Sparsity, variance and curvature in multi-armed bandits

Sébastien Bubeck SEBUBECK@MICROSOFT.COM
Microsoft Research

Michael B. Cohen
MIT

Yuanzhi Li YUANZHIL@PRINCETON.EDU
Princeton University

Editor: Mehryar Mohri and Karthik Sridharan

Abstract

In (online) learning theory the concepts of sparsity, variance and curvature are well-understood and are routinely used to obtain refined regret and generalization bounds. In this paper we further our understanding of these concepts in the more challenging limited feedback scenario. We consider the adversarial multi-armed bandit and linear bandit settings and solve several open problems pertaining to the existence of algorithms with favorable regret bounds under the following assumptions: (i) sparsity of the individual losses, (ii) small variation of the loss sequence, and (iii) curvature of the action set. Specifically we show that (i) for s-sparse losses one can obtain $\widetilde{O}(\sqrt{sT})$-regret (solving an open problem by Kwon and Perchet), (ii) for loss sequences with variation bounded by Q one can obtain $\widetilde{O}(\sqrt{Q})$-regret (solving an open problem by Kale and Hazan), and (iii) for linear bandit on an ℓ_p^n ball one can obtain $\widetilde{O}(\sqrt{nT})$-regret for $p \in [1, 2]$ and one has $\widetilde{\Omega}(n\sqrt{T})$-regret for $p > 2$ (solving an open problem by Bubeck, Cesa-Bianchi and Kakade). A key new insight to obtain these results is to use regularizers satisfying more refined conditions than general self-concordance.

1. Introduction

In this paper we resolve several open problems in multi-armed bandit theory. Let us first recall the general setting of bandit linear optimization on a compact set $\mathcal{K} \subset \mathbb{R}^n$ (the classical multi-armed bandit problem corresponds to $\mathcal{K} = \{e_1, \dots, e_n\}$, the canonical basis in $\mathbb{R}^n$). It can be described as the following sequential game: at each time step $t = 1, \dots, T$, a player selects an action $a_t \in \mathcal{K}$, and simultaneously an adversary selects a linear loss function $\ell_t : \mathcal{K} \to [-1, 1]$. The player's feedback is its suffered loss, $\ell_t(a_t)$. Equivalently we will view the loss function ℓ_t as a vector in the polar body $\mathcal{K}^\circ := \{h : \forall x \in \mathcal{K}, |h \cdot x| \leq 1\}$, and thus we write $\ell_t(x) = \ell_t \cdot x$. The player has access to external randomness, and can select her action a_t based on the history $H_t = (a_s, \ell_s(a_s))_{s<t}$. The player's perfomance at the end of the game is measured through the *pseudo-regret* (the expectation is with respect to the randomness in her strategy) :

$$R_T = \mathbb{E} \sum_{t=1}^{T} \ell_t(a_t) - \min_{x \in \mathcal{K}} \mathbb{E} \sum_{t=1}^{T} \ell_t(x), \tag{1}$$

which compares her cumulative loss to the smallest cumulative loss she could have obtained had she known the sequence of loss functions. We refer to Bubeck and Cesa-Bianchi (2012) for the history of this problem, and we simply mention that the minimax rate for the regret is known to be $\widetilde{\Theta}(n\sqrt{T})$ without further assumptions on $\mathcal{K}$, and for the special case where $\mathcal{K} = \{e_1, \dots, e_n\}$ (i.e., the multi-armed bandit problem) it is $\Theta(\sqrt{nT})$.

We consider three basic open problems in bandit theory (description below), each one part of a more general trend in learning theory/online learning, namely (i) exploiting sparsity, (ii) faster learning for "easy data", and (iii) interplay between curvature and learning[1]. In fact these problems are possibly the easiest at the intersection of bandit theory and topics (i), (ii), (iii). Thus, given the flurry of activity on these topics and on bandit theory in recent years, we believe that they epitomize the difficulty of adapting full information tools to limited feedback scenarios. In particular we hope that the tools we develop to resolve these problems will find broader applicability.

Sparse multi-armed bandit, Kwon and Perchet (2016). Consider the multi-armed bandit problem with the additional assumption that at each time step $t \in [T]$ the loss vector $\ell_t \in [-1,1]^n$ only has s non-zero entries. Trivially the best regret one can hope for in this setting is $\Omega(\sqrt{sT})$. Kwon and Perchet ask whether there is a strategy with regret matching this lower bound (possibly up to logarithmic factors). Surprisingly the state of the art for this problem is the standard $O(\sqrt{nT})$ bound, or in other words prior to this present work it was not known whether sparsity of the losses can be exploited in a bandit setting[2].

Small variation bound for multi-armed bandit, Hazan and Kale (2009). Consider again the multi-armed bandit problem with the additional assumption that the loss sequence $(\ell_1, \dots, \ell_T) \in ([-1,1]^n)^T$ has a *small variation* $Q := \sum_{t=1}^T \|\ell_t - \frac{1}{T}\sum_{s=1}^T \ell_s\|_2^2$ (note that $Q \leq nT$). The COLT 2011 open problem by Hazan and Kale ask whether there exists a strategy with regret $\widetilde{O}(\sqrt{Q})$ (Hazan and Kale (2011)). The current state of the art remains Hazan and Kale (2009) which gives a strategy with regret $\widetilde{O}(n^2\sqrt{Q})$. We also note that Gerchinovitz and Lattimore (2016) showed that for any fixed $Q > \log(T)$ one cannot obtain a regret smaller than $\Omega(\sqrt{Q})$ for all sequences with variation Q.

Linear bandit on ℓ_p^n balls, Bubeck et al. (2012). Consider the linear bandit problem on $\mathcal{K} = \{x \in \mathbb{R}^n : \|x\|_p \leq 1\}$. The general minimax rate show that for any $p \geq 1$ there exists a strategy with regret $\widetilde{O}(n\sqrt{T})$, and furthermore this is optimal for $p = \infty$. It is easy to see that for $p = 1$ the problem can be reduced to the classical multi-armed bandit (in dimension $2n$) and thus there exists a strategy with regret $\widetilde{O}(\sqrt{nT})$. In Bubeck et al. (2012) it is shown that the latter regret can also be achieved for $p = 2$. No other result is known for this problem, and a natural conjecture[3] would be that $\widetilde{O}(\sqrt{nT})$ is achievable for any $p \in [1,2]$, and that the minimax regret then degrades "smoothly" for $p > 2$ until $\widetilde{\Omega}(n\sqrt{T})$ for $p = \infty$.

We resolve all the above problems, constructing strategies with respective regret bounds $\widetilde{O}(\sqrt{sT})$, $\widetilde{O}(\sqrt{Q})$, and $\widetilde{O}(\sqrt{nT})$ for $p \in [1,2]$. Furthermore we show that in fact for $p > 2$ the minimax regret (for large T) is $\widetilde{\Theta}(n\sqrt{T})$. We also introduce the following more constrained version of bandit linear optimization, which we call *starved bandit*. In this model the player only observes feedback if she plays a_t from a *fixed* distribution $\mu \in \Delta(\mathcal{K})$, where μ is chosen by the player at the beginning of the game. Thus the player is "information starved". One can motivate such a setting in various ways, think for instance of applications where logging information on users is discouraged for privacy reasons. It is easy to see that one must have regret $\Omega(T^{2/3})$ for the starved multi-armed bandit game, and that the same lower bound also applies to starved linear bandit on ℓ_p^n unit ball with $p = 1$. Perhaps surprisingly we show that $\sqrt{T}$-type regret is achievable for the starved bandit

[1] Note that the terms sparsity and curvature in the paper's title apply respectively to the losses and the action set. They could also apply respectively to the action set and to the losses, see e.g. Langford et al. (2009) and Hazan and Levy (2014). We do not consider these (very different) settings here.

[2] We note however that for *non-negative* losses (which should intuitively be a much easier case than say sparse *non-positive* losses, a.k.a. sparse gains), Kwon and Perchet already answered positively the question, see Section 3.1.

