Regularized Contextual Bandits

Xavier Fontaine* Quentin Berthet† Vianney Perchet‡

Abstract

We consider the stochastic contextual bandit problem with additional regularization. The motivation comes from problems where the policy of the agent must be close to some baseline policy which is known to perform well on the task. To tackle this problem we use a nonparametric model and propose an algorithm splitting the context space into bins, and solving simultaneously — and independently — regularized multi-armed bandit instances on each bin. We derive slow and fast rates of convergence, depending on the unknown complexity of the problem. We also consider a new relevant margin condition to get problem-independent convergence rates, ending up in intermediate convergence rates interpolating between the aforementioned slow and fast rates.

1 Introduction and Related Work

In sequential optimization problems, an agent takes successive decisions in order to minimize an unknown loss function. An important class of such problems, nowadays known as bandit problems, has been mathematically formalized by Robbins in his seminal paper (Robbins, 1952). In the so-called stochastic multi-armed bandit problem, an agent chooses to sample (or "pull") among K arms returning random rewards. Only the rewards of the selected arms are revealed to the agent who does not get any additional feedback. Bandits problems naturally model the exploration/exploitation trade-offs which arise in sequential decision making under uncertainty. Various general algorithms have been proposed to solve this problem, following the work of Lai and Robbins (1985) who obtain a logarithmic regret for their sample-mean based policy. Further bounds have been obtained by Agrawal (1995) and Auer et al. (2002) who developed different versions of the well-known UCB algorithm.

The setting of classical stochastic multi-armed bandits is unfortunately too restrictive for real-world applications. The choice of the agent can and should be influenced by additional information (referred to as "context" or "covariate") that is revealed by the environment. It encodes features having an impact on the arms' rewards. For instance, in online advertising, the expected Click-Through-Rate depends on the identity, the profile and the browsing history of the customer. These problems of bandits with covariates have been initially introduced by Woodroofe (1979) and have attracted much attention since (Wang et al., 2005; Goldenshluger et al., 2009). This particular class of bandits problems is now known under the name of contextual bandits following Langford and Zhang (2008).

*CMLA, ENS Cachan, CNRS, Université Paris-Saclay — Email: fontaine@cmla.ens-cachan.fr

†Statistical Laboratory, DPMMS, University of Cambridge — Email: q.berthet@statslab.cam.ac.uk

‡CMLA, ENS Cachan, CNRS, Université Paris-Saclay & Criteo AI Lab, Paris — Email: perchet@cmla.ens-cachan.fr

Contextual bandits have been extensively studied in the last decades and several improvements upon multi-armed bandits algorithms have been applied to contextual bandits, including Thompson sampling (Agrawal and Goyal, 2013), Explore-Then-Commit strategies (Perchet and Rigollet, 2013), and policy elimination (Dudik et al., 2011). Contextual bandits are quite intricate to study as they borrow aspects from both supervised learning and reinforcement learning. Indeed they use features to encode the context variables, as in supervised learning but also require an exploration phase to discover all the possible choices, as in reinforcement learning algorithms. Applications of contextual bandits are numerous, ranging from online advertising (Tang et al., 2013), to news articles recommendation (Li et al., 2010) or decision-making in the health and medicine sectors (Tewari and Murphy, 2017; Bastani and Bayati, 2015).

Among the general class of stochastic multi-armed bandits different settings can be studied. One natural hypothesis that can be made is to consider that the arms' rewards are regular functions of the context, *i.e.* two close context values have similar expected rewards. This setting has been studied in Slivkins (2014) and Perchet and Rigollet (2013). A possible approach to this problem is to take inspiration from the regressograms used in nonparametric estimation (Tsybakov, 2008) and to divide the context space into several bins. This technique also used in online learning (Hazan and Megiddo, 2007) leads to the concept of UCBograms (Rigollet and Zeevi, 2010) in bandits.

In this article, we introduce regularization to the problem of stochastic multi-armed bandits. Regularization is a widely-used technique in machine learning to avoid overfitting or to solve ill-posed problems. Here, the regularization will force the solution of the contextual bandits problem to be close to an existing known policy. As an example of motivation, an online-advertiser or any decision-maker may wish not to diverge too much from a handcrafted policy that is known to perform relatively well. This has already motivated previous work such as Conservative Bandits (Wu et al., 2016). By adding regularization, typically under the form of a Kullback-Leibler divergence, the agent can be sure to end up close to the chosen policy. Within this regularized setting, the form of the objective function is not a classical bandit loss anymore, but contains a regularization term on the global policy. We fall therefore in the more general setting of online optimization and borrow tools from this field to build and analyze an algorithm on contextual multi-armed bandits. As a substitute of the UCB algorithm we use the recently introduced UCB-Frank Wolfe algorithm (Berthet and Perchet, 2017). Our main contribution consists in an algorithm with proven slow or fast rates of convergence, depending on the unknown complexity of the problem at hand. These rates are better than the ones obtained for classical nonparametric contextual bandits. We also derive intermediate convergence rates, independent of the parameters of the problems with relatively mild assumptions.

The remaining of this paper is organized as follows. We present the settings and the problem in Section 2. Our algorithm is described in Section 3. Sections 4 and 5 are devoted to establishing the convergence rates. Lower bounds are detailed in Section 6 and we present the results of experiments in Section 7. Section 8 concludes the paper.

2 Problem Setting and Definitions

2.1 Problem Description

We consider a stochastic contextual bandits problem with $K \in \mathbb{N}^*$ arms and a time horizon T. It is defined as follows. At each time $t \in \{1, \dots, T\}$, Nature draws a context variable

$X_t \in \mathcal{X} = [0,1]^d$ uniformly at random. This context is revealed to an agent who chooses an arm π_t amongst the K arms. Only the loss $Y_t^{(\pi_t)} \in [0,1]$ is revealed to the agent.

For each arm $k \in \{1, \dots, K\}$ we note $\mu_k(X) = \mathbb{E}(Y^{(k)}|X)$ the conditional expectation of the arm's loss given the context. We impose classical regularity assumptions on the functions μ_k borrowed from nonparametric estimation. Namely we suppose that the functions μ_k are (β, L_β)-Hölder, with $\beta \in (0,1]$. We note $\mathcal{H}_{\beta,L_\beta}$ this class of functions.

Assumption 1 (Smoothness)**.** *For all $k \in \{1, \dots, K\}$,*

$$\forall x, y \in \mathcal{X},\ |\mu_k(x) - \mu_k(y)| \leq L_\beta \|x - y\|_2^\beta .$$

We denote by $p : \mathcal{X} \to \Delta^K$ the proportion function of each arm (also called occupation measure), where Δ^K is the unit simplex of $\mathbb{R}^K$. In classical stochastic contextual bandits the goal of the agent is to minimize the following loss function

$$L(p) = \int_{\mathcal{X}} \langle \mu(x), p(x) \rangle \, \mathrm{d}x.$$

We consider an additional regularization term representing the constraint on the optimal proportion function $p^\star$. For example we may want to encourage $p^\star$ to be close to a chosen proportion function q, or to be far from the boundary of the simplex Δ^K. So we consider a convex regularization function $\rho : \Delta^K \times \mathcal{X} \to \mathbb{R}$, and a regularization parameter $\lambda : \mathcal{X} \to \mathbb{R}$. Both ρ and λ are assumed to be known and given to the agent, while the μ_k functions are unknown and must be learned.

We are interested in minimizing the loss function

$$L(p) = \int_{\mathcal{X}} \langle \mu(x), p(x) \rangle + \lambda(x) \rho(p(x), x) \, \mathrm{d}x.$$

This is the most general form of the loss function. We study first the case where the regularization does not depend on the context (*i.e.* when λ is a constant and when ρ is only a function of p).

The function λ modulates the weight of the regularization and is chosen to be regular enough. More precisely we make the following assumption.

Assumption 2. *λ is a $\mathcal{C}^\infty$ function and ρ is a $\mathcal{C}^1$ convex function.*

In order to establish some propositions, the convexity of ρ will not be enough and we will need to consider strongly-convex functions.

Definition 1. *ρ is a $\mathcal{C}^1$ ζ-strongly convex function with $\zeta > 0$ if*

$$\forall p, q \in \Delta^K,\ \rho(q) \geq \rho(p) + \langle \nabla \rho(p), q - p \rangle + \frac{\zeta}{2} \|p - q\|^2 .$$

We will also be led to consider S-smooth functions:

Definition 2. *A continuously differentiable function f defined on a set $\mathcal{D} \subset \mathbb{R}^K$ is S-smooth (with $S > 0$) if its gradient is S-Lipschitz continuous.*

The optimal proportion function is denoted by $p^\star$ and verifies $p^\star = \operatorname{arginf}_{p \in \{\mathcal{X} \to \Delta^K\}} L(p)$. If an algorithm aiming at minimizing the loss L returns a proportion function p_T we define the regret as follows.

Definition 3. *The regret of an algorithm outputting* $p_T \in \{p : \mathcal{X} \to \Delta^K\}$ *is*

$$R(T) = \mathbb{E}L(p_T) - L(p^\star).$$

In the previous definition the expectation is taken on the choices of the algorithm. The goal is to find after T samples a $p_T \in \{p : \mathcal{X} \to \Delta^K\}$ the closest possible to $p^\star$ in the sense of minimizing the regret.

