

Fast Rates for Bandit Optimization with Upper-Confidence Frank-Wolfe

QUENTIN BERTHET* AND VIANNEY PERCHET†

University of Cambridge

ENS Paris–Saclay & Criteo Research, Paris

Abstract. We consider the problem of bandit optimization, inspired by stochastic optimization and online learning problems with bandit feedback. In this problem, the objective is to minimize a global loss function of all the actions, not necessarily a cumulative loss. This framework allows us to study a very general class of problems, with applications in statistics, machine learning, and other fields. To solve this problem, we analyze the Upper-Confidence Frank-Wolfe algorithm, inspired by techniques for bandits and convex optimization. We give theoretical guarantees for the performance of this algorithm over various classes of functions, and discuss the optimality of these results.

Key words and phrases: Bandit optimization, Convex optimization, Bandit problems, Frank-Wolfe algorithm.

INTRODUCTION

In online optimization problems, a decision maker chooses at each round $t \geq 1$ an action π_t from some given action space, observes some information through a feedback mechanism in order to minimize a loss, function of the set of actions $\{\pi_1, \dots, \pi_T\}$. Traditionally, this objective is computed as a cumulative loss of the form $\sum_t \ell_t(\pi_t)$ [Haz12, SS11], or as a function thereof [AD14, ADL16, EDKMM09, RST11].

Examples include classical multi-armed bandit problems where the action space is finite with K elements, in stochastic or adversarial settings [BCB12]. In these problems, the loss at round t can be written as $\ell_t(e_{\pi_t})$ for a linear form ℓ_t on \mathbb{R}^K , and basis vectors e_i . More generally, this includes also bandit problems over a convex body \mathcal{C} , where the action at each round consists in picking $x_t \in \mathcal{C}$ and where the loss $\ell_t(x_t)$ is for some convex function ℓ_t (see, e.g. [BCB12, CBL06, HL14, BEL16]).

In this work, we consider the online learning problem of *bandit optimization*. Similarly to other problems of this type, a decision maker chooses at each round an action π_t from a set of size K , and observes information about an unknown convex

*Supported by an Isaac Newton Trust Early Career Support Scheme and by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

†Supported by the ANR grant ANR-13-JS01-0004-01, of the *FMJH Program Gaspard Monge in optimization and operations research* (supported in part by EDF) and from the Labex LMH.

loss function L . The difference is that the objective is to minimize a global convex loss $L(\frac{1}{T} \sum_{t=1}^T e_{\pi_t})$, not a cumulative one. At each round, choosing the i -th action increases the information about the local dependency of L on its i -th coefficient. This problem can be contrasted with the objective of minimizing the average pseudo-regret in a stochastic bandit problem, i.e. of minimizing $\frac{1}{T} \sum_{t=1}^T L(e_{\pi_t})$ with observation $\ell_t(e_{\pi_t})$, a noisy estimate of $L(e_{\pi_t})$. At the intersection of these frameworks, when L is a linear form, is the stochastic multi-armed bandit problem. Our problem is also related to maximization of known convex objectives [AD14, ADL16]. We compare our framework to these settings in Section 1.4.

Bandit optimization shares some similarities with stochastic optimization problems, where the objective is to minimize $f(x_T)$ for an unknown function f , while choosing at each round a variable x_t and observing some noisy information about the function f . Our problem can be seen as a stochastic optimization problem over the simplex, with the caveat that the list of actions π_1, \dots, π_T determines the variable, as $x_t = \frac{1}{t} \sum_{s=1}^t e_{\pi_s}$, as well as the manner in which additional information about the function can be gathered. This setting allows us to study a more general class of problems than multi-armed bandits, and to cover examples where there is not one optimal action, but rather an optimal global strategy, that is an optimal mix of actions. We describe several natural problems from machine learning, statistics, or economics that are cases of bandit optimization.

This problem draws inspiration from the world of multi-armed bandit problems and that of stochastic convex optimization, and our solution to it does as well. We analyze the Upper-Confidence Frank-Wolfe algorithm, a modification of the Frank-Wolfe algorithm [FW56] and of the UCB algorithm for bandits [ACBFS02]. The link with Frank-Wolfe is related to the choice of one action, and encourages exploitation, while the link with UCB encourages to chose rarely picked actions in order to increase knowledge about the function, encouraging exploration. This algorithm can be used for all convex functions L , and performs in a near-optimal manner over various classes of functions. Indeed, if it has been already proved that it achieves slow rates of convergence in some cases, i.e., the error decreases as $1/\sqrt{T}$, we are able to exhibit fast rates decreasing in $1/T$, up to logarithmic terms.

These fast rates are surprising, as they sometimes even hold for non-strongly convex functions, and in many problems with bandit feedback they cannot be reached [JNR12, Sha13]. As shown in our lower bounds, the main complexity of this problem is statistical and comes from the limited information available about the unknown function L . Usual results in optimization with a known function are not necessarily relevant to our problem. As an example, while linear rates in e^{-cT} are possible in deterministic settings with variants in the Frank-Wolfe algorithm, we are limited to fast rates in $1/T$ under similar assumptions. Interestingly, while linear functions are one of the settings in which the deterministic Frank-Wolfe algorithm is the most efficient, it is among the most complicated for bandit optimization, and only slow rates are possible in general (see theorems 3 and 8).

Our work is organized in the following manner: we describe in Section 1 the problem of bandit optimization. The main algorithm is introduced in Section 2, and its performance in various settings is studied in Section 3, 4, and 5. All proofs of the main results are in the supplementary material.

Notations: For any positive integer n , denote by $[n]$ the set $\{1, \dots, n\}$ and, for any positive integer K , by $\Delta_K := \{p \in \mathbb{R}^K : p_i \geq 0 \text{ and } \sum_{i \in [K]} p_i = 1\}$ the unit simplex of \mathbb{R}^K . Finally, e_i stands for the i -th vector of the canonical basis of \mathbb{R}^K . Notice that Δ_K is their convex hull.

1. BANDIT OPTIMIZATION

We describe the *bandit optimization* problem, generalizing multi-armed bandits. This stochastic optimization problem is doubly related to bandits: The decision variable cannot be chosen freely but is tied to the past actions, and information about the function is obtained via a bandit feedback.

1.1 Problem description

At each time step $t \geq 1$, a decision maker chooses an action $\pi_t \in [K]$ from K different actions with the objective of minimizing an unknown convex loss function $L : \Delta_K \rightarrow \mathbb{R}$. Unlike in traditional online learning problems, we do not assume that the overall objective of the agent is to minimize a cumulative loss $\sum_t L(e_{\pi_t})$ but rather to minimize the global loss $L(p_T)$, where $p_t \in \Delta_K$ is the vector of *proportions* of each action (also called occupation measure), i.e.,

$$p_t = (T_1(t)/t, \dots, T_K(t)/t) \text{ with } T_i(t) = \sum_{s=1}^t \mathbb{1}\{\pi_s = i\}.$$

Alternatively, $p_t = \frac{1}{t} \sum_{s=1}^t e_{\pi_s}$. As usual in stochastic optimization, the performance of a policy is evaluated by controlling the difference

$$r(T) := \mathbb{E}[L(p_T)] - \min_{p \in \Delta_K} L(p).$$

The information available to the policy is a feedback of *bandit type*: given the choice $\pi_t = i$, it is an estimate \hat{g}_t of $\nabla L(p_t)$. Its precision, with respect to each coefficient $i \in [K]$, is specified by a deviation function $\alpha_{t,i}$, meaning that for all $\delta \in (0, 1)$, it holds with probability $1 - \delta$ that

$$|\hat{g}_{t,i} - \nabla_i L(p_t)| \leq \alpha_{t,i}(T_i(t), \delta).$$

At each round, it is possible to improve the precision for one of the coefficients of the gradient but possibly at a cost of increasing the global loss. The most typical case, described in the following section, is of $\alpha_{t,i}(T_i, \delta) = \sqrt{2 \log(t/\delta)/T_i}$, when the information consists of observations from different distributions. In general, this type of feedback mechanism is indicative of a bandit feedback (and not of a full information setting), as motivated by the following parametric setting.

1.2 Bandit feedback and parametric setting

One of the motivations is the minimization of a loss function L belonging to a known class $\{L(\mu, \cdot), \mu \in \mathbb{R}^K\}$ with an unknown parameter μ . Choosing the i -th action provides information about μ_i , through an observation of some auxiliary distribution ν_i .

As an example, the classical stochastic multi-armed bandit problem [BCB12] falls within our framework. Denoting by μ_i the expected loss of arm $i \in [K]$, the average pseudo-regret \bar{R} can be expressed as

$$\bar{R}(t) = \frac{1}{t} \sum_{s=1}^t \mu_{\pi_s} - \mu^* = \sum_{i=1}^K \mu_i \frac{T_i(t)}{t} - \mu^* = p_t^\top \mu - p_*^\top \mu, \quad \text{with } p_* = e_{i^*},$$

Hence the choice of $L(\mu, p) = \mu^\top p$ corresponds to the problem of multi-armed bandits. Since $\nabla L(\mu, p) = \mu$, the feedback mechanism for \hat{g}_t is induced by having a sample X_t from ν_{π_t} at time step t , taking $\hat{g}_{t,i} = \bar{X}_{t,i}$, the empirical mean of the $T_i(t)$ observations ν_i . In this case, if ν_i is sub-Gaussian with parameter 1, we have $\alpha_{t,i}(T_i, \delta) = 2\sqrt{2\log(t/\delta)/T_i}$.

More generally, for any parametric model, we can consider the following observation setting: For all $i \in [K]$, let ν_i be a sub-Gaussian distribution with mean μ_i and tail parameter σ^2 . At time t , for an action $\pi_t \in [K]$, we observe a realization from ν_{π_t} . We estimate μ_i by the empirical mean $\hat{\mu}_{t,i}$ of the $T_i(t)$ draws from ν_i , and $\hat{g}_t = \nabla_p L(\hat{\mu}_t, p_t)$ as an estimate of the gradient of $L = L(\mu, \cdot)$ at p_t . The following bound on α_i under smoothness conditions on the parametric model is a direct application of Hoeffding's inequality.

