{t-1}$, p to denote p_t, and g to denote the true loss g_t. Note that by definition of Framework 1, $\hat{g}$ is a sparse vector with one non-zero and non-positive coordinate

$\hat{g}_{i_t} = g_{i_t}/p_{i_t} = -|g_{i_t}/p_{i_t}|$. Morever, conditioned on $i_{1:t-1}$, i_t takes value i with probability p_i. For any $i \in \text{supp}(p)$, let

$$h_i(r) = D_{\tilde{\Phi}}(\hat{G} - r\mathbf{e}_i, \hat{G}),$$

so that $h_i'(r) = -\nabla_i \tilde{\Phi}\left(\hat{G} - r\mathbf{e}_i\right) + \nabla_i \tilde{\Phi}\left(\hat{G}\right)$ and $h_i''(r) = \nabla_{ii}^2 \tilde{\Phi}\left(\hat{G} - r\mathbf{e}_i\right)$. Now we write:

$$\begin{aligned}
\mathbb{E}[D_{\tilde{\Phi}}(\hat{G} + \hat{g}, \hat{G})|i_{1:t-1}] &= \sum_{i \in \text{supp}(p)} p_i D_{\tilde{\Phi}}(\hat{G} + g_i/p_i \mathbf{e}_i, \hat{G}) = \sum_{i \in \text{supp}(p)} p_i D_{\tilde{\Phi}}(\hat{G} - |g_i/p_i|\, \mathbf{e}_i, \hat{G}) \\
&= \sum_{i \in \text{supp}(p)} p_i h_i(|g_i/p_i|) = \sum_{i \in \text{supp}(p)} p_i \int_0^{|g_i/p_i|} \int_0^s h_i''(r) dr\, ds \\
&= \sum_{i \in \text{supp}(p)} p_i \int_0^{|g_i/p_i|} \int_0^s \nabla_{ii}^2 \tilde{\Phi}\left(\hat{G} - r\mathbf{e}_i\right) dr\, ds \\
&= \sum_{i \in \text{supp}(p)} p_i \int_0^{|g_i/p_i|} \int_0^s \mathbb{E}_{\hat{G}_{-i}} f(\hat{G}_{-i} - \hat{G}_i + r) dr\, ds \\
&= \sum_{i \in \text{supp}(p_t)} p_{t,i} \int_0^{|g_i/p_i|} \mathbb{E}_{\hat{G}_{-i}} \left[\int_0^s f(\hat{G}_{-i} - \hat{G}_i + r) dr\right] ds.
\end{aligned}$$

□

Proof of Theorem 14. From Lemma 13, we have, with $\hat{G}_{-i} = \max_{j \neq i} \hat{G}_{t-1,j} + \eta Z_j$,

$$\begin{aligned}
&\mathbb{E}\left[D_{\tilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1})|i_{1:t-1}\right] \\
&= \sum_{i \in \text{supp}(p_t)} p_{t,i} \int_0^{\left|\frac{g_{t,i}}{p_{t,i}}\right|} \mathbb{E}_{\hat{G}_{-i}} \left[\int_0^s f_\eta(\hat{G}_{-i} - \hat{G}_{t-1,i} + r) dr\right] ds \\
&= \sum_{i \in \text{supp}(p_t)} p_{t,i} \int_0^{\left|\frac{g_{t,i}}{p_{t,i}}\right|} \mathbb{E}_{\hat{G}_{-i}} \left[\int_{\hat{G}_{-i} - \hat{G}_{t-1,i}}^{\hat{G}_{-i} - \hat{G}_{t-1,i} + s} f_\eta(z) dz\right] ds \\
&= \sum_{i \in \text{supp}(p_t)} p_{t,i} \int_0^{\left|\frac{g_{t,i}}{p_{t,i}}\right|} \Big(\mathbb{E}_{\hat{G}_{-i}} \Big[\underbrace{\int_{[\hat{G}_{-i} - \hat{G}_{t-1,i}, \hat{G}_{-i} - \hat{G}_{t-1,i} + s] \setminus [F_\eta^{-1}(1-\epsilon), \eta]} f_\eta(z) dz}_{(I)}\Big] \\
&\qquad + \underbrace{\int_{[F_\eta^{-1}(1-\epsilon), \eta]} f_\eta(z) dz}_{(II)}\Big) ds. \qquad (2.5)
\end{aligned}$$