2.2 Examples of Regularizations

The most natural regularization function considered throughout this paper is the (negative) entropy function defined as follows:

$$\rho(p) = \sum_{i=1}^{K} p_i \log(p_i) \quad \text{for } p \in \Delta^K.$$

Since $\nabla^2_{ii}\rho(p) = 1/p_i \geq 1$, ρ is 1-strongly convex. Using this function as a regularization forces p to go to the center of the simplex, which means that each arm will be sampled a linear amount of time.

We can consider instead the Kullback-Leibler divergence between p and a known proportion function q:

$$\rho(p) = D_{KL}(p||q) = \sum_{i=1}^{K} p_i \log\left(\frac{p_i}{q_i}\right) \quad \text{for } p \in \Delta^K.$$

Instead of pushing p to the center of the simplex, the KL divergence regularization will push p towards q. This is typically motivated by problems where the decision maker should not alter too much an existing policy q, known to perform well on the task.

Another way to force p to be close to a chosen policy q is to use the ℓ^2-regularization $\rho(p) = \|p - q\|_2^2$. These two last examples of regularization have an explicit dependency on the context x since q depends on the context values, which was not the case of the entropy (which only depends on x through p). We show that both the KL divergence and the ℓ^2-regularization have a special form that allows us to remove this explicit dependency on x. They can indeed be written as

$$\rho(p(x), x) = H(p(x)) + \langle p(x), k(x)\rangle + c(x)$$

with H a ζ-strongly convex function of p, k a β-Hölder function of x and c any function of x.

Indeed,

$$\begin{aligned} D_{KL}(p||q) &= \sum_{i=1}^{K} p_i(x) \log\left(\frac{p_i(x)}{q_i(x)}\right) \\ &= \underbrace{\sum_{i=1}^{K} p_i(x) \log p_i(x)}_{H(p(x))} + \langle p(x), \underbrace{-\log q(x)}_{k(x)}\rangle. \end{aligned}$$

And

$$\|p(x)-q(x)\|_2^2 = \underbrace{\|p(x)\|^2}_{H(p(x))} + \langle p(x), \underbrace{-2q(x)}_{k(x)}\rangle + \underbrace{\|q(x)\|^2}_{c(x)}.$$

With this specific form the loss function writes as

$$\begin{aligned} L(p) &= \int_{\mathcal{X}} \langle \mu(x), p(x)\rangle + \lambda(x)\rho(p(x), x)\,\mathrm{d}x \\ &= \int_{\mathcal{X}} \langle \mu(x) + \lambda(x)k(x), p(x)\rangle + \lambda(x)H(p(x))\,\mathrm{d}x \\ &\qquad + \int_{\mathcal{X}} \lambda(x)c(x)\,\mathrm{d}x. \end{aligned}$$

Since we aim at minimizing L with respect to p the last term $\int_{\mathcal{X}} \lambda(x)c(x)\,\mathrm{d}x$ is irrelevant to the minimization process. Let us now note $\tilde{\mu} = \mu + \lambda k$. We are therefore minimizing

$$\tilde{L}(p) = \int_{\mathcal{X}} \langle \tilde{\mu}(x), p(x)\rangle + \lambda(x)H(p(x))\,\mathrm{d}x.$$

This is actually the standard setting of Subsection 2.1 with a regularization function H independent of x. In order to preserve the regularity of $\tilde{\mu}$ we need $\lambda\rho$ to be β-Hölder which is the case if q is sufficiently regular. Nonetheless, we remark that the relevant regularity is the one of μ since λ and ρ are known by the agent.

As a consequence, from now on we will only consider regularization functions ρ that only depend on p.

3 Algorithm

3.1 Idea of the Algorithm

As the horizon is finite, even if we could use the doubling-trick, and the reward functions μ_k are smooth, we choose to split the context space $\mathcal{X}$ into B^d cubic bins of side size $1/B$. We are going to construct a (bin by bin) piece-wise constant solution $\tilde{p}_T$.

We denote by $\mathcal{B}$ the set of bins introduced. If $b \in \mathcal{B}$ is a bin we note $|b| = B^{-d}$ its volume and $\mathrm{diam}(b) = \sqrt{d}/B$ its diameter. Since $\tilde{p}_T$ is piece-wise constant on each bin $b \in \mathcal{B}$ (with value $\tilde{p}_T(b)$), we rewrite the loss function into

$$\begin{aligned} L(\tilde{p}_T) &= \int_{\mathcal{X}} \langle \mu(x), \tilde{p}_T(x)\rangle + \lambda(x)\rho(\tilde{p}_T(x))\,\mathrm{d}x \\ &= \sum_{b\in\mathcal{B}} \int_b \langle \mu(x), \tilde{p}_T(b)\rangle + \lambda(x)\rho(\tilde{p}_T(b))\,\mathrm{d}x \\ &= \frac{1}{B^d} \sum_{b\in\mathcal{B}} \langle \bar{\mu}(b), \tilde{p}_T(b)\rangle + \bar{\lambda}(b)\rho(\tilde{p}_T(b)) \\ &= \frac{1}{B^d} \sum_{b\in\mathcal{B}} L_b(\tilde{p}_T(b)) \end{aligned} \tag{1}$$

where $L_b(p) = \langle \bar{\mu}(b), p \rangle + \bar{\lambda}(b)\rho(p)$ and $\bar{\mu}(b) = \frac{1}{|b|}\int_b \mu(x)\,\mathrm{d}x$ and $\bar{\lambda}(b) = \frac{1}{|b|}\int_b \lambda(x)\,\mathrm{d}x$ are the mean values of μ and λ on the bin b.

Consequently we just need to minimize the unknown convex loss function L_b for each bin $b \in \mathcal{B}$. We fall precisely in the setting of Berthet and Perchet (2017) where the authors propose an Upper-Confidence Frank-Wolfe algorithm to minimize an unknown convex function. We propose consequently the following Algorithm 1: for each time step $t \geq 1$, given the context value X_t, we run one iteration of the Upper-Confidence Frank-Wolfe algorithm for the loss function L_b corresponding to the bin $b \ni X_t$. We note $p_T(b)$ the results of the algorithm on each bin b.

Algorithm 1 Regularized Contextual Bandits

Require: K number of arms, T time horizon
Require: $\mathcal{B} = \{1, \dots, B^d\}$ set of bins
Require: $\left(t \mapsto \alpha_k^{(b)}(t)\right)_{k\in[K]}^{b\in\mathcal{B}}$ pre-sampling functions
1: **for** b in $\mathcal{B}$ **do**
2: Sample arm k $\alpha_k^{(b)}(T/B^d)$ times for all $k \in [K]$
3: **end for**
4: **for** $t \geq 1$ **do**
5: Sample context variable X_t uniformly in $[0,1]^d$
6: $b_t \leftarrow$ bin of X_t
7: Perform one step of the UCB Frank-Wolfe algorithm for the L_{b_t} function on bin b_t
8: **end for**
9: **return** the proportion vector $(p_T(1), \dots, p_T(B^d))$

Line 2 of the algorithm consists in a pre-sampling stage where all arms are sampled a certain amount of time. We will see how this can be used to enforce constraints on the p_i and especially to force p to be far from the boundaries of Δ^K.

In the remaining of this paper, we derive slow or fast rates of convergence of this algorithm, depending on the complexity of the current instance of the problem.

3.2 Estimation and Approximation Errors

In order to obtain a bound on the regret, we decompose it into an estimation error and an approximation error.

We note for all bins $b \in \mathcal{B}$, $p_b^\star = \operatorname{arginf}_{p\in\Delta^\kappa} L_b(p)$ the minimum of L_b on the bin b. We note $\tilde{p}^\star$ the piece-wise constant function taking the values $p_b^\star$ on the bin b.

The approximation error is the minimal achievable error within the class of piece-wise constant functions.

Definition 4. *The approximation error $A(p)$ is the error between the best piece-wise constant function $\tilde{p}^\star$ and the optimal solution $p^\star$.*

$$A(p^\star) = L(\tilde{p}^\star) - L(p^\star).$$

The estimation error is due to the errors made by the algorithm.

Definition 5. *The estimation error $E(p_T)$ is the error between the result of the algorithm p_T and the best piece-wise constant function $\tilde{p}^\star$.*

$$E(p_T) = \mathbb{E}L(p_T) - L(\tilde{p}^\star) = \frac{1}{B^d}\sum_{b\in\mathcal{B}} \mathbb{E}L_b(p_T(b)) - L_b(p_b^\star)$$

where the last equality comes from (1).

We naturally have $R(T) = E(p_T) + A(p^\star)$. In order to bound $R(T)$ we want to obtain bounds on both the estimation and the approximation error terms.

4 Convergence rates for constant λ

In this section we consider the case where λ is constant. We derive slow or fast rates of convergence.

4.1 Slow Rates

To derive slow rates we bound the approximation error and the estimation error. The proofs of this Subsection are deferred to Appendix A.

As in Berthet and Perchet (2017), we obtain the following convergence rate

Proposition 1. *Let ρ be a S-smooth convex function on Δ^K. If p_T is the result of Algorithm 1 and $\tilde{p}^\star$ the best piece-wise constant function on the set of bins $\mathcal{B}$, then the following bound on the estimation error holds*

$$\mathbb{E}L(p_T) - L(\tilde{p}^\star) = \mathcal{O}\left(\sqrt{K}B^{d/2}\sqrt{\frac{\log(T)}{T}}\right).$$

The Landau notation $\mathcal{O}(\cdot)$ has to be understood with respect to T. The precise bound is given in the proof.

There exist regularization functions that are not S-smooth on Δ^K. This is for example the case of the entropy whose Hessian is not bounded on Δ^K. However the following proposition shows that the result of Proposition 1 still holds, at least for the entropy.

