Return of the bias: Almost minimax optimal high probability bounds for adversarial linear bandits

Julian Zimmert ZIMMERT@GOOGLE.COM
Google Research
Tor Lattimore LATTIMORE@GOOGLE.COM
DeepMind

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We introduce a modification of follow the regularised leader and combine it with the log determinant potential and suitable loss estimators to prove that the minimax regret for adaptive adversarial linear bandits is at most $O(d\sqrt{T\log(T)})$ where d is the dimension and T is the number of rounds. By using exponential weights, we improve this bound to $O(\sqrt{dT\log(kT)})$ when the action set has size k. These results confirms an old conjecture. We also show that follow the regularized leader with the entropic barrier and suitable loss estimators has regret against an adaptive adversary of at most $O(d^2\sqrt{T}\log(T))$ and can be implement in polynomial time, which improves on the best known bound for an efficient algorithm of $\mathcal{O}(d^{7/2}\sqrt{T}\operatorname{poly}(\log(T)))$ by Lee et al. (2020).

Keywords: Adversarial linear bandits, high probability bounds, adaptive adversary.

1. Introduction

Let $\mathcal{A}$ be a compact subset of $\mathbb{R}^d$ and assume its affine hull spans $\mathbb{R}^d$.[1] An agent and environment interact sequentially over T rounds. In each round t the agent and adversary act simultaneously. The agent chooses an action $a_t \in \mathcal{A}$ and the adversary chooses a vector $y_t \in \mathcal{X}^\circ = \{y \in \mathbb{R}^d : \max_{a\in\mathcal{A}} |\langle a, y\rangle| \le 1\}$. The agent observes $\ell_t = \langle a_t, y_t\rangle$ and the regret is

$$\operatorname{Reg}_T = \max_{u\in\mathcal{A}} \operatorname{Reg}_T(u)\,,$$

where $\operatorname{Reg}_T(u) = \sum_{t=1}^T \langle a_t - u, y_t\rangle$.

Contributions Our focus is on high probability and adaptive bounds.

1. We introduce follow the regularized leader with fixed point bias (FTRL-FB) that injects a linear bias into the objective of follow the regularized leader and solves a fixed point problem. The negative terms in the resulting regret bound are used to cancel terms that appear when controlling the variance of loss estimators when proving high probability bounds.

2. By lifting the linear bandit to the space of information matrices and instantiating FTRL-FB with the log determinant potential function we prove there exists an agent such that $\operatorname{Reg}_T = O(d\sqrt{T}\log(T/\delta))$ with probability at least $1-\delta$. This shows that for general action sets the minimax regret for adaptive adversaries is the same as for oblivious adversaries. One of the insights from our analysis is that by lifting to the space of positive definite matrices we introduces a kind of positivity that arises naturally for finite-armed (orthogonal) bandits and was the limiting factor in extending the high probability bounds in that setting to general linear bandits.

1. There is no loss of generality since otherwise, we can always project $\mathcal{A}$ into a lower dimensional subspace.

3. By combining FTRL-FB with the entropic barrier we design a polynomial time algorithm for which $\mathrm{Reg}_T = O(d^2\sqrt{T}\log(T/\delta)$ with probability at least $1-\delta$, which improves on the best known bound for a polynomial time algorithm by a factor of $d^{\frac{3}{2}}$ (Lee et al., 2020).

4. We improve the results by Bartlett et al. (2008) by using modern exploration techniques to show that when $|\mathcal{A}| = k$, then a suitable version of exponential weights on $\mathcal{A}$ has regret $\mathcal{O}(\sqrt{dT\log(k/\delta)})$ with high probability. Although this result is not especially novel, we are not aware of where it has appeared before. And, although it provides the strongest guarantees, we find it the least exciting because it does not introduce new ideas.

Related work Adversarial linear bandits by now have quite a long history (McMahan and Blum, 2004; Awerbuch and Kleinberg, 2004; Dani and Hayes, 2006). These early works set the scene for future work but provide suboptimal (not $\sqrt{T}$) regret in either the adaptive or oblivious setting. More recent works provide a web of results emphasizing different aspects of the problem, especially: computational efficiency, data-dependent bounds and high probability bounds (and adaptive adversaries). A summary of these results is given in Table 1.

High probability bounds for finite-armed adversarial bandits have been understood since near the beginning (Auer et al., 2002) with a number of more recent refinements providing alternative mechanisms (Abernethy and Rakhlin, 2009; Kocák et al., 2014; Neu, 2015). The ideas in these works strongly exploit the positivity of actions in the probability simplex. The only known way to apply these techniques to linear bandits is to lift the algorithm to play exponential weights on the space of actions. This was done by Bartlett et al. (2008), who use old exploration techniques to show this idea can achieve regret against an adaptive adversary of $\mathrm{Reg}_T = \tilde{\mathcal{O}}(d^{3/2}\sqrt{T})$. We improve the dependence on the dimension using modern exploration techniques. The disadvantage of this approach is that there is little hope for an efficient implementation.

There is very little work providing algorithms that are both efficient and where the regret is controlled with high probability. The best known bound is by Lee et al. (2020), who provide a polynomial time algorithm for which the regret is $\tilde{\mathcal{O}}(d^{7/2}\sqrt{T})$ with high probability.

One of our algorithms makes use of the entropic barrier, which was introduced by Bubeck and Eldan (2014) who proved it is $d(1+o(1))$-self concordant. The latter result was recently improved by Chewi (2021) who proved that the entropic barrier is d-self-concordant.

Notation The Dirac distribution on x is δ_x. Give set $A \subset \mathbb{R}^d$, the relative interior of A is denoted by $\mathrm{ri}(A)$ and its interior is $\mathrm{int}(A)$. The space of probability measures on $\mathcal{A}$ with the Borel σ-algebra is denoted by $\Delta(\mathcal{A})$. The convex hull of $\mathcal{A}$ is denoted by $\mathcal{X}$ and when $\mathcal{A}$ is finite we let $k = |\mathcal{A}|$ be the number of actions. For any distribution $p \in \Delta(\mathcal{A})$, denote the mean by $\mu(p) = \mathbb{E}_{a\sim p}[a]$, the covariance matrix by $\mathrm{Cov}(p) = \mathbb{E}_{a\sim p}[(a-\mu(p))(a-\mu(p))^\top]$. For any $a \in \mathcal{A}$, we further define the lifting $\boldsymbol{a} = \binom{a}{1}$ and the lifted covariance matrix $\widehat{\mathrm{Cov}}(p) = \mathbb{E}_{a\sim p}[\boldsymbol{a}\boldsymbol{a}^\top]$. Let $\mathcal{F}_t = \sigma(a_1, \ell_1, \ldots, a_t, \ell_t)$ be the σ-algebra generated by the interaction sequence observed by the learner until round t and let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_t]$. Bregman divergences with respect to a differentiable convex function $F : \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ is $D_F(x,y) = F(x) - F(y) - \nabla_{x-y}F(y)$ where $\nabla_v F(y)$ is the directional derivative of F at y in direction v. The domain of D_F is $\mathbb{R}^d \times \mathrm{dom}(F)$ and we adopt the convention that $D_F(x,y) = \infty$ whenever $x \notin \mathrm{dom}(F)$.

Paper	Action set	Regret	Efficient	Adaptive adversary
Auer et al. 2002, Neu 2015	simplex	$\sqrt{dT}$	YES	YES
Abernethy et al. 2008	continuous	$d^{3/2}\sqrt{T}$	YES	NO
Bartlett et al. 2008	finite	$d^{3/2}\sqrt{T}$	NO	YES
Audibert and Bubeck 2010	continuous	$d\sqrt{T}$	NO*	NO
Hazan and Karnin 2016	continuous	$d\sqrt{T}$	YES	NO
Ito et al. 2020	continuous	$d\sqrt{L_T^*}$	YES	NO
Lee et al. 2020	continuous	$d^{7/2}\sqrt{T}$	YES	YES
OUR WORK	continuous	$d\sqrt{T}$	NO	YES
	continuous	$d^2\sqrt{T}$	YES	YES
	finite	$d\sqrt{T}$	NO*	YES

Table 1: A history of results for adversarial linear bandits. Logarithms have been omitted from regret bounds. For algorithms designed for linear bandits with finitely many actions we have substituted $\log(k)$ for d as would be obtained by standard covering arguments. These algorithms are labelled as being inefficient with a star because their running time is at least linear in k and we are principally interested in the case where k is exponential in d.

2. Main techniques

We start by recalling the basic result about follow the regularized leader specialized to our situation. Let $\mathcal{K}$ be a closed, nonempty and convex subset of $\mathbb{R}^d$ and let $(\hat{y}_t)_{t=1}^T$ and $(b_t)_{t=1}^T$ be sequences of vectors in $\mathbb{R}^d$. The sequence $(\hat{y}_t)$ will later be replaced by estimated losses and the (b_t) will be bias terms introduced by the algorithm. Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be convex and differentiable on the interior of its domain and $\eta > 0$ and define a sequence $(x_t)_{t=1}^T$ by

$$x_t = \operatorname*{arg\,min}_{x \in \mathcal{K}} \left\langle x, \sum_{s=1}^{t-1} (\hat{y}_s - b_s) \right\rangle + \frac{F(x)}{\eta} \tag{1}$$

The following theorem is standard (Lattimore and Szepesvári, 2020, Theorem 28.5):

Theorem 1 *Suppose that (x_t) are chosen according to (1) and $u \in \mathcal{K}$, then*

$$\sum_{t=1}^T \langle x_t - u, \hat{y}_t \rangle \leq \frac{F(u) - F(x_1)}{\eta} + \sum_{t=1}^T \langle x_t - u, b_t \rangle + \sum_{t=1}^T \max_{x \in \mathcal{K}} \left[\langle x_t - x, \hat{y}_t - b_t \rangle - \frac{D_F(x, x_t)}{\eta} \right].$$

Notice that the bias terms (b_t) appear linearly in the first sum on the right-hand side and contribute to the stability term in the second sum on the right-hand side. The challenge when proving high probability bounds is to choose (b_t) in such a way that the linear terms cancel the standard deviation of the loss estimators while the contribution to the stability is small. Note that x_t depends on $(\hat{y}_s)_{s=1}^{t-1}$ and $(b_s)_{s=1}^{t-1}$, so there is no problem defining (b_t) adaptively.

FTRL-FB We also introduce a novel modification of FTRL that eliminates the bias term from the stability component of the bound. The modified algorithm needs to solve a fixed point problem that is likely not computationally efficient in general. Nevertheless, the analysis becomes more straightforward and the role of the bias is conceptually simpler. Suppose that $b : \cup_{t=0}^{\infty}(\mathcal{K} \times \mathbb{R}^d)^t \times \mathcal{K} \to \mathbb{R}^d$ is continuous. Then the modified algorithm computes b_t and x_t as solutions to the fixed point problem

$$b_t = b(x_1, \hat{y}_1, \dots, x_{t-1}, \hat{y}_{t-1}, x_t) \qquad x_t = \operatorname*{arg\,min}_{x \in \mathcal{K}} \left\langle x, \sum_{s=1}^{t-1} \hat{y}_s - \sum_{s=1}^{t} b_s \right\rangle + \frac{F(x)}{\eta} . \tag{2}$$

When F is strictly convex, the mapping $x \mapsto \arg\max_{x \in \mathcal{K}} \langle x, u\rangle + F(x)/\eta$ is continuous for any u, which means that Brouwer's theorem establishes the existence of a solution for x_t and b_t.

Theorem 2 *If $(x_t)_{t=1}^T$ and $(b_t)_{t=1}^T$ satisfy Eq. (2) and $u \in \mathcal{K}$, then*

$$\sum_{t=1}^{T} \langle x_t - u, \hat{y}_t\rangle \le \frac{F(u) - \min_{x \in \mathcal{K}} F(x)}{\eta} + \sum_{t=1}^{T} \langle x_t - u, b_t\rangle + \sum_{t=1}^{T} \underbrace{\max_{x \in \mathcal{K}} \left[\langle x_t - x, \hat{y}_t\rangle - \frac{D_F(x, x_t)}{\eta} \right]}_{\triangleq \mathrm{stability}_t} .$$

The proof is deferred to Appendix E. Comparing the bound in Theorem 2 to that in Theorem 1, we can see that in the latter the bias term only appears linearly.