We bound the two integrals above differently. For the first integral, we add the restriction $f_\eta(z) > 0$ by intersecting the integral interval with the support of the function $f_\eta(z)$, denoted as $I_{f_\eta(z)}$ so that $1 - F_\eta(z)$ is not 0 on the interval to be integrated. Thus, we get,

$$\begin{aligned}
(I) &= \int_{([\hat{G}_{-i}-\hat{G}_{t-1,i},\hat{G}_{-i}-\hat{G}_{t-1,i}+s]\setminus[F_\eta^{-1}(1-\epsilon),\eta])\cap I_{f_\eta(z)}} f_\eta(z)dz \\
&= \int_{([\hat{G}_{-i}-\hat{G}_{t-1,i},\hat{G}_{-i}-\hat{G}_{t-1,i}+s]\setminus[F_\eta^{-1}(1-\epsilon),\eta])\cap I_{f_\eta(z)}} (1 - F_\eta(z)) \cdot \frac{f_\eta(z)}{1 - F_\eta(z)} dz \\
&\leq \int_{([\hat{G}_{-i}-\hat{G}_{t-1,i},\hat{G}_{-i}-\hat{G}_{t-1,i}+s]\setminus[F_\eta^{-1}(1-\epsilon),\eta])\cap I_{f_\eta(z)}} (1 - F_\eta(z)) \cdot \frac{L}{\eta\epsilon} \\
&\leq (1 - F_\eta(\hat{G}_{-i} - \hat{G}_{t-1,i}))\frac{sL}{\eta\epsilon}. \qquad (2.6)
\end{aligned}$$

The first inequality holds because $f_\eta(z) \leq L/\eta$ and $(1 - F_\eta(z)) \geq \epsilon$ on the set of z's over which we are integrating. The second inequality holds because on the set under consideration $1 - F_\eta(z) \leq 1 - F_\eta(\hat{G}_{-i} - \hat{G}_{t-1,i})$ and the measure of the set is at most s.

For the second integral, we use the bound $f_\eta(z) \leq L/\eta$ again to get,

$$(II) = \int_{[F_\eta^{-1}(1-\epsilon),\eta]} f_\eta(z)dz \leq \frac{L}{\eta} \cdot (\eta - F_\eta^{-1}(1-\epsilon)). \qquad (2.7)$$

Plugging in (2.6) and (2.7) into (2.5), we can bound the divergence penalty by,

$$\begin{aligned}
&\leq \sum_{i\in\text{supp}(p_t)} p_{t,i} \int_0^{\left|\frac{g_{t,i}}{p_{t,i}}\right|} \Big(\mathbb{E}_{\hat{G}_{-i}}[1 - F_\eta(\hat{G}_{-i} - \hat{G}_{t-1,i})]\frac{sL}{\eta\epsilon} + \frac{L(\eta - F_\eta^{-1}(1-\epsilon))}{\eta}\Big)ds \\
&= \sum_{i\in\text{supp}(p_t)} p_{t,i} \int_0^{\left|\frac{g_{t,i}}{p_{t,i}}\right|} \Big(p_{t,i}\frac{sL}{\eta\epsilon} + L(1 - F^{-1}(1-\epsilon))\Big)ds \\
&= \sum_{i\in\text{supp}(p_t)} p_{t,i}\Big(p_{t,i}\frac{L}{\eta\epsilon}\frac{g_{t,i}^2}{2p_{t,i}^2} + L(1 - F^{-1}(1-\epsilon))\frac{|g_{t,i}|}{p_{t,i}}\Big) \\
&\leq \sum_{i\in\text{supp}(p_t)} \Big(\frac{L}{2\eta\epsilon} + L(1 - F^{-1}(1-\epsilon))\Big) \\
&\leq NL\Big(\frac{1}{2\eta\epsilon} + 1 - F^{-1}(1-\epsilon)\Big).
\end{aligned}$$

The second to last inequality holds because $|g_{t,i}| \leq 1$ and the last inequality holds because the sum over i is at most over all N arms. □