Proposition 2. *If ρ is the entropy function the following bound on the estimation error holds*

$$\mathbb{E}L(p_T(b)) - L(\tilde{p}^\star) \leq \mathcal{O}\left(B^{d/2}\frac{\log(T)}{\sqrt{T}}\right).$$

The idea of the proof is to force the result of the algorithm to be "inside" the simplex Δ^K (in the sense of the induced topology) by pre-sampling each arm.

In order to obtain a bound on the approximation error we notice that

$$\begin{aligned} L_b(p_b^\star) &= \inf_{p\in\Delta^K} L_b(p) = \inf_{p\in\Delta^K} \lambda\rho(p) - \langle -\bar{\mu}(b), p\rangle \\ &= -(\lambda\rho)^*(-\bar{\mu}(b)) = -\lambda\rho^*\left(-\frac{\bar{\mu}(b)}{\lambda}\right) \end{aligned}$$

where ρ^* is the Legendre-Fenchel transform of ρ.

Similarly,

$$\begin{aligned}
&\int_b \langle \mu(x), p^\star(x)\rangle + \lambda\rho(p^\star(x))\,\mathrm{d}x \\
&\qquad = \int_b \inf_{p\in\Delta^K} -\langle -\mu(x), p\rangle + \lambda\rho(p)\,\mathrm{d}x \\
&\qquad = \int_b -(\lambda\rho)^*(-\mu(x))\,\mathrm{d}x \\
&\qquad = \int_b -\lambda\rho^*\left(-\frac{\mu(x)}{\lambda}\right)\,\mathrm{d}x.
\end{aligned}$$

We want to bound

$$\begin{aligned}
A(p^\star) &= \sum_{b\in\mathcal{B}} \int_b \langle \mu(x), \tilde{p}^\star(x)\rangle + \lambda\rho(\tilde{p}^\star(x)) \\
&\qquad\qquad - \langle \mu(x), p^\star(x)\rangle - \lambda\rho(p^\star(x))\,\mathrm{d}x \\
&= \sum_{b\in\mathcal{B}} \int_b \langle \bar{\mu}(b), p_b^\star\rangle + \lambda\rho(p_b^\star) \\
&\qquad\qquad - \langle \mu(x), p^\star(x)\rangle - \lambda\rho(p^\star(x))\,\mathrm{d}x \\
&= \sum_{b\in\mathcal{B}} \Bigg(\int_b L_b(p_b^\star)\,\mathrm{d}x \\
&\qquad\qquad - \int_b \langle \mu(x), p^\star(x)\rangle + \lambda\rho(p^\star(x))\,\mathrm{d}x \Bigg) \\
&= \lambda \sum_{b\in\mathcal{B}} \int_b \rho^*(-\mu(x)/\lambda) - \rho^*(-\bar{\mu}(b)/\lambda)\,\mathrm{d}x. \qquad (2)
\end{aligned}$$

With this expression we prove the

Proposition 3. *If $\tilde{p}^\star$ is the piece-wise constant function on the set of bins $\mathcal{B}$ minimizing the loss function L, we have the following bound*

$$L(\tilde{p}^\star) - L(p^\star) \leq \sqrt{L_\beta K d^\beta} B^{-\beta}.$$

Combining Propositions 1 and 3 we get the

Theorem 1 (Slow rates)**.** *If ρ is a S-smooth convex function, applying Algorithm 1 with choice $B = \Theta\left((T/\log(T))^{1/(2\beta+d)}\right)$ gives*

$$R(T) \leq \mathcal{O}_{L_\beta,\beta,K,d}\left(\left(\frac{T}{\log(T)}\right)^{-\frac{\beta}{2\beta+d}}\right).$$

In this theorem we have used the notation $\mathcal{O}_{L_\beta,\beta,K,d}$ to mean that there is a hidden constant depending on L_β, β, K and d. For clarity purposes the constant is not explicit in the theorem but can be found in the proof in Appendix A.

Proposition 2 directly shows that the result of this theorem also holds when ρ is the entropy function.

The detailed proof of the theorem can be read in Appendix A and consists in choosing a value of B balancing between the estimation and the approximation errors. Since $\beta \in (0, 1]$, we can see that the exponent of the convergence rate is below $1/2$ and that the proposed rate is slower than $T^{-1/2}$, hence the denomination of *slow rate*.

4.2 Fast Rates

We now consider possible fast rates, *i.e.* convergence rates faster than $\mathcal{O}\left(T^{1/2}\right)$. The price to pay to obtain these quicker rates compared to the ones from Subsection 4.1 is to have problem-dependent bounds, *i.e.* convergence rates depending on the parameters of the problem, and especially on λ.

As in the previous section we can obtain a bound on the estimation error based on the convergence rates of Upper-Confidence Frank-Wolfe algorithm.

Proposition 4. *If ρ is ζ-strongly convex, S-smooth and if there exists $\eta > 0$ such that for all $b \in \mathcal{B}$, $\operatorname{dist}(p_b^\star, \partial\Delta^K) \geq \eta$, then running Algorithm 1 gives the estimation error*

$$\mathbb{E}L(p_T) - L(\tilde{p}^\star) = \mathcal{O}\left(B^d\left(S\lambda + \frac{K}{\lambda^2\zeta^2\eta^4}\right)\frac{\log^2(T)}{T}\right).$$

Note that the fast rates for the estimation error depend on several parameters of the problem: λ, distance η of the optimum to the boundary of the simplex, strong convexity and smoothness constants. Since λ can be arbitrarily small, η can be small as well and S arbitrarily large. Consequently the "constant" factor can explode despite the convergence rate being "fast": these terms describe only the dependency in T.

Similarly to the previous section we want to consider the case of regularization functions ρ whose gradient is not Lipschitz-continuous at the boundary of Δ^K. In order to include these functions in our analysis we have to force the vectors p to be inside the simplex by pre-sampling all arms at the beginning of the algorithm. The following lemma shows that this is indeed valid.

Lemma 1. *On a bin b if there exists $\alpha \in (0, 1/2)$ and $p^o \in \Delta^K$ such that $p_b^\star \succeq \alpha p^o$ (component-wise) then for all $i \in [K]$, the agent can safely sample arm i $\alpha p_i^o T$ times at the beginning of the algorithm without changing the convergence results.*

The intuition behind this lemma is that if all arms have to be sampled a linear amount of times to reach the optimum value, it is safe to pre-sample each of the arms linearly at the beginning of the algorithm. The goal is to ensure that the current proportion vector p_t will always be far from the boundary in order to leverage the smoothness of ρ in the interior of the simplex.

Proposition 5. *If ρ is the entropy function, sampling each arm $Te^{-1/\lambda}/K$ times during the presampling phase guarantees the same estimation error as in Proposition 4 with constant $S = Ke^{1/\lambda}$.*

In order to obtain faster rates for the approximation error we use Equation (2) and the fact that $\nabla\rho^*$ is $1/\zeta$-Lipschitz since ρ is ζ-strongly convex. The full proof of the following resultat is given in Appendix B.

Proposition 6. *If ρ is ζ-strongly convex and if $\tilde{p}^\star$ is the piece-wise constant function on the set of bins $\mathcal{B}$ minimizing the loss function L, the following bound on the approximation error holds*

$$L(\tilde{p}^\star) - L(p^\star) \leq \frac{L_\beta K d^\beta}{2\zeta\lambda} B^{-2\beta}.$$

Combining Propositions 4 and 6, we obtain fast rates for our problem.

Theorem 2 (Fast rates)**.** *If ρ is ζ-strongly convex and if there exists $\eta > 0$ such that for all $b \in \mathcal{B}$, $\operatorname{dist}(p_b^\star, \partial\Delta^K) \geq \eta$, applying Algorithm 1 with the choice $B = \Theta\left(T/\log^2(T)\right)^{1/(2\beta+d)}$ gives the regret*

$$R(T) \leq \mathcal{O}_{L_\beta,\beta,K,d,\lambda,\eta,\zeta,S}\left(\left(\frac{T}{\log^2(T)}\right)^{-\frac{2\beta}{2\beta+d}}\right).$$

This rate matches the rates obtained in nonparametric estimation (Tsybakov, 2008). However, as shown in the proof presented in Appendix B, this fast rate is obtained at the price of a factor involving λ, η and S, which can be arbitrarily large. It is the goal of the next section to see how to remove this dependency in the parameters of the problem.

Proposition 5 shows that the previous theorem can also be applied to the entropy regularization.

5 Convergence rates for non-constant λ

In this section, we study the case where λ is a function of the context value. This is quite interesting as agents might want to modulate the weight of the regularization term depending on the context. All the proofs of this Section can be found in Appendix C.