2.1. Motivation: high probability bounds

All our algorithms are based on the following elementary regret decomposition:

Lemma 3 *Let $x_t = \mathbb{E}_{t-1}[a_t]$. With probability at least $1 - \delta$, for any $u \in \mathcal{X}$,*

$$\mathrm{Reg}_T(u) \le \sqrt{2T \log(1/\delta)} + \sum_{t=1}^{T} \underbrace{\langle x_t - u, y_t - \hat{y}_t\rangle}_{\triangleq \mathrm{dev}_t(u)} + \langle x_t - u, \hat{y}_t\rangle ,$$

where $\hat{y}_t$ is an arbitrary sequence of vectors in $\mathbb{R}^d$.

In all our applications $\hat{y}_t$ will be a (conditionally) unbiased estimator of y_t so that $\mathbb{E}_{t-1}[\hat{y}_t] = y_t$. The sum of the last terms in Lemma 3 will be bounded using follow the regularized leader. By concentration of measure arguments, provided that $\mathrm{dev}_t(u)$ has suitable tails, we should expect

$$\sum_{t=1}^{T} \mathrm{dev}_t(u) = \mathcal{O}\left(\sqrt{\sum_{t=1}^{T} \mathbb{E}_{t-1}[\mathrm{dev}_t(u)^2] \log\left(\frac{1}{\delta}\right)} \right) .$$

Combining this with the regret bound in Theorem 2 yields

$$\mathrm{Reg}_T(u) \le \mathcal{O}\left(\sqrt{\sum_{t=1}^{T} \mathbb{E}_{t-1}[\mathrm{dev}_t(u)^2] \log\left(\frac{1}{\delta}\right)} \right) + \sum_{t=1}^{T} \langle x_t - u, b_t\rangle + \sum_{t=1}^{T} \mathrm{stability}_t .$$

Generally speaking the stability term is relatively easy to control with high probability. The challenge is that the first term can be of order T, so one needs to choose the biases so that the bias terms cancel the variation of the deviation terms for all u simultaneously. The same challenge has been faced by Lee et al. (2020), who generate negative regret via increasing learning rates. Foster et al. (2020) discussed the close relationship between increasing learning rate and adding linear biases. Generally speaking, any application of the negative regret by increasing learning rate (e.g. (Agarwal et al., 2017; Luo et al., 2018; Lee et al., 2020)) can also be solved via linear biases at improved logarithmic terms. Additionally, the increasing learning rate trick is limited to biases of the type $b_t \propto \nabla F(x_t)$, which does not work for our last algorithm.

3. Algorithms

We present three algorithms, all based on FTRL(-FB) with different potential functions, exploration mechanisms and biases. The first uses the entropic barrier (also known as continuous exponential weights). This algorithm can be implemented in polynomial time provided the learner has access to a representation of $\mathcal{X}$ that allows linear optimization. The second algorithm makes use of a novel lifting to the space of information matrices and the log determinant for regularization. The analysis of this algorithm is especially clean, but at the moment we do not know if it can be implemented efficiently. Finally we show that (discrete) exponential weights with John's exploration can also be modified to obtain high probability bounds, but it is very unlikely that any efficient implementation exists. The theorems are presented first, with the algorithms appearing in subsections afterwards. We present the proof of Theorem 5 in its subsection, while all other proofs are deferred to the appendix.

Theorem 4 (ENTROPIC BARRIER) *Assume that $\mathcal{A} = \mathcal{X}$. For any $\delta \in (0,1)$, with probability at least $1-\delta$ there exists a tuning of Algorithm 1 such that the regret against any adaptive adversary is at most*

$$\mathrm{Reg}_T = \mathcal{O}\left(d^2\sqrt{T}\log(T/\delta)\right) .$$

The required algorithm parameter to obtain this theorem are given in Appendix F. Note that this algorithm does not use the fixed-point bias of FTRL and is computationally comparable to the algorithm of Ito et al. (2020). Setting $\delta = 1/T$ yields the first efficient algorithm with $\mathcal{O}(d^2\sqrt{T}\log(T))$ regret for adversarial bandits, improving the recent result of Lee et al. (2020) by a factor of $d^{3/2}$ and several $\log(T)$ factors. The assumption that $\mathcal{A} = \mathcal{X}$ is for convenience only. The modification needed if this is not the case is to sample an 'action' from $\mathcal{X}$ and then play randomly from a barycentric spanner with the required mean. This introduces an easily controllable amount of noise and complicates the notation but otherwise does not change the results in a material way. Note that barycentric spanners can be computed using linear optimization on $\mathcal{X}$. For the algorithm using the log determinant we have the following bound, which is minimax optimal up to logarithmic factors.

Theorem 5 (LOG DETERMINANT) *For any $\delta \in (0,1)$, with probability at least $1-\delta$ the regret of Algorithm 2 against any adaptive adversary is at most*

$$\mathrm{Reg}_T = \mathcal{O}\left(d\sqrt{T\log(T/\delta)}\right) .$$

Theorem 5 shows that the minimax rate against adaptive adversaries is the same as against oblivious adversaries, resolving a long-standing open problem.

Our final algorithm is a combination of exponential weights with John's exploration (Bubeck et al., 2012) and the biasing technique by Bartlett et al. (2008). There is not much novelty in our analysis, which simply injects modern exploration techniques into an old algorithm. Nevertheless, we include it because as far as we know this has not been written anywhere.

Theorem 6 (EXPONENTIAL WEIGHTS) *For any $\delta \in (0,1)$, with probability at least $1-\delta$ the regret of Algorithm 3 against any adaptive adversary is at most*

$$\mathrm{Reg}_T = \mathcal{O}\left(\sqrt{dT\log(|\mathcal{A}|/\delta)}\right).$$

3.1. Entropic barrier

Recall that the entropic barrier F is defined in terms of its Fenchel dual F^*, which is

$$F^*(\theta) = \log\left(\int_{\mathcal{X}} \exp(\langle x, \theta\rangle) dx\right).$$

By definition of the Fenchel dual, $F(x) = \sup_\theta \langle x,\theta\rangle - F^*(\theta)$. The following facts have all been established by (Bubeck and Eldan, 2014). Let p_θ be the density of a probability measure on $\mathcal{X}$ defined by

$$p_\theta(x) = \mathbb{I}_{\mathcal{X}}(x)\exp\left(-\langle x,\theta\rangle - F^*(\theta)\right),$$

which is an exponential weights distribution on $\mathcal{X}$. The gradient of F^* is $x(\theta) \triangleq \nabla F^*(\theta) = \int_{\mathcal{X}} x p_\theta(x) dx$, which is an invertible function with inverse $x \mapsto \theta(x)$. The Hessian of the entropic barrier is

$$\nabla^2 F(x) = \left(\int_{\mathcal{X}} (z-x)(z-x)^\top p_{\theta(x)}(z) dz\right)^{-1}.$$

Algorithm 1 follows FTRL with the entropic barrier. To ensure that the loss estimates are bounded, we sample from a mixture p'_t of the exponential weights distribution p_t and the uniform distribution p_0. At any round, we add biases proportional to $\nabla F(x_t)$, where the factor depends on the lifted information matrices $\boldsymbol{G}_t = \widehat{\mathrm{Cov}}(p'_t)$. The reason we tune the scaling factor according to the lifted information matrix, is the property

$$\|\boldsymbol{u}\|^2_{\boldsymbol{G}_t^{-1}} = \left\|u - \mu(p'_t)\right\|^2_{G_t^{-1}} + 1,$$

where the right hand side is related to $\mathbb{E}_{t-1}[\mathrm{dev}_t(u)^2]$ which we like to cancel.

Computation Note that $\nabla F(x_t) = -\eta\sum_{s=1}^{t-1}(\hat{y}_s - b_s)$, which means that we do not require direct gradient access of F. Assuming that linear optimization over $\mathcal{X}$ can be solved efficiently, Lovász and Vempala (2007) show that sampling from p'_t as a mixture of two log-concave distributions admits a polynomial time (approximate) implementation. Similarly, G_t as the variance of p'_t can be ε-approximated with $(d/\varepsilon)^{\mathcal{O}(1)}$ samples (Lovász and Vempala, 2007, Corollary 4.2). The eigenvalues of G_t are lower bounded by $\Omega(1/\sqrt{T})$, hence with $\frac{1}{T}$-precision, the inverse G_t^{-1} is also $\mathcal{O}\left(\frac{1}{T}\right)$-approximated. This results in a Poly(dT) per-step computation.

Algorithm 1: FTRL-FB with entropic barrier

Input: Action set $\mathcal{X}$, entropic barrier F, η, λ, γ, p_0 uniform distribution over $\mathcal{X}$

for $t = 1, \dots$ **do**

$p_t(x) \propto \mathbb{I}_{\mathcal{X}}(x) \exp\left(-\eta \langle x, \sum_{s=1}^{t-1}(y_s - b_s)\rangle\right)$

$p'_t = p_t + \lambda(p_0 - p_t)$

Let $G_t, \boldsymbol{G}_t$ be the covariance and lifted covariance of p'_t and x'_t its mean

Sample a_t from probability measure with density p'_t

Observe ℓ_t and construct $\hat{y}_t = (G_t)^{-1}(a_t - x'_t)\ell_t$

$b_t = \gamma\sqrt{\mathrm{Tr}(\boldsymbol{G}_t^{-1} \cdot (\sum_{s=1}^{t} \boldsymbol{G}_s^{-1})^{-1})}\nabla F(x_t)$

end

Remark 7 *Readers might object that this is not* efficient *as claimed and we cannot compare with the result of Lee et al. (2020), which is based on the SCRiBLe algorithm. Note however, that SCRiBLe simply assumes oracle access to gradients and Hessians of the potential F. For general action sets, there is no self-concordant barrier known whose gradients and Hessians can be computed more efficiently than a ε-approximation in Poly($\frac{d}{\varepsilon}$) time. Finally, our proof does not use any special properties of the entropic barrier, besides that it is self-concordant and admits a sampling distribution with information matrix proportional to $\nabla^2 F(x)^{-1}$. It is an easy exercise to adapt the proof of Theorem 4 to a SCRiBLe style algorithm at a cost of an extra $\sqrt{d}$ factor in the regret.*

3.2. Log determinant barrier

The high probability bounds for finite-armed bandits heavily rely on the fact that the comparator class is the probability simplex and the positivity of the vectors there-in. To generalize those techniques we make use of a new lifting to the space of positive definite matrices using the negative log determinant as a potential function in combination with FTRL-FB.

Remark 8 *A comparable bound to Theorem 5 can be obtained using the standard FTRL algorithm and a slightly less elegant analysis. Even so, we do not know of an efficient implementation of this algorithm. The best known implementation for solving the minimization problem requires a runtime that is linear in the number of arms $|\mathcal{A}|$ (Foster et al., 2020), which makes this algorithm unpractical when the action set is exponentially large.*

Let $p_0 \in \Delta(\mathcal{A})$ be a distribution such that $\|x - \mu(p_0)\|^2_{\mathrm{Cov}(p_0)^{-1}} \leq d$ for all $x \in \mathcal{X}$, which exists by the theory minimum volume enclosing ellipsoids (Todd, 2016, Corollary 2.11). Let $\eta > 0$ be a learning rate and

$$\mathbb{R}^{(d+1)\times(d+1)} \supset \mathcal{H} = \left\{\widehat{\mathrm{Cov}}(p) : p \in \Delta(\mathcal{A}), \widehat{\mathrm{Cov}}(p) \succeq d\eta\widehat{\mathrm{Cov}}(p_0)\right\},$$

which is convex. Our learner will play on $\mathcal{H}$ using FTRL-FB. The losses and their estimates are lifted to the same space by

$$\gamma_t = \begin{pmatrix} 0 & y_t \\ 0 & 0 \end{pmatrix}, \text{ and } \hat{\gamma}_t = \begin{pmatrix} 0 & \hat{y}_t \\ 0 & 0 \end{pmatrix}.$$

By the definition of $\widehat{\mathrm{Cov}}(p)$ the lifted losses and their estimates satisfy $\langle \widehat{\mathrm{Cov}}(p), \boldsymbol{\gamma}_t \rangle = \langle \mu(p), y_t \rangle$ and $\langle \widehat{\mathrm{Cov}}(p), \hat{\boldsymbol{\gamma}}_t \rangle = \langle \mu(p), \hat{y}_t \rangle$.