[3] This conjecture was mentioned in talks related to Bubeck et al. (2012).

for any $p \in (1, 2]$ and *not* achievable for any $p > 2$.

A key feature of our work that enables these improved regret bounds is that we avoid resorting to "global" smoothness of the regularizers. Slightly more precisely, as we will recall shortly, an important step in the analysis of FTRL (Follow The Regularized Leader) is to show that the regularizer is well-conditioned. Since the groundbreaking work Abernethy et al. (2008) it has been realized that self-concordance (Nesterov and Nemirovski (1994)) exactly gives such a good conditioning *for all directions*. In this paper we use more refined properties of the regularizers, by noticing that one only needs the well-conditioning in directions (and magnitudes) *attainable with loss estimators*.

Next we describe more formally our main results.

1.1 Main results

The brief algorithms' description given in the theorem statements below use standard bandit theory terminology which is recalled in Section 2. Note also that in this paper we assume that the parameters of the game (such as the time horizon T, or the variation of the loss sequence) are known. Standard methodology (such as the doubling trick, or more sophisticated variants of it) can be used to circumvent this issue.

We start with a theorem resolving the sparse bandit open problem by Kwon and Perchet (notice that if $\|\ell_t\|_0 \leq s$ and $\|\ell_t\|_\infty \leq 1$ then $\sum_{t=1}^T \|\ell_t\|_2^2 \leq sT$).

Theorem 1 *There exists a multi-armed bandit strategy such that for any loss sequence satisfying $\sum_{t=1}^T \|\ell_t\|_2^2 \leq L$ (and $\ell_t \in [-1,1]^n$) one has*

$$R_T \leq 10\sqrt{L \log(n)} + 20n \log(T) .$$

In fact this can be achieved with the FTRL strategy (with standard unbiased loss estimator) with the regularizer $\Phi(x) = \sum_{i=1}^n x(i) \log x(i) - \gamma \sum_{i=1}^n \log x(i)$, learning rate $\eta = \min\left(\frac{1}{5}\sqrt{\frac{\log(T)}{L}}, \frac{1}{15n}\right)$, and soft-exploration parameter $\gamma = 2\eta$.

The difficulty in achieving a result such as Theorem 1 is that standard multi-armed bandit algorithms *explore too much*. In fact as was noted in Hazan and Kale (2011) for the variation bound open problem (the same observation holds for the sparse bound open problem): "We note that EXP3 itself has $\Omega(\sqrt{T})$ regret, since it mixes with the uniform distribution every iteration to enable sufficient exploration. Hence, the desired algorithm should be a little different from EXP3, incorporating just enough exploration proportional to the variation in the data." Our new idea to achieve this is to introduce *soft exploration*, by adding to the regularizer a little bit of the log-barrier for the positive orthant. This new hybrid regularizer and its analysis is one of our key contribution. We give detailed intuition for it in Section 3.2. It also allows to solve the variation bound open problem:

Theorem 2 *There exists a multi-armed bandit strategy and a numerical constant $C > 0$ such that for any loss sequence satisfying $\sum_{t=1}^T \|\ell_t - \frac{1}{T}\sum_{s=1}^T \ell_s\|_2^2 \leq Q$ (and $\ell_t \in [-1,1]^n$) one has*

$$R_T \leq C\sqrt{Q \log(n)} + Cn \log^2(T) .$$

In fact this can be achieved by combining the Hazan-Kale reservoir sampling idea with the strategy of Theorem 1

Next we give our main theorems for linear bandit on ℓ_p^n balls. Notice that the polar of the ℓ_p^n ball is the ℓ_q^n ball with $q = p/(p-1)$.

Theorem 3 *Let $p \in (1,2]$. There exists a linear bandit algorithm playing on the unit ball of ℓ_p^n such that*

$$R_T \leq 2^{\frac{6}{p-1}} \sqrt{nT\log(T)} \,.$$

Our lower bound construction for ℓ_p^n balls with $p > 2$ uses Gaussian losses which satisfy the constraint $\|\ell_t\|_q^q \leq 1$ only in expectation. Note that from standard Gaussian concentration the same bound (up to a logarithmic factor) then holds with high probability. We work with Gaussian losses mostly for clarity of exposition, and at the expense of technical complications one could use losses which satisfy the bound $\|\ell_t\|_q^q \leq 1$ almost surely. We also note that the lower bound is only valid in the large T regime, which is necessary since there exist intermediate regimes of (T, n) where a better regret than $n\sqrt{T}$ is achievable.

Theorem 4 *Let $p > 2$ and $T \geq n^{\max\left(2, \frac{p-1}{p-2}\right)}$. There exists a numerical constant $C > 0$ such that for any linear bandit algorithm playing on the unit ball of ℓ_p^n, there exists $(\ell_t)_{t\in[T]}$, i.i.d. Gaussian random variables in $\mathbb{R}^n$ such that*

$$\mathbb{E}\|\ell_t\|_q^q \leq 1 \,, \tag{2}$$

and

$$\mathbb{E}R_T \geq Cn\sqrt{T} \,.$$

We recall the starved bandit setting introduced above. At the beginning of the game the player chooses an exploration distribution $\mu \in \Delta(\mathcal{K})$. At any time t the player can choose to play a_t at random, either from μ or from an adaptive distribution p_t (where p_t depends on the observed feedback so far). The loss of the player is $\ell_t(a_t)$. The feedback is either (i) nothing if a_t was played from p_t, or (ii) the standard bandit feedback $\ell_t(a_t)$ if a_t was played from μ. For sake of simplicity we assume that if $\mathcal{K}$ contains the (signed) canonical basis then μ is uniform on the (signed) canonical basis.

We observe that Theorem 3 holds true for the starved linear bandit framework too (indeed the strategy we give to prove Theorem 3 is a starved bandit strategy). Our main additional result for this setting is to show that for any p not covered by Theorem 3 one cannot achieve $\sqrt{T}$-type regret:

Theorem 5 *For any strategy for the starved multi-armed bandit there exists a loss sequence such that $R_T \geq \frac{1}{20} n^{1/3} T^{2/3}$. The same lower bound holds for the starved linear bandit on the ℓ_1^n ball. Furthemore for any $p > 2$ there exists a constant $C > 0$ such that for any starved linear bandit algorithm playing on the unit ball of ℓ_p^n, there exists $(\ell_t)_{t\in[T]}$, i.i.d. Gaussian random variables in $\mathbb{R}^n$ satisfying* (2) *and such that*

$$\mathbb{E}R_T \geq Cn^{\frac{q}{2+q}} T^{\frac{2}{2+q}} \,.$$

1.2 Notation

We use the following (standard) notation: $\Delta(\mathcal{K})$ for the set of probability measures supported on $\mathcal{K}$, $\Delta = \{x \in \mathbb{R}_+^n : \sum_{i=1}^n x(i) = 1\}$ for the simplex, $\|x\|_p = \left(\sum_{i=1}^n |x(i)|^p\right)^{1/p}$ for the ℓ_p^n norm, $\Phi^*(\theta) = \sup_{x\in\mathbb{R}^n} \theta \cdot x - \Phi(x)$ for the Fenchel dual of $\Phi : \mathbb{R}^n \to \overline{\mathbb{R}}$, $D_\Phi(x,y) = \Phi(x) - \Phi(y) - \nabla\Phi(y)\cdot(y-x)$ for the Bregman divergence associated to Φ, $\|h\|_x = \sqrt{\nabla^2\Phi(x)[h,h]}$ for the local norm induced by Φ at x, $\|h\|_{x,*} = \sqrt{(\nabla^2\Phi(x))^{-1}[h,h]}$ for the dual local norm, $\odot$ for the Hadamard product (i.e., entrywise product of vectors), and $\succeq$ for the positive semi-definite ordering on matrices.

2. Bandit theory reminders

We give a few brief reminders of multi-armed bandit and linear bandit theory.