5.1 Estimation and Approximation errors

Equation (1) implies that the estimation errors obtained in Propositions 1 and 4 are still correct if λ is replaced by $\bar{\lambda}(b)$. This is unfortunately not the case for the approximation error propositions because Equation (2) does not hold anymore. Indeed the approximation

error becomes :

$$\begin{aligned}
A(p^\star) &= \sum_{b\in\mathcal{B}} \int_b \langle \mu(x), \tilde{p}^\star(x)\rangle + \lambda(x)\rho(\tilde{p}^\star(x)) \\
&\qquad - \langle \mu(x), p^\star(x)\rangle - \lambda(x)\rho(p^\star(x))\,\mathrm{d}x \\
&= \sum_{b\in\mathcal{B}} \int_b \langle \bar{\mu}(b), p_b^\star\rangle + \lambda(x)\rho(p_b^\star) \\
&\qquad - \langle \mu(x), p^\star(x)\rangle - \lambda(x)\rho(p^\star(x))\,\mathrm{d}x \\
&= \sum_{b\in\mathcal{B}} \Bigg(\int_b L_b(p_b^\star)\,\mathrm{d}x \\
&\qquad - \int_b \langle \mu(x), p^\star(x)\rangle + \lambda(x)\rho(p^\star(x))\,\mathrm{d}x \Bigg) \\
&= \sum_{b\in\mathcal{B}} \int_b -(\bar{\lambda}(b)\rho)^*(-\bar{\mu}(b)) + (\lambda(x)\rho)^*(-\mu(x))\,\mathrm{d}x \\
&= \sum_{b\in\mathcal{B}} \int_b \lambda(x)\rho^*\left(-\frac{\mu(x)}{\lambda(x)}\right) - \bar{\lambda}(b)\rho^*\left(-\frac{\bar{\mu}(b)}{\bar{\lambda}(b)}\right)\,\mathrm{d}x. \qquad (3)
\end{aligned}$$

From this expression we obtain the following slow and fast rates of convergence.

Proposition 7. *If ρ is a strongly convex function and λ a $\mathcal{C}^\infty$ integrable non-negative function whose inverse is also integrable, we have on a bin b:*

$$\begin{aligned}
&\int_b (\lambda(x)\rho)^*\left(-\mu(x)\right) - (\bar{\lambda}(b)\rho)^*\left(-\bar{\mu}(b)\right)\,\mathrm{d}x \\
&\qquad\qquad \leq \mathcal{O}(L_\beta d^{\beta/2} B^{-\beta-d}).
\end{aligned}$$

The important point of this proposition is that the bound does not depend on $\lambda_{\min}$, which is not the case when we want to obtain fast rates for the approximation error as in the following proposition:

Proposition 8. *If ρ is a ζ-strongly convex function and λ a $\mathcal{C}^\infty$ integrable non negative function whose inverse is also integrable, we have on a bin b:*

$$\begin{aligned}
&\int_b (\lambda(x)\rho)^*\left(-\mu(x)\right) - (\bar{\lambda}(b)\rho)^*\left(-\bar{\mu}(b)\right)\,\mathrm{d}x \\
&\qquad\qquad \leq \mathcal{O}\left(K d L_\beta^2 \left\|\nabla\lambda\right\|_\infty^2 \frac{B^{-2\beta-d}}{\zeta\lambda_{\min}^3} \right).
\end{aligned}$$

The rate in B is improved compared to Proposition 7 at the expense of the constant $1/\lambda_{\min}^3$ which can unfortunately be arbitrarily high.

5.2 Margin Condition

We begin by giving a precise definition of the function η, the distance of the optimum to the boundary of Δ^K.

Definition 6. *Let $x \in \mathcal{X}$ a context value. We define by $p^\star(x) \in \Delta^K$ the point where $p \mapsto \langle \mu(x), p\rangle + \lambda(x)\rho(p)$ attains its minimum, and*

$$\eta(x) := \operatorname{dist}(p^\star(x), \partial\Delta^K).$$

Similarly, if $p_b^\star$ is the point where $L_b : p \mapsto \langle \bar{\mu}(b), p\rangle + \bar{\lambda}(b)\rho(p)$ attains its minimum, we define

$$\eta(b) := \operatorname{dist}(p_b^\star, \partial\Delta^K).$$

We have obtained in Subsection 4.2 fast rates for Algorithm 1. These convergence rates provide good theoretical guarantees but may be useless in practice since they depend on a constant that can be arbitrarily large. We would like to discard the dependency on the parameters of the problem, and especially λ (that controls η and S).

Difficulties arise when λ and η take values that are very small, meaning for instance that we consider nearly no regularization. This is not likely to happen since we do want to study contextual bandits with regularization. To formalize that we make an additional assumption, which is common in nonparametric regression (Tsybakov, 2008) and is known as a *margin condition*:

Assumption 3 (Margin Condition)**.** *We assume that there exist $\delta_1 > 0$ and $\delta_2 > 0$ as well as $\alpha > 0$ and $C_m > 0$ such that*

$$\begin{aligned} &\forall \delta \in (0, \delta_1],\ \mathbb{P}_X(\lambda(x) < \delta) \leq C_m \delta^{6\alpha} \\ \textit{and}\quad &\forall \delta \in (0, \delta_2],\ \mathbb{P}_X(\eta(x) < \delta) \leq C_m \delta^{6\alpha}. \end{aligned}$$

The non-negative parameter α controls the importance of the margin condition.

The margin condition limits the number of bins on which λ or η can be small. Therefore we can split the bins of $\mathcal{B}$ into two categories, the "well-behaved bins" on which λ and η are not too small, and the "ill-behaved bins" where those functions can take arbitrarily small values. The idea is to use the fast rates on the "well-behaved bins" and the slow rates (that do not depend on λ and η) on the "ill-behaved bins". This is be the point of Subsection 5.3.

Let $C_L = \sqrt{\frac{K}{K-1}} \frac{\|\lambda\|_\infty + \|\nabla\lambda\|_\infty}{\zeta}$, $c_1 = 1 + \|\nabla\lambda\|_\infty d^{\beta/6}$ and $c_2 = 1 + C_L d^{\beta/2}$.

We define the set of "well-behaved bins" $\mathcal{WB}$ as

$$\begin{aligned} \mathcal{WB} = \{b \in \mathcal{B},\ &\exists\ x_1 \in b,\ \lambda(x_1) \geq c_1 B^{-\beta/3} \\ &\text{and } \exists\ x_2 \in b,\ \eta(x_2) \geq c_2 B^{-\beta/3}\}, \end{aligned}$$

and the set of "ill-behaved bins" as its complementary set in $\mathcal{B}$.

With the smoothness and regularity Assumptions 1 and 2, we derive lower bounds for λ and η on the "well-behaved bins".

Lemma 2. *If b is a well-behaved bin then*

$$\forall x \in b,\ \lambda(x) \geq B^{-\beta/3} \quad \textit{and} \quad \forall x \in b,\ \eta(x) \geq B^{-\beta/3}.$$

5.3 Intermediate Rates

We summarize the different error rates obtained in the previous sections.

Table 1: Slow and Fast Rates for Estimation and Approximation Errors on a Bin

Error	**Slow**	**Fast**
Estim.	$B^{-d/2}\sqrt{\frac{\log(T)}{T}}$	$\frac{\log^2(T)}{T}\left(S\lambda + \frac{1}{\eta^4\lambda^2}\right)$
Approx.	$B^{-d}B^{-\beta}$	$\frac{B^{-2\beta-d}}{\lambda^3}$
B	$\left(\frac{T}{\log(T)}\right)^{\frac{1}{2\beta+d}}$	$\left(\frac{T}{\log^2(T)}\right)^{\frac{1}{2\beta+d}}$
$R(T)$	$\left(\frac{T}{\log(T)}\right)^{\frac{-\beta}{2\beta+d}}$	$\left(\frac{T}{\log^2(T)}\right)^{\frac{-2\beta}{2\beta+d}}$

In this table we only kept the relevant constants that can be very small, *i.e.* λ and η, or very large, *i.e.* S. For the sake of clarity we remove the dependency on the bin, writing λ instead of $\bar{\lambda}(b)$.

Table 1 clearly shows that the slow rates do not depend on those constants. Therefore we can use the slow rates on the "ill-behaved bins".

Theorem 3 (Intermediate rates)**.** *Applying Algorithm 1 on the contextual bandits problem with an entropy regularization and margin condition with parameter $\alpha \in (0,1)$, the choice $B = \Theta\left(T/\log^2(T)\right)^{\frac{1}{2\beta+d}}$ leads to the regret*

$$R(T) = \mathcal{O}_{K,d,\alpha,\beta,L_\beta}\left(\frac{T}{\log^2(T)}\right)^{-\frac{\beta}{2\beta+d}(1+\alpha)}.$$

As explained in the proof (Appendix C), we use a pre-sampling stage on each bin to force the entropy to be smooth, as in the proofs of Propositions 2 and 5.

We consider now the extreme values of α. If $\alpha \to 0$, there is no margin condition and the speed obtainedis $T^{-\frac{\beta}{2\beta+d}}$ which is exactly the slow rate from Theorem 1. If $\alpha \to 1$, there is a strong margin condition and the rate of Theorem 3 tends to $T^{-\frac{2\beta}{2\beta+d}}$ which is the fast rate from Theorem 2. Consequently we get that the intermediate rates from Theorem 3 do interpolate between the slow and fast rates obtained previously.

6 Lower Bounds

The results in Theorems 1 and 2 have optimal exponents in the dependency in T. For the slow rate, since the regularization can be equal to 0, or a linear form, the lower bounds on contextual bandits in this setting apply (Audibert et al., 2007; Rigollet and Zeevi, 2010), matching this upper bound. For the fast rates, the following lower bound holds, based on a reduction to nonparametric regression (Tsybakov, 2008; Györfi et al., 2006).

Theorem 4. *For any algorithm with bandit input and output $\hat{p}_T$, for ρ that is 1-strongly convex, we have*

$$\inf_{\hat{p}} \sup_{\substack{\mu\in\mathcal{H}_\beta \\ \rho\in 1\text{-str. conv.}}} \left\{\mathbb{E}[L(\hat{p}_T)] - L(p^\star)\right\} \geq C\, T^{-\frac{2\beta}{2\beta+d}},$$

for a universal constant C.