PROPOSITION 1. *Let $L = L(\mu, \cdot)$ for some $\mu \in \mathbb{R}^K$ being μ -gradient-Lipschitz, i.e., such that*

$$\left| (\nabla_p L(\mu, p))_i - (\nabla_p L(\mu', p))_i \right| \leq |\mu_i - \mu'_i|, \quad \forall p \in \Delta([K]).$$

Under the sub-Gaussian observation setting above, $\hat{g}_t = \nabla_p L(\hat{\mu}_t, p_t)$ is a valid gradient feedback with deviation bounds $\alpha_{t,i}(T_i, \delta) = \sqrt{2\sigma^2 \log(t/\delta)/T_i}$.

This Lipschitz condition on the parameter μ gives a motivation for our gradient bandit feedback.

1.3 Examples

Stochastic multi-armed bandit: As noted above, the stochastic multi-armed bandit problem is a special case of our setting for a loss $L(p) = \mu^\top p$, and the bandit feedback allows to construct a proxy for the gradient \hat{g}_t with deviations α_i decaying in $1/\sqrt{T_i}$. The UCB algorithm used to solve this problem inspires our algorithm that generalizes to any loss function L , as discussed in Section 2.

Online experimental design: In the context of statistical estimation with heterogeneous data sources [BC16], consider the problem of allocating samples in order to minimize the variance of the final estimate. At time t , it is possible to sample from one of K distributions $\mathcal{N}(\theta_i, \sigma_i^2)$ for $i \in [K]$, the objective being to minimize the average variance of the simple unbiased estimator

$$\mathbb{E}[\|\hat{\theta} - \theta\|_2^2] = \sum_{i \in [K]} \sigma_i^2 / T_i \quad \text{equivalent to} \quad L(p) = \sum_{i \in [K]} \sigma_i^2 / p_i.$$

For unknown σ_i , this problem falls within our framework and the gradient with coordinates $-\sigma_i^2/p_i^2$ can be estimated by using the T_i draws from $\mathcal{N}(\theta_i, \sigma_i^2)$ to construct $\hat{\sigma}_i^2$. This function is only defined on the interior of the simplex and is unbounded, matters that we discuss further in Section 4.3. Other objective functions than the expected ℓ_2 norm of the error can be used, as in [CLG⁺15], who consider the ℓ_∞ norm of the actual estimated deviations, not its expectation.

Utility maximization: A classical model to describe the utility of an agent purchasing x_i units of K different goods is the Cobb-Douglas utility (see e.g. [MCWG95]) defined for parameters $\beta_i \in (0, 1)$ by

$$U(x_1, \dots, x_K) = \prod_{i \in [K]} x_i^{\beta_i}.$$

Maximizing this utility for unknown β_i under a budget constraint - where each price is assumed to be 1 for ease of notations - by buying one unit of one of K goods at each round, is therefore equivalent to minimizing in p_i (the proportion of good i in the basket) $L(p) = -\sum_{i \in [K]} \beta_i \log(p_i)$.

Other examples: More generally, the notion of bandit optimization can be applied to any situation where one optimizes a strategy through actions that are taken sequentially, with information gained at each round, and where the objective depends only on the proportions of actions. Other examples include a problem inspired by online Markovitz portfolio optimization, where the goal is to minimize $L(p) = p^\top \Sigma p - \lambda \mu^\top p$, with a known covariance matrix Σ and unknown returns μ , or several generalizations of bandit problems such as minimizing $L(p) = \sum_{i \in [K]} f_i(\mu_i) p_i$ when observations are drawn from a distribution with mean μ_i , for known f_i .

1.4 Comparison with other problems

As mentioned in the introduction, the problem of bandit optimization is different from online learning problems related to regret minimization [HAK07, AFH⁺11, BEL16], even in a stochastic setting. While the usual objective is to minimize a cumulative regret related to $\frac{1}{T} \sum_t \ell_t(x_t)$, we focus on $L(\frac{1}{T} \sum_t e_{\pi_t})$.

Problems related to online optimization of global costs or objectives have been studied in similar settings [AD14, ADL16, EDKMM09, RST11]. They are equivalent to minimizing a loss $L(p_T^\top V)$ where V is a $K \times d$ unknown matrix and $L(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$ is known. The feedback at stage t is a noisy evaluations of V_{π_t} . In the stochastic case [AD14, ADL16], this is close to our setting - even though none of them subsumes directly the other one. Only slow rates of convergence of order $1/\sqrt{T}$ are derived for the variant of Frank-Wolfe, while we aim at fast rates, which are optimal. In contrast, in the adversarial case [EDKMM09, RST11], there are instances of the problem where the average regret cannot decrease to zero [MPS14].

Using the Frank-Wolfe algorithm in a stochastic optimization problem has also already been considered, particularly in [LWM15], where the estimates of the gradients are increasingly precise in t , independently of the actions of the decision maker. This setting, where the action at each round is to pick x_t in the domain in order to minimize $f(x_T)$ is therefore closer to classical stochastic optimization than online learning problems related to bandits [BCB12, HL14, BEL16].

2. UPPER-CONFIDENCE FRANK-WOLFE ALGORITHM

With linear functions, as in multi-armed bandits, an estimate of the gradient can be established by using the past observations, as well as confidence intervals on each coefficient in $1/\sqrt{T_i}$. The UCB algorithm instructs to pick the action with the smallest lower confidence estimate $\underline{\mu}_{t,i}$ for the loss. This is equivalent to making a step of size $1/(t+1)$ in the direction of the corner of the simplex e that minimizes $e^\top \underline{\mu}_t$. Following this intuition, we introduce the UCB Frank-Wolfe algorithm that uses a proxy of the gradient, penalized by the size of confidence intervals.

Input: K , $p_0 = \mathbf{1}_{[K]}/K$, sequence $(\delta_t)_{t \geq 0}$;
for $t \geq 0$ **do**
 Observe \hat{g}_t , noisy estimate of $\nabla L(p_t)$;
 for $i \in [K]$ **do**
 | $\hat{U}_{t,i} = \hat{g}_{t,i} - \alpha_{t,i}(T_i(t), \delta_t)$
 end
 Select $\pi_{t+1} \in \operatorname{argmin}_{i \in [K]} \hat{U}_{t,i}$;
 Update $p_{t+1} = p_t + \frac{1}{t+1}(e_{\pi_{t+1}} - p_t)$
end

Notice that for any algorithm, the selection of an action $\pi_{t+1} \in [K]$ at time step $t+1$ updates the variable p with respect to the following dynamics

$$(2.1) \quad p_{t+1} = \left(1 - \frac{1}{t+1}\right) p_t + \frac{1}{t+1} e_{\pi_{t+1}} = p_t + \frac{1}{t+1} (e_{\pi_{t+1}} - p_t).$$

This is implied by the mechanism of the problem, and is not dependent on the choice of an algorithm. If the choice of $e_{\pi_{t+1}}$ is $e_{\star_{t+1}}$, the minimizer of $s^\top \nabla L(p_t)$ over all $s \in \Delta_K$, this would precisely be the Frank-Wolfe algorithm with step size $1/(t+1)$. Inspired by this similarity, our selection rule is driven by the same principle, using a proxy \hat{U}_t for $\nabla L(p_t)$ based on the information up to time t . Our selection rule is therefore driven by two principles, borrowing from tools in convex optimization (the Frank-Wolfe algorithm) and classical bandit problems (Upper-confidence bounds).

The choice of action π_{t+1} is equivalent to taking $e_{\pi_{t+1}} \in \operatorname{argmin}_{s \in \Delta_K} s^\top \hat{U}_t$. The computational cost of this procedure is very light, and apart from gradient computations, it is linear in K at each iteration, with a global cost of order KT .

3. SLOW RATES

In this section we show that when α_i is of order $1/\sqrt{T_i}$, as motivated by the parametric model of Section 1.2, our algorithm has an approximation error of order $\sqrt{\log(T)/T}$ over the very general class of smooth convex functions. We refer to this as the *slow rate*. Our analysis is based on the classical study of the Frank-Wolfe algorithm (see, e.g. [Jag11] and references therein). We consider the case of C -smooth convex functions on the unit simplex, for which we recall the definition.

DEFINITION 2 (Smooth functions). *For a set $\mathcal{D} \subset \mathbb{R}^n$, a function $f : \mathcal{D} \rightarrow \mathbb{R}$ is said to be a C -smooth function if it is differentiable and if its gradient is C -Lipshitz continuous, i.e. the following holds*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq C \|x - y\|_2, \quad \forall x, y \in \mathcal{D}.$$

We denote by $\mathcal{F}_{C,K}$ the set of C -smooth convex functions. They attain their minimum at a point $p_\star \in \Delta_K$ and their Hessian is uniformly bounded, i.e. $\nabla^2 L(p) \preceq CI_K$, if they are twice differentiable. We establish in this general setting a slow rate when α_i decreases like $1/\sqrt{T_i}$.

THEOREM 3 (Slow rate). *Let L be a C -smooth convex function over the unit simplex Δ_K . For any $T \geq 1$, after T steps of the UCB Frank-Wolfe algorithm with a bandit feedback such that $\alpha_{t,i}(T_i, \delta) = 2\sqrt{\log(t/\delta)/T_i}$ and the choice $\delta_t = 1/t^2$, it holds that*

$$\mathbb{E}[L(p_T)] - L(p_\star) \leq 4\sqrt{\frac{3K \log(T)}{T}} + \frac{C \log(eT)}{T} + \left(\frac{\pi^2}{6} + K\right) \frac{2\|\nabla L\|_\infty + \|L\|_\infty}{T}.$$

The proof draws inspiration from the analysis of the Frank-Wolfe algorithm with stepsize of $1/(t+1)$ and of the UCB algorithm. Notice that our algorithm is adaptive to the gradient Lipschitz constant C , and that the leading term of the error does not depend on it. We also emphasize the fact that the dependency in \sqrt{K} is expected, and optimal, in bandit setting.