2.1 Full information strategies

In this section we assume that $\mathcal{K}$ is a convex body in $\mathbb{R}^n$. We fix a learning rate $\eta > 0$ and a mirror map $\Phi : \mathbb{R}^n \to \overline{\mathbb{R}}$, that is a strictly convex and differentiable map with $\nabla\Phi(\mathbb{R}^n) = \mathbb{R}^n$ and diverging gradient as one approaches the boundary of its domain. The following theorem is a standard result on the mirror descent strategy for online linear optimization (with full information), see e.g., [Theorem 5.5, Bubeck and Cesa-Bianchi (2012)].

Theorem 6 *Let $\ell_1, \dots, \ell_T \in \mathbb{R}^n$ be a fixed sequence of loss vectors and let $x_1, \dots, x_T \in \mathcal{K}$ be defined by: $x_1 = \mathrm{argmin}_{x\in\mathcal{K}} \Phi(x)$ and*

$$x_{t+1} = \operatorname*{argmin}_{x\in\mathcal{K}} D_\Phi(x, \nabla\Phi^*(\nabla\Phi(x_t) - \eta\ell_t)). \tag{3}$$

Then one has for any $x \in \mathcal{K}$,

$$\sum_{t=1}^{T} \ell_t \cdot (x_t - x) \leq \frac{\Phi(x) - \Phi(x_1)}{\eta} + \frac{1}{\eta}\sum_{t=1}^{T} D_{\Phi^*}\Big(\nabla\Phi(x_t) - \eta\ell_t, \nabla\Phi(x_t)\Big) . \tag{4}$$

Futhermore assuming that the following implication holds true for any $y_t \in \mathbb{R}^n$,

$$\nabla\Phi(y_t) \in [\nabla\Phi(x_t), \nabla\Phi(x_t) - \eta\ell_t] \Rightarrow \nabla^2\Phi(y_t) \succeq c\nabla^2\Phi(x_t) \tag{5}$$

one obtains

$$\sum_{t=1}^{T} \ell_t \cdot (x_t - x) \leq \frac{\Phi(x) - \Phi(x_1)}{\eta} + \frac{\eta}{2c}\sum_{t=1}^{T} \|\ell_t\|_{x_t,*}^2 . \tag{6}$$

We will also use the lazy variant of mirror descent, also known as FTRL (Follow The Regularized Leader), and its corresponding "primal only" analysis. In particular while for mirror descent one has to check that Φ is "well-conditioned" on a "dual segment" (equation (5)) we will see below that for FTRL one needs to check the well-conditioning on a "primal segment" (equation (9)). Note also that mirror descent and FTRL give the same update equation when Φ is a barrier for $\mathcal{K}$ (see e.g., Bubeck (2015)), which is often the case in bandit scenario.

Theorem 7 *Let $\ell_1, \dots, \ell_T \in \mathbb{R}^n$ be a fixed sequence of loss vectors and let $x_1, \dots, x_T \in \mathcal{K}$ be defined by:*

$$x_t = \operatorname*{argmin}_{x\in\mathcal{K}} \eta \sum_{s=1}^{t-1} \ell_s \cdot x + \Phi(x). \tag{7}$$

Then one has for any $x \in \mathcal{K}$,

$$\sum_{t=1}^{T} \ell_t \cdot (x_t - x) \leq \frac{\Phi(x) - \Phi(x_1)}{\eta} + \sum_{t=1}^{T} \ell_t \cdot (x_t - x_{t+1}) . \tag{8}$$

Futhermore assuming that the following implication holds true for any $y_t \in \mathbb{R}^n$,

$$y_t \in [x_t, x_{t+1}] \Rightarrow \nabla^2\Phi(y_t) \succeq c\nabla^2\Phi(x_t) \tag{9}$$

then one has that (6) *holds true with the term $\frac{\eta}{2c}$ replaced by $\frac{2\eta}{c}$.*

Proof The proof of (8) is a classical one-line induction (sometimes referred to as the Be-The-Leader lemma). We turn to (6) and note that it suffices to show that $\|x_t - x_{t+1}\|_{x_t} \leq \frac{2\eta}{c}\|\ell_t\|_{x_t,*}$.

Observe that, using a Taylor expansion, for some $y_t \in [x_t, x_{t+1}]$ one has, with the notation $\Phi_t(x) := \eta \sum_{s=1}^t \ell_s \cdot x + \Phi(x)$ (thus $x_{t+1} \in \mathrm{argmin}\, \Phi_t$ and $x_t \in \mathrm{argmin}\, \Phi_t - \eta \ell_t$),

$$\begin{aligned}\frac{1}{2}\|x_t - x_{t+1}\|_{y_t}^2 = \Phi_t(x_t) - \Phi_t(x_{t+1}) - \nabla \Phi_t(x_{t+1}) \cdot (x_t - x_{t+1}) & \leq \Phi_t(x_t) - \Phi_t(x_{t+1}) \\ & \leq \eta \ell_t \cdot (x_t - x_{t+1}) \, .\end{aligned}$$

Using that $\nabla^2 \Phi(y_t) \succeq c \nabla^2 \Phi(x_t)$ one also has $\|x_t - x_{t+1}\|_{x_t}^2 \leq \frac{1}{c}\|x_t - x_{t+1}\|_{y_t}^2$ and thus

$$\|x_t - x_{t+1}\|_{x_t}^2 \leq \frac{2\eta}{c} \ell_t \cdot (x_t - x_{t+1}) \leq \frac{2\eta}{c} \|\ell_t\|_{x_t,*} \|x_t - x_{t+1}\|_{x_t} \, ,$$

which concludes the proof. ■

2.2 Bandit strategies

In addition to choosing a regularizer, a bandit strategy also rely on a sampling scheme, that is a map $p : \mathrm{conv}(\mathcal{K}) \to \Delta(\mathcal{K})$ such that $\mathbb{E}_{X \sim p(x)} X = x$. One then runs FTRL (or mirror descent), with the (unobserved) true losses ℓ_t replaced by estimators $\widetilde{\ell}_t$ (constructed based on the observed feedback). Moreover instead of playing the point x_t recommended by FTRL, i.e., $x_t = \mathrm{argmin}_{x \in \mathrm{conv}(\mathcal{K})} \sum_{s=1}^{t-1} \widetilde{\ell}_s \cdot x + \Phi(x)$, one plays at random $a_t \sim p(x_t)$ (where the sampling is done independently of the past given x_t). The key point is that if the loss estimator is unbiased, i.e., $\mathbb{E}_{a_t \sim p(x_t)} \widetilde{\ell}_t = \ell_t$, then one has for any $x \in \mathcal{K}$,

$$\mathbb{E} \sum_{t=1}^T \ell_t \cdot (a_t - x) = \mathbb{E} \sum_{t=1}^T \widetilde{\ell}_t \cdot (x_t - x) \, ,$$

and thus one can use Theorem 6 or Theorem 7 to bound the regret. In particular assuming that one can prove the well-conditioning condition (5) or (9), the key quantity to control is the “variance” of the loss estimator appearing in (6), namely $\mathbb{E} \, \|\widetilde{\ell}_t\|_{x_t,*}^2$.

