Scale-Free Adversarial Multi Armed Bandits

Sudeep Raja Putta SP3794@COLUMBIA.EDU
Columbia University

Shipra Agrawal SA3305@COLUMBIA.EDU
Columbia University

Editors: Sanjoy Dasgupta and Nika Haghtalab

Abstract

We consider the Scale-Free Adversarial Multi Armed Bandits(MAB) problem. At the beginning of the game, the player only knows the number of arms n. It does not know the scale and magnitude of the losses chosen by the adversary or the number of rounds T. In each round, it sees bandit feedback about the loss vectors $l_1, \ldots, l_T \in \mathbb{R}^n$. The goal is to bound its regret as a function of n and norms of $l_1, \ldots, l_T$. We design a bandit Follow The Regularized Leader (FTRL) algorithm, that uses an adaptive learning rate and give two different regret bounds, based on the exploration parameter used. With non-adaptive exploration, our algorithm has a regret of $\tilde{\mathcal{O}}(\sqrt{nL_2} + L_\infty\sqrt{nT})$ and with adaptive exploration, it has a regret of $\tilde{\mathcal{O}}(\sqrt{nL_2} + L_\infty\sqrt{nL_1})$. Here $L_\infty = \sup_t \|l_t\|_\infty$, $L_2 = \sum_{t=1}^T \|l_t\|_2^2$, $L_1 = \sum_{t=1}^T \|l_t\|_1$ and the $\tilde{\mathcal{O}}$ notation suppress logarithmic factors. These are the first MAB bounds that adapt to the $\|\cdot\|_2$, $\|\cdot\|_1$ norms of the losses. The second bound is the first data-dependent scale-free MAB bound as T does not directly appear in the regret. We also develop a new technique for obtaining a rich class of local-norm lower-bounds for Bregman Divergences. This technique plays a crucial role in our analysis for controlling the regret when using importance weighted estimators of unbounded losses. This technique could be of independent interest.

Keywords: Multi Armed Bandit, Scale-Free Algorithm, FTRL, Adaptive FTRL

1. Introduction

The Adversarial Multi Armed Bandit(MAB) problem proceeds as a sequential game of T rounds between a player and an adversary. In each round $t = 1, \ldots, T$, the player selects a distribution p_t over the n-arms and the adversary selects a loss vector l_t belonging to some set $\mathcal{L} \subseteq \mathbb{R}^n$. An action i_t is sampled from p_t and the player observes the loss $l_t(i_t)$. The (expected) regret of the player is:

$$R_T = \mathbb{E}\left[\sum_{t=1}^T l_t(i_t) - \min_{i\in[n]} \sum_{t=1}^T l_t(i)\right]$$

We assume that the adversary is oblivious, i.e., the loss vectors $l_1, \ldots, l_T$ are chosen before the game begins. So, the above expectation is with respect to the randomness in the player's strategy. The goal of the player is to sequentially select the distributions $p_1, \ldots, p_T$ such that R_T is minimized. The adversarial MAB problem has been studied extensively; we refer the reader to the texts of Bubeck and Cesa-Bianchi (2012); Lattimore and Szepesvári (2020); Slivkins (2019) for further details. Assuming that $\mathcal{L}$ is bounded, and the $\|\cdot\|_\infty$-Lipschitz constant G is known to the player in advance (i.e. $\sup_{l\in\mathcal{L}} \|l\|_\infty = G < \infty$), the minimax rate of regret is known to be $\Theta(G\sqrt{nT})$. The Exp3 algorithm (Auer et al., 2002) has a $\mathcal{O}(G\sqrt{nT\log(n)})$ regret bound whereas the Poly-INF algorithm (Audibert and Bubeck, 2009) removes the $\sqrt{\log(n)}$ factor, achieving the optimal $\mathcal{O}(G\sqrt{nT})$ regret bound.

Exp3 and Poly-INF use G in tuning the learning rate, which helps them achieve a linear dependence on G.

In this paper, we address the case when the player has no knowledge of $\mathcal{L}$. We consider *Scale-Free* bounds for MABs, which aim to bound the regret in terms of n and norms of the loss vectors $l_1, \ldots, l_T$ for any sequence of loss vectors chosen arbitrarily by adversary. Scale-free bounds have been studied in the *full-information* setting (where the player sees the complete vector l_t in each round). For the Experts problem, which is the full-information counterpart of adversarial MAB, the AdaHedge algorithm (de Rooij et al., 2014) has a scale-free regret bound of $\mathcal{O}(\sqrt{\log(n)(\sum_{t=1}^{T} \|l_t\|_\infty^2)})$. For the same problem, the Hedge algorithm (Freund and Schapire, 1997) has a regret bound of $\mathcal{O}(G\sqrt{T\log(n)})$ with knowledge of G. The scale-free bound is more general as it holds for any $l_1, \ldots, l_T \in \mathbb{R}^n$, whereas the bound achieved by the Hedge algorithm only holds provided that $\sup_t \|l_t\|_\infty < G$ where G needs to be known in advance.

1.1. Our Contributions

We present an algorithm for the scale-free MAB problem. By appropriately setting the parameters of this algorithm, we can achieve a scale-free regret upper-bound of either $\tilde{\mathcal{O}}(\sqrt{nL_2} + L_\infty\sqrt{nT})$, or $\tilde{\mathcal{O}}(\sqrt{nL_2} + L_\infty\sqrt{nL_1})$. Here $L_\infty = \sup_t \|l_t\|_\infty$, $L_2 = \sum_{t=1}^{T} \|l_t\|_2^2$, $L_1 = \sum_{t=1}^{T} \|l_t\|_1$ and the $\tilde{\mathcal{O}}$ notation suppress logarithmic factors. Our algorithm is also *any-time* as it does not need to know the number of rounds T in advance. Assuming $\sup_t \|l_t\|_\infty < G$, our first regret bound achieves linear dependence on G (sans the hidden logarithmic terms). This bound is only $\tilde{\mathcal{O}}(\sqrt{n})$ factor larger than Poly-INF's regret of $\mathcal{O}(G\sqrt{nT})$. The second bound is the first completely data-dependent scale-free regret bound for MABs as it has no direct dependence on T. Moreover, these are the first MAB bounds that adapt to the $\|\cdot\|_2$, $\|\cdot\|_1$ norms of the losses. The only previously known scale-free result for MABs was $\mathcal{O}(L_\infty\sqrt{nT\log(n)})$ by Hadiji and Stoltz (2020), which adapts to the $\|\cdot\|_\infty$ norm and is not completely data-dependent due to the T in their bound.

In the analysis, we present a novel and general technique to obtain *local-norm* lower-bounds for *Bregman divergences* induced by a special class of functions that are commonly used in online learning. These local-norm lower-bounds can be used to obtain regret inequalities as shown in Lattimore and Szepesvári (2020, Corollary 28.8). We use our technique to obtain a full-information regret inequality that holds for any arbitrary sequence of losses and is particularly useful in the bandit setting due to its local-norm structure. This technique could be of independent interest.

1.2. Related Work

Scale-Free Regret. As mentioned earlier, Scale-Free regret bounds were studied in the full information setting. The AdaHedge algorithm from de Rooij et al. (2014) gives a scale-free bound for the experts problem. The AdaFTRL algorithm from Orabona and Pál (2018) extends these bounds to the general online convex optimization problem. We rely on the analysis of AdaFTRL as presented in Koolen (2016). For the MAB problem, Hadiji and Stoltz (2020) show a scale-free bound of $\mathcal{O}(L_\infty\sqrt{nT\log(n)})$, which is close to the $\mathcal{O}(G\sqrt{nT\log(n)})$ bound of Exp3. Our scale-free bounds are more versatile as they are able to adapt to additional structure in the loss sequence, such as the case of sparse losses with large magnitude, i.e., when $L_2 << L_\infty^2 nT$ and $L_1 << L_\infty nT$. Even in the worst-case, our bounds are a factor of $\tilde{\mathcal{O}}(\sqrt{n})$ and $\tilde{\mathcal{O}}(\sqrt{nL_\infty})$ larger than their bound respectivley.

Data-dependent Regret. These bounds use a "measure of hardness" of the sequence of loss vectors instead of T. Algorithms that have a data-dependent regret bound perform better than the worst-case regret, when the sequence of losses is "easy" according to the measure of hardness used. For instance, First-order bounds (Allenberg et al., 2006; Foster et al., 2016; Pogodin and Lattimore, 2019), also known as small-loss or $L^\star$ bounds depend on $L^\star = \min_{i\in[n]} \sum_{t=1}^{T} l_t(i)$. Bounds that depend on the empirical variance of the losses were shown in Hazan and Kale (2011); Bubeck et al. (2018). Path length bounds that depend on $\sum_{t=1}^{T-1} \|l_t - l_{t+1}\|$ or a similar quantity appear in Wei and Luo (2018); Bubeck et al. (2019). Zimmert and Seldin (2021) give an algorithm that adapts to any stochastictiy present in the losses. Our bound is comparable to a result in Bubeck et al. (2018), where they derive a regret bound depending on $\sum_{t=1}^{T} \|l_t\|_2^2$. However, all these results assume either $\mathcal{L} = [0,1]^n$ or $\mathcal{L} = [-1,1]^n$.