For linear mappings $L(p) = p^\top \mu$, our analysis is equivalent to studying the UCB algorithm in multi-armed bandits. The slow rate in Theorem 3 corresponds to a regret of order $\sqrt{KT \log(T)}$, the distribution-independent (or worst case) performance of UCB. The extra dependency in $\sqrt{\log(T)}$ could be reduced to $\sqrt{\log(K)}$ or even optimally to 1 by using confidence intervals more carefully tailored, for instance by replacing the $\log(t)$ term appearing in the definition of the estimated gradients by $\log(T/T_i(t))$ or $\log(T/KT_i(t))$ if the horizon T is known in advance as in the algorithms MOSS or ETC (see [AB09, PR13, PRCS16]), but at the cost of a more involved analysis.

Thus, multi-armed bandits provide a lower bound for the approximation error $\mathbb{E}[L(p_T)] - L(p_\star)$ of order $\sqrt{K/T}$ for smooth convex functions. We discuss lower bounds further in Section 5.

For the sake of clarity, we state all our results when $\alpha_{t,i}(T_i, \delta) = 2\sqrt{\log(t/\delta)/T_i}$, but our techniques handle more general deviations as $\alpha_{t,i}(T_i, \delta) = (\theta \log(t/\delta)/T_i)^\beta$ where $\theta \in \mathbb{R}$ and $\beta > 0$ are some known parameters. More general results can be found in the supplementary material.

4. FAST RATES

In this section, we describe situations where the approximation error rate can be improved to a *fast rate* of order $\log(T)/T$, when we consider various classes of functions, with additional assumptions.

4.1 Stochastic multi-armed bandits and functions minimized on vertices

A very natural and well-known - yet illustrative - example of such a restricted class of functions is simply the case of classical bandits where $\Delta^{(i)} := \mu_i - \mu_\star$ is bounded away from 0 for $i \neq \star$. Our analysis of the algorithm can be adapted to this special case with the following result.

PROPOSITION 4. *Let L be the linear function $p \mapsto p^\top \mu$. After T steps of the UCB Frank-Wolfe algorithm with a bandit feedback such that $\alpha_{t,i}(T_i, \delta) = 2\sqrt{\log(t/\delta)/T_i}$, the choices of $\delta_t = 1/t^2$ hold the following*

$$\mathbb{E}[L(p_T)] - L(p_\star) \leq \frac{48 \log(T)}{T} \sum_{i \neq \star} \frac{1}{\Delta^{(i)}} + 3\left(\frac{\pi^2}{3} + K\right) \frac{\sqrt{K}\|\mu\|_\infty}{T}.$$

The constants of this proposition are sub-optimal (for instance the 48 can be reduced up to 2 using more careful but involved analysis). It is provided here to show that this classical bound on the pseudo-regret in stochastic multi-armed bandits (see e.g. [BCB12] and references therein) can be recovered with Frank-Wolfe type of techniques illustrating further the links between bandit problems and convex optimization [Haz12, SS11]. This result can actually be generalized to any convex functions which is minimized on a vertex of the simplex with a gradient whose component-wise differences are bounded away from 0.

PROPOSITION 5. *Let L be a convex mapping that attains its minimum on Δ_K at a vertex $p^* = e_{i^*}$ and such that $\Delta^{(i)}(L) := \nabla_i L(p^*) - \nabla_{i^*} L(p^*) > 0$ for all $i \neq i^*$. Then, after T steps of the UCB Frank-Wolfe algorithm with a bandit feedback such that $\alpha_{t,i}(T_i, \delta) = 2\sqrt{\log(t/\delta)/T_i}$, the choices of $\delta_t = 1/t^2$ hold the following*

$$\mathbb{E}[L(p_T)] - L(p_*) \leq \rho(L) \left(\frac{48 \log(T)}{T} \sum_{i \neq i^*} \frac{1}{\Delta^{(i)}(L)} + \frac{C \log(eT)}{T} + \left(\frac{\pi^2}{6} + K\right) \frac{2\|\nabla L\|_\infty + \|L\|_\infty}{T} \right),$$

where $\rho(L) = \left(1 + \frac{CK}{\Delta_{\min}(L)}\right)$ and $\Delta_{\min}(L) = \min_{i \neq i^*} \Delta^{(i)}(L)$.

The KKT conditions imply that $\Delta^{(i)}(L) \geq 0$ but the strict inequality is not always guaranteed. In particular, this result may not hold if p^* is the global minimum of L over \mathbb{R}^K . This type of condition has also been linked with rates of convergence in stochastic optimization problems [DR16].

The extra multiplicative factor $\rho(L)$ can be large, but it would be of the order of $1 + o(1)$ using variants of our algorithms with results that holds only with great probability (typically with confidence bounds of the form $2\sqrt{\log(1/\delta)/T_i}$).

4.2 Strongly convex functions

Another classical assumption in convex optimization is strong convexity, as recalled below. We denote by $\mathcal{S}_{\mu,K}$ the set of μ -strongly convex functions of Δ_K . This assumption usually improves the rates in errors of approximation in many settings, even in stochastic optimization or some settings of online learning (see, e.g. [PT90, Dip03, ST11, BM13, HKL14, HL14, BP16]). Interestingly enough though, strong convexity cannot be leveraged to improve rates of convergence in online convex optimization [Sha13, JNR12], where the $1/\sqrt{T}$ rate of convergence cannot be improved. Moreover, leveraging strong convexity usually implies to adapt step size of gradient descents or with linear search and/or away steps for classical Frank-Wolfe methods. Those techniques cannot be adapted to our setting where step sizes are fixed.

DEFINITION 6 (Strongly convex functions). *For a set $\mathcal{D} \subset \mathbb{R}^n$, a function $f : \mathcal{D} \rightarrow \mathbb{R}$ is said to be a μ -strongly convex if for all $x, y \in \mathcal{D}$, we have*

$$f(x) \geq f(y) + \nabla f(x)^\top (x - y) + \frac{\mu}{2} \|x - y\|_2^2.$$

We already covered the case where the convex functions are minimized outside the simplex. We will now assume that the minimum lies in its relative interior.

THEOREM 7. *Let $L : \Delta_K \rightarrow \mathbb{R}$ be a C -smooth, μ -strongly convex function such that its minimum p_\star satisfies $\text{dist}(p_\star, \partial\Delta_K) \geq \eta$, for some $\eta \in (0, 1/K]$. After T steps of the UCB Frank-Wolfe algorithm with a bandit feedback such that $\alpha_{t,i}(T_i, \delta) = 2\sqrt{\log(t/\delta)/T_i}$, it holds that, with the choice of $\delta_t = 1/t^2$,*

$$\mathbb{E}[L(p_T)] - L(p_\star) \leq c_1 \frac{\log^2(T)}{T} + c_2 \frac{\log(T)}{T} + c_3 \frac{1}{T},$$

for constants $c_1 = \frac{96K}{\mu\eta^2}$, $c_2 = \frac{24}{\mu\eta^3} + C$ and $c_3 = 24(\frac{20}{\mu\eta^2})^2 K + \frac{\mu\eta^2}{2} + C$.

The proof is based on an improvement in the analysis of the UCB Frank-Wolfe algorithm, based on a better control on the duality gap, possible in the strongly convex case. It is a consequence of an inequality due to Lacoste-Julien and Jaggi (Lemma 2 in [LJJ13]). In order to get the result, we adapt these ideas to a case of unknown gradient, with bandit feedback. We note that this approach is similar to the one in [LWM15] that focuses on stochastic optimization problems, as discussed in Section 1.4.

Our framework is more complicated in some aspects than typical settings in stochastic optimization, where strong assumptions can usually be made over the noisy gradient feedback. These include stochastic gradients that are independent unbiased estimates of the true gradient, or with error terms that are decreasing in t . Here, such properties do not hold: as an example, in a parametric setting, information is only obtained about one of the coefficients, and there are strong dependencies between successive gradients feedbacks. Dealing with these aspects, as well as the fact that our gradient proxy is penalized by the size of the confidence intervals, are some of the main challenges of the proof.

4.3 Interior-smooth functions

Many interesting examples of bandit optimization are not exactly covered by the case of functions that are C -smooth on the whole unit simplex. In particular, for several applications, the function diverges at its boundary, as in the examples of Cobb-Douglas utility maximization and variance minimization from Section 1.3. Recall the the loss was defined by

$$\mathbb{E}[\|\hat{\theta} - \theta\|_2^2] = \sum_{i \in [K]} \frac{\sigma_i^2}{T_i} = \frac{1}{T} L(p) = \frac{1}{T} \sum_{i \in [K]} \frac{\sigma_i^2}{p_i}.$$

The gradient Lipschitz constant is infinite but if we knew for instance that $\sigma_i \in [\underline{\sigma}_i, \bar{\sigma}_i]$, we could safely sample first each arm i a linear number of time because $p_i^* \geq \underline{p}_i := \underline{\sigma}_i / \sum_j \bar{\sigma}_j$. We would have $(p_t)_i \geq \underline{p}_i$ at all stages and our analysis holds with the constant $C = 2\sigma_{\max}^2 (\sum_j \bar{\sigma}_j)^3 / \underline{\sigma}_{\min}^3$.

Even without knowledge on σ_i^2 , it is possible to quickly have rough estimates, as illustrated by Lemma 13 in the appendix. Only a logarithmic number of sample of each action are needed. Once they are gathered, one can keep sampling each arm a linear number of times, as suggested when the lower/upper bounds are known beforehand. This leads to a Lipschitz constant $C = (9 \sum_j \sigma_j)^3 / \sigma_{\min}$, which is, up to to a multiplicative factor, the gradient Lipschitz constant at the minimum.

5. LOWER BOUNDS

The results shown in Sections 3 and 4 exhibit different theoretical guarantees for our algorithm depending on the class of function considered. We discuss here the optimality of these results.