Proof of Corollary 15. For $[0,1]$ uniform distribution, we have $L = 1$, $F^{-1}(1-\epsilon) = 1-\epsilon$ so the divergence penalty is upper bounded by

$$NT(\frac{1}{2\eta\epsilon} + \epsilon).$$

If we let $\epsilon = \frac{1}{\sqrt{2\eta}}$, we can see that the divergence penalty is upper bounded by $NT\sqrt{\frac{2}{\eta}}$. Together with the overestimation penalty which is trivially bounded by η and a non-positive underestimation penalty, we see that the final regret bound is

$$NT\sqrt{\frac{2}{\eta}} + \eta.$$

Setting $\eta = (NT)^{2/3}$ gives the desired result. □

Proof of Corollary 16. For a general distribution, let $\epsilon = \frac{1}{\sqrt{\eta}}$. Since the overestimation penalty is trivially bounded by η and the underestimation penalty is non-positive, the expected regret can be upper bounded by

$$LNT\Big(\frac{1}{2\sqrt{\eta}} + 1 - F^{-1}(1 - \frac{1}{\sqrt{\eta}})\Big) + \eta.$$

Setting $\eta = (NT)^{2/3}$ we see that the expected regret can be upper bounded by

$$(\frac{L}{2} + 1)(NT)^{2/3} + LNT(1 - F^{-1}(1 - \frac{1}{\sqrt{\eta}}).$$

Since

$$\lim_{T\to\infty} 1 - F^{-1}(1 - \frac{1}{\sqrt{\eta}}) = \lim_{\eta\to\infty} 1 - F^{-1}(1 - \frac{1}{\sqrt{\eta}}) = 1 - F^{-1}(1) = 0,$$

we conclude that

$$\lim_{T\to\infty} \frac{\mathbb{E}\text{Regret}_T}{T} = 0.$$

□

Proof of Theorem 17. Because of the unbounded support of Z, $\mathrm{supp}(p_t) = \{1, \ldots, N\}$. Lemma 13 gives us:

$$\begin{aligned}
\mathbb{E}[D_{\tilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1})|i_{1:t-1}] &= \sum_{i=1}^{N} p_{t,i} \int_0^{|g_{t,i}/p_{t,i}|} \mathbb{E}_{\tilde{G}_{-i}} \int_0^s f(\tilde{G}_{-i} - \hat{G}_{t-1,i} + r) dr ds \\
&= \sum_{i=1}^{N} p_{t,i} \int_0^{|g_{t,i}/p_{t,i}|} \mathbb{E}_{\tilde{G}_{-i}} \int_{\tilde{G}_{-i} - \hat{G}_{t-1,i}}^{\tilde{G}_{-i} - \hat{G}_{t-1,i} + s} f(z) dz ds \\
&\leq C \sum_{i=1}^{N} p_{t,i} \int_0^{|g_{t,i}/p_{t,i}|} \mathbb{E}_{\tilde{G}_{-i}} \int_{\tilde{G}_{-i} - \hat{G}_{t-1,i}}^{\tilde{G}_{-i} - \hat{G}_{t-1,i} + s} (1 - F(z))^{1-\alpha} |z|^{-\alpha} dz\, ds \\
&\leq C \sum_{i=1}^{N} p_{t,i} \int_0^{|g_{t,i}/p_{t,i}|} \mathbb{E}_{\tilde{G}_{-i}} (1 - F(\tilde{G}_{-i} - \hat{G}_{t-1,i}))^{1-\alpha} \int_{\tilde{G}_{-i} - \hat{G}_{t-1,i}}^{\tilde{G}_{-i} - \hat{G}_{t-1,i} + s} |z|^{-\alpha} dz\, ds.
\end{aligned}$$

Since the function $f(z) = |z|^{-\alpha}$ is symmetric, monotonically decreasing as $|z| \to \infty$, we have

$$\int_{\tilde{G}_{-i} - \hat{G}_{t-1,i}}^{\tilde{G}_{-i} - \hat{G}_{t-1,i} + s} |z|^{-\alpha} dz \leq \int_{-s/2}^{s/2} |z|^{-\alpha} dz = \frac{2^\alpha}{1-\alpha} s^{1-\alpha}.$$