The proof is in Appendix D. The upper and lower bound match up to logarithmic terms. This bound is obtained for $K = 2$, and the dependency of the rate in K is not analyzed here.

7 Empirical Results

We present in this section experiments and simulations for the regularized contextual bandits problem. The setting we consider uses $K = 3$ arms, with an entropy regularization and a fixed parameter $\lambda = 0.1$. We run successive experiments for values of T ranging from 1 000 to 100 000, and for different values of the smoothness parameter β.

The results presented in Figure 1 shows that $T \mapsto T \cdot R(T)$ growths as expected, and the lower β, the slower the convergence rate, as shown on the graph.

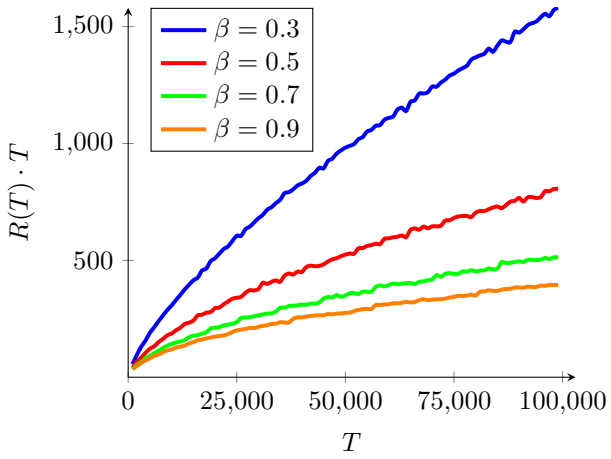

Figure 1: Regret as a Function of T

In order to verify that the fast rates proven in Subsection 4.2 are indeed reached, we plot on Figure 2 the ratio between the regret and the theoretical bound on the regret $\left(T/\log^2(T)\right)^{-\frac{2\beta}{2\beta+d}}$. We observe that this ratio is approximately constant as a function of T, which validates empirically the theoretical convergence rates.

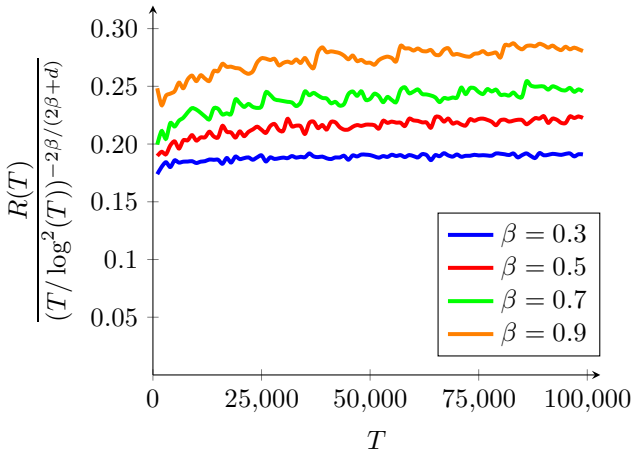

Figure 2: Normalized Regret as a Function of T

8 Conclusion

We proposed an algorithm for the problem of contextual bandits with regularization reaching fast rates similar to the ones obtained in nonparametric estimation. The conducted experiments validate the fast rates obtained and we can discard the parameters of the problem in the convergence rates by applying a margin condition that allows us to derive intermediate convergence rates interpolating perfectly between the slow and fast rates.

References

Agrawal, R. (1995). Sample mean based index policies by O(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078.

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135.

Audibert, J.-Y., Tsybakov, A. B., et al. (2007). Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

Bastani, H. and Bayati, M. (2015). Online decision-making with high-dimensional covariates. In *SSRN Electronic Journal*.

Berthet, Q. and Perchet, V. (2017). Fast rates for bandit optimization with upper-confidence Frank-Wolfe. In *Advances in Neural Information Processing Systems*, pages 2225–2234.

Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. (2011). Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 169–178, Arlington, Virginia, United States. AUAI Press.

Goldenshluger, A., Zeevi, A., et al. (2009). Woodroofe's one-armed bandit problem revisited. *The Annals of Applied Probability*, 19(4):1603–1633.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Hazan, E. and Megiddo, N. (2007). Online learning with prior knowledge. In *Learning Theory, 20th Annual Conference on Learning Theory, COLT 2007, San Diego, CA, USA, June 13-15, 2007, Proceedings*, pages 499–513.

Hiriart-Urruty, J.-B. and Lemaréchal, C. (2013a). *Convex analysis and minimization algorithms I*, volume 305. Springer science & business media.

Hiriart-Urruty, J.-B. and Lemaréchal, C. (2013b). *Convex analysis and minimization algorithms II*, volume 306. Springer science & business media.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Langford, J. and Zhang, T. (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM.

Nesterov, Y. (2013). *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.

Perchet, V. and Rigollet, P. (2013). The multi-armed bandit problem with covariates. *The Annals of Statistics*, pages 693–721.

Rigollet, P. and Zeevi, A. J. (2010). Nonparametric bandits with covariates. In *COLT*.

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

Slivkins, A. (2014). Contextual bandits with similarity information. *The Journal of Machine Learning Research*, 15(1):2533–2568.

Tang, L., Rosales, R., Singh, A., and Agarwal, D. (2013). Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1587–1594. ACM.

Tewari, A. and Murphy, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. In *Mobile Health - Sensors, Analytic Methods, and Applications*, pages 495–517.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.

Wang, C.-C., Kulkarni, S. R., and Poor, H. V. (2005). Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50(3):338–355.

Woodroofe, M. (1979). A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806.

Wu, Y., Shariff, R., Lattimore, T., and Szepesvári, C. (2016). Conservative bandits. In *International Conference on Machine Learning*, pages 1254–1262.

A Proofs of Slow Rates

We prove in this section the propositions and theorem of Subsection 4.1.

We begin by a lemma on the concentration of T_b, the number of context samples falling in a bin b.

Lemma 3. *For all $b \in \mathcal{B}$, let T_b the number of context samples falling in the bin b. We have*

$$\mathbb{P}\left(\exists b \in \mathcal{B},\ \left|T_b - \frac{T}{B^d}\right| \geq \frac{1}{2}\frac{T}{B^d}\right) \leq 2B^d \exp\left(-\frac{T}{12B^d}\right).$$

Proof. For a bin $b \in \mathcal{B}$ and $t \in \{1, \dots, T\}$, let $Z_t^{(b)} = \mathbb{1}_{\{X_t \in \mathcal{B}\}}$ which is a random Bernoulli variable of parameter $1/B^d$.

We have $T_b = \sum_{t=1}^{T} Z_t$ and $\mathbb{E}[T_b] = T/B^d$.

Using a multiplicative Chernoff's bound (Vershynin, 2018) we obtain:

$$\mathbb{P}\left(|T_b - \mathbb{E}[T_b]| \geq \frac{1}{2}\mathbb{E}[T_b]\right) \leq 2\exp\left(-\frac{1}{3}\left(\frac{1}{2}\right)^2 \frac{T}{B^d}\right) = 2\exp\left(-\frac{T}{12B^d}\right).$$

We conclude with an union bound on all the bins. □

Proof of Proposition 1. We have

$$E(p_T) = \mathbb{E}L(p_T) - L(\tilde{p}^\star) = \frac{1}{B^d}\sum_{b\in\mathcal{B}} \mathbb{E}L_b(p_T(b)) - L_b(p_b^\star)$$

Let us now consider a single bin $b \in \mathcal{B}$. We have run the UCB Frank-Wolfe (Berthet and Perchet, 2017) algorithm for the function L_b on the bin b with T_b iterations.

For all $p \in \Delta^K$, $L_b(p) = \langle \bar{\mu}(b), p\rangle + \lambda\rho(p)$, then for all $p \in \Delta^K$, $\nabla L_b(p) = \bar{\mu}(b) + \lambda\nabla\rho(p)$ and $\nabla^2 L_b(p) = \lambda\nabla^2\rho(p)$. Since ρ is a S-smooth convex function, L_b is a λS-smooth convex function.

We consider the event A:

$$A \doteq \left\{\forall b \in \mathcal{B},\ T_b \in \left[\frac{T}{2B^d}, \frac{3T}{2B^d}\right]\right\}.$$

Lemma 3 shows that $\mathbb{P}(A^\complement) \leq 2B^d \exp\left(-\frac{T}{12B^d}\right)$.

Theorem 3 of Berthet and Perchet (2017) shows that, on event A:

$$\begin{aligned}\mathbb{E}L_b(p_T(b)) - L_b(p_b^\star) &\leq 4\sqrt{\frac{3K\log(T_b)}{T_b}} + \frac{S\log(eT_b)}{T_b} + \left(\frac{\pi^2}{6} + K\right)\frac{2\left\|\nabla L_b\right\|_\infty + \left\|L_b\right\|_\infty}{T_b}\\ &\leq 4\sqrt{\frac{6K\log(T)}{T/B^d}} + \frac{2S\log(eT)}{T/B^d} + 2\left(\frac{\pi^2}{6} + K\right)\frac{2\left\|\nabla L_b\right\|_\infty + \left\|L_b\right\|_\infty}{T/B^d}.\end{aligned}$$

Since ρ is of class $\mathcal{C}^1$, ρ and $\nabla\rho$ are bounded on the compact set Δ^K. It is also the case for L_b and consequently $\left\|L_b\right\|_\infty$ and $\left\|\nabla L_b\right\|_\infty$ exist and are finite and can be expressed in function of $\left\|\rho\right\|_\infty$, $\left\|\nabla\rho\right\|_\infty$ and $\left\|\lambda\right\|_\infty$. On event $A^\complement$, $\mathbb{E}L_b(p_T(b)) - L_b(p_b^\star) \leq 2\left\|L_b\right\|_\infty \leq 2 + 2\left\|\lambda\rho\right\|_\infty$.