5.1 Slow rate lower bound

In Theorem 3, we show a slow rate of order $\sqrt{K \log(T)/T}$ for the error approximation of our algorithm over the class of C -smooth convex functions of \mathbb{R}^K . Up to the logarithmic term, this result is optimal: no algorithm based on the same feedback can significantly improve the rate of approximation. This is a consequence of the following theorem, a direct corollary of a result by [ACBFS02].

THEOREM 8. *For any algorithm based on a bandit feedback such that $\alpha_{t,i}(T_i, \delta) = \sqrt{2 \log(t/\delta)/T_i}$ and that outputs \hat{p}_T , we have over the class of linear forms \mathcal{L}_K that for some constant $c > 0$*

$$\inf_{\hat{p}_T} \sup_{L \in \mathcal{L}_K} \left\{ \mathbb{E}[L(\hat{p}_T)] - L(p_*) \right\} \geq c \sqrt{K/T}.$$

This result is established over the class of linear functions over the simplex (for which $C = 0$), when the feedback consists of a draw from a distribution with mean μ_i . As mentioned in Section 3, the extra logarithmic term in our upper bound comes from our algorithm, which has the same behavior as UCB. Nevertheless, as mentioned before, modifying our algorithm to recover the behavior of MOSS [AB09], or even ETC, (see e.g. [PR13, PRCS16]), would improve the upper bound and remove the logarithmic term.

5.2 Fast rate lower bound

We have shown that in the case of strongly convex smooth functions, there is an approximation error upper bound of order $(K/\eta^4) \log(T)/T$ for the performance of our algorithm, where $\eta \leq 1/K$. We provide a lower bound over this class of functions in the following theorem.

THEOREM 9. *For any algorithm with a bandit feedback such that $\alpha_{t,i}(T_i, \delta) = \sqrt{2 \log(t/\delta)/T_i}$ and output \hat{p}_T , we have over the class $\mathcal{S}_{1,K}$ of 1-strongly convex functions that for some constant $c > 0$*

$$\inf_{\hat{p}} \sup_{L \in \mathcal{S}_{1,K}} \left\{ \mathbb{E}[L(\hat{p}_T)] - L(p^*) \right\} \geq c K^2/T.$$

The proof relies on the complexity of minimizing quadratic functions $\frac{1}{2} \|p - \theta\|_2^2$ when observing a draw from distribution with mean θ_i . Our upper bound is in the best case of order $K^5 \log(T)/T$, as $\eta \leq 1/K$. Understanding more precisely the optimal rate is an interesting venue for future research.

5.3 Mixed feedbacks lower bound

In our analysis of this problem, we have only considered settings where the feedback upon choosing action i gives information about the i -th coefficient of the gradient. The two following cases show that even in simple settings, our upper bounds will not hold if the relationship between action and feedback is different, when the feedback corresponds to another coefficient.

PROPOSITION 10. For L in the class of 1-strongly convex functions on Δ_3 , we have in the case of a mixed bandit feedback that

$$\inf_{\hat{p}} \sup_{L \in \mathcal{S}_{1,3}} \left\{ \mathbb{E}[L(\hat{p}_T)] - L(p^*) \right\} \geq c/T^{2/3}.$$

For strongly convex functions, even with $K = 3$, there are therefore pathological mixed feedback settings where the error is at least of order $1/T^{2/3}$ instead of $1/T$. The case of smooth convex functions is covered by the existing lower bounds for the problem of *partial monitoring* [CBL06], and gives a lower bound of order $1/T^{1/3}$ instead of $1/\sqrt{T}$.

PROPOSITION 11. For L in the class of linear forms \mathcal{F}_3 on Δ_3 , with a mixed bandit feedback we have

$$\inf_{\hat{p}} \sup_{L \in \mathcal{F}_3} \left\{ \mathbb{E}[L(\hat{p}_T)] - L\theta(p^*) \right\} \geq c/T^{1/3}.$$

6. DISCUSSION

We study the online minimization of stochastic global loss with a bandit feedback. This is naturally motivated by many applications with a parametric setting, and tradeoffs between exploration and exploitation. The UCB Frank-Wolfe algorithm performs optimally in a generic setting.

The fast rates of convergence obtained for some classes of functions are a significant improvement over the slow rates that hold for smooth convex functions. In bandit-type problems similar to our problem, it is not always possible to leverage additional assumptions such as strong convexity: It has been proved impossible in the closely related setting of online convex optimization [JNR12, Sha13]. When it is possible, step sizes must usually depend on the strong convexity parameter, as in gradient descent [Nes03]. This is not the case here, where the step size is fixed by the mechanics of the problem. We have also shown that fast rates are possible without requiring strong convexity, with a gap condition on the gradient at an extreme point, more commonly associated with bandit problems.

We mention that several extensions of our models, motivated by heterogeneous estimations, are quite interesting but out of scope. For instance, assume an experimentalist can choose one of K known covariates X_i in order to estimate an unknown $\beta \in \mathbb{R}^K$, and observes $y_t = X_{\pi_t}^\top (\beta + \xi_t)$, where $\xi_t \sim \mathcal{N}(0, \Sigma)$. Variants of that problem with covariates or contexts [PR13] can also be considered. Assume for instance that $\mu_i(\cdot)$ and $\sigma_i^2(\cdot)$ are regular functions of covariates $\omega \in \mathbb{R}^d$. The objective is to estimate all the functions $\mu_i(\cdot)$.

REFERENCES

- [AB09] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2009.
- [ACBFS02] 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.

- [AD14] Shipra Agrawal and Nikhil R. Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation, EC '14*, pages 989–1006, New York, NY, USA, 2014. ACM.
- [ADL16] Shipra Agrawal, Nikhil R. Devanur, and Lihong Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2016.
- [AFH⁺11] Alekh Agarwal, Dean P. Foster, Daniel Hsu, Sham M. Kakade, and Alexander Rakhlin. Stochastic convex optimization with bandit feedback. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, 2011.
- [BC16] Quentin Berthet and Venkat Chandrasekaran. Resource allocation for statistical estimation. *Proceedings of the IEEE*, 104(1):115–125, 2016.
- [BCB12] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Machine Learning*, 5(1):1–122, 2012.
- [BEL16] Sébastien Bubeck, Ronen Eldan, and Yin Tat Lee. Kernel-based methods for bandit convex optimization. *CoRR*, abs/1607.03084, 2016.
- [BM13] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. In *Adv. NIPS*, 2013.
- [BP16] Francis Bach and Vianney Perchet. Highly-smooth zero-th order online optimization. *COLT 2016*, 2016.
- [CBL06] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- [CBLS06] Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Math. Oper. Res.*, 31(3):562–580, August 2006.
- [CLG⁺15] Alexandra Carpentier, Alessandro Lazaric, Mohammad Ghavamzadeh, Rémi Munos, and András Antos. Upper-confidence-bound algorithms for active learning in multi-armed bandits. *Preprint*, 2015.
- [Dip03] J. Dippon. Accelerated randomized stochastic optimization. *Ann. Statist.*, 31(4):1260–1281, 08 2003.
- [DR16] John Duchi and Feng Ruan. Local asymptotics for some stochastic optimization problems: Optimality, constraint identification, and dual averaging. *Arxiv Preprint*, 2016.
- [EDKMM09] E. Even-Dar, R. Kleinberg, S. Mannor, and Y. Mansour. Online learning for global cost functions. In *Proceedings of COLT*, 2009.
- [FW56] M Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Res. Logis. Quart.*, 3:95–110, 1956.
- [HAK07] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Mach. Learn.*, 69(2-3):169–192, 2007.
- [Haz12] Elad Hazan. The convex optimization approach to regret minimization. *Optimization for machine learning*, pages 287–303, 2012.

- [HKL14] E. Hazan, T. Koren, and K. Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Proc. Conference On Learning Theory (COLT)*, 2014.
- [HL14] E. Hazan and K. Levy. Bandit convex optimization: Towards tight bounds. In *Adv. NIPS*, 2014.
- [Jag11] Martin Jaggi. *Sparse Convex Optimization Methods for Machine Learning*. PhD thesis, ETH Zurich, 2011.
- [JNR12] Kevin G Jamieson, Robert Nowak, and Ben Recht. Query complexity of derivative-free optimization. *Advances in Neural Information Processing Systems*, 2012.
- [LJJ13] Simon Lacoste-Julien and Martin Jaggi. An affine invariant linear convergence analysis for frank-wolfe algorithms. *NIPS 2013*, 2013.
- [LWM15] Jean Lafond, Hoi-To Wai, and Eric Moulines. On the online frank-wolfe algorithms for convex and non-convex optimizations. *Arxiv Preprint*, 2015.
- [MCWG95] Andreu Mas-Colell, Michael Dennis Whinston, and Jerry R. Green. *Microeconomic theory*. Oxford University press, New York, 1995.
- [MPS14] Shie Mannor, Vianney Perchet, and Gilles Stoltz. Approachability in unknown games: Online learning meets multi-objective optimization. In *Proceedings of COLT*, 2014.
- [Nes03] Yuri Nesterov. *Introductory Lectures on Convex Optimization*. Springer, 2003.
- [PR13] Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *Ann. Statist.*, 41:693–721, 2013.
- [PRCS16] Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. *Ann. Statist.*, 44(2):660–681, 04 2016.
- [PT90] B. T. Polyak and A. B. Tsybakov. Optimal order of accuracy of search algorithms in stochastic optimization. *Problemy Peredachi Informatsii*, 26(2):45–53, 1990.
- [RST11] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Beyond regret. In Sham M. Kakade and Ulrike von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 559–594, Budapest, Hungary, 09–11 Jun 2011. PMLR.
- [Sha13] O. Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Proc. Conference on Learning Theory*, 2013.
- [SS11] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.
- [ST11] A. Saha and A. Tewari. Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

APPENDIX A: PROOFS

LEMMA 12. *Let L be a C -smooth convex function over the unit simplex Δ_K . For any $T \geq 1$, after T steps of the UCB Frank-Wolfe algorithm, it holds that*

$$L(p_T) - L(p_\star) \leq \frac{1}{T} \sum_{t=1}^T \varepsilon_t + \frac{C \log(eT)}{T},$$

where $\varepsilon_{t+1} = (e_{\pi_{t+1}} - e_{\star_{t+1}})^\top \nabla L(p_t)$ is the error compared to Frank-Wolfe with explicit, known and observed gradients, i.e., $e_{\star_{t+1}} = \operatorname{argmax}_{p \in \Delta_K} p^\top \nabla L(p_t)$.