Effective Range Regret. The effective range of the loss sequence is defined as $\sup_{t,i,j} |l_t(i) - l_t(j)|$. Gerchinovitz and Lattimore (2016) showed that it is impossible to adapt to the effective range in adversarial MAB. This result does not contradict the existence of scale-free bounds as the effective range could be much smaller than, for instance, the complete range $\sup_{t,s,i,j} |l_t(i) - l_s(j)|$. In fact, Hadiji and Stoltz (2020) already show a regret bound that adapts to the complete range. We do note that under some mild additional assumptions, Cesa-Bianchi and Shamir (2018) show that it is possible to adapt to the effective range.

1.3. Organization

In Section 2 we present the scale-free MAB algorithm (Algorithm 1) and its scale-free regret bound (Theorem 1). Section 3 introduces Potential functions, based on which we build our analysis. Section 4 shows a technique for obtaining local-norm lower-bounds for Bregman divergences. Section 5 briefly discusses full-information FTRL, AdaFTRL and in Theorem 8 we obtain a regret inequality for AdaFTRL with the log-barrier regularizer. Theorem 1 is proved in Section 6.

1.4. Notation

Let Δ_n be the probability simplex $\{p \in \mathbb{R}^n : \sum_{i=1}^{n} p(i) = 1, p(i) \geq 0, i \in [n]\}$. Let $\mathbf{1}^i$ be the vector with $\mathbf{1}^i(i) = 1$ and $\mathbf{1}^i(j) = 0$ for all $j \neq i$. For $\epsilon \in (0,1]$, let $\mathbf{1}^i_\epsilon = (1-\epsilon)\mathbf{1}^i + \epsilon/n$. The all ones and all zeros vector are denoted by $\mathbf{1}$ and $\mathbf{0}$ respectively. Let H_t be the history from time-step 1 to t, i.e., $H_t = \{l_1(i_1), l_2(i_2), \ldots, l_t(i_t)\}$.

2. Algorithm

Consider for a moment, full-information strategies on Δ_n. In the full information setting, in each round t, the player picks a point $p_t \in \Delta_n$. Simultaneously, the adversary picks a loss vector $l_t \in \mathbb{R}^n$. The player incurs a loss of $l_t^\top p_t$ and (unlike the bandit setting) *sees the entire vector* l_t. A full-information strategy $\mathcal{F}$ takes as input a sequence of loss vectors $l_1, \ldots, l_t$ and outputs the next iterate $p_{t+1} \in \Delta_n$. A MAB strategy $\mathcal{B}$ can be constructed from a full-information strategy $\mathcal{F}$ along with two other components as follows:

1. A sampling scheme $\mathcal{S}$, which constructs a sampling distribution p'_t from the current iterate p_t. An arm i_t is then sampled from p'_t and the loss $l_t(i_t)$ is revealed to the player.

2. An estimation scheme $\mathcal{E}$, that constructs an estimate $\tilde{l}_t$ of the loss vector l_t using $l_t(i_t)$ and p_t.

3. A full-information strategy $\mathcal{F}$, which computes the next iterate p_{t+1} using all the estimates $\tilde{l}_1, \ldots, \tilde{l}_t$.

In fact, most existing MAB strategies in the literature can be described in the above framework with different choices of $\mathcal{S}, \mathcal{E}, \mathcal{F}$.

A delicate balance needs to be struck between $\mathcal{S}, \mathcal{E}$ and $\mathcal{F}$ in order to achieve a good regret bound for $\mathcal{B}$. Suppose the best arm in hindsight is $i_\star = \arg\min_{i\in[n]} \sum_{t=1}^{T} l_t(i)$ The expected regret of MAB strategy $\mathcal{B}$ can be decomposed as follows:

$$\mathbb{E}\left[\sum_{t=1}^{T}(l_t(i_t) - l_t(i^\star))\right] = \mathbb{E}\left[\sum_{t=1}^{T} l_t^\top(p'_t - \mathbf{1}^{i^\star})\right] = \mathbb{E}\left[\sum_{t=1}^{T} l_t^\top(p'_t - p_t)\right] + \mathbb{E}\left[\sum_{t=1}^{T} l_t^\top(p_t - \mathbf{1}^{i^\star})\right]$$
$$= \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} l_t^\top(p'_t - p_t)\right]}_{(1)} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}(l_t - \tilde{l}_t^\top)(p_t - \mathbf{1}^{i^\star})\right]}_{(2)} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \tilde{l}_t^\top(p_t - \mathbf{1}^{i^\star})\right]}_{(3)}$$

Term (1) is due to the sampling scheme $\mathcal{S}$, term (2) is the effect of the estimation scheme $\mathcal{E}$ and term (3) is the expected regret of the full-information strategy $\mathcal{F}$ on the loss sequence $\tilde{l}_1, \ldots, \tilde{l}_T$ compared to playing the fixed strategy $\mathbf{1}^{i^\star}$.

Sampling Scheme. A commonly used sampling scheme mixes p_t with the uniform distribution using a parameter γ, i.e., $p'_t = (1-\gamma)p_t + \gamma/n$. Such schemes were first introduced in the seminal work of Auer et al. (2002) and have remained a mainstay in MAB algorithm design. We use a time-varying γ, i.e., we pick $p'_t = (1-\gamma_{t-1})p_t + \gamma_{t-1}/n$. Here γ_{t-1} could be any measurable function of H_{t-1}.

Estimation Scheme. We use the *Importance Weighted*(IW) estimator which was also introduced by Auer et al. (2002). It computes $\tilde{l}_t$ as:

$$\tilde{l}_t = \frac{l_t(i_t)}{p'_t(i_t)}\mathbf{1}^{i_t}$$

Since the sampling distribution is p'_t, the IW estimator is an unbiased estimate of l_t:

$$\mathbb{E}_{i_t\sim p'_t}[\tilde{l}_t] = \sum_{i_t=1}^{n} p'_t(i_t)\frac{l_t(i_t)}{p'_t(i_t)}\mathbf{1}^{i_t} = l_t$$

Note that p_t is a measurable function of H_{t-1}. Using the tower rule and the fact that $\mathbb{E}_{i_t\sim p'_t}[\tilde{l}_t] = l_t$, we can see that term (2) is 0.

Full-information strategy. For $\mathcal{F}$, there is a large variety of full-information algorithms that one could pick from. Most if not all of them belong to one of the two principle families of algorithms: Follow The Regularized Leader(FTRL) or Online Mirror Descent(OMD). Further, one also has to choose a suitable *regularizer* F within these algorithms for the particular application at hand. We refer to Cesa-Bianchi and Lugosi (2006); Shalev-Shwartz (2012); Hazan (2016); Orabona (2019); Joulani et al. (2017, 2020) for a detailed history and comparison of these algorithms. The particular

algorithm we use is FTRL with a H_t measurable, adaptive learning rate η_t that resembles the adaptive schemes in AdaHedge (de Rooij et al., 2014) and AdaFTRL (Orabona and Pál, 2018).

The regret of $\mathcal{F}$ has an component called the stability term $\Psi_p : \mathbb{R}^n \to \mathbb{R}$. In the bandit case, $\mathcal{F}$ receives the IW estimates $\tilde{l}_t$. So, it is important that the stability term be bounded with IW estimates. Without going into any technical details, we note that it is desirable to have a stability term bounded by $\Psi_p(l) \leq p^\top l^2$ as its expectation with IW estimates can be bounded.

Previous techniques to bound the stability term by $p^\top l^2$ relied on the assumptions on l, such as either $l \geq \mathbf{0}$ or $l \geq -\mathbf{1}$ (See (Lattimore and Szepesvári, 2019, Page 5)). For arbitrary $l \in \mathbb{R}^n$, we show that it is possible to bound the stability term by $p^\top l^2$ using the *log-barrier* regularizer. The procedure we develop to obtain this bound is the main technical contribution of our paper.

The complete algorithm for the scale-free MAB problem is described below. We give two choices for the exploration parameter γ_t. A simple non-adaptive scheme that is similar to the one in Hadiji and Stoltz (2020), where $\gamma_t \propto \frac{1}{\sqrt{t}}$ and an adaptive scheme that picks γ_t in a fashion that resembles the adaptive learning rate scheme η_t.

Algorithm 1: Scale-Free Multi Armed Bandit

Starting Parameters: $\eta_0 = n, \gamma_0 = 1/2$

Regularizer $F(q) = \sum_{i=1}^{n} (f(q(i)) - f(1/n))$, where $f(x) = -\log(x)$

First iterate $p_1 = (1/n, \ldots, 1/n)$

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

Sampling Scheme: $p'_t = (1 - \gamma_{t-1})p_t + \frac{\gamma_{t-1}}{n}$

Sample Arm $i_t \sim p'_t$ and see loss $l_t(i_t)$.

Estimation Scheme: $\tilde{l}_t = \frac{l_t(i_t)}{p'_t(i_t)} \mathbf{1}^{i_t}$

Compute γ_t for next step:

(Option 1) Non-adaptive $\gamma_t = \min(1/2, \sqrt{n/t})$

(Option 2) Adaptive $\gamma_t = \frac{n}{2n + \sum_{s=1}^{t} \Gamma_s(\gamma_{s-1})}$ where $\Gamma_t(\gamma) = \frac{\gamma |l_t(i_t)|}{(1-\gamma)p_t(i_t) + \gamma/n}$

Compute $\eta_t = \frac{n}{1 + \sum_{s=1}^{t} M_s(\eta_{s-1})}$ where $M_t(\eta) = \sup_{q \in \Delta_n} \left[\tilde{l}_t^\top (p_t - q) - \frac{1}{\eta} \text{Breg}_F(q \| p_t) \right]$

Find next iterate using FTRL: $p_{t+1} = \arg\min_{q \in \Delta_n} \left[F(q) + \eta_t \sum_{s=1}^{t} q^\top \tilde{l}_s \right]$

end

Our main result is the following regret bound for Algorithm 1.