Summing over all the bins in $\mathcal{B}$ we obtain:

$$\mathbb{E}L(p_T) - L(p^\star) \leq 4B^{d/2}\sqrt{\frac{6K\log(T)}{T}} + B^d\frac{2S\log(eT)}{T} + 4KB^d\frac{4 + 2\left\|\lambda\nabla\rho\right\|_\infty + \left\|\lambda\rho\right\|_\infty}{T} + 4B^d(1 + \left\|\lambda\rho\right\|_\infty)e^{-\frac{T}{12B^d}}. \tag{4}$$

The first term of Equation (4) dominates the others and we can therefore write that

$$\mathbb{E}L(p_T) - L(p^\star) = \mathcal{O}\left(\sqrt{K}B^{d/2}\sqrt{\frac{\log(T)}{T}}\right)$$

where the $\mathcal{O}$ is valid for $T \to \infty$.

□

Proof of Proposition 2. We consider a bin $b \in \mathcal{B}$ containing t samples.

Let $\mathcal{S} \doteq \left\{ p \in \Delta^K \mid \forall i \in [K],\ p_i \geq \frac{\lambda}{\sqrt{t}} \right\}$. In order to force all the successive estimations of $p_b^\star$ to be in $\mathcal{S}$ we sample each arm $\lambda\sqrt{t}$ times. Thus we have $\forall i \in [K],\ p_i \geq \lambda/\sqrt{t}$. Then we apply the UCB-Frank Wolfe algorithm on the bin b. Let

$$\hat{p}_b \doteq \min_{p \in \mathcal{S}} L_b(p) \quad \text{and} \quad p_b^\star \doteq \min_{p \in \Delta^K} L_b(p).$$

- **Case 1:** $\hat{p}_b = p_b^\star$, *i.e.* the minimum of L_b is in $\mathcal{S}$.

  For all $p \in \Delta^K$, $L_b(p) = \langle \bar{\mu}(b), p \rangle + \lambda\rho(p)$, then for all $p \in \Delta^K$, $\nabla L_b(p) = \bar{\mu}(b) + \lambda(1 + \log(p))$ and $\nabla^2_{ii} L_b(p) = \lambda/p_i$. Therefore on $\mathcal{S}$ we have

  $$\nabla^2_{ii} L_b(p) \leq \sqrt{t}.$$

  And consequently L_b is $\sqrt{t}$-smooth. And since $\nabla_i L_b(p) = 1 + \lambda \log(p_i)$, $\|\nabla L_b(p)\|_\infty \lesssim \log(t)$. We can apply the same steps as in the proof of Proposition 1 to find that

  $$\mathbb{E} L_b(p_t(b)) - L_b(p_b^\star) \leq 4\sqrt{\frac{3K\log(t)}{t}} + \frac{\sqrt{t}\log(et)}{t} + \left(\frac{\pi^2}{6} + K\right)\frac{2\log(t) + \log(K)}{t} = \mathcal{O}\left(\frac{\log(t)}{\sqrt{t}}\right).$$

- **Case 2:** $\hat{p}_b \neq p_b^\star$. By strong convexity of L_b, $\hat{p}_b$ cannot be a local minimum of L_b and therefore $\hat{p}_b \in \partial\Delta^K$.

  The Case 1 shows that

  $$\mathbb{E} L_b(p_t(b)) - L_b(\hat{p}_b) \leq \mathcal{O}\left(\frac{\log(t)}{\sqrt{t}}\right).$$

  Let $\pi = (\pi_1, \dots, \pi_K)$ with $\pi_i \doteq \max(\lambda/\sqrt{t}, \hat{p}_{b,i})$. We have $\|\pi - \hat{p}_b\|_2 \leq \sqrt{K}\lambda/\sqrt{t}$.

  Let us derive an explicit formula for $p_b^\star$ knowing the explicit expression of ρ. In order to find the optimal $\rho^\star$ value let us minimize $(p \mapsto L_b(p))$ under the constraint that p lies in the simplex Δ^K. The KKT equations give the existence of $\xi \in \mathbb{R}$ such that for each $i \in [K]$, $\bar{\mu}_i(b) + \lambda\log(p_i) + \lambda + \xi = 0$ which leads to $p^\star_{b,i} = e^{-\bar{\mu}_i(b)/\lambda}/Z$ where Z is a normalization factor. Since $Z = \sum_{i=1}^K e^{-\bar{\mu}_i(b)/\lambda}$ we have $Z \leq K$ and $p^\star_{b,i} \geq e^{-1/\lambda}/K$. Consequently for all p on the segment between π and $p_b^\star$ we have $p_i \geq e^{-1/\lambda}/K$ and therefore $\lambda(1 + \log(p_i)) \geq \lambda(1 - \log K) - 1$ and finally $|\nabla_i L_b(p)| \leq 4\|\lambda\|_\infty \log(K)$.

  Therefore L_b is $4\sqrt{K}\log(K)$-Lipschitz and

  $$\|L_b(p_b^\star) - L_b(\pi)\|_2 \leq 4\|\lambda\|_\infty \sqrt{K}\log(K)\|\pi - \hat{p}_b\|_2 \leq 4K\log(K)\|\lambda\|^2_\infty/\sqrt{t} = \mathcal{O}(1/\sqrt{t}).$$

  Finally, since $L_b(\pi) \geq L_b(\hat{p}_b)$ (because $\pi \in \mathcal{S}$), we have

  $$\mathbb{E} L_b(p_t(b)) - L_b(p_b^\star) \leq \mathbb{E} L_b(p_t(b)) - L_b(\hat{p}_b) + L_b(\hat{p}_b) - L_b(p_b^\star) \leq \mathcal{O}\left(\frac{\log(t)}{\sqrt{t}}\right) + L(\pi) - L(p_b^\star) = \mathcal{O}\left(\frac{\log(t)}{\sqrt{t}}\right).$$

  We conclude by summing on the bins and using that $t \in [T/2B^d, 3T/2B^d]$ with high probability, as in the proof of Proposition 1.

□

Proof of Proposition 3. We have to bound the quantity

$$L(\tilde{p}^\star) - L(p^\star) = \lambda \sum_{b\in\mathcal{B}} \int_b \rho^*(-\mu(x)/\lambda) - \rho^*(-\bar{\mu}(b)/\lambda)\,\mathrm{d}x.$$

Classical results on convex conjugates (Hiriart-Urruty and Lemaréchal, 2013a) give that $\nabla\rho^*(y) = \operatorname{argmin}_{x\in\Delta^K} \rho(x) - \langle x, y\rangle$ for all $y \in \mathbb{R}^K$. Consequently, $\nabla\rho^*(y) \in \Delta^K$ and for all $y \in \mathbb{R}^K$, $\|\nabla\rho^*(y)\| \leq 1$ showing that ρ^* is 1-Lipschitz continuous. This leads to

$$\begin{aligned} L(\tilde{p}^\star) - L(p^\star) &\leq \lambda \sum_{b\in\mathcal{B}} \int_b \left\| \frac{\mu(x) - \bar{\mu}(b)}{\lambda} \right\| \mathrm{d}x \\ &\leq \sum_{b\in\mathcal{B}} \int_b \sqrt{L_\beta K} \left(\frac{\sqrt{d}}{B}\right)^\beta \mathrm{d}x \\ &\leq \sqrt{L_\beta K d^\beta} B^{-\beta} \end{aligned}$$

because all the μ_k are (L_β, β)-Hölder. □

Proof of Theorem 1. We will denote by C_k with increasing values of k the constants. Since the regret is the sum of the approximation error and the estimation error we obtain

$$R(T) \leq \sqrt{L_\beta d^\beta K} B^{-\beta} + C_1 \sqrt{K} B^{d/2} \sqrt{\frac{\log(T)}{T}} + B^d \frac{2S\log(eT)}{T} + C_2 K \frac{B^d}{T} + 4B^d(1 + \|\lambda\rho\|_\infty) \exp\left(-\frac{T}{12B^d}\right).$$

With the choice of

$$B = \left(C_2 \beta \sqrt{L_\beta} d^{\beta/2-1}\right)^{1/(\beta+d/2)} \left(\frac{T}{\log(T)}\right)^{1/(2\beta+d)},$$

we find that the three last terms of the regret are negligible with respect to the first two. This gives

$$R(T) \leq \mathcal{O}\left(\left(3\sqrt{K} L_\beta^{d/(4\beta+2d)} d^{\beta(4+d)/(4\beta+2d)} (C_2\beta)^{-\beta/(2\beta+d)}\right) \left(\frac{T}{\log(T)}\right)^{-\beta/(2\beta+d)}\right).$$

□

B Proofs of Fast Rates

We prove now the propositions and theorem of Subsection 4.2.

Proof of Proposition 4. The proof is very similar to the one of Proposition 1. We decompose the estimation error on the bins:

$$\mathbb{E}L(p_T) - L(\tilde{p}^\star) = \frac{1}{B^d} \sum_{b\in\mathcal{B}} \mathbb{E}L_b(p_T(b)) - L_b(p_b^\star).$$

Let us now consider a single bin $b \in \mathcal{B}$. We have run the UCB Frank-Wolfe algorithm for the function L_b on the bin b with T_b samples.