To illustrate the above discussion let us briefly recall the classical multi-armed bandit setting (i.e., $\mathcal{K} = \{e_1, \dots, e_n\}$) with **nonnegative losses**. We use mirror descent with $\Phi(x) = \sum_{i=1}^n x(i) \log x(i)$, the sampling scheme $p : \Delta \to \Delta(e_1, \dots e_n)$ is simply the identity map (in the sense that $\mathbb{P}_{a \sim p(x)}(a = e_i) = x(i)$), and the unbiased loss estimator is

$$\widetilde{\ell}_t(i) = \frac{\ell_t(i)}{x_t(i)} \mathbb{1}\{a_t = e_i\} \, .$$

The key is to observe that since $\widetilde{\ell}_t$ has nonegative entries, one has that (5) is satisfied with $c = 1$, and thus (6) gives

$$R_T \leq \frac{\log(n)}{\eta} + \frac{\eta}{2} \sum_{t \in [T], i \in [n]} \mathbb{E} \, \|\widetilde{\ell}_t\|_{x_t,*}^2 \, .$$

The last thing to observe is that, since $\|h\|_x^2 = \sum_{i=1}^n \frac{h(i)^2}{x(i)}$, one has

$$\mathbb{E} \, \|\widetilde{\ell}_t\|_{x_t,*}^2 = \mathbb{E} \sum_{i=1}^n x_t(i) \widetilde{\ell}_t(i)^2 = \mathbb{E} \sum_{i=1}^n x_t(i) \frac{\ell_t(i)^2}{x_t(i)} \mathbb{1}\{a_t = e_i\} = \|\ell_t\|_2^2 \, .$$

Thus with an appropriate choice of η one gets

$$R_T \leq \sqrt{\frac{\log(n)}{2} \sum_{t=1}^{T} \|\ell_t\|_2^2} \,. \tag{10}$$

As a side note we observe that using the polynomial INF regularizer of Audibert and Bubeck (2009) (see Section 3.2 for a brief reminder on the INF regularizer), for any primal dual pair $p, q \geq 1$, one obtains an algorithm with a regret bound scaling in $\frac{q}{q-1}\sqrt{n^{1/q}\sum_{t=1}^{T}\|\ell_t\|_{2p}^2}$.

3. Sparsity and variation bounds for multi-armed bandit

We start first by describing some basic obstacles to obtain a sparsity type bound in Section 3.1. Then in Section 3.2 we give some intuition for our new "hybrid regularizer", $\sum_{i=1}^{n} x(i)\log(x(i)) - \gamma\sum_{i=1}^{n}\log(x(i))$, that is the weighted combination of the negentropy and the logarithmic barrier for the positive orthant[4]. The extra logarithmic barrier term can be understood as a soft way to encourage exploration (to the contrary of the usual forced exploration). Finally in Section 3.3 we prove Theorem 1 (this section is self-contained and does not require reading the two previous subsections).

3.1 Basic obstacles

The basic issue is that (10) only holds for nonnegative losses[5]. The reason nonnegativity was needed is that the well-conditioned assumption for the negentropy Φ, equation (5), crucially relies on the fact that (note that $\nabla\Phi = \log, \nabla^2\Phi = \mathrm{diag}(1/x)$) for $\log(y) = \log(x) - \ell$ with $\ell \geq 0$ one has $1/y \geq 1/x$. A standard fix to maintain the latter inequality approximately true for general losses is to ensure that the magnitude of the (estimated) loss is controlled. Indeed (5) is satisfied for some constant c provided that almost surely $\|\eta\widetilde{\ell}_t\|_\infty \leq \log(1/c)$. This almost sure control can be achieved by adding forced exploration, as was done in the original adversarial multi-armed bandit paper Auer et al. (2002), that is the sampling scheme is now $(1 - n\gamma)x_t + \gamma\mathbb{1}$, or in words explore uniformly at random with probability $n\gamma$ and otherwise play from x_t. Indeed in this case $\|\eta\widetilde{\ell}_t\|_\infty \leq \eta/\gamma$, and thus the well-conditioned assumption (5) is satisfied when $\gamma \simeq \eta$. However the added regret (with respect to $i^* \in [n]$) suffered by the extra exploration is exactly $\gamma\sum_{i,t}(\ell_t(i) - \ell_t(i^*))$. This latter term destroys the scaling with sparsity (for example if $\ell_t = -e_{i^*}$ then this term is of order $\gamma(n-1)T \simeq \eta n T$). More prosaically, the uniform exploration might make us miss out on a $n\gamma$ fraction of the "gains" of the best arm, which could be far too much. We also observe that the recently proposed implicit exploration by Kocák et al. (2014) (see also Neu (2015)) suffers from the exact same issue.

We also note that, without going into any technical details, the case of arbitrary losses seem harder than the case of nonnegative losses. Indeed the former contains the case of nonpositive losses, or equivalently nonnegative *gains*. Sparse nonnegative losses mean that most arms are performing well and only a handful are to be avoided. On the other hand sparse nonnegative gains mean that most arms are bad, and only a handful are performing well. Intuitively, finding this small set of good arms hiding in a sea of bad arms is harder than avoiding a small set of bad arms in a sea of good arms.

[4] The logarithmic barrier was recently used as a regularizer for bandits in Foster et al. (2016) to obtain first order regret bounds. We note however that the behavior of our hybrid regularizer is fundamentally different from using only the log-barrier term.

[5] Notice that one cannot simply shift the losses as this could potentially suppress sparsity.

3.2 Intuition for the hybrid regularizer

The intuition is divided in two parts: (i) the fact that the added regret for $\gamma > 0$ is controlled, and (ii) that the well-conditioning still holds.

For the first part we start with a slightly different point of view on extra (forced) exploration. It is easy to check that adding extra exploration exactly corresponds to taking the regularizer to be a "negatively shifted negentropy": $\sum_{i=1}^n (x(i) - \gamma) \log(x(i) - \gamma)$. For such a regularizer the range $\Phi(x) - \Phi(x_1)$ is controlled only for x's such that $\min_{i \in [n]} x(i) > \gamma$. In the worst case the gap between the regret with respect to such x's, and with respect to an arbitrary x can be as large as $n\gamma T$, and since the well-conditioned assumption requires $\gamma \simeq \eta$ this leads us to the extra term $\eta n T$. On the other hand for the hybrid barrier one can compare to x's with $\min_{i \in [n]} x(i) = 1/\mathrm{poly}(T)$, only at the expense of a term of the form $\frac{\gamma n \log(T)}{\eta}$. Thus provided that the well-conditioning assumption remains true for $\gamma \simeq \eta$ (this is the key part to verify) the hybrid regularizer could lead to a bound of the form (10) up to to an extra additive term of order $n \log(T)$.

For the well-conditioning intuition we first recall the INF parametrization of a regularizer (Audibert et al. (2014)): For $\psi : \mathbb{R} \to \mathbb{R}$, let Φ be defined by $\nabla \Phi^*(x) := (\psi(x_i))_{i \in [n]}$. The negentropy regularizer exactly corresponds to $\psi(s) = \exp(s)$ while adding forced extra exploration with probability $n\gamma$ can be achieved by taking $\psi(s) = \exp(s) + \gamma$. The hybrid regularizer essentially corresponds to taking $\psi(s)$ to be the exponential function when $\psi(s) \geq \gamma$, and otherwise to be roughly like $\frac{\gamma \log \gamma}{s}$. In particular we see that the well-conditioning is satisfied for $\gamma \simeq \eta$ when the played arm has probability greater than γ (since in this case everything behaves essentially as with forced exploration), and on the other hand when the played arm has probability smaller γ, its probability x is of the form $1/L$ and the updated probability is $1/(L + 1/x) \simeq x$, and thus the well-conditioning also holds in this case.

3.3 Proof of Theorem 1

Observe that the hybrid regularizer Φ is lower bounded by the negentropy in the sense that $\nabla^2 \Phi(x) \succeq \mathrm{diag}(1/x(i))$. Thus the standard argument of Section 2.2 shows that

$$\mathbb{E} \, \|\widetilde{\ell}_t\|_{x_t,*}^2 \leq \|\ell_t\|_2^2 \, .$$

In particular, using Theorem 7, it only remains to check (9). The next lemma is the key justification for our new regularizer.