Also, note that $z^{1-\alpha}$ is concave. Hence, by Jensen's inequality,

$$\mathbb{E}_{\tilde{G}_{-i}}[(1 - F(\tilde{G}_{-i} - \hat{G}_{t-1,i}))^{1-\alpha}] \leq (\mathbb{E}_{\tilde{G}_{-i}}[1 - F(\tilde{G}_{-i} - \hat{G}_{t-1,i}])^{1-\alpha} = p_{t,i}^{1-\alpha}.$$

Hence,

$$\begin{aligned}
\mathbb{E}[D_{\tilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1})|i_{1:t-1}] &\leq \frac{2^\alpha C}{1-\alpha} \sum_{i=1}^{N} p_{t,i} \int_0^{|g_{t,i}/p_{t,i}|} p_{t,i}^{1-\alpha} s^{1-\alpha} ds \\
&= \frac{2^\alpha C}{1-\alpha} \sum_{i=1}^{N} p_{t,i}^{2-\alpha} \int_0^{|g_{t,i}/p_{t,i}|} s^{1-\alpha} ds \\
&= \frac{2^\alpha C}{(1-\alpha)(2-\alpha)} \sum_{i=1}^{N} p_{t,i}^{2-\alpha} |g_{t,i}/p_{t,i}|^{2-\alpha} \\
&= \frac{2^\alpha C}{(1-\alpha)(2-\alpha)} \sum_{i=1}^{N} |g_{t,i}|^{2-\alpha} \\
&\leq \frac{2^\alpha C}{(1-\alpha)(2-\alpha)} N \leq \frac{2C}{1-\alpha} N.
\end{aligned}$$

□

Proof of Theorem 18. The divergence penalty can be controlled through Theorem 17 once we have bounded generalized hazard rate. It remains to control the overestimation and underestimation penalty. By Lemma 12, they are at most $\mathbb{E}_{Z_1,\ldots,Z_n}[\max_i Z_i]$ and $-\mathbb{E}[Z_1]$ respectively. Suppose we scale the perturbation Z by $\eta > 0$, i.e., we add ηZ_i to each coordinate. It is easy to see that $\mathbb{E}[\max_{i=1,\ldots,n} \eta Z_i] = \eta\mathbb{E}[\max_{i=1,\ldots,n} Z_i]$ and $\mathbb{E}[\eta Z_1] = \eta\mathbb{E}[Z_1]$. For the divergence penalty, observe that $F_\eta(t) = F(t/\eta)$ and thus $f_\eta(t) = \frac{1}{\eta} f(t/\eta)$. Hence, the constant in the assumption needs to scale by $\eta^{\alpha-1}$. Plugging new bounds for the scaled perturbations into Lemma 11 gives us

$$\mathbb{E}\text{Regret}_T \le \eta^{\alpha-1}\frac{2C}{1-\alpha} \times NT + \eta Q(N).$$

Setting $\eta = (\frac{2CNT}{(1-\alpha)Q(N)})^{1/(2-\alpha)}$ finishes the proof. □

Proof of Lemma 20. Since the numerator of the left hand side is an even function of z, and the denominator is a decreasing function, and the inequality is trivially true when $z = 0$, it suffices to prove for $z > 0$, which we assume for the rest of the proof. From Lemma 19 we can derive that

$$\frac{f(z)}{1-F(z)} < z + 1.$$

Therefore,

$$\begin{aligned}\frac{f(z)|z|^\alpha}{(1-F(z))^{1-\alpha}} &\le \frac{f(z)z^\alpha}{(\frac{f(z)}{z+1})^{1-\alpha}} = (f(z)z)^\alpha(z+1)^{1-\alpha} \\ &\le f(z)^\alpha(z+1) \le zf(z)^\alpha + 1 = \sqrt{\frac{1}{2\pi}}ze^{-\alpha z^2/2} + 1.\end{aligned}$$

Let $g(z) = ze^{-\alpha z^2/2}$, $g'(z) = (1-\alpha z^2)e^{-\alpha z^2/2}$. Therefore $g(z)$ is maximized at $z^* = \sqrt{\frac{1}{\alpha}}$. Therefore,

$$\frac{f(z)|z|^\alpha}{(1-F(z))^{1-\alpha}} \le \sqrt{\frac{1}{2\pi}}ze^{-\alpha z^2/2} + 1 \le \sqrt{\frac{1}{2\pi}}z^* + 1 \le z^* + 1 = \sqrt{\frac{1}{\alpha}} + 1 \le \frac{2}{\alpha}.$$