Remark: If we denote by $\|L\|_\infty = \sup_{p \in \Delta_K} L(p)$ and $\|\nabla L\|_\infty = \sup_{p \in \Delta_K} \|\nabla L(p)\|$, then the same statements would hold with $(t+1)\|L\|_\infty$ or $2\|\nabla L\|_\infty + \|L\|_\infty$ instead of ε_t .

PROOF OF LEMMA 12. We apply the update equation in (2.1) and follow the usual analysis of Frank-Wolfe in the absence of noise. We denote by ρ_t the approximation error at any time t

$$\rho_{t+1} := L(p_{t+1}) - L(p_\star) = L(p_t + \frac{1}{t+1}(e_{\pi_{t+1}} - p_t)) - L(p_\star)$$

By definition of $e_{\star t}$, C -smoothness and finally convexity of L , we obtain

$$\begin{aligned} \rho_{t+1} &= L(p_t) - L(p_\star) + \frac{1}{t+1} \nabla L(p_t)^\top (e_{\star_{t+1}} - p_t) + \frac{C}{(t+1)^2} + \frac{1}{t+1} \nabla L(p_t)^\top (e_{\pi_{t+1}} - e_{\star_{t+1}}) \\ &\leq (1 - \frac{1}{t+1}) [L(p_t) - L(p_\star)] + \frac{1}{t+1} \nabla L(p_t)^\top (e_{\pi_{t+1}} - e_{\star_{t+1}}) + \frac{C}{(t+1)^2} \end{aligned}$$

Finally, introducing the notation ε_t and multiplying by $(t+1)$, we get

$$(t+1)\rho_{t+1} \leq t\rho_t + \varepsilon_{t+1} + \frac{C}{(t+1)}.$$

Summing from 1 to T yields the desired result. \square

PROOF OF THEOREM 3. We use the result of Lemma 12, which yields

$$L(p_T) - L(p_\star) \leq \frac{1}{T} \sum_{t=1}^T \varepsilon_t + \frac{C \log(eT)}{T}.$$

We recall that ε_t is the error due to the lack of information on the gradient at step t , and we have

$$\varepsilon_{t+1} := (e_{\pi_{t+1}} - e_{\star_{t+1}})^\top \nabla L(p_t) = \nabla_{\pi_{t+1}} L(p_t) - \nabla_{\star_{t+1}} L(p_t).$$

The difference between the two coefficients of the gradient can be controlled by using the definition of our selection rule, and the relationship between $\nabla L(p_t)$ and \hat{g}_t , similarly to the analysis of the UCB algorithm for multi-armed bandit

problems. Indeed, by definition of \hat{g}_t , we have that with probability at least $1 - \delta_t$, conditionally to the history

$$\begin{aligned} \nabla_{\pi_{t+1}} L(p_t) &\leq \hat{g}_{t, \pi_{t+1}} + \alpha_{t, \pi_{t+1}}(T_{\pi_{t+1}}(t), \delta_{t+1}) \\ &\leq (\hat{g}_{t, \pi_{t+1}} - \alpha_{t, \pi_{t+1}}(T_{\pi_{t+1}}(t), \delta_{t+1})) + 2\alpha_{t, \pi_{t+1}}(T_{\pi_{t+1}}(t), \delta_{t+1}) \\ &\leq (\hat{g}_{t, \star_{t+1}} - \alpha_{t, \star_{t+1}}(T_{\pi_{t+1}}(t), \delta_{t+1})) + 2\alpha_{t, \pi_{t+1}}(T_{\pi_{t+1}}(t), \delta_{t+1}) \\ &\leq \nabla_{\star_{t+1}} L(p_t) + 2\alpha_{t, \pi_{t+1}}(T_{\pi_{t+1}}(t), \delta_{t+1}), \end{aligned}$$

as $\hat{U}_{\pi_{t+1}} \leq \hat{U}_{\star_{t+1}}$ by the definition of our selection rule. With probability δ_{t+1} , we also have that $\varepsilon_{t+1} \leq 2\|\nabla L\|_\infty + \|L\|_\infty$. As a consequence, this yields

$$\mathbb{E}\varepsilon_{t+1} \leq 2\alpha_{t, \pi_{t+1}}(T_{\pi_{t+1}}(t), \delta_{t+1}) + \delta_{t+1}(2\|\nabla L\|_\infty + \|L\|_\infty).$$

We now bound the approximation error as a function of the precision of the estimate \hat{g}_t . Using the above inequality, we get, by denoting $\lambda := 2\|\nabla L\|_\infty + \|L\|_\infty$,

$$\begin{aligned} \mathbb{E} \sum_{t=1}^T \varepsilon_t &\leq K\lambda + \mathbb{E} \sum_{t=K+1}^T 2\alpha_{t, \pi_t}(T_{\pi_t}(t-1), \delta_t) + \frac{\lambda}{t^2} \leq 2\mathbb{E} \sum_{t=K+1}^T \sqrt{\frac{6 \log(t)}{T_{\pi_t}(t-1)}} + \left(K + \frac{\pi^2}{6}\right)\lambda \\ &\leq 2\mathbb{E} \sum_{i=1}^K \sum_{s=1}^{T_i(T-1)} \sqrt{\frac{6 \log(T)}{s}} + \left(K + \frac{\pi^2}{6}\right)\lambda \leq 4\mathbb{E} \sum_{i=1}^K \sqrt{3T_i(T) \log(T)} + \left(K + \frac{\pi^2}{6}\right)\lambda. \end{aligned}$$

We used the fact that the algorithm necessarily select actions in a round robin fashion during the first K stages.

Applying Cauchy-Schwarz inequality and the fact that $\sum_i T_i(t) = t$ yield the desired result. \square

PROOF OF PROPOSITION 4. We adapt the proof of Theorem 3, using that $C = 0$ and that $\varepsilon_t = 0$ whenever $\pi_t = \star_t = \star$. We obtained that

$$\mathbb{E}T(L(p_T) - L(p_\star)) \leq 4\mathbb{E} \sum_{i=1}^K \sqrt{3T_i(T) \log(T)} + \left(\frac{\pi^2}{6} + K\right)\lambda$$

However, in this particular case, we have $T(L(p_T) - L(p_\star)) = \sum_{i \neq \star} \Delta_i T_i$. We therefore obtain

$$\mathbb{E} \sum_{i \neq \star} \Delta_i T_i \leq 4\sqrt{3}\sqrt{\log(T)}\mathbb{E} \sum_{i \neq \star} \sqrt{T_i} + \left(\frac{\pi^2}{6} + K\right)\lambda \leq \left(\sum_{i \neq \star} \frac{48 \log(T)}{\Delta_i}\right)^{1/2} \mathbb{E} \left(\sum_{i \neq \star} \Delta_i T_i\right)^{1/2} + \left(\frac{\pi^2}{6} + K\right)\lambda,$$

by Cauchy-Schwarz inequality. Standard algebra yields

$$\mathbb{E}L(p_T) - L(p_\star) = \mathbb{E} \frac{1}{T} \sum_{i \neq \star} \Delta_i T_i \leq \frac{48 \log(T)}{T} \sum_{i \neq \star} \frac{1}{\Delta_i} + 2\left(\frac{\pi^2}{6} + K\right)\lambda$$

\square

OF PROPOSITION 4. We adapt again the proof of Theorem 3. First of all, notice that $\varepsilon_t \leq 0$ whenever $\pi_t = i_\star$, so that we obtain the following equation

$$\mathbb{E}T(L(p_T) - L(p_\star)) \leq 4\mathbb{E} \sum_{i \neq i_\star} \sqrt{3T_i(T) \log(T)} + \left(K + \frac{\pi^2}{6}\right)\lambda + C \log(eT).$$

Using the fact that L is Lipschitz and that $p_\star = e_{i_\star}$, it also holds that

$$T(L(p_T) - L(p_\star)) \geq T(p_T - p_\star)^\top \nabla L(p_\star) = \sum_{i \neq i_\star} T_i \Delta_i(L)$$

Cauchy-Schwartz inequality yields again that

$$\mathbb{E} \sum_{i \neq i_\star} T_i \Delta_i(L) \leq 48 \sum_{i \neq i_\star} \frac{\log(T)}{\Delta_i(L)} + 2 \left(K + \frac{\pi^2}{6} \right) \lambda + 2C \log(eT)$$

It remains to lower bound the lhs by the regret. Since L is C -smooth, we get also that

$$\begin{aligned} T(L(p_T) - L(p_\star)) &\leq T(p_T - p_\star)^\top \nabla L(p_\star) + CT \|p_T - p_\star\|^2 \\ &= \sum_{i \neq i_\star} T_i \Delta_i(L) + \frac{C}{T} \sum_{i \neq i_\star} T_i^2 + \frac{C}{T} \left(\sum_{i \neq i_\star} T_i \right)^2 \\ &\leq \sum_{i \neq i_\star} T_i \Delta_i(L) + \frac{CK}{T} \sum_{i \neq i_\star} T_i^2. \end{aligned}$$

As a consequence, it remains to compute a quantity γ such that

$$\mathbb{E} \sum_{i \neq i_\star} T_i \Delta_i(L) + \frac{CK}{T} \sum_{i \neq i_\star} T_i^2 \leq \gamma \mathbb{E} \sum_{i \neq i_\star} T_i \Delta_i(L)$$

or at least that for all $i \neq i_\star$

$$\frac{CK}{T} \mathbb{E} T_i^2 \leq (\gamma - 1) \mathbb{E} T_i \Delta_i(L),$$

in particular this is ensured for $\gamma = 1 + \frac{c(1+K)}{\min \Delta_i(L)}$ which gives the result. \square