Lemma 8 *Let Φ be the hybrid regularizer, $\eta > 0$, $L \in \mathbb{R}^n$, $\xi \in \mathbb{R}$, $L' := L + \xi e_1$,*

$$x := \operatorname*{argmin}_{y \in \Delta} \eta L \cdot y + \Phi(y) \textit{ and } x' := \operatorname*{argmin}_{y \in \Delta} \eta L' \cdot y + \Phi(y) \, .$$

Assuming that $|\xi| \leq C/x(1)$ for some $C > 0$ and that $\gamma \geq \eta C$, one has for any $i \in [n]$, and any $u \in (0,1)$,

$$\max\left(\frac{x'(i)}{x(i)}, \frac{x(i)}{x'(i)}\right) \leq \max\left(\exp\left(\frac{1}{\frac{\gamma}{\eta C} - 1}\right), \frac{1}{1 - \gamma - u} \exp(\gamma n / u)\right) \, .$$

For example with $C = 1$, $u = 1/2$, $\gamma = 2\eta$, and $\eta \leq \frac{1}{15n}$ one obtains

$$\max\left(\frac{x'(i)}{x(i)}, \frac{x(i)}{x'(i)}\right) \leq 3,$$

which means in particular (notice that $\nabla^2 \Phi(x) = \mathrm{diag}(1/x(i) + \gamma/x(i)^2)$) that for any $y_t \in [x_t, x_{t+1}]$ one has

$$\nabla^2 \Phi(x_t) \preceq 9 \nabla^2 \Phi(y_t) \, ,$$

which finishes the proof of Theorem 1 up to straightforward calculations.
Proof First note that the KKT conditions for x and x' show that there exist $\lambda, \lambda' \in \mathbb{R}$ such that

$$\eta L + \nabla \Phi(x) = \lambda \mathbb{1},\ \eta L' + \nabla \Phi(x') = \lambda' \mathbb{1}\ . \tag{11}$$

Also note that $\nabla^2 \Phi(x)$ is diagonal with positive entries.

Step 1: We show that λ' and $x'(i)$ for $i \neq 1$ are increasing with ξ, while $x'(1)$ is decreasing with ξ. By differentiating (11) one gets

$$\frac{d\lambda'}{d\xi}\ \mathbb{1} = \eta e_1 + \nabla^2 \Phi(x)\ \frac{dx'}{d\xi}\quad . \tag{12}$$

By multiplying the above equation with $(\nabla^2 \Phi(x))^{-1}$ and summing over the coordinates (recall that $\sum_{i=1}^n \frac{dx'(i)}{d\xi} = 0$) one obtains $\frac{d\lambda'}{d\xi} > 0$. In particular using this in (12) one obtains for any $i \neq 1$, $\frac{dx'(i)}{d\xi} > 0$, and thus $\frac{dx'(1)}{d\xi} < 0$.

Step 2: We now show that the first coordinate has a small multiplicative change. Substracting the two identities in (11) one obtains, since $\nabla \Phi(x) = (1 + \log x(i) - \gamma / x(i))_{i \in [n]}$,

$$\lambda' - \lambda + \log \frac{x(1)}{x'(1)} + \gamma \left(\frac{1}{x'(1)} - \frac{1}{x(1)} \right) = \eta \xi\ . \tag{13}$$

Observe that that by Step 1 all the terms on the lhs have the same sign and thus

$$|\lambda' - \lambda| + \left| \log \frac{x(1)}{x'(1)} \right| + \gamma \left| \frac{1}{x'(1)} - \frac{1}{x(1)} \right| = \eta |\xi|\ . \tag{14}$$

In particular we have

$$\left| \frac{1}{x'(1)} - \frac{1}{x(1)} \right| \leq \frac{\eta C/\gamma}{x(1)} \Leftrightarrow \frac{x(1)}{x'(1)} \in [1 - \eta C/\gamma, 1 + \eta C/\gamma]\ .$$

Also note that that for any $s \in (0,1)$, $\max\left(1+s, \frac{1}{1-s}\right) \leq \exp\left(\frac{1}{\frac{1}{s}-1}\right)$.

Step 3: Assuming that $x(1) \geq \gamma - \eta C$ we show that all the other coordinates also have a small multiplicative change (the case $x(1) < \gamma - \eta C$ is dealt with in the next step). Substracting the two identities in (11) one obtains for any $i \neq 1$,

$$\log \frac{x(i)}{x'(i)} + \gamma \left(\frac{1}{x'(i)} - \frac{1}{x(i)} \right) = \lambda - \lambda'\ . \tag{15}$$

In particular since the two terms on the left hand side in (15) have the same sign one has

$$\left| \log \frac{x(i)}{x'(i)} \right| + \gamma \left| \frac{1}{x'(i)} - \frac{1}{x(i)} \right| = |\lambda - \lambda'|\ . \tag{16}$$

Next we also observe that thanks to (14):

$$|\lambda - \lambda'| \leq \eta |\xi| \leq \frac{\eta C}{x(1)}\ .$$

In particular together with (16) we proved that if $x(1) \geq \gamma - \eta C$ then one has

$$\left| \log \frac{x(i)}{x'(i)} \right| \leq \frac{1}{\frac{\gamma}{\eta C} - 1}\ .$$

Step 4: Finally we show that if $x(1) \leq \gamma - \eta C$ one also has that all the other coordinates have a small multiplicative change. Let $I := \{i \neq 1 \text{ s.t. } \min(x(i), x'(i)) \geq u/n\}$ (notice that, by Step 1, the minimum is attained uniformly either at x or x'). Then thanks to (16) one has for any $i \in I$,

$$\left|\log \frac{x(i)}{x'(i)}\right| \geq |\lambda - \lambda'| - \gamma n/u ,$$

and thus

$$1 \geq \sum_{i \in I} \min(x(i), x'(i)) \exp(|\lambda - \lambda'| - \gamma n/u) .$$

Observe that if $\min(x(i), x'(i)) = x(i)$ for some $i \in I$ then one has

$$\sum_{i \in I} \min(x(i), x'(i)) = \sum_{i \in I} x(i) \geq 1 - (\gamma - \eta C) - u ,$$

while if $\min(x(i), x'(i)) = x'(i)$ for some $i \in I$ then one has (thanks to Step 2)

$$\sum_{i \in I} \min(x(i), x'(i)) = \sum_{i \in I} x'(i) \geq 1 - \frac{\gamma - \eta C}{1 - \frac{\eta C}{\gamma}} - u = 1 - \gamma - u .$$

Thus we have

$$1 \geq (1 - \gamma - u) \exp(|\lambda - \lambda'| - \gamma n/u) ,$$

which concludes the proof (recall that by (16) one has for any $i \neq 1$, $\left|\log \frac{x(i)}{x'(i)}\right| \leq |\lambda - \lambda'|$). ■

3.4 Variation bound for multi-armed bandit

We only give a brief sketch of proof of Theorem 2, as it is essentially a straightforward combination of the proof of Theorem 1 together with the arguments of Hazan and Kale (2009). In particular we ignore explicit numerical constants with the notation O.

First note that it is easy to see from (8) that the following bound holds for full information FTRL under the well-conditioning assumption (9): for any sequence $m_1, \dots, m_T \in \mathbb{R}^n$ and with $m_{T+1} = 0$ one has

$$\sum_{t=1}^{T} \ell_t \cdot (x_t - x) \leq \frac{\Phi(x) - \Phi(x_1)}{\eta} + \frac{2\eta}{c} \sum_{t=1}^{T} \|\ell_t - m_t\|_{x_t,*}^2 + \sum_{t=1}^{T+1} \|m_t - m_{t-1}\|_2 . \tag{17}$$

The strategy of Hazan and Kale is to use a small portion of "exploration" rounds to estimate $\mu_t = \frac{1}{t} \sum_{s=1}^{t} \ell_s$ by some $\widetilde{\mu}_t$ and then use it to center the loss estimator (for the non-"exploration" rounds) by setting for any $i \in [n]$:

$$\widetilde{\ell}_t(i) = \frac{(\ell_t - \widetilde{\mu}_t)(i)}{x_t(i)} \mathbb{1}\{a_t = e_i\} + \widetilde{\mu}_t(i) .$$