□

Proof of Corollary 21. It is known that for standard Gaussian random variable, we have $\mathbb{E}[Z_1] = 0$ and

$$\mathbb{E}_{Z_1,\ldots,Z_n}[\max_i Z_i] \leq \sqrt{2\log N}.$$

Plug in to Theorem 18 gives the result. □

Proof of Theorem 22. From Corollary 21 we see that the expected regret can be upper bounded by

$$2(C_1C_2NT)^{1/(2-\alpha)}(\sqrt{2\log N})^{(1-\alpha)/(2-\alpha)}$$

where $C_1 = \frac{2}{\alpha}$ and $C_1 = \frac{2}{1-\alpha}$. Note that

$$\begin{aligned}
&2(C_1C_2NT)^{1/(2-\alpha)}(\sqrt{2\log N})^{(1-\alpha)/(2-\alpha)} \\
\leq &4(C_1C_2)^{1/(2-\alpha)}N^{1/(2-\alpha)}\sqrt{\log N}^{(1-\alpha)/(2-\alpha)}T^{1/(2-\alpha)} \\
= &4N^{1/(2-\alpha)}\sqrt{\log N}^{(1-\alpha)/(2-\alpha)}T^{1/2} \times (C_1C_2)^{1/(2-\alpha)}T^{\alpha/(4-2\alpha)} \\
\leq &4N^{1/2}N^{\alpha/(4-2\alpha)}\sqrt{\log N}T^{1/2} \times (\frac{4}{\alpha(1-\alpha)})^{1/(2-\alpha)}T^{\alpha/(4-2\alpha)} \\
\leq &4N^{1/2}N^{\alpha}\sqrt{\log N}T^{1/2} \times \frac{4T^{\alpha}}{\alpha(1-\alpha)} \\
\leq &16\sqrt{NT}N^{\alpha}\sqrt{\log N} \times \frac{T^{\alpha}}{\alpha(1-\alpha)}.
\end{aligned}$$

□

If we let $\alpha = \frac{1}{\log T}$, then $T^{\alpha} = T^{1/\log T} = e < 3$. Then, we have, for $T > 4$,

$$\frac{T^{\alpha}}{\alpha(1-\alpha)} \leq \frac{3\log T}{1-\frac{1}{\log T}} = \frac{3\log^2 T}{\log T - 1} \leq 6\log T.$$

Putting things together finishes the proof.

Proof of Lemma 24. Let $g(z) = \overline{F}(z)e^{\lambda z}$, then $g'(z) = e^{\lambda z}\overline{F}(z)(\lambda - \frac{f(z)}{\overline{F}(z)})$. Since $\frac{f(z)}{\overline{F}(z)}$ is eventually monotone by assumption (d), $g'(z)$ is eventually positive, negative or zero. The lemma immediately follows. □

Proof of Theorem 25. If the distribution is heavy-tailed, we have

$$\limsup_{z\to\infty} \overline{F}(z)e^{\lambda z} = \infty \quad \text{for all } \lambda > 0.$$

By Lemma 24, we can erase the supremum operator and just write

$$\lim_{z\to\infty} \overline{F}(z)e^{\lambda z} = \infty \quad \text{for all } \lambda > 0.$$

Hence,

$$\lim_{z\to\infty} \overline{F}(z)e^{\lambda z} = \lim_{x\to\infty} e^{-R(z)+\lambda z} = \infty \text{ for all } \lambda > 0 \Rightarrow \limsup_{z\to\infty} \frac{R(z)}{z} = 0.$$

Note that $R'(z) = \frac{f(z)}{\overline{F}(z)}$, which is eventually monotone by assumption. Therefore, we can conclude that

$$\limsup_{z\to\infty} R'(z) < \infty \Rightarrow \sup_z \frac{f(z)}{\overline{F}(z)} < \infty.$$

□

Proof of Theorem 26. If the distribution is light-tailed, we have

$$\lim_{z\to\infty} \overline{F}(z)e^{\lambda^* z} < \infty \quad \text{for some } \lambda^* > 0. \tag{2.8}$$

This immediately implies that

$$\lim_{z\to+\infty} \overline{F}(z)^a z^b = 0 \quad \forall a, b > 0. \tag{2.9}$$