Algorithm 2: FTRL-FB with logdet barrier

Input: log determinant barrier $F(\boldsymbol{H}) = -\log\det(\boldsymbol{H})$, $\eta = \sqrt{\frac{\log(T/\delta)}{T}}$

for $t = 1, \dots$ **do**

Find $\boldsymbol{H}_t$ as the solution to the fixed point problem

$$\boldsymbol{H}_t = \operatorname*{arg\,min}_{\boldsymbol{H} \in \mathcal{H}} \left\langle \boldsymbol{H}, \sum_{s=1}^{t-1} \hat{\boldsymbol{\gamma}}_s - \eta \sum_{s=1}^{t} \boldsymbol{H}_s^{-1} \right\rangle + \frac{F(\boldsymbol{H})}{\eta} .$$

Select $p_t \in \Delta(\mathcal{A})$ such that $\boldsymbol{H}_t = \widehat{\mathrm{Cov}}(p_t)$ and let $x_t = \mu(p_t)$ and $H_t = \mathrm{Cov}(p_t) and$

Sample $a_t \sim p_t$

Observe ℓ_t and construct $\hat{y}_t = \mathrm{Cov}(p_t)^{-1}(a_t - x_t)\ell_t$ and $\hat{\boldsymbol{\gamma}}_t$

end

Proof [THEOREM 5] We start by decomposing the regret relative to a fixed comparator into a deviation term and the regret with estimated losses. The latter is bounded using Theorem 2 in step 2. In step 3 we handle the deviation term and random terms that appeared in step 2. In the last step we put together the pieces and take a union bound over a suitable finite cover of possible comparators. Let $x_t = \mathbb{E}_{t-1}[a_t] = \mu(p_t)$ be the mean of the learner's action distribution in round t and $H_t = \mathrm{Cov}(p_t)$ the covariance matrix.

Step 1: Decomposing the regret Let p be such that $\boldsymbol{U} = \widehat{\mathrm{Cov}}(p) \in \mathcal{H}$ and $u = \mu(p)$ be the mean of p. Since $\boldsymbol{U} \in \mathcal{H}$, by the definition of $\mathcal{H}$, $\boldsymbol{U} \succeq d\eta \boldsymbol{H}_0$. By the Courant-Fischer-Weyl min-max principle, $\boldsymbol{U}$ has larger eignevalues than $d\eta\boldsymbol{H}_0$ and hence

$$F(\boldsymbol{U}) - F(\boldsymbol{H}_0) = \log\left(\frac{\det(\boldsymbol{H}_0)}{\det(\boldsymbol{U})}\right) \leq (d+1)\log\left(\frac{1}{d\eta}\right) . \tag{3}$$

By Lemma 3, then with probability at least $1-\delta$,

$$\mathrm{Reg}_T(u) \leq \sqrt{2T\log(1/\delta)} + \sum_{t=1}^{T} \langle \boldsymbol{H}_t - \boldsymbol{U}, \boldsymbol{\gamma}_t - \hat{\boldsymbol{\gamma}}_t \rangle + \langle \boldsymbol{H}_t - \boldsymbol{U}, \hat{\boldsymbol{\gamma}}_t \rangle . \tag{4}$$

Step 2: Controlling the empirical regret Note that F is strictly convex (Boyd and Vandenberghe, 2004, p.74) so that the fixed point problem defining $\boldsymbol{H}_t$ is guaranteed to have a solution by

Brouwer's theorem. The second sum in (4) is bounded using Theorem 2 and Lemma 16:

$$\begin{aligned}
&\sum_{t=1}^{T} \langle \boldsymbol{H}_t - \boldsymbol{U}, \hat{\boldsymbol{\gamma}}_t \rangle \\
&\leq \frac{F(\boldsymbol{U}) - F(\boldsymbol{H}_0)}{\eta} + \sum_{t=1}^{T} \left[\max_{\boldsymbol{H} \in \mathcal{H}} \langle \boldsymbol{H}_t - \boldsymbol{H}, \hat{\boldsymbol{\gamma}}_t \rangle - \frac{D_F(\boldsymbol{H}, \boldsymbol{H}_t)}{\eta} + \eta \langle \boldsymbol{H}_t - \boldsymbol{U}, \boldsymbol{H}_t^{-1} \rangle \right] \quad \text{(Lem. 2)} \\
&\leq \frac{(d+1) \log\left(\frac{1}{d\eta}\right)}{\eta} + \sum_{t=1}^{T} \left[\max_{\boldsymbol{H} \in \mathcal{H}} \langle \boldsymbol{H}_t - \boldsymbol{H}, \hat{\boldsymbol{\gamma}}_t \rangle - \frac{D_F(\boldsymbol{H}, \boldsymbol{H}_t)}{\eta} + \eta \langle \boldsymbol{H}_t - \boldsymbol{U}, \boldsymbol{H}_t^{-1} \rangle \right] \quad \text{(by (3))} \\
&\leq \frac{(d+1) \log\left(\frac{1}{d\eta}\right)}{\eta} + \frac{\eta}{4} \sum_{t=1}^{T} \|a_t - x_t\|^2_{H_t^{-1}} + \eta \sum_{t=1}^{T} \langle \boldsymbol{H}_t - \boldsymbol{U}, \boldsymbol{H}_t^{-1} \rangle \quad \text{(Lem. 16)} \\
&= \frac{(d+1) \log\left(\frac{1}{d\eta}\right)}{\eta} + \frac{\eta}{4} \sum_{t=1}^{T} \|a_t - x_t\|^2_{H_t^{-1}} + \eta dT - \eta \|u - x_t\|^2_{H_t^{-1}} \,, \qquad (5)
\end{aligned}$$

where in the last line we used the definition of the lifting. Notice the negative term that will be used to cancel the variation of the sum of deviations.

Step 3: Concentration The first sum in (4) vanishes in expectation since $\hat{\boldsymbol{\gamma}}_t$ is a conditionally unbiased estimate of $\boldsymbol{\gamma}_t$. For a high probability bound we make use of Exercise 5.15 by Lattimore and Szepesvári (2020). Using the fact that $\boldsymbol{H}_t \in \mathcal{H}$ so that for any $x \in \mathcal{X}$,

$$\|x - x_t\|^2_{H_t^{-1}} \leq \frac{1}{d\eta} \|x - x_t\|^2_{\mathrm{Cov}(p_0)^{-1}} \leq \frac{1}{\eta} \,.$$

Therefore, by Exercise 5.15 by Lattimore and Szepesvári (2020), with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} \eta \|a_t - x_t\|^2_{H_t^{-1}} \leq \eta(d+1)T + \frac{1}{\eta} \log\left(\frac{1}{\delta}\right) \,, \qquad (6)$$

where we used the fact that $\mathbb{E}_{t-1}\left[\|a_t - x_t\|^2_{H_t^{-1}}\right] = d$. Furthermore,

$$\eta |\langle u - x_t, \hat{y}_t \rangle| = \eta |\langle u - x_t, H_t^{-1}(a_t - x_t)\ell_t \rangle| \leq 1 \,.$$

Then, by the same exercise, with probability at least $1 - \delta$,

$$\begin{aligned}
\sum_{t=1}^{T} \langle u - x_t, \hat{y}_t \rangle &\leq \sum_{t=1}^{T} \langle u - x_t, y_t \rangle + \eta \sum_{t=1}^{T} \mathbb{E}_{t-1}[\langle u - x_t, \hat{y}_t \rangle^2] + \frac{1}{\eta} \log\left(\frac{1}{\delta}\right) \\
&\leq \sum_{t=1}^{T} \langle u - x_t, y_t \rangle + \eta \sum_{t=1}^{T} \|u - x_t\|^2_{H_t^{-1}} + \frac{1}{\eta} \log\left(\frac{1}{\delta}\right) \,.
\end{aligned}$$

Therefore, with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} \langle \boldsymbol{H}_t - \boldsymbol{U}, \boldsymbol{\gamma}_t - \hat{\boldsymbol{\gamma}}_t \rangle \leq \eta \sum_{t=1}^{T} \|u - x_t\|^2_{H_t^{-1}} + \frac{1}{\eta} \log\left(\frac{1}{\delta}\right) \,. \qquad (7)$$

Step 4: Finishing up Combining (5), (6) and (7), for any $\boldsymbol{U} \in \mathcal{H}$, with probability at least $1 - 2\delta$,

$$\sum_{t=1}^{T} \langle \boldsymbol{H_t} - \boldsymbol{U}, \boldsymbol{\gamma}_t \rangle \leq \frac{(d+1)\log\left(\frac{1}{\eta d}\right)}{\eta} + \eta(2d+1)T + \frac{2}{\eta}\log\left(\frac{1}{\delta}\right) .$$

We complete the proof with a covering argument and a union bound to control the regret compared to an adaptive adversary. Let $\|x\|_{\mathcal{X}^\circ} = \sup_{y \in \mathcal{X}^\circ} \langle x, y \rangle$ and $\mathcal{C}$ be a finite subset of $\mathcal{X}$ such that

$$\max_{x \in \mathcal{X}} \min_{x' \in \mathcal{C}} \left\| x - x' \right\|_{\mathcal{X}^\circ} \leq \frac{d}{\sqrt{T}} .$$

The covering set $\mathcal{C}$ can be chosen so that $\log|\mathcal{C}| \leq d\log(6\sqrt{T})$ (Lattimore and Szepesvári, 2020, Exercise 27.6). Let $\mathcal{U} = \{(1-d\eta)\widehat{\mathrm{Cov}}(\delta_x) + d\eta \boldsymbol{H}_0 : x \in \mathcal{C}\} \subset \mathcal{H}$. By a union bound, with probability at least $1 - 2\delta$ the following holds for all $\boldsymbol{U} \in \mathcal{U}$,

$$\sum_{t=1}^{T} \langle \boldsymbol{H_t} - \boldsymbol{U}, \boldsymbol{\gamma}_t \rangle \leq \frac{(d+1)\log\left(\frac{1}{\eta d}\right)}{\eta} + \eta(2d+1)T + \frac{2}{\eta}\log\left(\frac{|\mathcal{C}|}{\delta}\right) .$$

Let $\boldsymbol{U} \in \mathcal{U}$ be such that

$$\sum_{t=1}^{T} \langle \boldsymbol{U} - \boldsymbol{U}^*, y_t \rangle \leq d\eta T + d\sqrt{T} .$$

Then, with probability at least $1 - 2\delta$,

$$\mathrm{Reg}_T \leq \sqrt{2T\log\left(\frac{1}{\delta}\right)} + d\sqrt{T} + \frac{(d+1)\log\left(\frac{1}{\eta d}\right)}{\eta} + \eta(3d+1)T + \frac{2}{\eta}\log\left(\frac{|\mathcal{C}|}{\delta}\right) .$$

Substituting $\eta = \sqrt{\log\left(\frac{T}{\delta}\right)/T}$ completes the proof. ■

3.3. Exponential weights

Our last algorithm is a combination of exponential weights with Kiefer-Wolfowitz exploration (Audibert and Bubeck, 2009) and arguments by Bartlett et al. (2008) and Auer et al. (2002). Nothing here is particularly remarkable but as far as we know this has not been written down anywhere. The drawback of this approach is that there seems very little hope for an efficient implementation when the number of actions is large. Nevertheless, it provides the strongest results when the action set is small. The algorithm makes use of an exploration distribution $p_0 \in \Delta(\mathcal{A})$ such that when $G_0 = \sum_{a \in \mathcal{A}} p(a) a a^\top$ is the design matrix of p_0, $\|a\|^2_{G_0^{-1}} \leq d$. The existence of such a distribution is guaranteed by Kiefer-Wolfowitz theorem (Kiefer and Wolfowitz, 1960).