More precisely by doing an exploration round with probability kn/t at round t (the so-called "reservoir sampling", here $k > 0$ is a parameter of the algorithm) one can obtain an estimator $\widetilde{\mu}_t$ such that $\mathbb{E}\, \widetilde{\mu}_t = \mu_t$ and $\mathrm{Var}(\widetilde{\mu}_t) \leq \frac{Q}{kt}$. Moreover the added regret from those rounds is $O(kn \log(T))$. Thus using the bound (17) with $m_t = \mu_t$ it only remains to bound the terms $\eta \sum_{t=1}^{T} \|\widetilde{\ell}_t - \mu_t\|_{x_t,*}^2$

and $\sum_{t=1}^{T+1} \|\mu_t - \mu_{t-1}\|_2$. The latter term is easily controlled by $O(\sqrt{n}\log(Q))$, see Lemma 12 in Hazan and Kale (2009). On the other hand for the former term one gets

$$\mathbb{E}\,\|\widetilde{\ell}_t - \mu_t\|_{x_t,*}^2 \leq 2\mathbb{E}\,\|\widetilde{\ell}_t - \widetilde{\mu}_t\|_{x_t,*}^2 + 2\mathbb{E}\,\|\widetilde{\mu}_t - \mu_t\|_{x_t,*}^2 = 2\mathbb{E}\|\ell_t - \mu_t\|_2^2 + 2\mathrm{Var}(\widetilde{\mu}_t)\ ,$$

and thus $\eta\mathbb{E}\sum_{t=1}^{T} \|\widetilde{\ell}_t - \mu_t\|_{x_t,*}^2 = O(\eta Q(1+\log(T)/k))$, which easily concludes the proof up to straigthforward computations.

References

J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, 2008.

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

J.Y. Audibert, S. Bubeck, and G. Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39:31–45, 2014.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

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

S. Bubeck, N. Cesa-Bianchi, and S.M. Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, 2012.

D. Foster, Z. Li, T. Lykouris, K. Sridharan, and E. Tardos. Learning in games: Robustness of fast convergence. In *Advances in Neural Information Processing Systems 29*, pages 4734–4742. 2016.

S. Gerchinovitz and T. Lattimore. Refined lower bounds for adversarial bandits. In *Advances in Neural Information Processing Systems 29*, pages 1198–1206. 2016.

E. Hazan and S. Kale. Better algorithms for benign bandits. In *SODA*, 2009.

E. Hazan and S. Kale. A simple multi-armed bandit algorithm with optimal variation-bounded regret. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 817–820. PMLR, 2011.

E. Hazan and K. Levy. Bandit convex optimization: Towards tight bounds. In *Advances in Neural Information Processing Systems (NIPS)*. 2014.

T. Kocák, G. Neu, M. Valko, and R. Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 613–621, 2014.

J. Kwon and V. Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. *Journal of Machine Learning Research*, 17(229):1–32, 2016.

J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10:777–801, 2009.

Y. Nesterov and A. Nemirovski. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

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

Appendix A. Regular and starved linear bandits on ℓ_p^n balls

In this appendix we prove the results related to linear bandits on ℓ_p^n balls. Recall that $q = p/(p-1)$.

A.1 Proof of Theorem 3

Let $p \in (1, 2]$. We first describe a new strategy to play on ℓ_p^n balls based on a non-self-concordant barrier (when $p \neq 2$). Let $d(x) = 1 - \|x\|_p^p$, and $\Phi(x) = -\log d(x)$ (notice that for $p \neq 2$ the Hessian of Φ blows up at 0, and thus Φ cannot be self-concordant). We play FTRL with regularizer Φ and with sampling scheme given by: with probability $\max(d(x), \gamma)$ play uniformly in $\{e_1, -e_1, \dots, e_n, -e_n\}$, and otherwise play $x/\|x\|_p$. Note that this not unbiased, but rather "γ-biased", which adds a γT term to the regret. The estimator is defined by $\widetilde{\ell}_t = n \frac{\ell_t \cdot \widetilde{x}_t}{1 - \|x_t\|_p, \gamma)} \widetilde{x}_t$ if played uniformly in $\{e_1, -e_1, \dots, e_n, -e_n\}$, and $\widetilde{\ell}_t = 0$ otherwise.

While Φ is not self-concordant, the next lemma shows that one still has some form of well-conditioning (though not (5)) that will turn out to be sufficient to control the regret.

Lemma 9 *Let $x, \ell \in \mathbb{R}^n$ such that $\|x\|_p < 1$, $\|\ell\|_0 = 1$ and $\|\ell\|_2 \leq 1$. Let $y \in \mathbb{R}^n$ such that $\nabla\Phi(y) \in [\nabla\Phi(x), \nabla\Phi(x) + \ell]$. Then one has for $p \in [1, 2]$,*

$$\|\ell\|_{y,*}^2 \leq \frac{2^{\frac{3}{p-1}} d(x)}{p(p-1)} \sum_{i=1}^{n} (|x(i)|^{2-p} + |\ell(i)|^{\frac{2-p}{p-1}}) \ell(i)^2 .$$

Before moving to the proof of Lemma 9 we show how to use it to control the variance of the loss estimator. The proof of Theorem 3 is then straightforward from (4) and Lemma 10.

Lemma 10 *The above strategy satisfies for any $y_t \in \mathbb{R}^n$ such that $\nabla\Phi(y_t) \in [\nabla\Phi(x_t), \nabla\Phi(x) - \eta\widetilde{\ell}_t]$*

$$\mathbb{E}_{a_t} \|\widetilde{\ell}_t\|_{y_t,*}^2 \leq \frac{2^{\frac{4}{p-1}}}{p-1} n .$$

Proof Note that $\|\eta\widetilde{\ell}_t\|_2 \leq n\eta/\gamma$. Thus by Lemma 9 we have, provided that $\gamma \geq n\eta$,

$$\|\widetilde{\ell}_t\|_{y_t,*}^2 \leq \frac{2^{\frac{3}{p-1}} d(x_t)}{p(p-1)} \mathbb{E} \sum_{i=1}^{n} (|x_t(i)|^{2-p} + |\eta\widetilde{\ell}_t(i)|^{\frac{2-p}{p-1}}) \widetilde{\ell}_t(i)^2 .$$

We now bound separately the two terms. For the first one we have (note that $1 - \|x\|_p \geq \frac{1}{p}(1 - \|x\|_p^p)$ and thus $d(x_t) \leq p \max(1 - \|x_t\|_p, \gamma)$)

$$d(x_t) \mathbb{E}_{a_t} \sum_{i=1}^{n} |x_t(i)|^{2-p} \widetilde{\ell}_t(i)^2 \leq pn \sum_{i=1}^{n} |x_t(i)|^{2-p} \ell_t(i)^2 \leq pn ,$$

where the second inequality follows from Holder's inequality with $\frac{2}{q} + \frac{2-p}{p} = 1$. Now we bound the second term (note that $\frac{2-p}{p-1} + 2 = q$)

$$d(x_t) \mathbb{E}_{a_t} \sum_{i=1}^{n} |\eta\widetilde{\ell}_t(i)|^{\frac{2-p}{p-1}} \widetilde{\ell}_t(i)^2 \leq pn \sum_{i=1}^{n} |\ell_t(i)\eta n/\gamma|^{\frac{2-p}{p-1}} \ell_t(i)^2 \leq pn \sum_{i=1}^{n} \ell_t(i)^q \leq pn ,$$

which concludes the proof. ■

We give now a few preliminary results before proving Lemma 9.