Theorem 1 *For any $l_1, \ldots, l_T \in \mathbb{R}^n$, the expected regret of Algorithm 1 is at most:*

1. $\tilde{\mathcal{O}}(\sqrt{nL_2} + L_\infty \sqrt{nT})$ *if γ_t is non-adaptive (Option 1) and $T \geq 4n$*
2. $\tilde{\mathcal{O}}(\sqrt{nL_2} + L_\infty \sqrt{nL_1})$ *if γ_t is adaptive (Option 2)*

Where $L_\infty = \max_t \|l_t\|_\infty$, $L_2 = \sum_{t=1}^{T} \|l_t\|_2^2$, $L_1 = \sum_{t=1}^{T} \|l_t\|_1$.

3. Preliminaries

We begin by recalling a few definitions.

Definition 2 (Legendre function) *A continuous function $F : \mathcal{D} \to \mathbb{R}$ is Legendre if F is strictly convex, continuously differentiable on Interior$(\mathcal{D})$ and* $\lim_{x \to \mathcal{D}/Interior(\mathcal{D})} \|\nabla F(x)\| = +\infty$.

For instance, the function $x\log(x) - x$, $-\sqrt{x}$, $-\log(x)$ are all Legendre on $(0, \infty)$

Definition 3 (Bregman Divergence) *The Bregman Divergence of function F is:*

$$Breg_F(x\|y) = F(x) - F(y) - \nabla F(y)^\top (x - y).$$

Definition 4 (Potential Function) *A function $\psi : (-\infty, a) \to (0, +\infty)$ for some $a \in \mathbb{R} \cup \{+\infty\}$ is called a Potential if it is convex, strictly increasing, continuously differentiable and satisfies:*

$$\lim_{x \to -\infty} \psi(x) = 0 \quad \textit{and} \quad \lim_{x \to a} \psi(x) = +\infty$$

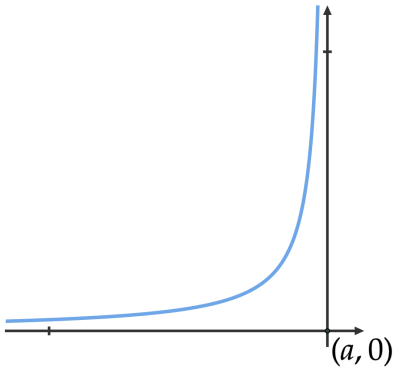

Figure 1: Potential Function

For instance, $\exp(x)$ is a potential with $a = \infty$ and $-1/x$ is a potential with $a = 0$. A potential function typically looks like Figure 1. Potentials were introduced in Audibert and Bubeck (2009); Audibert et al. (2011, 2014) for analyzing the Implicitly Normalized Forecaster(INF) algorithm, of which Poly-INF is a specific case.

Associated with a potential ψ, we define a function f_ψ as the indefinite integral $f_\psi(z) = \int \psi^{-1}(z)dz + C$. Since the domain of ψ^{-1} is $(0, \infty)$, the domain of f_ψ is also $(0, \infty)$. For instance, if $\psi(x) = -1/x$ on the domain $(-\infty, 0)$, the associated function is $f_\psi(x) = -\log(x) + C$.

Observe that $f'_\psi(z) = \psi^{-1}(z)$ and $f''_\psi(z) = \left[\psi'(\psi^{-1}(z))\right]^{-1}$. Since ψ is strictly convex and increasing, $\psi' > 0$ and thus $f''_\psi > 0$, making f_ψ strictly convex. Moreover, $\lim_{z \to 0} \mid f'_\psi(z) \mid = \lim_{z \to 0} \mid \psi^{-1}(z) \mid = +\infty$. Thus f_ψ is a Legendre function on $(0, \infty)$. Define the function $F_\psi : \mathbb{R}^n \to \mathbb{R}$ as $F_\psi(x) = \sum_{i=1}^n [f_\psi(x(i)) - f_\psi(1/n)]$. This function is Legendre on $(0, \infty)^n$.

Given a potential $\psi : (-\infty, a) \to (0, +\infty)$ and its associated function f_ψ, the Legendre-Fenchel dual of f_ψ is $f^\star_\psi : (-\infty, a) \to \mathbb{R}$ defined as $f^\star_\psi(u) = \sup_{z>0}(zu - f_\psi(z))$. The supremum is achieved

at $z = {f'_\psi}^{-1}(u) = \psi(u)$. So we have that $f^\star_\psi(u) = u\psi(u) - f_\psi(\psi(u))$. This implies $f^{\star\prime}_\psi(u) = \psi(u)$ and $f^{\star\prime\prime}_\psi(u) = \psi'(u)$. Further, using integration by parts on $\int \psi(u)du$ and substituting $\psi(u) = s$:

$$\int \psi(u)du = u\psi(u) - \int u\psi'(u)du = u\psi(u) - \int \psi^{-1}(s)ds = u\psi(u) - f_\psi(\psi(u)) + C = f^\star_\psi(u) + C$$

Thus $f^\star_\psi(u) = \int \psi(u)du - C$. Here C is the same constant of integration picked when defining $f_\psi(z) = \int \psi^{-1}(z)dz + C$. We have the following property (proof in Appendix A):

Lemma 5 *Let x, y be such that $x = \psi(u)$ and $y = \psi(v)$. Then $\text{Breg}_{f_\psi}(y\|x) = \text{Breg}_{f^\star_\psi}(u\|v)$*

4. New local-norm lower-bounds for Bregman divergences

Let ψ be a potential and $x, y \in \mathbb{R}_+$. We show a general way of obtaining lower-bounds using potential functions, that are of the form:

$$\text{Breg}_{f_\psi}(y\|x) \geq \frac{1}{2w(x)}(x-y)^2$$

Where w is some positive function.

Lemma 6 *Let ψ be a potential and $x \in \mathbb{R}_+$ such that $x = \psi(u)$ for some u. Let ϕ be a non-negative function such that $\psi(u + \phi(u))$ exists. Define the function $m(z) = \frac{\psi(z+\phi(z)) - \psi(z)}{\phi(z)}$. For all $0 < y \leq \psi(u + \phi(u))$ we have the lower bound: $\text{Breg}_{f_\psi}(y\|x) \geq \frac{1}{2}\frac{(x-y)^2}{m(\psi^{-1}(x))}$*

Proof Let v be such that $y = \psi(v)$. Using Lemma 5, we have $\text{Breg}_{f_\psi}(y\|x) = \text{Breg}_{f^\star_\psi}(u\|v)$. Using the fact that $f^\star_\psi(u) = \int \psi(u)du - C$, we have:

$$\text{Breg}_{f^\star_\psi}(u\|v) = f^\star_\psi(u) - f^\star_\psi(v) - f^{\star\prime}_\psi(v)(u-v) = \int_v^u \psi(s) - y(u-v)$$

We can visualize $\text{Breg}_{f^\star_\psi}(u\|v)$ using the potential function. When $v \leq u$, it is the area with green borders in Figure 2 and when $u \leq v$, it is the area with green borders in Figure 3.

Consider the line passing through (u, x) and $(u + \phi(u), \psi(u + \phi(u)))$. Its slope is $m(u) \geq \psi'(u) > 0$. In both cases, the height of the red triangle is $|x - y|$ and its base is $\frac{|x-y|}{m(u)}$. So, the area of the red triangle will be $\frac{1}{2}\frac{(x-y)^2}{m(u)}$. Since the triangle is always smaller than $\text{Breg}_{f^\star_\psi}(u\|v)$, we have the lower bound $\text{Breg}_{f_\psi}(y\|x) \geq \frac{1}{2}\frac{(x-y)^2}{m(\psi^{-1}(x))}$. ■

In the context of online learning, local-norm lower-bounds have been studied before, see for example Orabona (2019). However, these relied upon Taylor's theorem to show that $\text{Breg}_{f_\psi}(y\|x) = \frac{1}{2}(x-y)^2 f''_\psi(z)$ for some $z \in [x, y]$. Then, they used further conditions on x, y to argue that $c f''_\psi(x) \leq f''_\psi(z)$ for some positive constant c and thus arrive at $\text{Breg}_{f_\psi}(y\|x) \geq \frac{c}{2}(x-y)^2 f''_\psi(x)$. We generalize this argument in Lemma 6, through which we are able to generate a more rich class of lower-bounds. We illustrate with an example below:

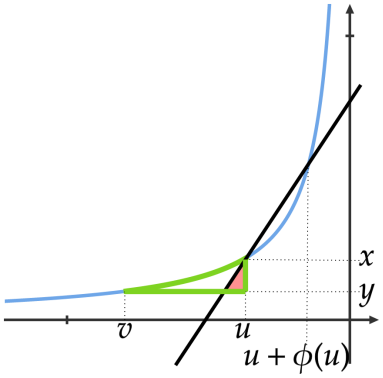

Figure 2: $v \leq u$

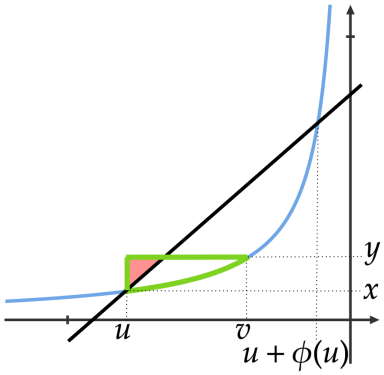

Figure 3: $u \leq v \leq u + \phi(u)$

Corollary 7 *Let $\psi(u) = -1/u$ in the domain $(-\infty, 0)$. For $x, y \in (0, 1]$, we have the lower-bound*

$$Breg_{f_\psi}(y\|x) = \frac{y}{x} - 1 - \ln\left(\frac{y}{x}\right) \geq \frac{1}{2}\frac{(x-y)^2}{x}$$