Algorithm 3: Exp3 with Kiefer-Wolfowitz exploration

Input: Exploration distribution p_0, learning rate $\eta = \sqrt{\log(k/\delta)/(dT)}$ and exploration rate $\lambda = 2\eta d$, exploration distribution p_0

for $t = 1, \dots$ **do**

Compute an exponential weights distribution $p_t \in \Delta(\mathcal{A})$

$$p_t(a) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} (\hat{y}_s(a) - b_s(a))\right)}{\sum_{b \in \mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} (\hat{y}_s(b) - b_s(b))\right)} .$$

Sample a_t from $p'_t = (1-\lambda)p_t + \lambda p_0$ and observe ℓ_t

Estimate losses $\hat{y}_s(a) = \langle a, G_t^{-1} a_t \rangle \ell_t$ with $G_t = \sum_{a \in \mathcal{A}} p'_t(a) a a^\top$ for all $a \in \mathcal{A}$

Compute bias $b_t(a) = \eta \left\|a\right\|^2_{G_t^{-1}}$ for all $a \in \mathcal{A}$

end

4. Discussion

We have derived a novel modification of FTRL that allows to insert an arbitrary adaptive bias sequence to the regret without changing the other moving parts of the analysis. We have resolved the question of the minimax rate of regret against adaptive linear bandits via two adaptations of the FTRL-FB framework. We improved the state-of-the-art regret bound against adaptive adversaries for efficiently implementable algorithms by factor $d^{\frac{3}{2}}$ and several $\log(T)$ factors.

Open problems for future work are the question of whether there are action sets of interest for which the logdet barrier admits efficient implementation. Finally, there is still a gap of factor d between efficient and inefficient algorithms.

References

Jacob Abernethy and Alexander Rakhlin. Beating the adaptive bandit with high probability. In *2009 Information Theory and Applications Workshop*, pages 280–289. IEEE, 2009.

Jacob Abernethy, Elad E Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *21st Annual Conference on Learning Theory, COLT 2008*, pages 263–273, 2008.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. 2017.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. 2009.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. 11(Oct), 2010.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1), 2002.

Baruch Awerbuch and Robert D Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 45–53, 2004.

Peter Bartlett, Varsha Dani, Thomas Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory-COLT 2008*, pages 335–342. Omnipress, 2008.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

Sébastien Bubeck and Ronen Eldan. The entropic barrier: a simple and optimal universal self-concordant barrier. *arXiv preprint arXiv:1412.1587*, 2014.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Sham M Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, pages 41–1. JMLR Workshop and Conference Proceedings, 2012.

Sinho Chewi. The entropic barrier is n-self-concordant. *arXiv preprint arXiv:2112.10947*, 2021.

Varsha Dani and Thomas P Hayes. Robbing the bandit: Less regret in online geometric optimization against an adaptive adversary. In *SODA*, volume 6, pages 937–943, 2006.

Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

Elad Hazan and Zohar Karnin. Volumetric spanners: an efficient exploration basis for learning. *Journal of Machine Learning Research*, 2016.

Shinji Ito, Shuichi Hirahara, Tasuku Soma, and Yuichi Yoshida. Tight first-and second-order regret bounds for adversarial linear bandits. *Advances in Neural Information Processing Systems*, 33: 2028–2038, 2020.

J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12(5):363–365, 1960.

Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 1*, pages 613–621, 2014.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. *Advances in Neural Information Processing Systems*, 33, 2020.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

Haipeng Luo, Chen-Yu Wei, and Kai Zheng. Efficient online portfolio with logarithmic regret. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8245–8255, 2018.

H Brendan McMahan and Avrim Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *International Conference on Computational Learning Theory*, pages 109–123. Springer, 2004.

Yurii Nesterov. *Introductory lectures on convex optimization : a basic course*. Mathematics and its applications ; v. 564. Kluwer Academic Publishers, 2004.

Yurii Nesterov. Constructing self-concordant barriers for convex cones. 2006.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances on Neural Information Processing Systems 28 (NIPS 2015)*, pages 3150–3158, 2015.

Michael J Todd. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM, 2016.

Appendix A. Proof of Lemma 3

Since $\langle a_t, y_t\rangle \in [0,1]$ and $\mathbb{E}_{t-1}[\langle a_t - x_t, y_t\rangle] = 0$, by Azuma's inequality,

$$\begin{aligned}
\mathrm{Reg}_T(u) &= \sum_{t=1}^{T} \langle a_t - u, y_t\rangle \\
&\le \sum_{t=1}^{T} \langle a_t - x_t, y_t\rangle + \sum_{t=1}^{T} \langle x_t - u, y_t\rangle \\
&\le \sqrt{2T\log(1/\delta)} + \sum_{t=1}^{T} \langle x_t - u, y_t\rangle \\
&= \sqrt{2T\log(1/\delta)} + \sum_{t=1}^{T} \underbrace{\langle x_t - u, y_t - \hat{y}_t\rangle}_{\mathrm{dev}_t(u)} + \langle x_t - u, \hat{y}_t\rangle\,.
\end{aligned}$$

Appendix B. A strengthened Freedman's inequality

This theorem is an improvement of Lee et al. (2020, Theorem 2.2), which contains an error in its proof.

Theorem 9 (Strengthened Freedman's inequality) *Let $X_1, X_2, \ldots$ be a martingale difference sequence with respect to a filtration $F_1 \subseteq F_2 \subseteq \ldots$ such that $\mathbb{E}[X_t \,|\, F_t] = 0$ and assume $\mathbb{E}[|X_t| \,|\, F_t] < \infty$ a.s. Then with probability at least $1-\delta$*

$$\sum_{t=1}^{T} X_t \le 3\sqrt{V_T \log\left(\frac{2\max\{U_T, \sqrt{V_T}\}}{\delta}\right)} + 2U_t \log\left(\frac{2\max\{U_T, \sqrt{V_T}\}}{\delta}\right),$$

where $V_T = \sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t^2]$, $U_T = \max\{1, \max_{s\in[T]} X_s\}$.

Proof Define $Z_t^{(i)} = X_t \cdot \mathbb{I}\left(U_t \le 2^i\right)$, then $Z_t^{(i)}$ is a sequence of random variables adapted to $(\mathcal{F}_t)_t$, such that $Z_t^{(i)} 2^{-i} \le 1$ almost surely. Hence by Exercise 5.15 by Lattimore and Szepesvári (2020) with probability at least $1 - \delta/2^i$, we have

$$\sum_{t=1}^{T} Z_t^{(i)} \le \sum_{t=1}^{T} (Z_t^{(i)} - \mathbb{E}_{t-1}[Z_t^{(i)}]) \le 2^{-i} \sum_{t=1}^{T} \mathbb{E}_{t-1}[{Z_t^{(i)}}^2] + 2^i \log\left(\frac{2^i}{\delta}\right).$$

By a union bound, this holds with probability $1-\delta$ uniformly over all i. Note that $\sum_{t=1}^{T} \mathbb{E}_{t-1}[{Z_t^{(i)}}^2] \le \sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t^2] = V_T$ and for any i such that $2^i \le U_T$, we have $\sum_{t=1}^{T} Z_t^{(i)} = \sum_{t=1}^{T} X_t$. Hence with probability $1 - \delta$

$$\begin{aligned}
\sum_{t=1}^{T} X_t &\le \min_{i: 2^i \ge U_T} 2^{-i} V_t + 2^i \log\left(\frac{2^i}{\delta}\right) \\
&\le \min_{i: 2\max\{U_T, \sqrt{V_t}\} \ge 2^i \ge U_T} 2^{-i} V_t + 2^i \log\left(\frac{2\max\{U_T, \sqrt{V_T}\}}{\delta}\right) \\
&\le 3\sqrt{V_T \log\left(\frac{2\max\{U_T, \sqrt{V_T}\}}{\delta}\right)} + 2U_t \log\left(\frac{2\max\{U_T, \sqrt{V_T}\}}{\delta}\right).
\end{aligned}$$

■

Appendix C. Properties of self-concordant barriers

In this section we collect the basic definitions and properties of self-concordant barriers. Let $f : \mathrm{int}(\mathcal{X}) \to \mathbb{R}$ be a C^3 smooth convex function. f is called a self-concordant barrier on $\mathcal{X}$ if it satisfies:

- $\mathcal{X}(x_i) \to \infty$ as $i \to \infty$ for any sequence $x_1, x_2, \ldots \in \mathrm{int}(\mathcal{X}) \subset \mathbb{R}^d$ converging to the boundary of $\mathcal{X}$;
- for all $x \in \mathrm{int}(\mathcal{X})$ and $h \in \mathbb{R}^d$, the following inequality always holds:

$$\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d} \frac{\partial^3 f(x)}{\partial x_i \partial x_j \partial x_k} h_i h_j h_k \le 2\|h\|^3_{\nabla^2 f(x)}.$$

We further call f is a ν-self-concordant barrier if it satisfies the conditions above and also

$$\langle \nabla f(x), h \rangle \le \sqrt{\nu} \|h\|_{\nabla^2 f(x)}$$

for all $x \in \mathrm{int}(\mathcal{X})$ and $h \in \mathbb{R}^d$.

Lemma 10 (Theorem 2.1.1 in (Nesterov and Nemirovskii, 1994)) *If f is a self-concordant barrier on $\mathcal{X}$, then the Dikin ellipsoid centered at $x \in \mathrm{int}(\mathcal{X})$, defined as $\{v : \|v - w\|_{\nabla^2 f(w)} \le 1\}$, is always within $\mathcal{X}$. Moreover,*

$$\|h\|_{\nabla^2 f(v)} \ge \|h\|_{\nabla^2 f(w)} \left(1 - \|v - w\|_{\nabla^2 f(w)}\right)$$

holds for any $h \in \mathbb{R}^d$ and any v with $\|v - w\|_{\nabla^2 f(w)} \le 1$.

Lemma 11 (Corollary 2.3.1 in (Nesterov and Nemirovskii, 1994)) *Let f be a self-concordant barrier for $\mathcal{X} \subset \mathbb{R}^d$. Then for any $x \in \operatorname{int}(\mathcal{X})$ and any $u \in \mathcal{X}$ such that $x + tu \in \mathcal{X}$ for all $t \geq 0$, we have*

$$\|u\|_{\nabla^2 f(x)} \leq -\langle u, \nabla f(x) \rangle.$$

Next, we show the definition of Minkowsky functions, which is used to define the shrunk decision domain similar to the clipped simplex in multi-armed bandit setting.

Minkowsky functions. The Minkowsky function of a convex body $\mathcal{X}$ with the pole at $w \in \operatorname{int}(\mathcal{X})$ is a function $\pi_w : \mathcal{X} \to \mathbb{R}$ defined as

$$\pi_w(u) = \inf \left\{ t > 0 \,\middle|\, w + \frac{u - w}{t} \in \mathcal{X} \right\}.$$

The last lemma shows several useful properties using the Minkowsky function.

Lemma 12 (Proposition 2.3.2 in Nesterov and Nemirovskii (1994)) *Let f be a ν-self-concordant barrier on $\mathcal{X} \subseteq \mathbb{R}^d$ and $u, w \in \operatorname{int}(\mathcal{X})$. Then for any $h \in \mathbb{R}^d$, we have*

$$\begin{aligned}
\|h\|_{\nabla^2 f(u)} &\leq \left(\frac{1 + 3\nu}{1 - \pi_w(u)} \right) \|h\|_{\nabla^2 f(w)}, \\
|\langle \nabla f(u), h \rangle| &\leq \left(\frac{\nu}{1 - \pi_w(u)} \right) \|h\|_{\nabla^2 f(w)}, \\
f(u) - f(w) &\leq \nu \ln \left(\frac{1}{1 - \pi_w(u)} \right).
\end{aligned}$$

Throughout this section, we assume that f is a ν-self-concordant barrier for $\mathcal{X}$.