Consider $\lim_{z\to\infty} \frac{f(z)}{\overline{F}(z)} = \lim_{z\to\infty} R'(z)$. If $\lim_{z\to\infty} R'(z) < \infty$ we can immediately conclude that $\sup_z \frac{f(z)}{1-F(z)} < \infty$. If $\lim_{z\to\infty} R'(z) = \infty$ instead, note that

$$\lim_{z\to\infty} \int_{-z}^{z} R'(t)e^{-\delta R(t)}dt = -\frac{1}{\delta}e^{-\delta R(z)}|_{z=-\infty}^{z=+\infty} = \frac{1}{\delta} < \infty.$$

Moreover, since $\lim_{z\to\infty} R'(z) = \infty$, $R'(z)e^{-\delta R(z)}$ is strictly positive for all $z > z_0$ for some z_0. Furthermore, $R'(z)e^{-\delta R(z)} = \frac{f(z)}{(\overline{F}(z))^{1-\delta}}$ is eventually monotone by assumption (e),

Therefore, we can conclude that

$$\lim_{z\to\infty} R'(z)e^{-\delta R(z)} = \frac{f(z)}{(\overline{F}(z))^{1-\delta}} = 0.$$

$\forall\alpha \in (\delta, 1)$, from Equation (2.9) we have $\lim_{z\to+\infty} z^{\alpha}\overline{F}(z)^{\alpha-\delta} = 0$, so

$$\lim_{z\to+\infty} \frac{f(z)z^{\alpha}}{(\overline{F}(z))^{1-\alpha}} = \lim_{z\to+\infty} \frac{f(z)}{\overline{F}(z)^{1-\delta}} \times z^{\alpha}\overline{F}(z)^{\alpha-\delta} = 0.$$

and hence

$$\sup_{z} \frac{f(z)z^{\alpha}}{(1-F(z))^{1-\alpha}} < \infty \quad \forall\alpha \in (\delta, 1).$$

□

Proof of Corollary 27. For a light-tailed distribution $\mathcal{D}$, we have

$$\lim_{z\to\infty} \overline{F}_{\mathcal{D}}(z)e^{\lambda^* z} < \infty \quad \text{for some } \lambda^* > 0.$$

This implies that

$$\overline{F}_{\mathcal{D}}(z) \leq Ce^{-\lambda^* z} \text{ for some } C > 0, z > z_0.$$

Let random variable Z follows distribution $\mathcal{D}$. Since Z might take negative values, we define a new distribution $\mathcal{D}'$ that only takes non-negative value by

$$f_{D'}(z) = \begin{cases} \frac{1}{p_{\mathcal{D}+}} f_D(z) & \text{if } z \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

where $p_{\mathcal{D}+} = \mathbb{P}(Z \geq 0) > 0$ by right unbounded support assumption. Clearly, with this definition of $\mathcal{D}'$ we see that $\mathbb{E}_{Z_1,\ldots,Z_N\sim\mathcal{D}}[\max_i Z_i] \leq \mathbb{E}_{Z_1,\ldots,Z_N\sim\mathcal{D}'}[\max_i Z_i]$ and for

$z > z_0$, we have $\overline{F}_{\mathcal{D}'}(z) = \frac{\overline{F}_{\mathcal{D}}(z)}{p_{\mathcal{D}+}} \leq C' e^{-\lambda^* z}$ where $C' = \frac{C}{p_{\mathcal{D}+}}$. Note that

$$\begin{aligned}
\mathbb{E}_{Z_1,\ldots,Z_N\sim\mathcal{D}}[\max_i Z_i] &\leq \mathbb{E}_{Z_1,\ldots,Z_N\sim\mathcal{D}'}[\max_i Z_i] \\
&= \int_0^\infty \mathbb{P}(\max_i Z_i > x)dx \\
&\leq u + \int_u^\infty \mathbb{P}(\max_i Z_i > z)dz \\
&\leq u + N\int_u^\infty \mathbb{P}(Z_i > z)dz \\
&\leq u + N\int_u^\infty C' e^{-\lambda^* z}dz \quad \text{assuming } u > z_0 \\
&= u + \frac{C'N}{\lambda^*} e^{-\lambda^* u}.
\end{aligned}$$