Proof For any $x \in (0, 1]$, let $u \in (-\infty, -1]$ be such that $\psi(u) = x$. Let $\phi(u) = -1 - u$. Clearly, $\phi(u) \geq 0$ and $\psi(u + \phi(u)) = \psi(-1) = 1$. We have

$$m(u) = \frac{\psi(u + \phi(u)) - \psi(u)}{\phi(u)} = \frac{1 + \frac{1}{u}}{-1 - u} = \frac{-1}{u} = \psi(u) = x$$

Applying Lemma 6, we have the lower-bound for all $0 < y \leq 1$:

$$\text{Breg}_{f_\psi}(y\|x) = \frac{y}{x} - 1 - \ln\left(\frac{y}{x}\right) \geq \frac{1}{2}\frac{(x-y)^2}{m(\psi^{-1}(x))} = \frac{1}{2}\frac{(x-y)^2}{x}$$

■

The result of Corollary 7 is illustrated in Figure 4. The shaded region is $\{(x, y) : x \geq 0, y \geq 0, \frac{y}{x} - 1 - \ln\left(\frac{y}{x}\right) \geq \frac{1}{2}\frac{(x-y)^2}{x}\}$. Clearly the region $\{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$ is within the shaded region.

5. Full-Information FTRL and AdaFTRL

The iterates of FTRL with the regularizer $F_\psi(x) = \sum_{i=1}^n [f_\psi(x(i)) - f_\psi(1/n)]$ for some potential function ψ and positive learning rates $\{\eta_t\}_{t=0}^T$, are of the form:

$$p_{t+1} = \arg\min_{q \in \Delta_n} \left[F_\psi(q) + \eta_t \sum_{s=1}^t l_s^\top q \right]$$

Since F_ψ is Legendre, the point p_{t+1} always exists strictly inside Δ_n. Orabona (2019) and Joulani et al. (2017, 2020) provide general purpose regret analysis of FTRL. For the sake of completeness,

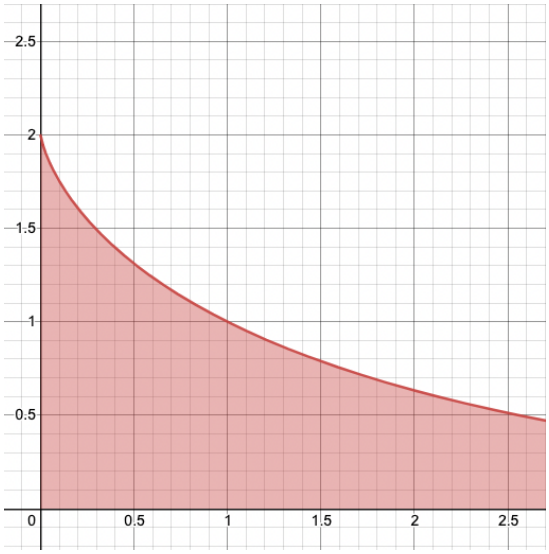

Figure 4: $\frac{y}{x} - 1 - \ln\left(\frac{y}{x}\right) \geq \frac{1}{2}\frac{(x-y)^2}{x}$

we show a simple way of analyzing FTRL when the action set is Δ_n and the regularizer chosen is of the form $F_\psi(x) = \sum_{i=1}^n [f_\psi(x(i)) - f_\psi(1/n)]$ in Appendix C.

The AdaFTRL strategy picks a specific sequence of learning rate η_t based on the history H_t. This strategy was analyzed in Orabona and Pál (2018) and a simpler analysis was given by Koolen (2016). Our analysis is adapted from Hadiji and Stoltz (2020, Section E.2.1). We consider the adaptive learning rate:

$$\eta_t = \frac{\alpha}{\beta + \sum_{s=1}^t M_s(\eta_{s-1})}$$

Where $M_t(\eta) = \sup_{q \in \Delta_n} \left[l_t^\top (p_t - q) - \frac{1}{\eta} \text{Breg}_{F_\psi}(q \| p_t) \right]$, is the *Mixability Gap* and $\alpha, \beta > 0$. Since $q = p_t$ is a feasible solution for this optimization problem, we have $M_t(\eta) \geq 0$. Let $p_t^\star$ be the optimal value of q in the optimization. We have the upper bound

$$M_t(\eta) = l_t^\top (p_t - p_t^\star) - \frac{1}{\eta} \text{Breg}_{F_\psi}(p_t^\star \| p_t) \leq l_t^\top (p_t - p_t^\star) \leq 2 \| l_t \|_\infty$$

Since $M_t(\eta)$ are non-negative and bounded, the sequence η_t is non-increasing.

Theorem 8 *If the regularizer is the log-barrier $F_\psi(x) = \sum_{i=1}^n [\log(1/n) - \log(x(i))]$ then for any $i \in [n]$, $\epsilon \in (0, 1]$ and any sequence of losses $l_1, \dots, l_T$, the iterates of AdaFTRL satisfy the regret inequality $\sum_{t=1}^T l_t^\top (p_t - \boldsymbol{I}_\epsilon^i)$:*

$$\leq n \log(1/\epsilon) \left(\frac{\beta}{\alpha} + \frac{2 \sup_t \| l_t \|_\infty}{\alpha} \right) + 2 \sup_t \| l_t \|_\infty + \sqrt{\sum_{t=1}^T p_t^\top l_t^2} \left(\frac{n \log(1/\epsilon)}{\sqrt{\alpha}} + \sqrt{\alpha} \right)$$

Proof The log-barrier regularizer $F_\psi(x) = \sum_{i=1}^n [\log(1/n) - \log(x(i))]$ is obtained by using the potential $\psi(u) = -1/u$ on the domain $(-\infty, 0)$. Using Corollary 7, we have the lower-bound:

$$\text{Breg}_{F_\psi}(p_t^\star \| p_t) = \sum_{i=1}^n \text{Breg}_{f_\psi}(p_t^\star(i) \| p_t(i)) \geq \sum_{i=1}^n \frac{1}{2} \frac{(p_t(i) - p_t^\star(i))^2}{p_t(i)}$$

This gives us the upper-bound:

$$\begin{aligned} M_t(\eta) &= l_t^\top (p_t - p_t^\star) - \frac{1}{\eta} \text{Breg}_{F_\psi}(p_t^\star \| p_t) \leq \sum_{i=1}^n \left[l_t(i)(p_t(i) - p_t^\star(i)) - \frac{(p_t(i) - p_t^\star(i))^2}{2\eta p_t(i)} \right] \\ &\leq \sum_{i=1}^n \sup_{s \in \mathbb{R}} \left[l_t(i) s - \frac{1}{2\eta} \frac{s^2}{p_t(i)} \right] \leq \frac{\eta}{2} \sum_{n=1}^n p_t(i) l_t(i)^2 = \frac{\eta}{2} p_t^\top l_t^2 \end{aligned}$$

Thus, we have

$$\frac{M_t(\eta_{t-1})}{\eta_{t-1}} \leq \frac{1}{2} p_t^\top l_t^2$$

Applying Theorem 13(Appendix C), for any $i \in [n]$ and $\epsilon \in (0, 1]$ we have that $\sum_{t=1}^T l_t(p_t - \mathbf{1}_\epsilon^i)$:

$$\leq F_\psi(\mathbf{1}_\epsilon^i) \left(\frac{\beta}{\alpha} + \frac{2 \sup_t \|l_t\|_\infty}{\alpha} \right) + 2 \sup_t \|l_t\|_\infty + \sqrt{\sum_{t=1}^T p_t^\top l_t^2} \left(\frac{F_\psi(\mathbf{1}_\epsilon^i)}{\sqrt{\alpha}} + \sqrt{\alpha} \right)$$

The term $F_\psi(\mathbf{1}_\epsilon^i)$ can be bounded as:

$$\begin{aligned} F_\psi(\mathbf{1}_\epsilon^i) &= n \log(1/n) - (n-1) \log(\epsilon/n) - \log((1-\epsilon) + \epsilon/n) \\ &\leq n \log(1/n) - n \log(\epsilon/n) = n \log(1/\epsilon) \end{aligned}$$

■

For $p \in \Delta_n$ and regularizer F_ψ, the stability term Ψ is defined as

$$\Psi_p(l) = \sup_{q \in \Delta_n} \left[l^\top (p - q) - \text{Breg}_{F_\psi}(q \| p) \right]$$

Observe that $\eta M_t(\eta) = \Psi_{p_t}(\eta l_t)$. For the log-barrier regularizer, we have $M_t(\eta) \leq \eta p_t^\top l_t^2$. Thus, $\Psi_p(l) \leq p^\top l^2$ for all $l \in \mathbb{R}^n$. Previously, the only known way to achieve $\Psi_p(l) \leq p^\top l^2$ was by using the negative-entropy regularizer along with the assumption $l \geq -\mathbf{1}$ (See Lattimore and Szepesvári (2019, Eq. 6) or Lattimore and Szepesvári (2020, Eq. 37.15)).