Lemma 11 *One has for any $x \in \mathbb{R}^n$ such that $\|x\|_p < 1$,*

$$\nabla^2 \Phi^*(\nabla \Phi(x)) \preceq \frac{d(x)}{p(p-1)} \textit{diag}(|x|^{2-p}) \; .$$

Proof Straightforward derivations show that

$$\nabla \Phi(x) = \frac{p \cdot \mathsf{sign}(x) \odot |x|^{p-1}}{1 - \|x\|_p^p} \; , \tag{18}$$

$$\begin{aligned} \nabla^2 \Phi(x) &= \frac{p(p-1)\mathsf{diag}(|x|^{p-2})}{1 - \|x\|_p^p} + \frac{p^2 \left(\mathsf{sign}(x) \odot |x|^{p-1}\right)^{\otimes 2}}{(1 - \|x\|_p^p)^2} \\ &\succeq \frac{p(p-1)\mathsf{diag}(|x|^{p-2})}{1 - \|x\|_p^p} \; , \end{aligned}$$

which directly implies the lemma. ■

Lemma 12 *Let $v \in \mathbb{R}^n$ and $\ell \in \mathbb{R}^n$ such that $\|\ell\|_0 = 1$ and $\|\ell\|_2 \leq 1$. Denote $x = \nabla\Phi^*(v)$ and $y = \nabla\Phi^*(v + \ell)$. Then one has*

$$d(y) \leq 4d(x) \; , \tag{19}$$

$$|y(i)| \leq 2^{\frac{3}{p-1}} |x(i)| + |2\ell(i)|^{\frac{1}{p-1}} \; . \tag{20}$$

Proof Observe that by definition (recall (18)) one has

$$|x(i)| = \left(\frac{|v(i)| d(x)}{p}\right)^{\frac{1}{p-1}} \; , \quad |y(i)| = \left(\frac{|v(i) + \ell(i)| d(y)}{p}\right)^{\frac{1}{p-1}} \; .$$

In particular we immediately see that (19) implies (20) by the triangle inequality (also $d(y) \leq 1$ and $p \geq 1$) as follows:

$$\begin{aligned} |y(i)| = \left(\frac{|v(i) + \ell(i)| d(y)}{p}\right)^{\frac{1}{p-1}} \quad &\leq \quad \left(\frac{2\max(|v(i)|, |\ell(i)|) d(y)}{p}\right)^{\frac{1}{p-1}} \\ &\leq \quad \max\left(\left(\frac{2d(y)}{d(x)}\right)^{\frac{1}{p-1}} |x(i)|, |2\ell(i)|^{\frac{1}{p-1}}\right) \\ &\leq \quad 8^{\frac{1}{p-1}} |x(i)| + |2\ell(i)|^{\frac{1}{p-1}} \; . \end{aligned}$$

We now move to the proof of (19). We first note that (19) is trivially true for $d(x) \geq 1/4$ and thus without loss of generality one can assume $\|x\|_p^p \geq 3/4$. Crucially we now consider two cases, depending on whether the non-zero coordinate of ℓ is a "light" or "heavy" coordinate in x. Let us assume $\ell(1) \neq 0$. If $x(1) \leq (1/2)^{1/p}$ (i.e., "light") then $\sum_{i \geq 2} |x(i)|^p \geq 1/4$ and thus

$$\|y\|_p^p \geq \sum_{i \geq 2} |y(i)|^p = \sum_{i \geq 2} |x(i)|^p \left(\frac{d(y)}{d(x)}\right)^{\frac{p}{p-1}} \geq \frac{1}{4}\left(\frac{d(y)}{d(x)}\right)^{\frac{p}{p-1}} \; ,$$

which implies $d(y) \leq 4d(x)$ (since $\|y\|_p \leq 1$). On the other hand if $x(1) \geq (1/2)^{1/p}$ (i.e., "heavy") then one has

$$|v(1)| = \frac{p}{d(x)}|x(1)|^{p-1} \geq 2 \,,$$

and thus $|v(1) + \ell(1)| \geq \frac{1}{2}|v(1)|$ (since $|\ell(1)| \leq 1$) which implies

$$1 \geq |y(1)| \geq |x(1)| \left(\frac{d(y)}{2d(x)}\right)^{\frac{1}{p-1}} \geq \left(\frac{d(y)}{4d(x)}\right)^{\frac{1}{p-1}} \,.$$

■

Finally we have:

Proof [of Lemma 9] Using successively Lemma 11, (19), (20), and the fact that $p \in [1, 2]$, one has

$$\begin{aligned} \|\ell\|_{y,*}^2 \leq \frac{d(y)}{p(p-1)} \sum_{i=1}^n |y(i)|^{2-p}\ell(i)^2 &\leq \frac{4d(x)}{p(p-1)} \sum_{i=1}^n |y(i)|^{2-p}\ell(i)^2 \\ &\leq \frac{4d(x)}{p(p-1)} \sum_{i=1}^n (2^{\frac{3}{p-1}}|x(i)| + |2\ell(i)|^{\frac{1}{p-1}})^{2-p}\ell(i)^2 \\ &\leq \frac{2^{\frac{3}{p-1}}d(x)}{p(p-1)} \sum_{i=1}^n (|x(i)|^{2-p} + |\ell(i)|^{\frac{2-p}{p-1}})\ell(i)^2 \,. \end{aligned}$$

■

A.2 Proof of Theorem 4

For sake of clarity we write $\mathcal{K} = \{(x, y) \in \mathbb{R} \times \mathbb{R}^n : |x|^p + \|y\|_p^p \leq 1\}$ and the losses as $\ell_t = (w_t, z_t) \in \mathbb{R} \times \mathbb{R}^n$. Let $\varepsilon > 0$ to be such that $\varepsilon^q = C/\sqrt{T}$ for some small enough universal constant $C \in (0, 1)$ (in particular since $T > n^2$ one has $\varepsilon^q n < 1$). We now define i.i.d. Gaussian losses as follows. For $\xi \in \{-1, 1\}^n$ let $\ell_t^\xi = (w_t, z_t^\xi)$ where $w_t \sim \mathcal{N}(-1, 1)$ and $z_t^\xi \sim \mathcal{N}(\varepsilon\xi, \frac{1}{n^{2/q}} I_n)$. We show that

$$\mathbb{E}_\xi \mathbb{E}_{\ell_t^\xi} R_T = \Omega(n\sqrt{T}) \,,$$

which clearly concludes the proof (notice since $T > n^2$ one has $\mathbb{E}\|\ell_t\|_q^q = O(1)$ and thus by rescaling by a constant one can also get (2)).

The key idea of the proof is to distinguish between "exploration rounds" and "exploitation rounds", depending on whether the played action $(x_t, y_t) \in \mathcal{K}$ satisfies $x_t \leq 1/4$ or $x_t \geq 1/4$. Exploration rounds suffer constant regret because the optimal action (x^*, y^*) has x^* close to 1. On the other hand exploitation rounds give little information about ξ because of the constant variance induced by the x component. Furthermore low-regret exploitation rounds should actually have the x component close to 1 which means that even less information about ξ is gathered. We make this tradeoff more precise below, but first in Lemma 13 we formalize the fact that identifying ξ matters for low-regret and in Lemma 14 we formalize the previous sentence.

Let us define $(\bar{x}, \bar{y}) = \frac{1}{T}\sum_{t=1}^T \mathbb{E}[(x_t, y_t)]$ and $(x^*, y^*) = \mathrm{argmin}_{(x,y)\in\mathcal{K}}\, x + \varepsilon\xi \cdot y$. In particular one has

$$\mathbb{E}_{\ell_t^\xi} \frac{R_T}{T} \geq -(\bar{x} - x) + \varepsilon\xi \cdot (\bar{y} - y^*) \,. \tag{21}$$

We say a coordinate $i \in [n]$ is wrong if $\bar{y}(i)\xi(i) \geq 0$.

Lemma 13 *Let s be the number of wrong coordinates, then $\mathbb{E}_{\ell_t^\xi} R_T \geq \varepsilon^q sT/4$.*

Proof Let us assume that the first s coordinates are wrong. A straightforward calculation shows that $-x^* + \varepsilon\xi \cdot y^* = -(1+\varepsilon^q n)^{1/q}$, and thus by (21) it suffices to show that

$$-\bar{x} + \varepsilon \sum_{i=s+1}^{n} \bar{y}(i)\xi(i) \geq \varepsilon^q s/4 - (1+\varepsilon^q n)^{1/q} .$$

Since $\|(\bar{x}, \bar{y}(s+1), \cdots, \bar{y}(n))\|_p \leq 1$, by Holder's inequality we know that

$$\bar{x} - \varepsilon \sum_{i=s+1}^{n} \bar{y}(i)\xi(i) \leq (1+\varepsilon^q(n-s))^{1/q} .$$

This concludes the proof since $(1+\varepsilon^q(n-s))^{1/q} \leq (1+\varepsilon^q n)^{1/q} - \frac{1}{2q}\varepsilon^q s$. ■

Lemma 14 $\bar{x} \leq 1 - 4\varepsilon^q n \Rightarrow \mathbb{E}_{\ell_t^\xi} R_T \geq \varepsilon^q nT$.