If we let $u = \frac{\log(N)}{\lambda^*}$, obviously $u > z_0$ if N is sufficiently large. Thus, we see that

$$\mathbb{E}_{Z_1,\ldots,Z_N\sim\mathcal{D}}[\max_i Z_i] \leq \frac{\log(N)}{\lambda^*} + C' = O(N^\xi) \quad \forall \xi > 0. \tag{2.10}$$

From Theorem 26 we see that $\forall \alpha \in (\delta, 1)$,

$$\frac{f(z)z^\alpha}{(1-F(z))^{1-\alpha}} \leq C_\alpha \quad \forall z \in \mathbb{R}. \tag{2.11}$$

Plug 2.10 and 2.11 into Theorem 18 gives the desired result. □

Proof of Corollary 29. By Corollary 27 we only need to check that assumptions (a)-(d) hold for distribution $\mathcal{D}_\beta$, exponential power family is light-tailed, and assumption (e) also holds for any $\delta \in (0, 1)$. By observing the density function f_β we can trivially see that assumptions (a)-(c) hold and that the subbotin family is light-tailed. Therefore, define

$$g_{\delta,\beta}(z) = \frac{f_\beta(z)}{(\overline{F}_\beta(z))^{1-\delta}} = \frac{f_\beta(z)}{(1-F_\beta(z))^{1-\delta}},$$

it suffices to show that $\forall \delta \in [0, 1), g_{\delta,\beta}(z)$ is eventually monotone. Note that

$$\begin{aligned}
g'_{\delta,\beta}(z) &= \frac{f'_\beta(z)(1-F_\beta(z))^{1-\delta} + (1-\delta)(1-F_\beta(z))^{-\delta} f_\beta^2(z)}{(1-F_\beta(z))^{2-2\delta}} \\
&= \frac{C_\beta^2 e^{-z^\beta}}{(1-F_\beta(z))^{2-\delta}} \times \Big((1-\delta)e^{-z^\beta} - \beta z^{\beta-1}\int_z^\infty e^{-t^\beta}dt\Big).
\end{aligned}$$

It further suffices to show that

$$m_{\delta,\beta}(z) = (1-\delta)e^{-z^\beta} - \beta z^{\beta-1}\int_z^\infty e^{-t^\beta}dt$$

is eventually non-negative or non-positive $\forall \beta > 1, \delta \in [0,1)$. Note that since $\beta > 1$,

$$\beta z^{\beta-1}\int_z^\infty e^{-t^\beta}dt = \int_z^\infty \beta z^{\beta-1}e^{-t^\beta}dt < \int_z^\infty \beta t^{\beta-1}e^{-t^\beta}dt = e^{-z^\beta}. \tag{2.12}$$

Therefore, $m_{0,\beta}(z) > 0$ for all $z \geq 0$, i.e, the hazard rate is always increasing and assumption (d) is satisfied. Now, we are left to show that $m_{\delta,\beta}(z)$ is eventually non-negative or non-positive for any $\delta \in (0,1)$. Note that

$$\begin{aligned}
\beta z^{\beta-1}\int_z^\infty e^{-t^\beta}dt &= \beta(\frac{z}{z+1})^{\beta-1}(z+1)^{\beta-1}\int_z^\infty e^{-t^\beta}dt \\
&\geq \beta(\frac{z}{z+1})^{\beta-1}(z+1)^{\beta-1}\int_z^{z+1} e^{-t^\beta}dt \\
&\geq (\frac{z}{z+1})^{\beta-1}\int_z^{z+1} \beta t^{\beta-1}e^{-t^\beta}dt \\
&= (\frac{z}{z+1})^{\beta-1}\Big(e^{-z^\beta} - e^{-(z+1)^\beta}\Big).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\liminf_{z\to\infty}\frac{\beta z^{\beta-1}\int_z^\infty e^{-t^\beta}dt}{e^{-z^\beta}} &\geq \liminf_{z\to\infty}\frac{(\frac{z}{z+1})^{\beta-1}\Big(e^{-z^\beta} - e^{-(z+1)^\beta}\Big)}{e^{-z^\beta}} \\
&= \lim_{z\to\infty}(\frac{z}{z+1})^{\beta-1} - \lim_{z\to\infty}(\frac{z}{z+1})^{\beta-1}e^{z^\beta-(z+1)^\beta} \\
&= 1.
\end{aligned}$$