6. Scale-free bandit regret bounds

Theorem 1 *For any $l_1, \ldots, l_T \in \mathbb{R}^n$, the expected regret of Algorithm 1 is at most:*

1. $\tilde{\mathcal{O}}(\sqrt{nL_2} + L_\infty \sqrt{nT})$ *if γ_t is non-adaptive (Option 1) and $T \geq 4n$*
2. $\tilde{\mathcal{O}}(\sqrt{nL_2} + L_\infty \sqrt{nL_1})$ *if γ_t is adaptive (Option 2)*

Where $L_\infty = \max_t \|l_t\|_\infty$, $L_2 = \sum_{t=1}^T \|l_t\|_2^2$, $L_1 = \sum_{t=1}^T \|l_t\|_1$.

Proof Suppose the best arm in hindsight is $i_\star = \arg\min_{i\in[n]} \sum_{t=1}^T l_t(i)$. Let $\mathbf{1}^{i_\star}$ be the vector with $\mathbf{1}^{i_\star}(i_\star) = 1$ and $\mathbf{1}^{i_\star}(i) = 0$ for all $i \neq i_\star$. Let $\mathbf{1}_\epsilon^{i_\star} = (1-\epsilon)\mathbf{1}^{i_\star} + \epsilon/n$. The expected regret of Algorithm 1 is:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T l_t(i_t) - l_t(i^\star)\right] &= \mathbb{E}\left[\sum_{t=1}^T l_t^\top(p_t' - \mathbf{1}^{i_\star})\right] = \mathbb{E}\left[\sum_{t=1}^T l_t^\top(\mathbf{1}_\epsilon^{i_\star} - \mathbf{1}^{i_\star})\right] + \mathbb{E}\left[\sum_{t=1}^T l_t^\top(p_t' - \mathbf{1}_\epsilon^{i_\star})\right] \\ &= \underbrace{\mathbb{E}\left[\sum_{t=1}^T l_t^\top(\mathbf{1}_\epsilon^{i_\star} - \mathbf{1}^{i_\star})\right]}_{(1)} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T l_t^\top(p_t - \mathbf{1}_\epsilon^{i_\star})\right]}_{(2)} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T l_t^\top(p_t' - p_t)\right]}_{(3)}\end{aligned}$$

Define $S_\infty = \|\sum_{t=1}^T l_t\|_\infty$. For term (1), we have:

$$\mathbb{E}\left[\sum_{t=1}^T l_t^\top(\mathbf{1}_\epsilon^{i_\star} - \mathbf{1}^{i_\star})\right] = \sum_{t=1}^T l_t^\top(\mathbf{1}_\epsilon^{i_\star} - \mathbf{1}^{i_\star}) \leq 2\epsilon \left\|\sum_{t=1}^T l_t\right\|_\infty = 2\epsilon S_\infty$$

For term (2), we use the fact that $\mathbb{E}[\tilde{l}_t] = l_t$:

$$\mathbb{E}\left[\sum_{t=1}^T l_t^\top(p_t - \mathbf{1}_\epsilon^{i_\star})\right] = \mathbb{E}\left[\sum_{t=1}^T \tilde{l}_t^\top(p_t - \mathbf{1}_\epsilon^{i_\star})\right]$$

Since Algorithm 1 runs log-barrier regularized AdaFTRL with the loss sequence $\tilde{l}_1, \ldots, \tilde{l}_T$, we can bound the sum inside the expectation using Theorem 8 as $\sum_{t=1}^T \tilde{l}_t^\top(p_t - \mathbf{1}_\epsilon^{i_\star})$:

$$\leq \log(1/\epsilon)\left(1 + 2\sup_t \|\tilde{l}_t\|_\infty\right) + 2\sup_t \|\tilde{l}_t\|_\infty + \sqrt{n\sum_{t=1}^T p_t^\top \tilde{l}_t^2}\,(\log(1/\epsilon) + 1) \qquad (*)$$

Consider the term $\sup_t \|\tilde{l}_t\|_\infty$:

$$\sup_t \|\tilde{l}_t\|_\infty = \sup_t \frac{|l_t(i_t)|}{p_t'(i_t)} = \sup_t \frac{|l_t(i_t)|}{(1-\gamma_{t-1})p_t(i_t) + \gamma_{t-1}/n} \leq n \sup_t \frac{|l_t(i_t)|}{\gamma_{t-1}}$$

Since γ_t is a positive, non-increasing sequence:

$$\sup_t \|\tilde{l}_t\|_\infty \leq n\frac{\sup_t |l_t(i_t)|}{\gamma_T} \leq \frac{nL_\infty}{\gamma_T}$$

Finally, consider the term $p_t^\top \tilde{l}_t^2$:

$$p_t^\top \tilde{l}_t^2 = p_t(i_t)\frac{l_t(i_t)^2}{p_t'(i_t)^2} = p_t(i_t)\frac{l_t(i_t)^2}{((1-\gamma_{t-1})p_t(i_t) + \frac{\gamma_{t-1}}{n})p_t'(i_t)} \leq \frac{l_t(i_t)^2}{(1-\gamma_{t-1})p_t'(i_t)}$$

Since $0 \leq \gamma_{t-1} \leq 1/2$, we have $1 \leq (1-\gamma_{t-1})^{-1} \leq 2$. Thus:

$$p_t^\top \tilde{l}_t^2 \leq 2\frac{l_t(i_t)^2}{p_t'(i_t)}$$

Substituting these bounds in the regret inequality $(*)$, we have $\sum_{t=1}^{T} \widetilde{l}_t^\top (p_t - \mathbf{1}_\epsilon^{i_\star})$:

$$\leq \log(1/\epsilon) + \sqrt{2n \sum_{t=1}^{T} \frac{l_t(i_t)^2}{p'_t(i_t)}} \left(\log(1/\epsilon) + 1\right) + \frac{2nL_\infty}{\gamma_T} \left(\log(1/\epsilon) + 1\right)$$

Applying expectation, we have $\mathbb{E}\left[\sum_{t=1}^{T} \widetilde{l}_t^\top (p_t - \mathbf{1}_\epsilon^{i_\star})\right]$:

$$\leq \log(1/\epsilon) + \mathbb{E}\left[\sqrt{2n \sum_{t=1}^{T} \frac{l_t(i_t)^2}{p'_t(i_t)}}\right] \left(\log(1/\epsilon) + 1\right) + 2nL_\infty \left(\log(1/\epsilon) + 1\right) \mathbb{E}\left[\frac{1}{\gamma_T}\right]$$

For the expectation in the second term, we apply Jensen's inequality:

$$\mathbb{E}\left[\sqrt{2n \sum_{t=1}^{T} \frac{l_t(i_t)^2}{p'_t(i_t)}}\right] \leq \sqrt{2n\mathbb{E} \sum_{t=1}^{T} \left[\frac{l_t(i_t)^2}{p'_t(i_t)}\right]} = \sqrt{2n \sum_{t=1}^{T} \sum_{i=1}^{n} l_t(i)^2} = \sqrt{2nL_2}$$

Thus term (2) can be bounded as $\mathbb{E}\left[\sum_{t=1}^{T} l_t^\top (p_t - \mathbf{1}_\epsilon^{i_\star})\right]$:

$$\leq \log(1/\epsilon) + \sqrt{2nL_2} \left(\log(1/\epsilon) + 1\right) + 2nL_\infty \left(\log(1/\epsilon) + 1\right) \mathbb{E}\left[\frac{1}{\gamma_T}\right]$$

6.1. Non-Adaptive Exploration

First, we present a simple way to bound term (3):

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} l_t^\top (p'_t - p_t)\right] &= \mathbb{E}\left[\sum_{t=1}^{T} l_t^\top ((1-\gamma_{t-1})p_t + \gamma_{t-1}/n - p_t)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \gamma_{t-1} l_t^\top (1/n - p_t)\right] \\
&\leq \mathbb{E}\left[2 \sum_{t=1}^{T} \gamma_{t-1} \|l_t\|_\infty\right] \leq 2L_\infty \mathbb{E}\left[\sum_{t=1}^{T} \gamma_{t-1}\right]
\end{aligned}$$

Combining the upper-bounds for term (1), (2) and (3), we have $\mathbb{E}\left[\sum_{t=1}^{T} l_t(i_t) - l_t(i^\star)\right]$:

$$\leq 2\epsilon S_\infty + \log(1/\epsilon) + \sqrt{2nL_2} \left(\log(1/\epsilon) + 1\right) + 2nL_\infty \left(\log(1/\epsilon) + 1\right) \mathbb{E}\left[\frac{1}{\gamma_T}\right] + 2L_\infty \mathbb{E}\left[\sum_{t=1}^{T} \gamma_{t-1}\right]$$

Pick $\epsilon = (1 + S_\infty)^{-1}$ and the exploration rate $\gamma_t = \min(1/2, \sqrt{n/t})$. If $T \geq 4n$, the regret of Algorithm 1 with non-adaptive exploration is bounded by:

$$\begin{aligned}
&\leq 2 + \log(1 + S_\infty) + \sqrt{2nL_2}(1 + \log(1 + S_\infty)) + 2L_\infty \sqrt{nT}(2 + \log(1 + S_\infty)) \\
&\leq (2 + \log(1 + S_\infty)) \left(1 + \sqrt{2nL_2} + 2L_\infty \sqrt{nT}\right) \\
&= \tilde{\mathcal{O}}(\sqrt{nL_2} + L_\infty \sqrt{nT})
\end{aligned}$$