PROOF OF THEOREMS 7. Recall that we assumed than on top of being smooth (C -Lipschitz gradient), the mapping L is μ -strongly convex and minimized in the relative interior of the simplex. Let η be the distance of p_\star to the relative boundary of the simplex then the following Lemma due to [LJJ13] yields

$$\nabla L(p_t)^\top (p_t - e_{\star_{t+1}}) \geq \sqrt{2\mu\eta^2} \sqrt{L(p_t) - L(p_\star)} \quad \text{and} \quad C \geq \mu\eta^2$$

This implies that

$$\begin{aligned} L(p_{t+1}) - L(p_\star) &= L(p_t) - L(p_\star) + \frac{1}{t+1} \nabla L(p_t)^\top (e_{\star_{t+1}} - p_t) + \frac{C}{(t+1)^2} + \frac{1}{t+1} \nabla L(p_t)^\top (e_{\pi_{t+1}} - e_{\star_{t+1}}) \\ &\leq L(p_t) - L(p_\star) - \frac{\sqrt{2\mu\eta^2}}{t+1} \sqrt{L(p_t) - L(p_\star)} + \frac{C}{(t+1)^2} + \frac{\varepsilon_{t+1}}{t+1} \end{aligned}$$

To ease up reading, we introduce the notations, $\alpha = \sqrt{2\mu\eta^2}$ and and $\rho_t = L(p_t) - L(p_\star)$ so that the previous equation rewrites in

$$\rho_{t+1} \leq \rho_t - \alpha \frac{\sqrt{\rho_t}}{t+1} + \frac{C}{(t+1)^2} + \frac{\varepsilon_{t+1}}{t+1},$$

which rewrites again, using the function $\psi(x) = x^2 - \alpha x$, into

$$(t+1)\rho_{t+1} \leq t\rho_t + \left[\psi(\sqrt{\rho_t}) - \psi\left(\frac{\varepsilon_{t+1}}{\alpha}\right) \right] + \frac{\varepsilon_{t+1}^2}{\alpha^2} + \frac{C}{t+1}.$$

Recall that we still have the guarantee that $\rho_t \leq \frac{\sum_{s=1}^t \varepsilon_s + \frac{C}{s+1}}{t}$, but we aim at proving some fast rates of convergence, of the type

$$\mathbb{E} \rho_T \leq O\left(\frac{\sum_s \mathbb{E} \varepsilon_s^2}{T}\right).$$

Assume for the moment that $\rho_T \geq \frac{\alpha^2}{4}$, then Cauchy-Schwarz inequality implies that

$$\begin{aligned} \left(\sum_{s=1}^T \varepsilon_s + \frac{C}{s+1}\right)^2 &\leq T \sum_{s=1}^T \left(\varepsilon_s + \frac{C}{s+1}\right)^2 \leq \frac{4}{\alpha^2} \sum_{s=1}^T \left(\varepsilon_s + \frac{C}{s+1}\right) \sum_{s=1}^T \left(\varepsilon_s + \frac{C}{s+1}\right)^2 \\ &\leq \sum_{s=1}^T \left(\varepsilon_s + \frac{C}{s+1}\right) \frac{8}{\alpha^2} \left(\sum_{s=1}^T \varepsilon_s^2 + \frac{C^2 \pi^2}{6}\right), \end{aligned}$$

and thus

$$(A.1) \quad \rho_T \leq \frac{\sum_{s=1}^T \varepsilon_s + \frac{C}{s+1}}{T} \leq \frac{8}{\alpha^2} \frac{\sum_{s=1}^T \varepsilon_s^2}{T} + \frac{14C^2}{\alpha^2} \frac{1}{T}$$

As a consequence, the claim holds if $\rho_T \geq \alpha^2/4$ and we will, from now on, assume that $\rho_T \leq \alpha^2/4$.

We denote by τ_0 the last time before T where $\rho_\tau \geq \alpha^2/4$ and we now consider several cases for the remaining of the proof.

Case1. If we can prove that $\frac{\sum_{s=1}^t \varepsilon_s^2}{t} \geq \varepsilon_{t+1}^2$, for example if ε_t is guaranteed to decrease

Then, for any $t \geq \tau_0$, we get that if $\rho_t \geq \frac{1}{\alpha^2} \frac{\sum_{s=1}^t \varepsilon_s^2}{t} \geq \frac{\varepsilon_{t+1}}{\alpha^2}$ (by assumption), then

$$(t+1)\rho_{t+1} \leq t\rho_t + \frac{\varepsilon_{t+1}^2}{\alpha^2} + \frac{C}{t+1}.$$

Thus, if we denote by τ_1 the last time where $\rho_\tau < \frac{1}{\alpha^2} \frac{\sum_{s=1}^\tau \varepsilon_s^2}{\tau}$, we obtain that, as long as $\tau_1 \geq \tau_0$,

$$\begin{aligned} T\rho_T &\leq \tau_1 \rho_{\tau_1} + \varepsilon_{\tau_1+1} + \frac{1}{\alpha^2} \sum_{s=\tau_1+2}^T \varepsilon_s^2 + \frac{C}{s+1} \\ &\leq \frac{1}{\alpha^2} \sum_{s=1}^T \varepsilon_s^2 + \varepsilon_{\tau_1+1} - \frac{\varepsilon_{\tau_1+1}^2}{\alpha^2} + C \log(eT) \end{aligned}$$

which gives the result we wanted as

$$(A.2) \quad T\rho_T \leq \frac{1}{\alpha^2} \sum_{s=1}^T \varepsilon_s^2 + \frac{\alpha^2}{4} + C \log(eT)$$

On the contrary, if $\tau_0 \geq \tau_1$, then the same computations give

$$T\rho_T \leq \tau_0\rho_{\tau_0} + \frac{\alpha^2}{4} + \frac{1}{\alpha^2} \sum_{s=\tau_0+1}^T \varepsilon_s^2 + \frac{C}{s+1}$$

Using the fact that $\delta_{\tau_0} \geq \alpha^2/4$, we also have that

$$\tau_0\delta_{\tau_0} \leq \frac{8}{\alpha^2} \sum_{s=1}^{\tau_0} \varepsilon_s^2 + \frac{14C^2}{\alpha^2}$$

thus, combining the two cases $\tau_1 \geq \tau_0$ and $\tau_0 \leq \tau_1$, we now obtain that

$$(A.3) \quad T\rho_T \leq \frac{8}{\alpha^2} \sum_{s=1}^T \varepsilon_s^2 + \frac{14C^2}{\alpha^2} + \frac{\alpha^2}{4} + C \log(eT)$$

Case 2. If it is not necessarily true that $\frac{\sum_{s=1}^t \varepsilon_s^2}{t} \geq \varepsilon_t^2$, for example if ε_t does not necessarily decrease or can make big jumps

Notice first that if $\frac{\varepsilon_t^2}{\alpha^2} \leq \rho_t \leq \frac{\alpha^2}{4}$, the latter holding because of $t \geq \tau_0$, then one has

$$(t+1)\rho_{t+1} \leq t\rho_t + \frac{\varepsilon_{t+1}^2}{\alpha^2} + \frac{C}{t+1} .$$

As a consequence, denoting by τ_2 the last stage before T such that $\rho_\tau < \frac{\varepsilon_\tau^2}{\alpha^2}$ and assuming that $\tau_2 \geq \tau_0$, we obtain following the same computations as before that

$$(A.4) \quad t\rho_t \leq \frac{\tau_2\varepsilon_{\tau_2}^2}{\alpha^2} + \frac{1}{\alpha^2} \sum_{s=\tau_2+1}^t \varepsilon_s^2 + \frac{\alpha^2}{4} + C \log(et) .$$

If $\tau_2 \leq \tau_0$, then we get that

$$T\rho_T \leq \tau_0\rho_{\tau_0} + \frac{1}{\alpha^2} \sum_{s=\tau_0+1}^T \varepsilon_s^2 + \frac{\alpha^2}{4} + C \log(eT)$$

thus

$$(A.5) \quad T\rho_T \leq \frac{8}{\alpha^2} \sum_{s=1}^T \varepsilon_s^2 + \frac{14C^2}{\alpha^2} + \frac{\alpha^2}{4} + C \log(eT),$$

which was our objective. Hence it only remains to upper-bound $\tau_2\varepsilon_{\tau_2}^2$ in Equation (A.4), i.e., when $\tau_2 \geq \tau_0$. To do that, we are going to use a second time the assumptions on L .

Since we assumed that L was μ -strongly convex and minimized in the interior of the simplex, it holds that

$$\|p_t - p_*\|^2 \leq \frac{1}{\mu} (L(p_t) - L(p_*)) \leq \frac{1}{\mu} \frac{\sum_{s=1}^t \varepsilon_s}{t}$$

As a consequence, this yields that

$$p_*^i - \sqrt{\frac{1}{\mu} \frac{\sum_{s=1}^t \varepsilon_s}{t}} \leq p_t^i \leq p_*^i + \sqrt{\frac{1}{\mu} \frac{\sum_{s=1}^t \varepsilon_s}{t}}$$

We are now going to make the assumption that the horizon T is known in advance, and that $\varepsilon_s \leq \left(\frac{\log(T/\delta)}{T\tau_s(s-1)}\right)^\beta$ with probability at least $1 - \delta^\gamma$, for some $\beta \leq 1/2$ and $\gamma > 0$. This implies, by the union bound, that with probability at least $1 - TK\delta^\gamma$,

$$\frac{1}{t} \sum_{s=1}^t \varepsilon_s \leq \frac{1}{1-\beta} \left(\frac{K \log(T/\delta)}{t}\right)^\beta,$$

hence

$$tp_*^i - t\sqrt{\frac{1}{\mu} \frac{1}{1-\beta} \left(\frac{K \log(T/\delta)}{t}\right)^\beta} \leq T_i(t) \leq tp_*^i + t\sqrt{\frac{1}{\mu} \frac{1}{1-\beta} \left(\frac{K \log(T/\delta)}{t}\right)^\beta}$$

in particular, if $\frac{K \log(T/\delta)}{t} \leq \left(\mu(1-\beta)\frac{\eta^2}{4}\right)^{1/\beta}$, i.e., if $t \geq \tau_\delta := \frac{K \log(T/\delta)}{(\mu(1-\beta)\frac{\eta^2}{4})^{1/\beta}}$,

$$\frac{t\delta}{2} \leq \frac{tp_*^i}{2} \leq T_i(t) \leq \frac{3tp_*^i}{2}$$

and thus,

$$t\varepsilon_t^2 \leq t\left(\frac{\log(T/\delta)}{t\eta/2}\right)^{2\beta} \leq \left(\frac{2\log(T/\delta)}{\eta}\right)^{2\beta} t^{1-2\beta} \leq \left(\frac{2\log(T/\delta)}{\eta}\right)^{2\beta} T^{1-2\beta}, \quad \forall t \geq \tau_\delta.$$

Concluding.