From Equation (2.12) we know that

$$\limsup_{z\to\infty}\frac{\beta z^{\beta-1}\int_z^\infty e^{-t^\beta}dt}{e^{-z^\beta}} \leq 1.$$

Hence, we conclude that

$$\lim_{z\to\infty}\frac{\beta z^{\beta-1}\int_z^\infty e^{-t^\beta}dt}{e^{-z^\beta}} = 1,$$

which implies that $m_{\delta,\beta}(z)$ is eventually non-positive for any $\delta \in (0,1)$, i.e, assumption (e) holds for any $\delta \in (0,1)$. □

Bibliography

Jacob Abernethy, Chansoo Lee, Abhinav Sinha, and Ambuj Tewari. Online linear optimization via smoothing. In *COLT*, pages 807–823, 2014.

Jacob Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems 28*, pages 2188–2196, 2015.

Jacob Abernethy, Chansoo Lee, and Ambuj Tewari. *Perturbation Techniques in Online Learning and Optimization*. MIT Press, 2016.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Minimax policies for combinatorial prediction games. In *COLT*, 2011.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The non-stochastic multi-armed bandit problem. *SIAM J. Comput.*, 32:48–77, 2002.

Mark Bagnoli and Ted Bergstrom. Log-concave probability and its applications. *Economic Theory*, 26(2):445–469, 2005.

Árpád Baricz. Mills' ratio: Monotonicity patterns and functional inequalities. *J. Math. Anal. Appl.*, 340(2):1362–1370, 2008.

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

Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(Jun):1307–1324, 2007.

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

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Richard A. Chechile. Mathematical tools for hazard function analysis. *J. Math. Psychol.*, 47:478–494, 2003.

Richard A. Chechile. Corrigendum to: mathematical tools for hazard function analysis [j. math. psychol. 47 (2003) 478494]. *J. Math. Psychol.*, 53:298–299, 2009.

Luc Devroye, Gábor Lugosi, and Gergely Neu. Prediction by random-walk perturbation. In *COLT*, pages 460–473, 2013.

Paul Erdős. On a lemma of Littlewood and Offord. *Bulletin of the American Mathematical Society*, 51:898–902, 1945.

Sergey Foss, Dmitry Korshunov, and Stan Zachary. *An Introduction to Heavy-tailed and Subexponential Distributions*. Springer, 2009.

Drew Fudenberg and David K. Levine. *The Theory of Learning in Games*. MIT Press, 1998.

Dennis Gilliland and Inha Jung. Play against the random past for matching binary bits. *Journal of Statistical Theory and Application*, 5(3):282–291, 2006.

James Hannan. Approximation to Bayes risk in repeated play. *Contributions to the theory of games*, 3(39):97–139, 1957.

Sergiu Hart and Andreu Mas-Colell. *Simple Adaptive Strategies: From Regret Matching to Uncoupled Dynamics*, volume 4 of *World Scientific Series in Economic Theory*. World Scientific Publishing, 2013.

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

Yuri M. Kaniovski and H. Peyton Young. Learning dynamics in games with stochastic perturbations. *Games and Economic Behavior*, 11(2):330–363, 1995.

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

Theodore J. Lambert III, Marina A. Epelman, and Robert L. Smith. A fictitious play approach to large-scale optimization. *Operation Research*, 53(3):477–489, 2005.

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

Jan Poland. FPL analysis for adaptive bandits. In Oleg B. Lupanov, Oktay M. Kasim-Zade, Alexander V. Chaskin, and Kathleen Steinhöfel, editors, *Stochastic Algorithms: Foundations and Applications: Third International Symposium, SAGA 2005, Moscow, Russia, October 20-22, 2005. Proceedings*, pages 58–69. Springer Berlin Heidelberg, 2005.

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

Gilles Stoltz and Gábor Lugosi. Internal regret in on-line portfolio selection. *Machine Learning*, 59(1-2):125–159, 2005.

Ewart A. C. Thomas. Sufficient conditions for monotone hazard rate an application to latency-probability curves. *J. Math. Psychol.*, 8:303–332, 1971.

Tim van Erven, Wojciech Kotlowski, and Manfred K. Warmuth. Follow the leader with dropout perturbations. In *Proceedings of Conference on Learning Theory (COLT)*, pages 949–974, 2014.