6.2. Adaptive Exploration

An alternate way to bound term (3) is:

$$\mathbb{E}\left[\sum_{t=1}^{T} l_t^\top (p'_t - p_t)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \tilde{l}_t^\top (p'_t - p_t)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \gamma_{t-1} \frac{l_t(i_t)}{p'_t(i_t)} (1/n - p_t(i_t))\right]$$
$$\leq \mathbb{E}\left[\sum_{t=1}^{T} \gamma_{t-1} \frac{|l_t(i_t)|}{p'_t(i_t)}\right]$$

Combining the upper-bounds for term (1), (2) and (3), we have $\mathbb{E}\left[\sum_{t=1}^{T} l_t(i_t) - l_t(i^\star)\right]$:

$$\leq 2\epsilon S_\infty + \log(1/\epsilon) + \sqrt{2nL_2}\left(\log(1/\epsilon) + 1\right) + \mathbb{E}\left[\frac{2nL_\infty\left(\log(1/\epsilon)+1\right)}{\gamma_T} + \sum_{t=1}^{T} \gamma_{t-1} \frac{|l_t(i_t)|}{p'_t(i_t)}\right]$$

Consider the expression inside the expectation. Let

$$\Gamma_t(\gamma) = \frac{\gamma |l_t(i_t)|}{(1-\gamma)p_t(i_t) + \gamma/n}$$

When $0 \leq \gamma \leq 1/2$, we have $0 \leq \Gamma_t(\gamma) \leq n|l_t(i_t)| \leq nL_\infty$. Moreover, we have

$$\frac{\Gamma_t(\gamma_{t-1})}{\gamma_{t-1}} = \frac{|l_t(i_t)|}{p'_t(i_t)}$$

Pick

$$\gamma_t = \frac{n}{2n + \sum_{s=1}^{t} \Gamma_s(\gamma_{s-1})}$$

We satisfy $0 \leq \gamma_t \leq 1/2$. Applying Lemma 10, we have:

$$\mathbb{E}\left[\frac{2nL_\infty\left(\log(1/\epsilon)+1\right)}{\gamma_T} + \sum_{t=1}^{T} \gamma_{t-1} \frac{|l_t(i_t)|}{p'_t(i_t)}\right] = \mathbb{E}\left[\frac{2nL_\infty\left(\log(1/\epsilon)+1\right)}{\gamma_T} + \sum_{t=1}^{T} \Gamma_t(\gamma_{t-1})\right]$$

$$\leq 2nL_\infty(2+L_\infty)\left(\log(1/\epsilon)+1\right) + nL_\infty + \left(2L_\infty\left(\log(1/\epsilon)+1\right)+1\right)\mathbb{E}\left[\sqrt{2n\sum_{t=1}^{T}\frac{|l_t(i_t)|}{p'_t(i_t)}}\right]$$

For the expectation above, we apply Jensen's inequality:

$$\mathbb{E}\left[\sqrt{2n\sum_{t=1}^{T}\frac{|l_t(i_t)|}{p'_t(i_t)}}\right] \leq \sqrt{2n\mathbb{E}\sum_{t=1}^{T}\left[\frac{|l_t(i_t)|}{p'_t(i_t)}\right]} = \sqrt{2n\sum_{t=1}^{T}\sum_{i=1}^{n}|l_t(i)|} = \sqrt{2nL_1}$$

Pick $\epsilon = (1+S_\infty)^{-1}$. The regret of Algorithm 1 with adaptive exploration is bounded by:

$$\leq 2 + \log(1+S_\infty) + \sqrt{2nL_2}\left(\log(1+S_\infty)+1\right)$$
$$+ 2nL_\infty(2+L_\infty)\left(\log(1+S_\infty)+1\right) + nL_\infty + \left(2L_\infty\left(\log(1+S_\infty)+1\right)+1\right)\sqrt{2nL_1}$$
$$= \tilde{\mathcal{O}}(\sqrt{nL_2} + L_\infty\sqrt{nL_1})$$

■

References

Chamy Allenberg, Peter Auer, László Györfi, and György Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In José L. Balcázar, Philip M. Long, and Frank Stephan, editors, *Algorithmic Learning Theory, 17th International Conference, ALT 2006, Barcelona, Spain, October 7-10, 2006, Proceedings*, volume 4264 of *Lecture Notes in Computer Science*, pages 229–243. Springer, 2006. doi: 10.1007/11894841_20. URL https://doi.org/10.1007/11894841_20.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009. URL http://www.cs.mcgill.ca/%7Ecolt2009/papers/022.pdf#page=1.

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Minimax policies for combinatorial prediction games. In Sham M. Kakade and Ulrike von Luxburg, editors, *COLT 2011 - The 24th Annual Conference on Learning Theory, June 9-11, 2011, Budapest, Hungary*, volume 19 of *JMLR Proceedings*, pages 107–132. JMLR.org, 2011. URL http://proceedings.mlr.press/v19/audibert11a/audibert11a.pdf.

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Math. Oper. Res.*, 39(1):31–45, 2014. doi: 10.1287/moor.2013.0598. URL https://doi.org/10.1287/moor.2013.0598.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002. doi: 10.1137/S0097539701398375. URL https://doi.org/10.1137/S0097539701398375.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends Mach. Learn.*, 5(1):1–122, 2012. doi: 10.1561/2200000024. URL https://doi.org/10.1561/2200000024.

Sébastien Bubeck, Michael B. Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. In Firdaus Janoos, Mehryar Mohri, and Karthik Sridharan, editors, *Algorithmic Learning Theory, ALT 2018, 7-9 April 2018, Lanzarote, Canary Islands, Spain*, volume 83 of *Proceedings of Machine Learning Research*, pages 111–127. PMLR, 2018. URL http://proceedings.mlr.press/v83/bubeck18a.html.

Sébastien Bubeck, Yuanzhi Li, Haipeng Luo, and Chen-Yu Wei. Improved path-length regret bounds for bandits. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 508–528. PMLR, 2019. URL http://proceedings.mlr.press/v99/bubeck19b.html.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006. ISBN 978-0-521-84108-5. doi: 10.1017/CBO9780511546921. URL https://doi.org/10.1017/CBO9780511546921.

Nicolò Cesa-Bianchi and Ohad Shamir. Bandit regret scaling with the effective loss range. In Firdaus Janoos, Mehryar Mohri, and Karthik Sridharan, editors, *Algorithmic Learning Theory, ALT 2018, 7-9 April 2018, Lanzarote, Canary Islands, Spain*, volume 83 of *Proceedings of Machine Learning Research*, pages 128–151. PMLR, 2018. URL http://proceedings.mlr.press/v83/cesa-bianchi18a.html.

Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *J. Mach. Learn. Res.*, 15(1):1281–1316, 2014. URL http://dl.acm.org/citation.cfm?id=2638576.

Dylan J. Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, and Éva Tardos. Learning in games: Robustness of fast convergence. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4727–4735, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/b3f61131b6eceeb2b14835fa648a48ff-Abstract.html.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997. doi: 10.1006/jcss.1997.1504. URL https://doi.org/10.1006/jcss.1997.1504.

Sébastien Gerchinovitz and Tor Lattimore. Refined lower bounds for adversarial bandits. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1190–1198, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/2f37d10131f2a483a8dd005b3d14b0d9-Abstract.html.

Hédi Hadiji and Gilles Stoltz. Adaptation to the range in k-armed bandits. *CoRR*, abs/2006.03378, 2020. URL http://arxiv.org/abs/2006.03378.

Elad Hazan. Introduction to online convex optimization. *Found. Trends Optim.*, 2(3-4):157–325, 2016. doi: 10.1561/2400000013. URL https://doi.org/10.1561/2400000013.

Elad Hazan and Satyen Kale. Better algorithms for benign bandits. *J. Mach. Learn. Res.*, 12:1287–1311, 2011. URL http://dl.acm.org/citation.cfm?id=2021042.

Pooria Joulani, András György, and Csaba Szepesvári. A modular analysis of adaptive (non-)convex optimization: Optimism, composite objectives, and variational bounds. In Steve Hanneke and Lev Reyzin, editors, *International Conference on Algorithmic Learning Theory, ALT 2017, 15-17 October 2017, Kyoto University, Kyoto, Japan*, volume 76 of *Proceedings of Machine Learning Research*, pages 681–720. PMLR, 2017. URL http://proceedings.mlr.press/v76/joulani17a.html.

Pooria Joulani, András György, and Csaba Szepesvári. A modular analysis of adaptive (non-)convex optimization: Optimism, composite objectives, variance reduction, and variational bounds. *Theor. Comput. Sci.*, 808:108–138, 2020. doi: 10.1016/j.tcs.2019.11.015. URL https://doi.org/10.1016/j.tcs.2019.11.015.

Wouter M. Koolen. Adaftrl, 2016. URL http://blog.wouterkoolen.info/AdaFTRL/post.html.

Tor Lattimore and Csaba Szepesvári. Exploration by optimisation in partial monitoring. *CoRR*, abs/1907.05772, 2019. URL http://arxiv.org/abs/1907.05772.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Francesco Orabona. A modern introduction to online learning. *CoRR*, abs/1912.13213, 2019. URL http://arxiv.org/abs/1912.13213.