Lemma 13 (Nesterov (2006), Theorem 1) *Let f be a self-concordant barrier for $\mathcal{X} \subset \mathbb{R}^d$. Then the function*

$$F(x, r) = \gamma(f(x/r) - 4\nu \ln(r)),$$

for $\gamma = \frac{(16\sqrt{\nu} + 7^{\frac{3}{2}})^2}{27 \cdot 4\nu}$ is a self-concordant barrier on the cone $\mathcal{X}_C := \{(x, r) \,|\, x/r \in \mathcal{X}\}$.

Lemma 14 *Let f be a self-concordant barrier for $\mathcal{X} \subset \mathbb{R}^d$. Then for any $u, x \in \mathcal{X}$,*

$$\|u - x\|_{\nabla^2 f(x)} \leq -\gamma' \langle u - x, \nabla f(x) \rangle + 4\gamma' \nu + 2\sqrt{\nu},$$

where $\gamma' = \frac{8}{3\sqrt{3}} + \frac{7^{\frac{3}{2}}}{6\sqrt{3\nu}}$ ($\gamma' \in [1, 4]$ for $\nu \geq 1$).

Proof Let F be the self-concordant function of Lemma 13. Note that any $(u, 1)$ is a recessive direction of $\mathcal{X}_C$ at $(x, 1)$ for any $u, x \in \mathcal{X}$. Hence we can apply Lemma 11 obtaining

$$\|(u, 1)\|_{\nabla^2 F((x,1))} \leq -\langle (u, 1), \nabla F((x, 1)) \rangle.$$

Computing the gradient and Hessian explicitly yields

$$\begin{aligned}
\nabla F(x, t) &= \gamma \begin{pmatrix} \frac{1}{t} \nabla f(\frac{x}{t}) \\ -\frac{1}{t^2} x^\top \nabla f(\frac{x}{t}) - \frac{4\nu}{t} \end{pmatrix} \\
\nabla^2 F(x, t) &= \gamma \begin{pmatrix} \frac{1}{t^2} \nabla^2 f(\frac{x}{t}) & -\frac{1}{t^2} \nabla f(\frac{x}{t}) - \frac{1}{t^3} \nabla^2 f(\frac{x}{t}) x \\ -\frac{1}{t^2} \nabla f(\frac{x}{t})^\top - \frac{1}{t^3} x^\top \nabla^2 f(\frac{x}{t}) & 2x^\top \nabla f(\frac{x}{t}) + \frac{1}{t^4} x^\top \nabla^2 f(\frac{x}{t}) x + \frac{4\nu}{t^2} \end{pmatrix}.
\end{aligned}$$

Hence we have

$$\|(u,1)\|_{\nabla^2 F((x,1))} = \sqrt{\gamma}\sqrt{\|u-x\|^2_{\nabla^2 f(x)} - 2\langle u-x, \nabla f(x)\rangle + 4\nu}$$
$$-\langle (u,1), \nabla F((x,1))\rangle = -\gamma\langle u-x, \nabla f(x)\rangle + 4\gamma\nu\,.$$

Combining these

$$\|u-x\|^2_{\nabla^2 f(x)} \leq \gamma(4\nu - \langle u-x, \nabla f(x)\rangle)^2 - 4\nu + 2\langle u-x, \nabla f(x)\rangle$$
$$= \gamma(4\nu - 1/\gamma - \langle u-x, \nabla f(x)\rangle)^2 - 1/\gamma^2 + 4\nu\,,$$

hence

$$\|u-x\|_{\nabla^2 f(x)} \leq -\sqrt{\gamma}\langle u-x, \nabla f(x)\rangle + 4\sqrt{\gamma}\nu + 2\sqrt{\nu}\,.$$

■

Lemma 15 *Let λ be the uniform measure on convex body $K \subset \mathbb{R}^d$, $d \geq 2$, and $\mu = \int x\lambda(dx)$ and $\Sigma = \int xx^\top \lambda(dx) - \mu\mu^\top$ be its covariance. Then for all $a \in K$,*

$$\|a-\mu\|^2_{\Sigma^{-1}} \leq (d + 2\sqrt{d})^2 \leq 6d^2$$

Proof If F is the entropic barrier on K, then $\nabla F(\mu) = 0$ and the result follows from Theorem 4.2.6 in the book by Nesterov (2004) and the fact that F is d-self-concordant and $\nabla^2 F(\mu) = \Sigma^{-1}$. ■

Appendix D. Stability of Log Determinant

The purpose of this section is to bound the stability term of follow the regularized leader for the negative log determinant potential function. Throughout we let D be the Bregman divergence with respect to $\boldsymbol{H} \mapsto -\log\det(\boldsymbol{H})$ and $\hat{\gamma}_t$ and other quantities be as defined by Algorithm 2.

Lemma 16 $\max_{\boldsymbol{H}\in\mathcal{H}} \langle \boldsymbol{H}_t - \boldsymbol{H}, \hat{\gamma}_t\rangle - \dfrac{D(\boldsymbol{H}, \boldsymbol{H}_t)}{\eta} \leq \dfrac{\eta}{4}\|a_t - x_t\|^2_{H_t^{-1}}$ *for all $t \in [T]$.*

Proof We start with an identity involving the Bregman divergence and then apply the standard local norm argument.

Step 1: An identity Suppose that $\boldsymbol{G}$ and $\boldsymbol{H}$ are matrices of the form

$$\boldsymbol{G} = \begin{pmatrix} G + gg^\top & g \\ g^\top & 1\,. \end{pmatrix} \qquad \boldsymbol{H} = \begin{pmatrix} H + hh^\top & h \\ h^\top & 1 \end{pmatrix},$$

where G and H are both invertible. By the definition of the Bregman divergence,

$$\begin{aligned}
D(\boldsymbol{G}, \boldsymbol{H}) &= F(\boldsymbol{G}) - F(\boldsymbol{H}) - \langle \nabla F(\boldsymbol{H}), \boldsymbol{G} - \boldsymbol{H} \rangle \\
&= \log\left(\frac{\det(\boldsymbol{H})}{\det(\boldsymbol{G})}\right) + \operatorname{Tr}\left(\boldsymbol{H}^{-1}(\boldsymbol{G} - \boldsymbol{H})\right) && \text{(Jacobi's formula)} \\
&= \log\left(\frac{\det(\boldsymbol{H})}{\det(\boldsymbol{G})}\right) + \operatorname{Tr}\left(\boldsymbol{H}^{-1}\boldsymbol{G}\right) - d - 1 \\
&= \log\left(\frac{\det(\boldsymbol{H})}{\det(\boldsymbol{G})}\right) + \operatorname{Tr}\left(H^{-1}G\right) + \|g - h\|^2_{H^{-1}} - d \\
&= \log\left(\frac{\det(H)}{\det(G)}\right) + \operatorname{Tr}\left(H^{-1}G\right) + \|g - h\|^2_{H^{-1}} - d \\
&= D(G, H) + \|g - h\|^2_{H^{-1}} \\
&\geq \|g - h\|^2_{H^{-1}} \ ,
\end{aligned}$$

where in the final equality we have abused notation by writing $D(G, H)$ as the Bregman divergence with respect to $-\log\det(\cdot)$ on the space of $d \times d$ matrices instead of $(d+1) \times (d+1)$ as in the lemma statement.

Step 2: Local norms and Cauchy-Schwarz Let $\boldsymbol{H} = \widehat{\operatorname{Cov}}(p) \in \mathcal{H}$ and $\boldsymbol{H}_t = \widehat{\operatorname{Cov}}(p_t)$. By the previous step,

$$\begin{aligned}
\langle \boldsymbol{H}_t - \boldsymbol{H}, \hat{\gamma}_t \rangle - \frac{D(\boldsymbol{H}, \boldsymbol{H}_t)}{\eta} &\leq \langle \boldsymbol{H}_t - \boldsymbol{H}, \hat{\gamma}_t \rangle - \frac{\|\mu(p) - x_t\|^2_{H_t^{-1}}}{\eta} \\
&= \langle x_t - \mu(p), \hat{y}_t \rangle - \frac{\|\mu(p) - x_t\|^2_{H_t^{-1}}}{\eta} \\
&\leq \|x_t - \mu(p)\|_{H_t^{-1}} \|\hat{y}_t\|_{H_t} - \frac{\|\mu(p) - x_t\|^2_{H_t^{-1}}}{\eta} \\
&\leq \frac{\eta}{4} \|\hat{y}_t\|^2_{H_t} \\
&\leq \frac{\eta}{4} \|a_t - x_t\|^2_{H_t^{-1}} \ ,
\end{aligned}$$

where in the final inequality we used the definition $\hat{y}_t = H_t^{-1}(a_t - x_t)\ell_t$ and the fact that the losses are in $[-1, 1]$. ■

Appendix E. Proof of Theorem 2

We prove a slightly more general result where η is replaced by η_t in the optimization problem and $(\eta_t)_{t=1}^T$ is non-increasing, which means that

$$x_t = \operatorname*{arg\,min}_{x \in \mathcal{K}} \left\langle x, \sum_{s=1}^{t-1} \hat{y}_s - \sum_{s=1}^{t} b_t \right\rangle + \frac{F(x)}{\eta_t} \, .$$

We adopt also the convention that $\eta_0 = \infty$ and $\eta_{T+1} = \eta_T$. Let $Y_t = \sum_{s=1}^t \hat{y}_s$ and $B_t = \sum_{s=1}^t b_s$ and

$$x_{T+1} = \arg\min_{x \in \mathcal{K}} \langle x, Y_T - B_T \rangle + \frac{F(x)}{\eta_{T+1}} \,.$$

Note that B_T not B_{T+1} appears in the objective here, in contrast to x_t for $1 \leq t \leq T$. By the first order optimality conditions for x_t we have for any $x \in \mathcal{K}$,

$$\frac{\langle \nabla F(x_t), x - x_t \rangle}{\eta_t} \geq \langle B_t - Y_{t-1}, x - x_t \rangle \,. \tag{8}$$

Our plan is to bound $\sum_{t=1}^T \langle x_t - u, \hat{y}_t - b_t \rangle$, which decomposes as

$$\sum_{t=1}^T \langle x_t - u, \hat{y}_t - b_t \rangle = \sum_{t=1}^T \langle x_t, \hat{y}_t - b_t \rangle - \langle u, Y_T - B_T \rangle \,. \tag{9}$$

The first term is bounded in a now-standard way:

$$\begin{aligned}
&\sum_{t=1}^T \langle x_t, \hat{y}_t - b_t \rangle = \sum_{t=1}^T \langle x_t - x_{t+1}, \hat{y}_t \rangle + \sum_{t=1}^T \left(\langle x_{t+1}, \hat{y}_t \rangle + \langle x_t, -b_t \rangle \right) \\
&= \sum_{t=1}^T \left(\langle x_t - x_{t+1}, \hat{y}_t \rangle - \frac{D(x_{t+1}, x_t)}{\eta_t} \right) + \sum_{t=1}^T \left(\langle x_{t+1}, \hat{y}_t \rangle + \langle x_t, -b_t \rangle + \frac{D(x_{t+1}, x_t)}{\eta_t} \right) \\
&\leq \sum_{t=1}^T \left(\max_{x \in \mathcal{K}} \langle x_t - x, \hat{y}_t \rangle - \frac{D(x, x_t)}{\eta_t} \right) + \underbrace{\sum_{t=1}^T \left(\langle x_{t+1}, \hat{y}_t \rangle + \langle x_t, -b_t \rangle + \frac{D(x_{t+1}, x_t)}{\eta_t} \right)}_{\text{A}} \,.
\end{aligned} \tag{10}$$