As in the proof of Proposition 1 we consider the event A.

Theorem 7 of Berthet and Perchet (2017), applied to L_b which is a λS-smooth $\lambda\zeta$-strongly convex function, shows that on event A:

$$\mathbb{E}L(p_T) - L(p^\star) \leq 2\tilde{c}_1 \frac{\log^2(T)}{T/B^d} + 2\tilde{c}_2 \frac{\log(T)}{T/B^d} + \tilde{c}_3 \frac{2}{T/B^d}$$

with $\tilde{c}_1 = \frac{96K}{\zeta\lambda\eta^2}$, $\tilde{c}_2 = \frac{24}{\zeta\lambda\eta^3} + \lambda S$ and $\tilde{c}_3 = 24\left(\frac{20}{\zeta\lambda\eta^2}\right)^2 K + \frac{\lambda\zeta\eta^2}{2} + \lambda S$. Consequently

$$\mathbb{E}L(p_T) - L(p^\star) \leq 2\tilde{c}_1\frac{\log^2(T)}{T/B^d} + 2\tilde{c}_2\frac{\log(T)}{T/B^d} + \tilde{c}_3\frac{2}{T/B^d} + 4B^d(1+\|\lambda\rho\|_\infty)\exp\left(-\frac{T}{12B^d}\right).$$

In order to have a simpler expression we can use the fact that λ and η are constants that can be small while S can be large. Consequently $\tilde{c}_3$ is the largest constant among $\tilde{c}_1$, $\tilde{c}_2$ and $\tilde{c}_3$ and we obtain

$$\mathbb{E}L(p_T) - L(p^\star) \leq \mathcal{O}\left(\left(\frac{K}{\lambda^2\zeta^2\eta^4} + S\lambda\right)B^d\frac{\log^2(T)}{T}\right),$$

because the other terms are negligible. □

Proof of Lemma 1. We consider a single bin $b \in \mathcal{B}$. Let us consider the function

$$\hat{L}_b : p \mapsto L_b(\alpha p^o + (1-\alpha)p).$$

Since for all i, $p^\star_{b,i} \geq \alpha p^o_i$ and since Δ^K is convex we know that $\min_{p\in\Delta^K}\hat{L}_b(p) = L_b(p^\star_b)$.

If p is the frequency vector obtained by running the UCB-Frank Wolfe algorithm for function $\hat{L}_b$ with $(1-\alpha)T$ samples then minimizing $\hat{L}_b$ is equivalent to minimizing L with a presampling stage.

Consequently the whole analysis on the regret still holds with T replaced by $(1-\alpha)T$. Thus fast rates are kept with a constant factor $1/(1-\alpha) \leq 2$. □

Proof of Proposition 5. For the entropy regularization, we have

$$p^\star_{b,i} = \frac{\exp(-\bar{\mu}(b)_i/\lambda)}{\sum_{j=1}^K \exp(-\bar{\mu}(b)_j/\lambda)} \leq \frac{\exp(-1/\lambda)}{K}.$$

We apply Lemma 1 with $p^o = \left(\frac{1}{K},\dots,\frac{1}{K}\right)$ and $\alpha = \exp(-1/\lambda)$. Consequently each arm is presampled $T\exp(-1/\lambda)/K$ times and finally we have

$$\forall i \in [K], p_i \geq \frac{\exp(-1/\lambda)}{K}.$$

Therefore we have

$$\forall i \in [K],\ \nabla_{ii}\rho(p) = \frac{1}{p_i} \leq K\exp(1/\lambda),$$

showing that ρ is $K\exp(1/\lambda)$-smooth. □

In order to prove the Proposition 6 we will need the following lemma which is a direct consequence of a result on smooth convex functions.

Lemma 4. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function of class $\mathcal{C}^1$ and $L > 0$. Let $g : \mathbb{R}^d \ni x \mapsto \frac{L}{2}\|x\|^2 - f(x)$. Then g is convex if and only if ∇f is L-Lipschitz continuous.*

Proof. Since g is continuously differentiable we can write

$$\begin{aligned}
g \text{ convex } &\Leftrightarrow \forall x,y \in \mathbb{R}^d,\ g(y) \geq g(x) + \langle \nabla g(x), y-x\rangle \\
&\Leftrightarrow \forall x,y \in \mathbb{R}^d, \frac{L}{2}\|y\|^2 - f(y) \geq \frac{L}{2}\|x\|^2 - f(x) + \langle Lx - \nabla f(x), y-x\rangle \\
&\Leftrightarrow \forall x,y \in \mathbb{R}^d, f(y) \leq f(x) + \langle \nabla f(x), y-x\rangle + \frac{L}{2}\left(\|y\|^2 + \|x\|^2 - 2\langle x,y\rangle\right) \\
&\Leftrightarrow \forall x,y \in \mathbb{R}^d, f(y) \leq f(x) + \langle \nabla f(x), y-x\rangle + \frac{L}{2}\|x-y\|^2 \\
&\Leftrightarrow \nabla f \text{ is } L\text{-Lipschitz continuous.}
\end{aligned}$$

where the last equivalence comes from Theorem 2.1.5 of Nesterov (2013). □

Proof of Proposition 6. Since ρ is ζ-strongly convex then $\nabla\rho^*$ is $1/\zeta$-Lipschitz continuous (see for example Theorem 4.2.1 at page 82 in Hiriart-Urruty and Lemaréchal (2013b)). Since ρ^* is also convex, Lemma 4 shows that $g : x \mapsto \frac{1}{2\zeta}\|x\|^2 - \rho^*(x)$ is convex.

Let us now consider the bin b and the function $\mu = (\mu_1, \dots, \mu_k)$. Jensen's inequality gives:

$$\frac{1}{|b|}\int_b g(-\mu(x)/\lambda)\,\mathrm{d}x \geq g\left(\frac{1}{|b|}\int_b -\frac{\mu(x)}{\lambda}\,\mathrm{d}x\right).$$

This leads to

$$\begin{aligned}
\int_b g(-\mu(x)/\lambda)\,\mathrm{d}x &\geq \int_b g(-\bar{\mu}(b)/\lambda)\,\mathrm{d}x \\
\int_b \frac{1}{2\zeta}\|-\mu(x)\|^2/\lambda^2 - \rho^*(-\mu(x)/\lambda)\,\mathrm{d}x &\geq \int_b \frac{1}{2\zeta}\|-\bar{\mu}(b)\|^2/\lambda^2 - \rho^*(-\bar{\mu}(b)/\lambda)\,\mathrm{d}x \\
\int_b \rho^*(-\mu(x)/\lambda) - \rho^*(-\bar{\mu}(b)/\lambda)\,\mathrm{d}x &\leq \frac{1}{2\zeta\lambda^2}\int_b \|\mu(x)\|^2 - \|\bar{\mu}(b)\|^2\,\mathrm{d}x.
\end{aligned}$$

We use the fact that $\int_b \|\mu(x) - \bar{\mu}(b)\|^2\,\mathrm{d}x = \int_b \|\mu(x)\|^2 + \|\bar{\mu}(b)\|^2 - 2\langle\mu(x), \bar{\mu}(b)\rangle\,\mathrm{d}x = \int_b \|\mu(x)\|^2 + \|\bar{\mu}(b)\|^2\,\mathrm{d}x - 2\langle\bar{\mu}(b), \int_b \mu(x)\,\mathrm{d}x\rangle = \int_b \|\mu(x)\|^2 + \|\bar{\mu}(b)\|^2\,\mathrm{d}x - 2\langle\bar{\mu}(b), |b|\bar{\mu}(b)\rangle = \int_b \left\|\mu(x)^2\right\| - \|\bar{\mu}(b)\|^2\,\mathrm{d}x$ and we get finally

$$\int_b \rho^*(-\mu(x)/\lambda) - \rho^*(-\bar{\mu}(b)/\lambda)\,\mathrm{d}x \leq \frac{1}{2\zeta\lambda^2}\int_b \|\mu(x) - \bar{\mu}(b)\|^2\,\mathrm{d}x.$$

Equation (2) shows that

$$\begin{aligned}
L(\tilde{p}^\star) - L(p^\star) &\leq \frac{1}{2\zeta\lambda}\sum_{b\in\mathcal{B}}\int_b \|\bar{\mu}(b) - \mu(x)\|^2\,\mathrm{d}x \\
&\leq \sum_{b\in\mathcal{B}}\int_b \frac{L_\beta K}{2\zeta\lambda}\left(\frac{\sqrt{d}}{B}\right)^{2\beta}\mathrm{d}x \\
&\leq \frac{L_\beta K d^\beta}{2\zeta\lambda}\left(\frac{1}{B}\right)^{2\beta}
\end{aligned}$$

because each μ_k is (L_β, β)-Hölder.

□

Proof of Theorem 2. We denote again by C_k the constants. We sum the approximation and the estimation errors (given in Propositions 6 and 4) to obtain the following bound on the regret:

$$R(T) \leq C_1\frac{L_\beta K d^\beta}{\zeta\lambda}B^{-2\beta} + C_2\frac{\log^2(T)}{T}B^d\left(\frac{1}{\zeta\lambda\eta^3} + \frac{K}{\zeta^2\lambda^2\eta^4} + \lambda\zeta\eta^2 + \lambda S\right) + 4B^d(1 + \|\lambda\rho\|_\infty)\exp\left(-\frac{T}{12B^d}\right).$$

For the sake of clarity let us note $\xi_1 \doteq C_1\frac{L_\beta K d^\beta}{\zeta\lambda}$ and $\xi_2 \doteq C_2\left(\frac{1}{\zeta\lambda\eta^3} + \frac{K}{\zeta^2\lambda^2\eta^4} + \lambda\zeta\eta^2 + \lambda S\right)$.