To wrap things up, we consider the three different cases. With probability at least $1 - TK\delta^\gamma$,

If $\tau_2 \leq \tau_0$ then: we have proved that

$$T\rho_T \leq \frac{8}{\alpha^2} \sum_{s=1}^T \varepsilon_s^2 + \frac{\alpha^2}{4} + C \log(eT)$$

If $\tau_0 \leq \tau_\delta \leq \tau_2$ then: using the above upper-bound on $\tau_2\varepsilon_{\tau_2}^2$, we get

$$T\rho_T \leq \frac{1}{\alpha^2} \left(\frac{2\log(T/\delta)}{\eta}\right)^{2\beta} T^{1-2\beta} + \frac{1}{\alpha^2} \sum_{s=1}^T \varepsilon_s^2 + \frac{\alpha^2}{4} + C \log(eT).$$

If $\tau_0 \leq \tau_2 \leq \tau_\delta$ then: going back to the original induction yields

$$T\rho_T \leq \tau_\delta \rho_{\tau_\delta} + \frac{1}{\alpha^2} \sum_{s=1}^T \varepsilon_s^2 + \frac{\alpha^2}{4} + C \log(eT).$$

Taking the maximum of all those terms gives that, with probability at least $1 - TK\delta^\gamma$,

(A.6)

$$\rho_T \leq \frac{\log(T/\delta)}{T} \frac{K\|L\|_\infty}{(\mu(1-\beta)\frac{\eta^2}{4})^{1/\beta}} + \frac{1}{\alpha^2} \left(\frac{2\log(T/\delta)}{\eta T}\right)^{2\beta} + \frac{8}{\alpha^2} \frac{1}{T} \sum_{s=1}^T \varepsilon_s^2 + \frac{\alpha^2}{4T} + \frac{C \log(eT)}{T}$$

A simple summation over t yields that,

$$\frac{1}{T} \sum \varepsilon_s^2 \leq \frac{1}{T} \sum \left(\frac{\log(T/\delta)}{T_i(s)} \right)^{2\beta} \leq \frac{1}{1-2\beta} \left(\frac{K \log(T/\delta)}{T} \right)^{2\beta} \quad \text{if } \beta < \frac{1}{2}$$

and

$$\frac{1}{T} \sum \varepsilon_s^2 \leq \frac{1}{T} \sum \frac{\log(T/\delta)}{T_i(s)} \leq \frac{K \log(T/\delta) \log(T)}{T} \quad \text{if } \beta = \frac{1}{2}$$

As a consequence, if $\beta < 1/2$

$$\mathbb{E} \rho_T \leq \delta^\gamma T K \|L\|_\infty + \frac{\log(T/\delta)}{T} \frac{K \|L\|_\infty}{(\mu(1-\beta)\frac{\eta^2}{4})^{1/\beta}} + \frac{1}{\alpha^2} \left(\frac{\log(T/\delta)}{T} \right)^{2\beta} \left[\frac{2^{2\beta}}{\eta^{2\beta}} + \frac{8K^{2\beta}}{1-2\beta} \right] + \frac{\alpha^2}{4T} + \frac{C \log(eT)}{T}$$

and if $\beta = 1/2$

$$\mathbb{E} \rho_T \leq \delta^\gamma T K \|L\|_\infty + \frac{\log(T/\delta)}{T} \left[\frac{K \|L\|_\infty}{(\mu\frac{\eta^2}{8})^2} + \frac{2}{\alpha^2 \eta} \right] + \frac{1}{\alpha^2} \frac{K \log(T/\delta) \log(T)}{T} + \frac{\alpha^2}{4T} + \frac{C \log(eT)}{T}$$

choosing $\delta^\gamma = T^{-(2\beta+1)}$ yields that, if $\beta < \frac{1}{2}$,

$$\mathbb{E} L(p_T) - L(p^*) \leq c_{1,\beta} \frac{\log(T)}{T^{2\beta}} + c_{2,\beta} \left(\frac{\log(T)}{T} \right)^{2\beta} + \frac{c_{3,\beta}}{T^{2\beta}}$$

where

$$c_{1,\beta} = \frac{2(\beta+1)K \|L\|_\infty}{\gamma(\mu(1-\beta)\frac{\eta^2}{4})^{1/\beta}} + C, \quad c_{2,\beta} = \frac{1}{\alpha^2} \left(\frac{2(\beta+1)}{\gamma} \right)^{2\beta} \left[\frac{2^{2\beta}}{\eta^{2\beta}} + \frac{8K^{2\beta}}{1-2\beta} \right], \quad c_{3,\beta} = K \|L\|_\infty + \frac{\alpha^2}{4} + C.$$

For $\beta = 1/2$, the choice of $\delta^\gamma = T^{-2}$ yields

$$\mathbb{E} L(p_T) - L(p^*) \leq c_1 \frac{\log^2(T)}{T} + c_2 \frac{\log(T)}{T} + c_3 \frac{1}{T}$$

where

$$c_1 = \frac{3K}{\gamma\alpha^2}, \quad c_2 = \frac{3}{\gamma\alpha^2} \left[\frac{K \|L\|_\infty}{(\mu\frac{\eta^2}{8})^2} + \frac{2}{\alpha^2 \eta} \right] + C, \quad c_3 = K \|L\|_\infty + \frac{\alpha^2}{4} + C.$$

Remark: We assumed that the horizon T was known. If it is not the case, there are two possible ways to deal with that issue to get an anytime algorithm

Use the Doubling Trick in the algorithm: The doubling trick is rather classical in online learning, and it consists in running several successive and independent instances of the same algorithm on block of stages of length that increases sufficiently fast enough (so that the error incurred on the first blocks disappears while averaging), but not too fast enough (so that the error during the last block is compensated by the small error cumulated so far on the previous blocks). Its main advantages are that it is simple to describe, to analyze and that it gives the same guarantees of the known horizon, up to some multiplicative constant. The latter depends on the speed of convergence achieved in the known horizon, and it might require careful tuning. The main drawback of the doubling trick is that it regularly discards all the past data and forgets the learning done so far.

In our setting, the correct size of blocks are proportional to $T_j = e^{(\frac{1}{1-\beta})^j}$.

Use the Doubling Trick in the analysis. Instead of using the doubling trick in the algorithm, we will prove in the following that we can somehow use it in the analysis of the anytime variant of the algorithm. We first consider the case where $\beta = 1/2$, and we assume that it holds that, for some fixed $\theta > 0$ and for every $s \in \mathbb{N}$, $\varepsilon_s \leq \left(\frac{\theta \log(s)}{T_i(s)}\right)^\beta$ with probability at least $1 - \frac{1}{s^6}$. The immediate consequence of that property is that

$$\frac{1}{T} \mathbb{E} \sum_{s=1}^T \varepsilon_s^2 \leq \frac{1}{T} \mathbb{E} \sum_{s=1}^T \left(\frac{\theta \log(T)}{T_i(s)}\right)^{2\beta} + \frac{1}{s^6} \leq \frac{2}{1-2\beta} \left(\frac{K\theta \log(T)}{T}\right)^{2\beta} \frac{1}{T}$$

where the last inequality is loose for $\beta < 1/2$ and

$$\frac{1}{T} \mathbb{E} \sum_{s=1}^T \varepsilon_s^2 \leq \frac{1}{T} \mathbb{E} \sum_{s=1}^T \frac{\theta \log(T)}{T_i(s)} + \frac{1}{s^6} \leq 2 \frac{K\theta \log^2(T)}{T}$$

for $\beta = 1/2$.

In order to upper-bound $\mathbb{E} \tau_2 \varepsilon_{\tau_2}^2$, we are going to decompose the set of stages in blocks $\mathcal{B}_j = \{t \in [T_{j-1} + 1, T_j]\}$ where $T_j = \lfloor e^{(\frac{1}{1-\beta})^j} \rfloor$. As a consequence:

$$\mathbb{P}\left\{\forall k \leq K, \forall s \in \mathcal{B}_j, \varepsilon_s^k \leq \left(\frac{\theta \log(s)}{T_k(s)}\right)^\beta\right\} \geq 1 - K \sum_{s \in \mathcal{B}_j} \frac{1}{s^5} =: 1 - p_j.$$

Hence, with probability at least $1 - (p_j + p_{j+1})$, it holds that for all $t \in \mathcal{B}_{j+1}$

$$\frac{1}{t} \sum_{s=1}^t \varepsilon_s \leq \frac{T_{j-1}}{t} + \frac{1}{t} \sum_{s=T_{j-1}+1}^t \left(\frac{\theta \log(s)}{T_i(s)}\right)^\beta \leq \frac{1}{t^\beta} + \frac{2}{1-\beta} \left(\frac{K\theta \log(t)}{t}\right)^\beta,$$

since $t \geq T_j + 1 \geq T_{j-1}^{\frac{1}{1-\beta}}$.