The first sum on the right-hand side now has the desired form. For the second sum we need to use the first order optimality conditions in (8) and the definition of the Bregman divergence so that

$$
\begin{aligned}
(\mathrm{A}) &= \sum_{t=1}^{T} \left(\langle x_{t+1}, \hat{y}_t \rangle + \langle x_t, -b_t \rangle + \frac{F(x_{t+1})}{\eta_t} - \frac{F(x_t)}{\eta_t} - \frac{\langle \nabla F(x_t), x_{t+1} - x_t \rangle}{\eta_t} \right) \\
&\leq \sum_{t=1}^{T} \left(\langle x_{t+1}, \hat{y}_t \rangle + \langle x_t, -b_t \rangle + \frac{F(x_{t+1})}{\eta_t} - \frac{F(x_t)}{\eta_t} + \langle B_t - Y_{t-1}, x_t - x_{t+1} \rangle \right) \qquad \text{(by (8))} \\
&= \sum_{t=1}^{T} \left(\frac{F(x_{t+1})}{\eta_t} - \frac{F(x_t)}{\eta_t} \right) + \langle x_{T+1}, Y_T - B_T \rangle \\
&\leq \sum_{t=1}^{T} \left(\frac{F(x_{t+1})}{\eta_t} - \frac{F(x_t)}{\eta_t} \right) + \frac{F(u)}{\eta_{T+1}} - \frac{F(x_{T+1})}{\eta_{T+1}} + \langle u, Y_T - B_T \rangle \\
&= \sum_{t=1}^{T+1} (F(u) - F(x_t)) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) + \langle u, Y_T - B_T \rangle \\
&\leq \sum_{t=1}^{T+1} (F(u) - F(\tilde{x}_1)) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) + \langle u, Y_T - B_T \rangle \\
&= \frac{F(u) - F(\tilde{x}_1)}{\eta_T} + \langle u, Y_T - B_T \rangle \,, \qquad (11)
\end{aligned}
$$

where in the first inequality we used the fact that x_{T+1} minimizes $x \mapsto \langle x, Y_T - B_T \rangle + F(x)/\eta_{T+1}$ on $\mathcal{K}$ and in the second we used the fact that $\tilde{x}_1$ minimizes F on $\mathcal{K}$ as well as the assumption that the learning rates are non-increasing. By combining (9), (10) and (11) we obtain

$$
\sum_{t=1}^{T} \langle x_t - u, \hat{y}_t - b_t \rangle \leq \sum_{t=1}^{T} \left(\max_{x \in \mathcal{K}} \langle x_t - x, \hat{y}_t \rangle - \frac{D(x, x_t)}{\eta_t} \right) + \frac{F(u) - F(\tilde{x}_1)}{\eta_T} \,.
$$

Rearranging completes the result.

Appendix F. Proof of Theorem 4

We begin by defining a few constants used in the proofs in this section and by fixing the tuning parameters of the algorithm:

$$\gamma' = \frac{8}{3\sqrt{3}} + \frac{7^{\frac{3}{2}}}{6\sqrt{3d}} \in [1,4]$$
$$\iota = \log\left(\frac{48d^2T^2}{\delta}\right) + d\log(6\sqrt{T}) = \mathcal{O}(d\log(T) + \log(1/\delta))$$
$$\lambda = \min\left\{1 - (1/2)^{\frac{2}{3}}, d\sqrt{\frac{\iota}{T}}\right\}$$
$$\gamma = \frac{9\gamma'\sqrt{\iota}}{(1-\lambda)^{\frac{3}{2}}} \leq 72\sqrt{\iota}$$
$$c_t = \sqrt{\mathrm{Tr}\left(\boldsymbol{G}_t^{-1}\left(\sum_{s=1}^{t}\boldsymbol{G}_s^{-1}\right)^{-1}\right)}$$
$$\eta = \min\left\{\frac{1}{16\left(\sqrt{\frac{24}{\lambda}}d + 72d\sqrt{\iota}\right)}, \frac{\log(T)}{\sqrt{T\iota}}\right\}.$$

Recall $b_t = \gamma c_t \nabla F(x_t)$, the agent samples actions from $p_t = (1-\lambda)p_{\theta_t} + \lambda p_0$, where p_0 is the uniform probability measure on $\mathcal{X}$ and p_{θ_t} is the exponential weights distribution with mean x_t and $\theta_t = \nabla F(x_t)$. G_t is the covariance of the p_t, x'_t its mean and $\hat{y}_t = G_t^{-1}(a_t - x'_t)\ell_t$ the loss estimator. $H_t = \nabla^2 F(x_t)^{-1}$ is the covariance of p_{θ_t} and

$$\boldsymbol{G}_t = \begin{pmatrix} G_t + x'_t {x'_t}^\top & x'_t \\ {x'_t}^\top & 1 \end{pmatrix}$$

is the lifted version of covariance matrix G_t.

Basic bounds Recall that H_0 is the covariance of p_0 and that $p_t = (1-\lambda)p_{\theta_t} + \lambda p_0$. Using this with Lemma 15 and the triangle inequality,

$$\|\hat{y}_t\|_{H_t}^2 = \left\|a_t - x'_t\right\|_{G_t^{-1}H_tG_t^{-1}}^2 \leq \frac{\|a_t - x'_t\|_{G_t^{-1}}^2}{1-\lambda} \leq \frac{\|a_t - x'_t\|_{H_0^{-1}}^2}{\lambda(1-\lambda)} \leq \frac{24d^2}{\lambda(1-\lambda)}.$$

Furthermore, for $u, v \in \mathcal{X}$,

$$\begin{aligned}
\eta|\langle u - v, \hat{y}_t\rangle| &= \eta|\langle u - v, G_t^{-1}(a_t - x'_t)\ell_t\rangle| \\
&\leq \eta \left\|u - v\right\|_{G_t^{-1}} \left\|a_t - x'_t\right\|_{G_t^{-1}} \\
&\leq \frac{\eta \left\|u - v\right\|_{H_0^{-1}} \left\|a_t - x'_t\right\|_{H_0^{-1}}}{\lambda} \\
&\leq \frac{24d^2\eta}{\lambda},
\end{aligned}$$

where we used the definition of $\hat{y}_t$, Cauchy-Schwarz and Lemma 15 again. Note that $c_t \leq \sqrt{d}$ and by the definition of ν-self-concordance and the fact that the entropic barrier is d-self-concordant,

$$c_t \left\|\nabla F(x_t)\right\|_{\nabla^2 F(x_t)^{-1}} \leq d\,.$$

Step 1: Decomposing the regret Let $u = u^* + \frac{1}{T}(x_0 - u^*)$, where x_0 is the mean of p_0 (the centroid of $\mathcal{X}$) and $u^* = \arg\min_{x\in\mathcal{C}} \sum_{t=1}^T \langle x, y_t\rangle$ and $\mathcal{C}$ is a $1/\sqrt{T}$-covering of $\mathcal{X}$ with respect to the norm $\|\cdot\|_{\mathcal{X}^\circ}$, which is sufficient to bound the regret up to $\sqrt{T}$. See the proof of Theorem 5 for details.

$$\begin{aligned}
\mathrm{Reg}_T &= \sum_{t=1}^T \langle x'_t - u^*, y_t\rangle \\
&\leq 2 + \sum_{t=1}^T \langle x'_t - x_t, y_t\rangle + \sum_{t=1}^T \langle x_t - u, y_t - \hat{y}_t - \gamma c_t \nabla F(x_t)\rangle + \sum_{t=1}^T \langle x_t - u, \hat{y}_t + \gamma c_t \nabla F(x_t)\rangle \\
&\leq 2 + \lambda T + \sum_{t=1}^T \langle x_t - u, y_t - \hat{y}_t - \gamma c_t \nabla F(x_t)\rangle + \sum_{t=1}^T \langle x_t - u, \hat{y}_t + \gamma c_t \nabla F(x_t)\rangle\,. \qquad (12)
\end{aligned}$$

Step 2: Bounding the empirical regret The first sum in (12) is the deviation term, which we control in a moment. The second sum is bounded using the standard FTRL analysis. We can apply Lemma 17 due to the following computation:

$$\begin{aligned}
\left\|\hat{y}_t + \gamma c_t \nabla F(x_t)\right\|_{\nabla^2 F(x_t)^{-1}} &\leq \left\|\hat{y}_t\right\|_{\nabla^2 F(x_t)^{-1}} + \gamma c_t \left\|\nabla F(x_t)\right\|_{\nabla^2 F(x_t)^{-1}} \\
&\leq \sqrt{\frac{24 d^2}{\lambda(1-\lambda)}} + \gamma d && \text{(by basic bounds)} \\
&\leq \sqrt{\frac{48}{\lambda}} d + 72 d\sqrt{\iota} \leq \frac{1}{16\eta}\,. && \text{(by constraint on } \eta)
\end{aligned}$$

Hence, by Theorem 1 and Lemma 17,

$$\begin{aligned}
\widehat{\mathrm{Reg}}_T(u) &\triangleq \sum_{t=1}^T \langle x_t - u, \hat{y}_t + \gamma c_t \nabla F(x_t)\rangle \\
&\leq \frac{F(u) - \min_{x\in\mathcal{X}} F(x)}{\eta} + \sum_{t=1}^T \max_{x\in\mathcal{X}} \left[\langle x - x_t, \hat{y}_t + \gamma c_t \nabla F(x_t)\rangle - \frac{1}{\eta} D(x, x_t)\right] \\
&\leq \frac{d\log(T)}{\eta} + 4\eta \sum_{t=1}^T \left(\left\|\hat{y}_t\right\|_{H_t}^2 + \gamma^2 c_t^2 d\right) && \text{(Lemmas 12 and 17)} \\
&\leq \frac{d\log(T)}{\eta} + 4\cdot 72^2 \eta d(d+1)\iota \log\left(\frac{eT}{\lambda}\right) + 4\eta \sum_{t=1}^T \left\|\hat{y}_t\right\|_{H_t}^2\,, && \text{(Lemma 18)}
\end{aligned}$$

where in the second last inequality we also used the fact that F is d-self-concordant so that $\left\|\nabla F(x_t)\right\|^2_{\nabla^2 F(x_t)^{-1}} \leq d$.

Step 3: Concentration Note that $\mathbb{E}_{t-1}[\|\hat{y}_t\|_{H_t}^2] \leq \mathrm{Tr}(G_t^{-1} H_t) \leq \frac{d}{1-\lambda}$ and by the basic bounds,

$$\eta \|\hat{y}_t\|_{H_t}^2 \leq \frac{24 d^2 \eta}{\lambda(1-\lambda)} .$$

Given $\xi X_t \in [0,1]$ for some $\xi > 0$, a sequence of non-negative random variables X_t and rearranging the second part of Exercise 5.15 by Lattimore and Szepesvári (2020), we have w.h.p.

$$\begin{aligned}\sum_{t=1}^T X_t &\leq \sum_{t=1}^T E_{t-1}[X_t] + \xi \sum_{t=1}^T E_{t-1}[X_t^2] + \xi^{-1} \log(\delta'^{-1}) \\ &\leq 2 \sum_{t=1}^T E_{t-1}[X_t] + \xi^{-1} \log(\delta'^{-1}) .\end{aligned}$$

Taking probability at least $1 - \delta/3$, setting $\xi = \lambda(1-\lambda)/(24d^2)$ and $X_t = \eta \|\hat{y}_t\|_{H_t}^2$ yields

$$\sum_{t=1}^T \eta \|\hat{y}_t\|_{H_t}^2 \leq 2 \frac{\eta d T}{1-\lambda} + \frac{24 d^2}{\lambda(1-\lambda)} \log\left(\frac{3}{\delta}\right) .$$

Also by the basic bounds $|\langle x_t - u, \hat{y}_t \rangle| \leq 24 d^2 / \lambda$ and

$$\mathbb{E}_{t-1}[\langle x_t - u, \hat{y}_t \rangle^2] \leq \|x_t - u\|_{G_t^{-1}}^2 \leq \frac{24 d^2}{\lambda} .$$