Francesco Orabona and Dávid Pál. Scale-free online learning. *Theor. Comput. Sci.*, 716:50–69, 2018. doi: 10.1016/j.tcs.2017.11.021. URL https://doi.org/10.1016/j.tcs.2017.11.021.

Roman Pogodin and Tor Lattimore. On first-order bounds, variance and gap-dependent bounds for adversarial bandits. In Amir Globerson and Ricardo Silva, editors, *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, volume 115 of *Proceedings of Machine Learning Research*, pages 894–904. AUAI Press, 2019. URL http://proceedings.mlr.press/v115/pogodin20a.html.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Found. Trends Mach. Learn.*, 4(2):107–194, 2012. doi: 10.1561/2200000018. URL https://doi.org/10.1561/2200000018.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Found. Trends Mach. Learn.*, 12(1-2):1–286, 2019. doi: 10.1561/2200000068. URL https://doi.org/10.1561/2200000068.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 1263–1291. PMLR, 2018. URL http://proceedings.mlr.press/v75/wei18a.html.

Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *J. Mach. Learn. Res.*, 22:28:1–28:49, 2021. URL http://jmlr.org/papers/v22/19-753.html.

Appendix A. Basic results on potentials

Consider a function $g : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}_+$ defined as $g(\theta, \lambda) = \sum_{i=1}^n \psi(\theta(i) + \lambda)$ for some potential ψ.

Lemma 9 *For every $\theta \in \mathbb{R}^n$, there exists a unique λ such that $g(\theta, \lambda) = 1$*

Proof For every $\theta \in \mathbb{R}^n$, we have that $\lim_{\lambda \to -\infty} g(\theta, \lambda) = 0$ and $\lim_{\lambda \to a - \min_i(\theta(i))} g(\theta, \lambda) = +\infty$. As g is monotonically increasing and continuous, by the intermediate value theorem, for every $\theta \in \mathbb{R}^n$ there exists a unique λ such that $g(\theta, \lambda) = 1$. ■

Using Lemma 9, we can define a function $\lambda(\theta)$ such that $g(\theta, \lambda(\theta)) = \sum_{i=1}^{n} \psi(\theta(i)+\lambda(\theta)) = 1$. Since $\psi(\theta(i) + \lambda(\theta)) \geq 0$ and $\sum_{i=1}^{n} \psi(\theta(i) + \lambda(\theta)) = 1$, we can see that the vector $\psi(\theta + \lambda(\theta)) \equiv \{\psi(\theta(i) + \lambda(\theta))\}_{i=1}^{n} \in \Delta_n$ forms a probability distribution.

Lemma 5 *Let x, y be such that $x = \psi(u)$ and $y = \psi(v)$. Then $Breg_{f_\psi}(y\|x) = Breg_{f_\psi^\star}(u\|v)$*

Proof Use the fact that $f_\psi^\star(u) = u\psi(u) - f(\psi(u))$.

$$\begin{aligned}\text{Breg}_{f_\psi}(y\|x) &= \text{Breg}_{f_\psi}(\psi(v)\|\psi(u)) = f_\psi(\psi(v)) - f_\psi(\psi(u)) - f'_\psi(\psi(u))(\psi(v) - \psi(u)) \\ &= v\psi(v) - f_\psi^\star(v) - (u\psi(u) - f_\psi^\star(u)) - u(\psi(v) - \psi(u)) \\ &= f_\psi^\star(u) - f_\psi^\star(v) - f^{\star\prime}{}_\psi(v)(u - v) = \text{Breg}_{f_\psi^\star}(u\|v)\end{aligned}$$

■

Appendix B. A useful summation

Lemma 10 *Let $A > 0$ and $0 \leq M_t(a) \leq L$ for all $t = 1, \ldots T$ and $a \in \mathcal{A} \subseteq (0, \infty)$. Consider the expression*

$$\frac{A}{a_T} + \sum_{t=1}^{T} M_t(a_{t-1})$$

Where

$$a_t = \frac{\alpha}{\beta + \sum_{s=1}^{t} M_s(a_{s-1})}$$

Constants $\alpha, \beta > 0$ are chosen such that $a_t \in \mathcal{A}$. If $\frac{M_t(a_{t-1})}{a_{t-1}} \leq g_t$, then we have the upper bound:

$$\frac{A}{a_T} + \sum_{t=1}^{T} M_t(a_{t-1}) \leq A\left(\frac{\beta}{\alpha} + \frac{L}{\alpha}\right) + L + \sqrt{2\sum_{t=1}^{T} g_t}\left(\frac{A}{\sqrt{\alpha}} + \sqrt{\alpha}\right)$$

Proof Substituting for a_T in the above expression, we have:

$$\frac{A}{a_T} + \sum_{t=1}^{T} M_t(a_{t-1}) = \frac{A\beta}{\alpha} + \left(\frac{A}{\alpha} + 1\right)\sum_{t=1}^{T} M_t(a_{t-1})$$

Consider $\left(\sum_{t=1}^{T} M_t(a_{t-1})\right)^2$

$$\begin{aligned}\left(\sum_{t=1}^{T} M_t(a_{t-1})\right)^2 &= \sum_{t=1}^{T} M_t(a_{t-1})^2 + 2\sum_{t=1}^{T} M_t(a_{t-1}) \sum_{s=1}^{t-1} M_s(a_{s-1}) \\ &= \sum_{t=1}^{T} M_t(a_{t-1})^2 + 2\sum_{t=1}^{T} M_t(a_{t-1}) \left(\frac{\alpha}{a_{t-1}} - \beta\right) \\ &\leq \sum_{t=1}^{T} M_t(a_{t-1})^2 + 2\alpha \sum_{t=1}^{T} \frac{M_t(a_{t-1})}{a_{t-1}} \\ &\leq L\sum_{t=1}^{T} M_t(a_{t-1}) + 2\alpha \sum_{t=1}^{T} g_t\end{aligned}$$

Using the fact that $x^2 \leq a + bx$ implies that $x \leq \sqrt{a} + b$ for all $a, b, x \geq 0$, we have:

$$\sum_{t=1}^{T} M_t(a_{t-1}) \leq \sqrt{2\alpha \sum_{t=1}^{T} g_t} + L$$

Thus, we get:

$$\begin{aligned}\frac{A}{a_T} + \sum_{t=1}^{T} M_t(a_{t-1}) = \frac{A\beta}{\alpha} + \left(\frac{A}{\alpha} + 1\right) \sum_{t=1}^{T} M_t(a_{t-1}) &\leq \frac{A\beta}{\alpha} + \left(\frac{A}{\alpha} + 1\right) \left(\sqrt{2\alpha \sum_{t=1}^{T} g_t} + L\right) \\ &= A\left(\frac{\beta}{\alpha} + \frac{L}{\alpha}\right) + L + \sqrt{2\sum_{t=1}^{T} g_t} \left(\frac{A}{\sqrt{\alpha}} + \sqrt{\alpha}\right)\end{aligned}$$

■

Appendix C. FTRL and AdaFTRL regret bound

Recall the FTRL update:

$$p_{t+1} = \arg\min_{q \in \Delta_n} \left[F_\psi(q) + \eta_t \sum_{s=1}^{t} l_s^\top q\right]$$

The iterate p_{t+1} can be expressed in a simple closed form using ψ. Let $\theta_t = -\eta_t \sum_{s=1}^{t} l_s$. The Lagrangian of the above optimization problem is $L(q, \alpha) = F_\psi(q) - \theta_t^\top q - \alpha(1 - \mathbf{1}^\top q)$, where $\mathbf{1}$ is the all ones vector. Taking its derivative with respect to $q(i)$ and equating to 0, we get:

$$\psi^{-1}(q(i)) = \theta_t(i) + \alpha \implies q(i) = \psi(\theta_t(i) + \alpha)$$

To compute α, we use the fact that $\sum_{i=1}^{n} q(i) = 1$ along with Lemma 9 to show that $\alpha = \lambda(\theta_t)$. Thus, p_{t+1} can be written as:

$$p_{t+1} = \psi(\theta_t + \lambda(\theta_t)) \quad \text{where} \quad \theta_t = -\eta_t \sum_{s=1}^{t} l_s$$

We introduce the Mixed Bregman in order to simplifies our analysis of FTRL.

Definition 11 (Mixed Bregman) *For $\alpha, \beta > 0$ the (α, β)-Mixed Bregman of function F is:*

$$Breg_F^{\alpha,\beta}(x\|y) = \frac{F(x)}{\alpha} - \frac{F(y)}{\beta} - \frac{\nabla F(y)}{\beta}^\top (x - y).$$

The Mixed Bregman is not a divergence as $\text{Breg}_F^{\alpha,\beta}(x\|x)$ may not be zero. However, we do have the relation $\alpha\text{Breg}_F^{\alpha,\alpha}(x\|y) = \text{Breg}_F(x\|y)$.