Following the same argument as in the case where the horizon was known, this implies that with probability at least $1 - (p_i + p_{i+1})$, for all $t \in \mathcal{B}_{i+1}$,

$$T_t^i \geq t p_\star^i - t \sqrt{\frac{1}{\mu} \left(\frac{1}{t^\beta} + \frac{2}{1-\beta} \left(\frac{K\theta \log(t)}{t}\right)^\beta\right)}.$$

In particular, let $\tau_{\beta, \star}$ be such that $\sqrt{\frac{1}{\mu} \left(\frac{1}{t^\beta} + \frac{2}{1-\beta} \left(\frac{K\theta \log(t)}{t}\right)^\beta\right)} \leq \frac{\eta}{2}$ for all $t \geq \tau_{\beta, \star}$ and $j_{\beta, \star}$ be the index of the block to which $\tau_{\beta, \star}$ belongs. Then we have that

$$\forall j \geq j_{\beta, \star}, \mathbb{P}\left\{\forall t \in \mathcal{B}_{j+1}, t \varepsilon_t^2 \leq \left(\frac{2\theta \log(t)}{\delta}\right)^{2\beta} t^{1-2\beta}\right\} \geq 1 - (p_j + p_{j+1})$$

It follows that

$$\mathbb{E} \tau_2 \varepsilon_{\tau_2}^2 \mathbf{1}\{\tau_2 \geq T_{j^\star}\} \leq \left(\frac{2\theta \log(t)}{\eta}\right)^{2\beta} t^{1-2\beta} + \sum_{j=j^\star} T_{j+1} (p_j + p_{j+1}) \leq 2 \left(\frac{2\theta \log(t)}{\eta}\right)^{2\beta} t^{1-2\beta},$$

where, again, the last inequality is loose but compact. This yields the any-time version of the previous theorem, that, for $\beta < 1/2$

$$\forall t \in \mathbb{N}, \quad \mathbb{E}L(p_t) - L(p^*) \leq c'_{1,\beta} \left(\frac{\log(t)}{t} \right)^{2\beta} + c'_{2,\beta} \frac{\log(t)}{t} + c'_{3,\beta} \frac{1}{t}$$

with $c'_{1,\beta} = \frac{2}{\alpha^2} \left(\frac{2\theta}{\eta} \right)^{2\beta} + \frac{8}{\alpha^2} \frac{2}{1-2\beta} (K\theta)^{2\beta}$, $c'_{2,\beta} = C$ and $c'_{3,\beta} = T_{j_{\beta,*}} \|L\|_\infty + \frac{\alpha^2}{4} + C$.

For $\beta = 1/2$, we get

$$\forall t \in \mathbb{N}, \quad \mathbb{E}L(p_t) - L(p^*) \leq c'_1 \frac{\log^2(t)}{t} + c'_2 \frac{\log(t)}{t} + c'_3 \frac{1}{t},$$

where $c'_1 = \frac{16K\theta}{\alpha^2}$, $c'_2 = \frac{4\theta}{\eta\alpha^2} + C$ and $c'_3 = T_{j_{1/2,*}} \|L\|_\infty + \frac{\alpha^2}{4} + C$.

□

LEMMA 13. *Let Z_s , $s \in \{1, \dots, T\}$ be i.i.d. random variable in $[0, 1]$ of expectation $\mathbb{E}Z_s = Z$, then, with probability at least $1 - \delta$, $Z \geq \bar{Z}_\tau/2$ where the random stage $\tau \leq T$ is the first such that $\bar{Z}_\tau \geq \sqrt{2 \log(2T/\delta)}/\tau$. As, it also holds that $\bar{Z}_\tau \geq Z - \sqrt{\frac{2 \log(T/\delta)}{2t}}$, thus $3\bar{Z}_\tau/2 \geq Z$, we get that*

$$\bar{Z}_\tau/2 \leq Z \leq 3\bar{Z}_\tau/2, \quad \text{for some random } \tau \leq 9 \log(2T/\delta)/(2Z^2) + 1$$

This lemma is a direct consequence of Hoeffding's inequality.

PROOF OF THEOREM 9. Let $\nu \in (0, 1/29)$, $K > 64 \log(2)/\nu$ and $T > 4\nu^2 K^4$. We assume for simplicity that K is even. For $\theta \in \Delta_K$, we consider $L_\theta(p) = \frac{\mu}{2} \|p - \theta\|^2$. We treat first the case of $\mu = 1$. For all $\varepsilon \in \{-1, 1\}^{K/2}$, we consider the vector θ_ε such that for all $i \in [K/2]$

$$\theta_{\varepsilon, 2i-1} = \frac{1}{K} + \varepsilon_i \sqrt{\frac{\nu K}{T}} \quad \text{and} \quad \theta_{\varepsilon, 2i} = \frac{1}{K} - \varepsilon_i \sqrt{\frac{\nu K}{T}}.$$

Note that for all $\varepsilon \in \{-1, 1\}^{K/2}$, $p_\varepsilon^* = \theta_\varepsilon \in \Delta_K$ and that $\nabla L_\theta(p) = p - \theta$, so that an observation from $\mathcal{N}(\theta_i, 1)$ for the i -th action constitutes a bandit feedback for the i -th coefficient of the gradient with deviation bound $\alpha(T_i, \delta) = \sqrt{2 \log(1/\delta)/T_i}$.

Let \mathcal{M} be a subset of $\{-1, 1\}^{K/2}$ such that for all $\varepsilon, \varepsilon' \in \mathcal{M}$, we have $\rho(\varepsilon, \varepsilon') \geq K/8$ and for which $\log(|\mathcal{M}|) \geq K/64$, whose existence is guaranteed by the Varshamov-Gilbert lemma. We have for $\varepsilon, \varepsilon' \in \mathcal{M}$ that $\nu/4 \cdot K^2 T \leq \|\theta_\varepsilon - \theta_{\varepsilon'}\|_2^2 \leq \nu K^2/T$. We consider the subsets \mathcal{C}_ε of the unit simplex defined by

$$\mathcal{C}_\varepsilon = \left\{ p \in \Delta_K : \|p - \theta_\varepsilon\|_2^2 < \frac{\nu}{16} \frac{K^2}{T} \right\}.$$

By construction of \mathcal{M} , these sets are disjoint.

For any algorithm, on the events where $T_j(T) > 2T/K > T/K + 2\sqrt{\nu K^2 T}$ for some $j \in K$ we have that $\hat{p}_T \notin \mathcal{C}_\varepsilon$. On the events for which $T_j(T) \leq 2T/K$ for all $j \in [K]$, we have that \hat{p}_T can only depend (possibly in a random manner) on an observation from $\mathcal{N}^{\otimes N}(\theta_\varepsilon, I_K)$, where $N \leq 2T/K$. We have that

$\text{KL}(\mathcal{N}^{\otimes N}(\theta_\varepsilon, I_K), \mathcal{N}^{\otimes N}(\theta_{\varepsilon'}, I_K)) = N\|\theta_\varepsilon - \theta_{\varepsilon'}\|_2^2 \leq \nu NK^2/T \leq 2\nu K$. Considering together these two events, we obtain as a consequence of Fano's inequality that

$$\inf_{\hat{p}} \max_{\varepsilon \in \mathcal{M}} \mathbf{P}_\varepsilon(\hat{p}_T \notin \mathcal{C}_\varepsilon) \geq 1 - \frac{2\nu K + \log(2)}{K/64} \geq 1 - 129\nu,$$

As a consequence, we have that

$$\inf_{\hat{p}} \max_{\varepsilon \in \mathcal{M}} \left\{ \mathbb{E}[L_{\theta_\varepsilon}(\hat{p}_T)] - L_{\theta_\varepsilon}(p_\varepsilon^*) \right\} \geq \frac{1}{2}(1 - 129\nu) \frac{\nu K^2}{16 T},$$

which yields the desired result. □

PROOF OF PROPOSITION 11. For $\theta \in [1/3, 2/3]$, take the class of functions $L_\theta : \mathbb{R}^3 \rightarrow \mathbb{R}$

$$L_\theta(p) = \frac{1}{2}(p_1 - \theta)^2 + \frac{1}{2}(p_2 - (1 - \theta))^2 + \frac{1}{2}p_3^2.$$

Consider the case where the mixed feedback for the three actions are drawings from respectively $\mathcal{N}(0, 1)$, $\mathcal{N}(0, 1)$, and $\mathcal{N}(\theta, 1)$. We consider the set

$$\mathcal{C}_\theta = \left\{ p \in \Delta_3 : \|p - p_\theta^*\|_2^2 \leq \frac{c}{T^{2/3}} \right\}.$$

For any algorithm, on the event where $T_3(T) > T^{2/3}$, we have $\|p_T - p_\theta^*\|_2^2 \geq 1/T^{2/3}$ and $p_T \notin \mathcal{C}_\theta$. On the event where $T_3 \leq T^{2/3}$, we have that \hat{p}_T can only depend on a drawing from $\mathcal{N}^{\otimes N}(\theta, 1)$, where $N \leq T^{2/3}$. In this case, we have that

$$\inf_{\hat{p}} \sup_{\theta \in [1/3, 2/3]} \mathbb{E}_\theta[(\hat{p}_{T,1} - \theta)^2] \geq \frac{c'}{N}.$$

Overall this yields the desired result. □

QUENTIN BERTHET
 DEPARTMENT OF PURE MATHEMATICS
 AND MATHEMATICAL STATISTICS,
 UNIVERSITY OF CAMBRIDGE
 WILBEFORCE ROAD
 CAMBRIDGE, CB3 0WB, UK
 (q.berthet@statslab.cam.ac.uk)

VIANNEY PERCHET
 CMLA
 ENS PARIS-SACLAY
 61, AVENUE DU PRÉSIDENT WILSON
 94230 CACHAN, FRANCE
 (vianney.perchet@normalesup.org)