Therefore by Theorem 9 with probability at least $1 - \delta/3$ for all u in a covering set $\mathcal{C}$ of size $\log|\mathcal{C}| \leq d \log(6\sqrt{T})$ simultaneously,

$$\begin{aligned}\sum_{t=1}^T \langle x_t - u, y_t - \hat{y}_t \rangle &\leq 3 \sqrt{\sum_{t=1}^T \frac{\|x_t - u\|_{G_t^{-1}}^2}{1-\lambda} \iota} + \frac{48 d^2}{\lambda} \iota \\ &\leq 3 \sqrt{\sum_{t=1}^T \frac{\|\boldsymbol{u}\|_{\boldsymbol{G}_t^{-1}}^2 + 24 d^2}{1-\lambda} \iota} + \frac{48 d^2}{\lambda} \iota .\end{aligned}$$

Step 4: Controlling the bias The bias component of the deviation term is bounded using Lemma 14:

$$\begin{aligned}\sum_{t=1}^T \langle u - x_t, \gamma c_t \nabla F(x_t) \rangle &\leq -\frac{\gamma}{\gamma'} \sum_{t=1}^T c_t \|u - x_t\|_{H_t^{-1}} + \gamma \left(4d + \frac{2\sqrt{d}}{\gamma'}\right) \sum_{t=1}^T c_t \\ &= -\frac{\gamma}{\gamma'} \sum_{t=1}^T c_t \|\boldsymbol{u}\|_{\boldsymbol{H}_t^{-1}} + \gamma \left(4d + \frac{2\sqrt{d}+1}{\gamma'}\right) \sum_{t=1}^T c_t .\end{aligned}$$

The positive term is bounded by

$$\begin{aligned}\gamma \left(4d + \frac{2\sqrt{d}+1}{\gamma'}\right) \sum_{t=1}^T c_t &\leq \gamma \left(4d + \frac{2\sqrt{d}+1}{\gamma'}\right) \sqrt{(d+1) T \log\left(\frac{eT}{\lambda}\right)} \qquad \text{(Lemma 18)} \\ &= \mathcal{O}(d^{\frac{3}{2}} \sqrt{T \log(T) \iota}) .\end{aligned}$$

The negative term is

$$
\begin{aligned}
-\frac{\gamma}{\gamma'}\sum_{t=1}^{T} c_t \left\|\boldsymbol{u}\right\|_{\boldsymbol{H}_t^{-1}} &\le -\frac{(1-\lambda)\gamma}{\gamma'}\sum_{t=1}^{T} c_t \left\|\boldsymbol{u}\right\|_{\boldsymbol{G}_t^{-1}} \\
&\le -9\sqrt{\frac{\iota}{1-\lambda}}\sum_{t=1}^{T}\frac{\|\boldsymbol{u}\|^2_{\boldsymbol{G}_t^{-1}}}{\sqrt{\sum_{s=1}^{t}\|\boldsymbol{u}\|^2_{\boldsymbol{G}_s^{-1}}}} \\
&\le -3\sqrt{\sum_{t=1}^{T}\frac{\|\boldsymbol{u}\|^2_{\boldsymbol{G}_t^{-1}}}{1-\lambda}\iota}\,. \quad \text{(by the max ratio } \|\boldsymbol{u}\|_{\boldsymbol{G}_{t+1}^{-1}} / \|\boldsymbol{u}\|_{\boldsymbol{G}_t^{-1}} \text{ of Lemma 19)}
\end{aligned}
$$

Step 5: Combining everything We have

$$
\mathrm{Reg}_T = \mathcal{O}\left(\lambda T + \frac{d\log(T)}{\eta} + \eta d T + \frac{d^2\iota}{\lambda} + d^{\frac{3}{2}}\sqrt{T\log(T)\iota}\right).
$$

By the choice of λ and η

$$
\mathrm{Reg}_T = \mathcal{O}\left(d^{\frac{3}{2}}\sqrt{T\log(T)}\left(\sqrt{d\log(T)+\log\left(\frac{1}{\delta}\right)}\right)\right).
$$

Appendix G. Proof of Theorem 6

Recall that p_0 is a distribution on $\mathcal{A}$ such that for all $x \in \mathcal{A}$,

$$
\|x\|^2_{H_0^{-1}} \le d\,,
$$

where $H_0 = \sum_{a\in\mathcal{A}} p_0(a) a a^\top$ is the design matrix of p_0. Exponential weights samples a_t from distribution $p'_t = \lambda p_0 + (1-\lambda)p_t$, where

$$
p_t(a) = \frac{\exp\left(-\eta\sum_{s=1}^{t-1}(\hat{y}_s(a) - b_s(a))\right)}{\sum_{b\in\mathcal{A}}\exp\left(-\eta\sum_{s=1}^{t-1}(\hat{y}_s(a) - b_s(a))\right)}\,.
$$

where $\hat{y}_t(a) = a^\top G_t^{-1} a_t \ell_t$ with $G_t = \sum_{a\in\mathcal{A}} p'_t(a) a a^\top$ and $b_t(a) = \eta\|a\|^2_{G_t^{-1}}$. For the sake of consistent notation let $y_t(a) = \langle a, y_t\rangle$.

Step 1: Basic Bounds We start by providing elementary uniform bounds on the loss estimators and the bias terms.

$$
\eta|\hat{y}_t(a)| = \eta|a^\top G_t^{-1} a_t \ell_t| \le \eta\|a\|_{G_t^{-1}}\|a_t\|_{G_t^{-1}} \le \frac{\eta}{\lambda}\|a\|_{H_0^{-1}}\|a_t\|_{H_0^{-1}} \le \frac{\eta d}{\lambda}\,.
$$

Similarly,

$$
|b_t(a)| = \eta\|a\|^2_{G_t^{-1}} \le \frac{\eta d}{\lambda}\,.
$$

Based on this, if $\lambda = 2\eta d$, then $\eta|\hat{y}_t(a) - b_t(a)| \le 1$.

Step 2: Bounding the regret Using the boundedness of $\eta|\hat{y}_t(a)-b_t(a)|$ and applying the standard analysis of exponential weights and a decomposition of the regret yields

$$\begin{aligned}
\mathrm{Reg}_T &= \sum_{t=1}^{T}\left(\sum_{a\in\mathcal{A}} p'_t(a)y_t(a) - y_t(a^*)\right) \\
&\le \lambda T + \sum_{t=1}^{T}\left(\sum_{a\in\mathcal{A}} p_t(a)y_t(a) - \langle a^*, y_t\rangle\right) \\
&\le \lambda T + \sum_{t=1}^{T}\sum_{a\in\mathcal{A}} p_t(a)\left(y_t(a) - \hat{y}_t(a) + \hat{y}_t(a^*) - y_t(a^*) + b_t(a) - b_t(a^*)\right) \\
&\qquad + \frac{\log(k)}{\eta} + \eta\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} p_t(a)(\hat{y}_t(a) - b_t(a))^2\,.
\end{aligned}$$

The last term (the stability) is bounded by

$$\begin{aligned}
\eta\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} p_t(a)(\hat{y}_t(a) - b_t(a))^2 &\le 2\eta\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} p_t(a)(\hat{y}_t(a)^2 + b_t(a)^2) \\
&\le \frac{2\eta^2 dT}{\lambda} + \frac{2\eta}{1-\lambda}\sum_{t=1}^{T}\|a_t\|^2_{G_t^{-1}}\,,
\end{aligned}$$

where we used $(x+y)^2 \le 2x^2 + 2y^2$ and $|b_t(a)| \le 1$ for all a and

$$\sum_{a\in\mathcal{A}} p_t(a)\hat{y}_t(a)^2 = a^\top G_t^{-1}\sum_{a\in\mathcal{A}} p_t(a)aa^\top G_t^{-1}a \le \frac{\|a\|^2_{G_t^{-1}}}{1-\lambda}\,.$$

We also have

$$\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} p_t(a)b_t(a) = 2\eta\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} p_t(a)\,\|a\|^2_{G_t^{-1}} \le \frac{2\eta dT}{1-\lambda}\,.$$

Step 3: Concentration Since $\mathbb{E}_{t-1}[\hat{y}_t(a)] = y_t(a)$, with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} p_t(a)(y_t(a) - \hat{y}_t(a)) &\le \eta\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} p_t(a)\mathbb{E}_{t-1}[\hat{y}_t(a)^2] + \frac{1}{\eta}\log\left(\frac{1}{\delta}\right) \\
&\le \frac{d\eta T}{1-\lambda} + \frac{1}{\eta}\log\left(\frac{1}{\delta}\right)\,.
\end{aligned}$$

Similarly, with probability $1-\delta$,

$$\begin{aligned}
\sum_{t=1}^{T}(\hat{y}_t(a^*) - y_t(a^*)) &\le \eta\sum_{t=1}^{T}\mathbb{E}_{t-1}[\hat{y}_t(a^*)^2] + \frac{1}{\eta}\log\left(\frac{1}{\delta}\right) \\
&\le \eta\sum_{t=1}^{T}\|a^*\|^2_{G_t^{-1}} + \frac{1}{\eta}\log\left(\frac{1}{\delta}\right)\,.
\end{aligned}$$

Lastly, with probability $1-\delta$, using the fact that $\mathbb{E}_{t-1}[\|a_t\|^2_{G_t^{-1}}] = d$ and $\eta \|a_t\|^2_{G_t^{-1}} \le 1$,

$$\frac{2\eta}{1-\lambda}\sum_{t=1}^{T}\|a_t\|^2_{G_t^{-1}} \le \frac{2\eta(d+1)T}{1-\lambda}\,.$$

Step 4: Combining By collecting all the pieces and taking a union bound over all $a^* \in \mathcal{A}$, with probability at least $1-\delta$,

$$\mathrm{Reg}_T \le \lambda T + 6\eta(d+1)T + \frac{3}{\eta}\log\left(\frac{3k}{\delta}\right) + \frac{\log(k)}{\eta}\,.$$

The result follows by choosing $\lambda = 2\eta d$ and $\eta = \sqrt{\frac{\log(k/\delta)}{dT}}$.

Appendix H. Support Lemmas for Entropic Barrier

In this section present a simple lemma for bounding the stability of FTRL with the entropic barrier and crucial properties of the bias factors.