Theorem 12 *For any $p \in \Delta_n$ and any sequence of losses $l_1, \ldots, l_T$, the iterates of FTRL satisfy the regret equality $\sum_{t=1}^T l_t^\top (p_t - p)$*

$$= \frac{1}{\eta_T}\left[Breg_{F_\psi}(p\|p_1) - Breg_{F_\psi}(p\|p_{T+1})\right] + \sum_{t=1}^T \left[l_t^\top (p_t - p_{t+1}) - Breg_{F_\psi}^{\eta_t,\eta_{t-1}}(p_{t+1}\|p_t)\right]$$

Further, if the sequence $\{\eta_t\}_{t=0}^T$ is non-decreasing, we have the regret inequality $\sum_{t=1}^T l_t^\top (p_t - p)$:

$$\leq \frac{F_\psi(p)}{\eta_T} + \sum_{t=1}^T \left[l_t^\top (p_t - p_{t+1}) - \frac{1}{\eta_{t-1}} Breg_{F_\psi}(p_{t+1}\|p_t)\right]$$

Proof Note that $\nabla F_\psi(p_{t+1}) = \psi^{-1}(p_{t+1}) = \theta_t + \lambda(\theta_t)$. We also have that $l_t = \frac{\theta_{t-1}}{\eta_{t-1}} - \frac{\theta_t}{\eta_t}$. For any $p \in \Delta_n$, we have $l_t^\top (p_t - p)$:

$$= l_t^\top (p_{t+1} - p) + l_t^\top (p_t - p_{t+1}) = \left(\frac{\theta_{t-1}}{\eta_{t-1}} - \frac{\theta_t}{\eta_t}\right)^\top (p_{t+1} - p) + l_t^\top (p_t - p_{t+1})$$
$$= \left(\frac{\nabla F_\psi(p_t) - \lambda(\theta_{t-1})}{\eta_{t-1}} - \frac{\nabla F_\psi(p_{t+1}) - \lambda(\theta_t)}{\eta_t}\right)^\top (p_{t+1} - p) + l_t^\top (p_t - p_{t+1})$$
$$= \left(\frac{\nabla F_\psi(p_t)}{\eta_{t-1}} - \frac{\nabla F_\psi(p_{t+1})}{\eta_t}\right)^\top (p_{t+1} - p) + \left(\frac{\lambda(\theta_t)}{\eta_t} - \frac{\lambda(\theta_{t-1})}{\eta_{t-1}}\right)^\top (p_{t+1} - p) + l_t^\top (p_t - p_{t+1})$$
$$= \left(\frac{\nabla F_\psi(p_t)}{\eta_{t-1}} - \frac{\nabla F_\psi(p_{t+1})}{\eta_t}\right)^\top (p_{t+1} - p) + l_t^\top (p_t - p_{t+1})$$

Note that $\frac{\lambda(\theta_t)}{\eta_t} - \frac{\lambda(\theta_{t-1})}{\eta_{t-1}}$ is a constant vector. So, $\left(\frac{\lambda(\theta_t)}{\eta_t} - \frac{\lambda(\theta_{t-1})}{\eta_{t-1}}\right)^\top (p_{t+1} - p) = 0$. Let α be any number. Observe that:

$$\left(\frac{\nabla F_\psi(p_t)}{\eta_{t-1}} - \frac{\nabla F_\psi(p_{t+1})}{\eta_t}\right)^\top (p_{t+1} - p) = \text{Breg}_{F_\psi}^{\alpha,\eta_{t-1}}(p\|p_t) - \text{Breg}_{F_\psi}^{\alpha,\eta_t}(p\|p_{t+1}) - \text{Breg}_{F_\psi}^{\eta_t,\eta_{t-1}}(p_{t+1}\|p_t)$$

Taking summation over t, we have $\sum_{t=1}^T l_t^\top (p_t - p)$:

$$= \sum_{t=1}^T \left[\text{Breg}_{F_\psi}^{\alpha,\eta_{t-1}}(p\|p_t) - \text{Breg}_{F_\psi}^{\alpha,\eta_t}(p\|p_{t+1})\right] + \sum_{t=1}^T \left[l_t^\top (p_t - p_{t+1}) - \text{Breg}_{F_\psi}^{\eta_t,\eta_{t-1}}(p_{t+1}\|p_t)\right]$$
$$= \text{Breg}_{F_\psi}^{\alpha,\eta_0}(p\|p_1) - \text{Breg}_{F_\psi}^{\alpha,\eta_T}(p\|p_{T+1}) + \sum_{t=1}^T \left[l_t^\top (p_t - p_{t+1}) - \text{Breg}_{F_\psi}^{\eta_t,\eta_{t-1}}(p_{t+1}\|p_t)\right]$$

Since $p_1 = (1/n, \ldots, 1/n)$, we have $F_\psi(p_1) = 0$ and $\nabla F_\psi(p_1)$ is a constant vector. We see that $\nabla F_\psi(p_1)^\top (p - p_1) = 0$, so the first term is:

$$\begin{aligned}\mathrm{Breg}_{F_\psi}^{\alpha,\eta_0}(p\|p_1) - \mathrm{Breg}_{F_\psi}^{\alpha,\eta_T}(p\|p_{T+1}) &= \frac{F_\psi(p_{T+1})}{\eta_T} + \frac{\nabla F_\psi(p_{T+1})^\top (p - p_{T+1})}{\eta_T} \\ &= \frac{1}{\eta_T}\left[\mathrm{Breg}_{F_\psi}(p\|p_1) - \mathrm{Breg}_{F_\psi}(p\|p_{T+1})\right]\end{aligned}$$

This completes the proof of the first part.

As $F_\psi(p_{t+1}) \geq 0$ and η_t are non-increasing we have:

$$\begin{aligned}\mathrm{Breg}_{F_\psi}^{\eta_t,\eta_{t-1}}(p_{t+1}\|p_t) &= \frac{F_\psi(p_{t+1})}{\eta_t} - \frac{F_\psi(p_t)}{\eta_{t-1}} - \frac{\nabla F_\psi(p_t)}{\eta_{t-1}}^\top (p_{t+1} - p_t) \\ &\geq \frac{F_\psi(p_{t+1})}{\eta_{t-1}} - \frac{F_\psi(p_t)}{\eta_{t-1}} - \frac{\nabla F_\psi(p_t)}{\eta_{t-1}}^\top (p_{t+1} - p_t) = \frac{1}{\eta_{t-1}}\mathrm{Breg}_{F_\psi}(p_{t+1}\|p_t)\end{aligned}$$

Thus, we have $\sum_{t=1}^T l_t^\top (p_t - p)$:

$$\begin{aligned}&= \frac{1}{\eta_T}\left[\mathrm{Breg}_{F_\psi}(p\|p_1) - \mathrm{Breg}_{F_\psi}(p\|p_{T+1})\right] + \sum_{t=1}^T \left[l_t^\top (p_t - p_{t+1}) - \mathrm{Breg}_{F_\psi}^{\eta_t,\eta_{t-1}}(p_{t+1}\|p_t)\right] \\ &\leq \frac{1}{\eta_T}\mathrm{Breg}_{F_\psi}(p\|p_1) + \sum_{t=1}^T \left[l_t^\top (p_t - p_{t+1}) - \mathrm{Breg}_{F_\psi}^{\eta_t,\eta_{t-1}}(p_{t+1}\|p_t)\right] \\ &\leq \frac{F_\psi(p)}{\eta_T} + \sum_{t=1}^T \left[l_t^\top (p_t - p_{t+1}) - \frac{1}{\eta_{t-1}}\mathrm{Breg}_{F_\psi}(p_{t+1}\|p_t)\right]\end{aligned}$$

This completes the proof. ■

Recall that the AdaFTRL strategy picks learning rate:

$$\eta_t = \frac{\alpha}{\beta + \sum_{s=1}^t M_s(\eta_{s-1})}$$

Where

$$M_t(\eta) = \sup_{q \in \Delta_n}\left[l_t^\top (p_t - q) - \frac{1}{\eta}\mathrm{Breg}_{F_\psi}(q\|p_t)\right]$$

Theorem 13 *If $M_t(\eta_{t-1})/\eta_{t-1} \leq g_t$, then for any $p \in \Delta_n$ and any sequence of losses $l_1, \ldots, l_T$, the iterates of AdaFTRL satisfy the regret inequality $\sum_{t=1}^T l_t^\top (p_t - p)$*

$$\leq F_\psi(p)\left(\frac{\beta}{\alpha} + \frac{2\sup_t \|l_t\|_\infty}{\alpha}\right) + 2\sup_t \|l_t\|_\infty + \sqrt{2\sum_{t=1}^T g_t}\left(\frac{F_\psi(p)}{\sqrt{\alpha}} + \sqrt{\alpha}\right)$$

Proof When using non-increasing η_t, the regret of FTRL is bounded by Theorem 12:

$$\begin{aligned}\sum_{t=1}^{T} l_t^\top (p_t - p) &\le \frac{F_\psi(p)}{\eta_T} + \sum_{t=1}^{T} \left[l_t^\top (p_t - p_{t+1}) - \frac{1}{\eta_{t-1}} \mathrm{Breg}_{F_\psi}(p_{t+1} \| p_t) \right] \\ &\le \frac{F_\psi(p)}{\eta_T} + \sum_{t=1}^{T} M_t(\eta_{t-1})\end{aligned}$$

Using the fact that $0 \le M_t(\eta) \le 2 \sup_t \|l_t\|_\infty$ and applying Lemma 10, we have $\sum_{t=1}^{T} l_t^\top (p_t - p)$

$$\le F_\psi(p) \left(\frac{\beta}{\alpha} + \frac{2 \sup_t \|l_t\|_\infty}{\alpha} \right) + 2 \sup_t \|l_t\|_\infty + \sqrt{2 \sum_{t=1}^{T} g_t} \left(\frac{F_\psi(p)}{\sqrt{\alpha}} + \sqrt{\alpha} \right)$$

■