Proof It suffices to show that $-\bar{x} + \varepsilon\xi \cdot \bar{y} \geq \varepsilon^q n - (1+\varepsilon^q n)^{1/q}$ (see beginning of previous proof). Observe that

$$-\bar{x} + \varepsilon\xi \cdot \bar{y} \geq -|\bar{x}| - \varepsilon\|\xi\|_q\|\bar{y}\|_p \geq -|\bar{x}| - (1-|\bar{x}|^p)^{1/p}\varepsilon n^{1/q} .$$

Observe that $x \mapsto x + (1-x^p)^{1/p}\varepsilon n^{1/q}$ is a nondecreasing function for $x \in [0, 1-\varepsilon^q n]$ since

$$\frac{1}{p}\varepsilon n^{1/q}(1-(1-\varepsilon^q n)^p)^{1/p-1} \leq \varepsilon n^{1/q}(\varepsilon^q n)^{1/p-1} = 1 .$$

Therefore we have

$$-\bar{x} + \varepsilon\xi \cdot \bar{y} \geq -(1-4\varepsilon^q n) - (1-(1-4\varepsilon^q n)^p)^{1/p}\varepsilon n^{1/q} ,$$

and thus the proof is concluded by $1 + (1-(1-4\varepsilon^q n)^p)^{1/p}(\varepsilon^q n)^{1/q} \leq (1+\varepsilon^q n)^{1/q} + 3\varepsilon^q n$. ■

Observe now that the observed feedback at round t is exactly

$$f_t^\xi := x_t w_t + y_t \cdot z_t^\xi \sim \mathcal{N}(x_t + \varepsilon y_t \cdot \xi, \sigma_t^2), \text{ where } \sigma_t^2 = x_t^2 + \|y_t\|_2^2/n^{2/q} .$$

Denote $\mathcal{L}_\xi$ for the law of the observed feedback up to time T, i.e., the law of $(f_1^\xi, \ldots, f_T^\xi)$. Standard calculations show that for ξ and ξ' differing only in coordinate $i \in [n]$ one has

$$\mathrm{TV}(\mathcal{L}(\xi), \mathcal{L}(\xi')) \leq \sqrt{\sum_{t=1}^{T} \mathbb{E}_{\ell_t^\xi} \frac{\varepsilon^2 y_t(i)^2}{\sigma_t^2}} .$$

Another standard calculation show that the above inequality implies

$$\mathbb{E}_{\xi,\ell_t^\xi} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbb{1}\{y_t(i)\xi(i) < 0\} \geq \frac{n}{2} - \sqrt{n \sum_{t=1}^{T} \mathbb{E}_{\xi,\ell_t^\xi} \frac{\varepsilon^2 \|y_t\|_2^2}{\sigma_t^2}} .$$

Note that the left hand side in the above inequality is exactly the average (over time) number of wrongly guessed coordinates for ξ, which we know controls the regret thanks to Lemma 13. In particular it only remains to show that

$$\sum_{t=1}^{T} \mathbb{E}_{\xi,\ell_t^\xi} \frac{\varepsilon^2 \|y_t\|_2^2}{\sigma_t^2} \le cn \, , \tag{22}$$

for some universal constant $c < 1/2$.

Note that one always has $\sigma_t^2 \ge \|y_t\|_2^2 / n^{2/q}$ and furthermore $x_t \ge 1/4 \Rightarrow \sigma_t^2 \ge 1/2^4$. Recall also that $\|y_t\|_2 \le n^{1/2-1/p}\|y_t\|_p \le n^{1/2-1/p}(1-|x_t|^p)^{1/p}$. Thus

$$\mathbb{E}\sum_{t=1}^{T} \frac{\varepsilon^2\|y_t\|_2^2}{\sigma_t^2} \le n^{2/q}\varepsilon^2 \mathbb{E}\sum_{t=1}^{T} \mathbb{1}\{x_t \le 1/4\} + 2^4 \varepsilon^2 n^{1-2/p} \sum_{t:x_t \ge 1/4} \mathbb{E}(1-|x_t|^p)^{2/p} \, . \tag{23}$$

Observe that one clearly has $\mathbb{E}R_T = \Omega(\mathbb{E}\sum_{t=1}^T \mathbb{1}\{x_t \le 1/4\})$ and thus without loss of generality we can assume $\mathbb{E}\sum_{t=1}^T \mathbb{1}\{x_t \le 1/4\} = O(n\sqrt{T})$, which means that the first term on the right hand side in (23) is smaller than $n^{1+2/q}\varepsilon^2\sqrt{T} = C^{2/q}n^{1+2/q}T^{1/2-1/q}$. This is smaller than n for $T \ge n^{\frac{2}{1-q/2}}$ and C small enough. For the second term we use that

$$\begin{aligned}
\sum_{t:x_t \ge 1/4} \mathbb{E}(1-|x_t|^p)^{2/p} &\le p^2 \sum_{t=1}^{T} \mathbb{E}(1-|x_t|)^{2/p} \\
&\le p^2 T \left(\mathbb{E}\left(1 - \frac{1}{T}\sum_{t=1}^{T} |x_t|\right)\right)^{2/p} ,
\end{aligned}$$

and because of Lemma 14 one can assume $\frac{1}{T}\mathbb{E}[\sum_{t=1}^T |x_t|] \ge 1 - 4\varepsilon^q n$ which means that the second term in (23) is smaller than $\varepsilon^2 n^{1-2/p} T(\varepsilon^q n)^{2/p} = \varepsilon^{2q} nT = C^2 n$. This concludes the proof of (22), and thus also concludes the proof of Theorem 4.

A.3 Proof of Theorem 5

We only give a brief proof sketch. The starved multi-armed bandit lower bound is standard and can be written succintly as follows. Consider random losses, where say action 1's loss is a Bernoulli of parameter $1/2$ plus or minus ε, action 2 is a Bernoulli of parameter $1/2$, and all the other actions always give a loss of 1. Denote by E the expected number of exploration rounds, i.e. rounds where the player plays from μ. It is a standard calculation that if $E/n \le c/\varepsilon^2$ for some sufficiently small constant c, then the regret is at least εT. On the other hand the regret is always larger than $\frac{n-2}{n}E/2$. Thus by setting $\varepsilon^2 = cn/E$ we have a regret lower bounded by (up to constant), with a such that $a = (1-a)\frac{1}{2}$ (i.e., $a = 1/3$):

$$\max\left(E, \left(\frac{n}{E}\right)^{1/2} T\right) \ge n^a T^{1-a} \, .$$

Essentially the same argument applies to the ℓ_1^n ball, we omit the details. We now turn to the case of ℓ_p^n balls with $p > 2$.

We see from (22) (observe that in the starved setting the sum over all $t \in [T]$ in this equation is replaced by the sum over rounds t where one plays from μ) that if $n^{2/q}\varepsilon^2 E \le cn$ for some sufficiently small constant c, then the regret is at least $\varepsilon^q nT$ (per Lemma 13). Moreover the regret is also always larger than E. Thus by setting $\varepsilon^2 = cn^{1-2/q}/E$ (i.e., $\varepsilon^q n = C(n/E)^{q/2}$) we have a regret lower bounded by (up to a constant), with a such that $a = (1-a)q/2$,

$$\max\left(E, \left(\frac{n}{E}\right)^{q/2} T\right) \geq n^a T^{1-a} \ ,$$

which concludes the proof.