Lemma 17 *Suppose that* $\|w\|_{\nabla^2 F(x_t)^{-1}} \le 1/(16\eta)$*, then*

$$\max_{x\in\mathcal{X}}\langle x_t - x, w\rangle - \frac{D_F(x,x_t)}{\eta} \le 2\eta\|w\|^2_{\nabla^2 F(x_t)^{-1}}\,.$$

Additionally, for $x_{t+1} = \arg\max_{x\in\mathcal{X}}\langle x_t - x, w\rangle - \frac{D_F(x,x_t)}{\eta}$*, we have* $\nabla^2 F(x_{t+1}) \succeq \frac{1}{4}\nabla^2 F(x_t)$

Proof Define $\varphi : \mathcal{X} \to \mathbb{R}$ by

$$\varphi(x) = \langle x_t - x, \hat{y}_t\rangle - \frac{D(x,x_t)}{\eta}\,.$$

We start by showing that $\varphi(x) \le 0$ for all $x \in \mathcal{X}$ with $\|x - x_t\|_{H_t^{-1}} = 1/2$. By Taylor's theorem, there exists ξ on the chord connecting x and x_t such that $D(x,x_t) = \frac{1}{2}\|x-x_t\|^2_{\nabla^2 F(\xi)}$. Since $H_t^{-1} = \nabla^2 F(x_t)$, our assumption that $\|x - x_t\|_{H_t^{-1}} = 1/2$ means that x is in the Dikin ellipsoid of F centered at x_t. Hence, by Lemma 10,

$$\begin{aligned}
D(x,x_t) &= \frac{1}{2}\|x-x_t\|^2_{\nabla^2 F(\xi)} \\
&\ge \frac{1}{2}\|x-x_t\|^2_{\nabla^2 F(x_t)}\left(1 - \|\xi - x_t\|_{\nabla^2 F(x_t)}\right)^2 \\
&\ge \frac{1}{8}\|x-x_t\|^2_{\nabla^2 F(x_t)} = \frac{1}{32},
\end{aligned}$$

where we use the identity $H_t^{-1} = \nabla^2 F(x_t)$. Combining this with Cauchy-Schwarz shows that

$$\begin{aligned}
\varphi(x) &= \langle x_t - x, w\rangle - \frac{D(x,x_t)}{\eta} \\
&\le \|x - x_t\|_{\nabla^2 F(x_t)}\|w\|_{\nabla^2 F(x_t)^{-1}} - \frac{1}{32\eta} \\
&= \frac{1}{2}\|w\|_{\nabla^2 F(x_t)^{-1}} - \frac{1}{32\eta} \\
&\le 0\,.
\end{aligned}$$

Therefore $\varphi(x) \le 0$ for all x with $\|x - x_t\|_{\nabla^2 F(x_t)} = 1/2$. Note that φ is concave and $\varphi(x_t) = 0$. Hence, $\varphi(x) \le 0$ for all x with $\|x - x_t\|_{\nabla^2 F(x_t)} \ge 1/2$. Suppose now that $\|x - x_t\|_{\nabla^2 F(x_t)} \le 1/2$. By repeating the argument above,

$$\begin{aligned}
\varphi(x) &= \langle x_t - x, w\rangle - \frac{D(x, x_t)}{\eta} \\
&\le \|x - x_t\|_{\nabla^2 F(x_t)} \|w\|_{\nabla^2 F(x_t)^{-1}} - \frac{1}{8\eta}\|x - x_t\|^2_{\nabla^2 F(x_t)} \\
&\le 2\eta \|w\|^2_{\nabla^2 F(x_t)^{-1}} \ .
\end{aligned}$$

Combining the cases completes the result. The second part follows by observing that x_{t+1} satisfies $\|x_{t+1} - x_t\|_{\nabla^2 F(x_t)} \ge 1/2$, as we have shown above. ■

Lemma 18 *Let*

$$c_t := \sqrt{\mathrm{Tr}\left(\boldsymbol{G}_t^{-1}\left(\sum_{s=1}^{t} \boldsymbol{G}_s^{-1}\right)^{-1}\right)} .$$

Then it holds that

$$\begin{aligned}
\sum_{t=1}^{T} c_t^2 &\le (d+1)\log\left(\frac{eT}{\lambda}\right) \\
\sum_{t=1}^{T} c_t &\le \sqrt{(d+1)T\log\left(\frac{eT}{\lambda}\right)} \\
c_t^2 &\ge \frac{\|\boldsymbol{u}\|^2_{\boldsymbol{G}_t^{-1}}}{\sum_{s=1}^{t}\|\boldsymbol{u}\|^2_{\boldsymbol{G}_s^{-1}}} .
\end{aligned}$$

Proof Due to convexity, we have

$$c_t^2 = \mathrm{Tr}\left(\boldsymbol{G}_t^{-1}\left(\sum_{s=1}^{t} \boldsymbol{G}_s^{-1}\right)^{-1}\right) \le \log\det\left(\sum_{s=1}^{t} \boldsymbol{G}_s^{-1}\right) - \log\det\left(\sum_{s=1}^{t-1} \boldsymbol{G}_s^{-1}\right) ,$$

hence

$$\sum_{t=1}^{T} c_t^2 \le d + 1 + \log\det\left(\sum_{s=1}^{T} \boldsymbol{G}_s^{-1}\right) - \log\det\left(\boldsymbol{G}_1\right) .$$

The distribution generating G_1 is the uniform sampling distribution, hence $\boldsymbol{G}_s^{-1} \preceq \frac{1}{\lambda}\boldsymbol{G}_1^{-1}$.

$$\sum_{t=1}^{T} c_t^2 \le (d+1)\log\left(\frac{eT}{\lambda}\right) .$$

The second statement follows by Cauchy-Schwarz:

$$\sum_{t=1}^{T} c_t \leq \sqrt{T \sum_{t=1}^{T} c_t^2}\,.$$

Finally, the last equation follows by

$$\begin{aligned}
\frac{\|\boldsymbol{u}\|^2_{\boldsymbol{G}_t^{-1}}}{\sum_{s=1}^t \|\boldsymbol{u}\|^2_{\boldsymbol{G}_s^{-1}}} &\leq \max_{u' \in \mathbb{R}^d} \frac{\|\boldsymbol{u}'\|^2_{\boldsymbol{G}_t^{-1}}}{\sum_{s=1}^t \|\boldsymbol{u}'\|^2_{\boldsymbol{G}_s^{-1}}} = \max_{\boldsymbol{u}' \in \mathbb{R}^d} \frac{\left\|(\sum_{s=1}^t \boldsymbol{G}_s^{-1})^{-1/2}\boldsymbol{u}'\right\|^2_{\boldsymbol{G}_t^{-1}}}{\|u'\|^2} \\
&= \lambda_{\max}\left(\left(\sum_{s=1}^t \boldsymbol{G}_s^{-1}\right)^{-1/2} \boldsymbol{G}_t^{-1} \left(\sum_{s=1}^t \boldsymbol{G}_s^{-1}\right)^{-1/2}\right) \\
&\leq \mathrm{Tr}\left(\left(\sum_{s=1}^t \boldsymbol{G}_s^{-1}\right)^{-1} \boldsymbol{G}_t^{-1}\right)\,.
\end{aligned}$$

■

Lemma 19 *For any $u \in \mathcal{X}$ it holds $\|u - x_t\|^2_{G_t^{-1}} \leq \|\boldsymbol{u}\|^2_{\boldsymbol{G}_t^{-1}} + 24d^2$, if further x_{t+1} is in the $\frac{1}{2}$-Dikin ellipsoid, then $\frac{\|\boldsymbol{u}\|^2_{\boldsymbol{G}_{t+1}^{-1}}}{\|\boldsymbol{u}\|^2_{\boldsymbol{G}_t^{-1}}} \leq 8$.*

Proof The first part follows from.

$$\begin{aligned}
\|u - x_t\|^2_{G_t^{-1}} &= \left\|u - x_t'\right\|^2_{G_t^{-1}} + 2\lambda \left\langle \frac{x_t + x_t'}{2} - u, G_t^{-1}(x_t - x_0)\right\rangle \\
&\leq \|\boldsymbol{u}\|^2_{\boldsymbol{G}_t^{-1}} + 2\lambda \left\|\frac{x_t + x_t'}{2} - u\right\|_{G_t^{-1}} \|x_t - x_0\|_{G_t^{-1}} \\
&\leq \|\boldsymbol{u}\|^2_{\boldsymbol{G}_t^{-1}} + 2 \left\|\frac{x_t + x_t'}{2} - u\right\|_{H_0^{-1}} \|x_t - x_0\|_{H_0^{-1}} \\
&\leq \|\boldsymbol{u}\|^2_{\boldsymbol{G}_t^{-1}} + 24d^2\,. \qquad \text{(Lemma 15)}
\end{aligned}$$

For the second part, observe that because x_{t+1} is in the $\frac{1}{2}$-Dikin ellipsoid and by Lemma 10 we have $H_{t+1}^{-1} \preceq 4H_t^{-1}$.

For any $u \in \mathbb{R}^d$, it holds that

$$
\begin{aligned}
\|\boldsymbol{u}\|_{\boldsymbol{G}_t}^2 &= \|u\|_{G_t}^2 + (1 + \langle u, x_t' \rangle)^2 \\
&= (1-\lambda)\|u\|_{H_t}^2 + \lambda\|u\|_{H_0}^2 + \lambda(1-\lambda)\langle u, x_t - x_0\rangle^2 + (1 + \langle u, x_t'\rangle)^2 \\
&\le (1-\lambda)\|u\|_{H_t}^2 + \lambda\|u\|_{H_0}^2 + \lambda(1-\lambda)\langle u, x_{t+1} - x_0\rangle^2 + (1 + \langle u, x_{t+1}'\rangle)^2 \\
&\quad + \lambda(1-\lambda)\langle u, x_t - x_{t+1}\rangle\langle u, x_t + x_{t+1} - 2x_0\rangle + 2\langle u, x_t' - x_{t+1}'\rangle + \langle u, x_t' - x_{t+1}'\rangle\langle u, x_t' + x_{t+1}'\rangle \\
&= (1-\lambda)\|u\|_{H_t}^2 + \lambda\|u\|_{H_0}^2 + \lambda(1-\lambda)\langle u, x_{t+1} - x_0\rangle^2 + (1 + \langle u, x_{t+1}'\rangle)^2 \\
&\quad + (1-\lambda)\langle u, x_t - x_{t+1}\rangle(2 + \langle u, \lambda(x_t + x_{t+1} - 2x_0) + x_t' + x_{t+1}'\rangle) \\
&= (1-\lambda)\|u\|_{H_t}^2 + \lambda\|u\|_{H_0}^2 + \lambda(1-\lambda)\langle u, x_{t+1} - x_0\rangle^2 + (1 + \langle u, x_{t+1}'\rangle)^2 \\
&\quad + (1-\lambda)\langle u, x_t - x_{t+1}\rangle(2 + \langle u, x_t + x_{t+1}\rangle) \\
&= (1-\lambda)\|u\|_{H_t}^2 + \lambda\|u\|_{H_0}^2 + \lambda(1-\lambda)\langle u, x_{t+1} - x_0\rangle^2 + (1 + \langle u, x_{t+1}'\rangle)^2 \\
&\quad + 2(1-\lambda)\langle u, x_t - x_{t+1}\rangle(1 + \langle u, x_{t+1}\rangle) + (1-\lambda)\langle u, x_t - x_{t+1}\rangle^2 \\
&\le \frac{5(1-\lambda)}{4}\|u\|_{H_t}^2 + \lambda\|u\|_{H_0}^2 + \lambda(1-\lambda)\langle u, x_{t+1} - x_0\rangle^2 + (1 + \langle u, x_{t+1}'\rangle)^2 \\
&\quad + 2(1-\lambda)\langle u, x_t - x_{t+1}\rangle(1 + \langle u, x_{t+1}'\rangle) + 2\lambda(1-\lambda)\langle u, x_t - x_{t+1}\rangle\langle u, x_{t+1} - x_0\rangle \\
&\le \frac{5(1-\lambda)}{4}\|u\|_{H_t}^2 + \lambda\|u\|_{H_0}^2 + \lambda(1-\lambda)\langle u, x_{t+1} - x_0\rangle^2 + (1 + \langle u, x_{t+1}'\rangle)^2 \\
&\quad + (1-\lambda)(\langle u, x_t - x_{t+1}\rangle^2 + (1 + \langle u, x_{t+1}'\rangle)^2) + \lambda(1-\lambda)(\langle u, x_t - x_{t+1}\rangle^2 + \langle u, x_{t+1} - x_0\rangle^2) \\
&= \frac{(6+\lambda)(1-\lambda)}{4}\|u\|_{H_t}^2 + \lambda\|u\|_{H_0}^2 + 2\lambda(1-\lambda)\langle u, x_{t+1} - x_0\rangle^2 + (3 - 2\lambda)(1 + \langle u, x_{t+1}'\rangle)^2 \\
&\le (6+\lambda)(1-\lambda)\|u\|_{H_{t+1}}^2 + \lambda\|u\|_{H_0}^2 + 2\lambda(1-\lambda)\langle u, x_{t+1} - x_0\rangle^2 + (3 - 2\lambda)(1 + \langle u, x_{t+1}'\rangle)^2 \\
&\le 8\|\boldsymbol{u}\|_{\boldsymbol{G}_{t+1}}^2 \ .
\end{aligned}
$$

We have shown $\boldsymbol{G}_t \preceq *\boldsymbol{G}_{t+1}$, which directly implies $\boldsymbol{G}_{t+1}^{-1} \preceq *\boldsymbol{G}_t^{-1}$. ■