We have

$$R(T) \leq \xi_1 B^{-2\beta} + \xi_2 B^d\frac{\log^2(T)}{T} + 4B^d(1 + \|\lambda\rho\|_\infty)\exp\left(-\frac{T}{12B^d}\right).$$

Taking

$$B = \left(\frac{2\xi_1\beta}{\xi_2}\right)^{1/(2\beta+d)}\left(\frac{T}{\log^2(T)}\right)^{1/(d+2\beta)},$$

we notice that the third term is negligible and we conclude that

$$R(T) \leq \mathcal{O}\left(2\xi_1 \left(\frac{2\xi_1\beta}{\xi_2}\right)^{-2\beta/(2\beta+d)} \left(\frac{T}{\log^2(T)}\right)^{-2\beta/(2\beta+d)}\right).$$

□

C Proofs of Intermediate Rates

We begin with a lemma on convex conjugates.

Lemma 5. *Let $\lambda, \mu > 0$ and let $y \in \mathbb{R}^n$ and ρ a non-negative bounded convex function. Then*

$$(\lambda\rho)^*(y) - (\mu\rho)^*(y) \leq |\lambda - \mu| \, \|\rho\|_\infty .$$

Proof. $(\lambda\rho)^*(y) = \sup_x \langle x, y\rangle - \lambda\rho(x) = \langle x_\lambda, y\rangle - \lambda\rho(x_\lambda)$.
And $(\mu\rho)^*(y) = \sup_x \langle x, y\rangle - \mu\rho(x) = \langle x_\mu, y\rangle - \mu\rho(x_\mu) \geq \langle x_\lambda, y\rangle - \mu\rho(x_\lambda)$.
Then, $(\lambda\rho)^*(y) - (\mu\rho)^*(y) \leq \langle x_\lambda, y\rangle - \lambda\rho(x_\lambda) - (\langle x_\lambda, y\rangle - \mu\rho(x_\lambda)) = (\mu - \lambda)\rho(x_\lambda)$.
Finally $(\lambda\rho)^*(y) - (\mu\rho)^*(y) \leq |\lambda - \mu| \, \|\rho\|_\infty$. □

Proof of Proposition 7. There exists $x_0 \in b$ such that $\bar{\lambda}(b) = \lambda(x_0)$ and $x_1 \in b$ such that $\bar{\mu}(b) = \mu(x_1)$. We use Lemma 5 to derive a bound for the approximation error.

$$\begin{aligned}
&\int_b (\lambda(x)\rho)^*\left(-\mu(x)\right) - (\bar{\lambda}(b)\rho)^*\left(-\bar{\mu}(b)\right) \mathrm{d}x \\
&= \int_b (\lambda(x)\rho)^*\left(-\mu(x)\right) - (\lambda(x)\rho)^*\left(-\bar{\mu}(b)\right) \mathrm{d}x + \int_b (\lambda(x)\rho)^*\left(-\bar{\mu}(b)\right) - (\bar{\lambda}(b)\rho)^*\left(-\bar{\mu}(b)\right) \mathrm{d}x \\
&\leq \int_b \lambda(x)\left(\rho^*\left(-\frac{\mu(x)}{\lambda(x)}\right) - \rho^*\left(-\frac{\bar{\mu}(b)}{\lambda(x)}\right)\right) \mathrm{d}x + \int_b |\lambda(x) - \bar{\lambda}(b)| \, \|\rho\|_\infty \, \mathrm{d}x \\
&\leq \int_b \lambda(x)\left|\frac{\mu(x)}{\lambda(x)} - \frac{\bar{\mu}(b)}{\lambda(x)}\right| \mathrm{d}x + \|\rho\|_\infty \int_b |\lambda(x) - \lambda(x_0)| \, \mathrm{d}x \\
&\leq \int_b L_\beta |x - x_1|^\beta \, \mathrm{d}x + \|\rho\|_\infty \int_b \|\lambda'\|_\infty |x - x_0| \, \mathrm{d}x \\
&\leq B^{-d}\left(L_\beta d^{\beta/2} B^{-\beta} + \|\rho\|_\infty \|\lambda'\|_\infty \sqrt{d} B^{-1}\right) = \mathcal{O}(B^{-\beta-d}).
\end{aligned}$$

□

Proof of Proposition 8. As in the proof of Proposition 6 we consider a bin $b \in \mathcal{B}$ and the goal is to bound

$$\int_b \lambda(x)\rho^*\left(-\frac{\mu(x)}{\lambda(x)}\right) - \bar{\lambda}(b)\rho^*\left(-\frac{\bar{\mu}(b)}{\bar{\lambda}(b)}\right) \mathrm{d}x.$$

We use a similar method and we apply Jensen inequality with density $\dfrac{\lambda(x)}{|b|\bar{\lambda}(b)}$ to the function $g : x \mapsto$

$\frac{1}{2\zeta}\|x\|^2 - \rho^*(x)$ which is convex.

$$
\begin{aligned}
g\left(\int_b -\frac{\mu(x)}{\lambda(x)}\frac{\lambda(x)}{|b|\bar{\lambda}(b)}\,\mathrm{d}x\right) &\leq \int_b g\left(-\frac{\mu(x)}{\lambda(x)}\right)\frac{\lambda(x)}{|b|\bar{\lambda}(b)}\,\mathrm{d}x \\
g\left(-\frac{\bar{\mu}(b)}{\bar{\lambda}(b)}\right) &\leq \int_b g\left(-\frac{\mu(x)}{\lambda(x)}\right)\frac{\lambda(x)}{|b|\bar{\lambda}(b)}\,\mathrm{d}x \\
\frac{1}{2\zeta}\left\|-\frac{\bar{\mu}(b)}{\bar{\lambda}(b)}\right\|^2 - \rho^*\left(-\frac{\bar{\mu}(b)}{\bar{\lambda}(b)}\right) &\leq \frac{1}{|b|\bar{\lambda}(b)}\int_b\left[\frac{1}{2\zeta}\left\|-\frac{\mu(x)}{\lambda(x)}\right\|^2 - \rho^*\left(-\frac{\mu(x)}{\lambda(x)}\right)\right]\lambda(x)\,\mathrm{d}x \\
\int_b \lambda(x)\rho^*\left(-\frac{\mu(x)}{\lambda(x)}\right) - \bar{\lambda}(b)\rho^*\left(-\frac{\bar{\mu}(b)}{\bar{\lambda}(b)}\right)\,\mathrm{d}x &\leq \frac{1}{2\zeta}\int_b \frac{\|\mu(x)\|^2}{\lambda(x)} - \frac{\|\bar{\mu}(b)\|^2}{\bar{\lambda}(b)}\,\mathrm{d}x.
\end{aligned}
$$

Consequently we have proven that

$$
\begin{aligned}
\int_b \lambda(x)\rho^*\left(-\frac{\mu(x)}{\lambda(x)}\right) - \bar{\lambda}(b)\rho^*\left(-\frac{\bar{\mu}(b)}{\bar{\lambda}(b)}\right)\,\mathrm{d}x &\leq \frac{1}{2\zeta}\int_b \frac{\|\mu(x)\|^2}{\lambda(x)} - \frac{\|\bar{\mu}(b)\|^2}{\bar{\lambda}(b)}\,\mathrm{d}x \\
&\leq \frac{1}{2\zeta}\sum_{k=1}^{K}\int_b \frac{\mu_k(x)^2}{\lambda(x)} - \frac{\bar{\mu}_k(b)^2}{\bar{\lambda}(b)}\,\mathrm{d}x.
\end{aligned}
$$

Therefore we have to bound, for each k, $I = \displaystyle\int_b \frac{\mu_k(x)^2}{\lambda(x)} - \frac{\bar{\mu}_k(b)^2}{\bar{\lambda}(b)}\,\mathrm{d}x$.
Let us omit the subscript k and consider a β-Hölder function μ.
We have

$$
\begin{aligned}
I &= \int_b \frac{\mu(x)^2}{\lambda(x)} - \frac{\bar{\mu}(b)^2}{\bar{\lambda}(b)}\,\mathrm{d}x \\
&= \int_b \frac{\mu(x)^2}{\lambda(x)} - \frac{\mu(x)^2}{\bar{\lambda}(b)} + \frac{\mu(x)^2}{\bar{\lambda}(b)} - \frac{\bar{\mu}(b)^2}{\bar{\lambda}(b)}\,\mathrm{d}x \\
&= \underbrace{\int_b \left(\mu(x)^2 - \bar{\mu}(b)^2\right)\left(\frac{1}{\lambda(x)} - \frac{1}{\bar{\lambda}(b)}\right)\,\mathrm{d}x}_{I_1} + \underbrace{\int_b \bar{\mu}(b)^2\left(\frac{1}{\lambda(x)} - \frac{1}{\bar{\lambda}(b)}\right)\,\mathrm{d}x}_{I_2} + \underbrace{\int_b \frac{1}{\bar{\lambda}(b)}\left(\mu(x)^2 - \bar{\mu}(b)^2\right)\,\mathrm{d}x}_{I_3}.
\end{aligned}
$$

We now have to bound these three integrals.

Bounding I_1:

$$
\begin{aligned}
I_1 &= \int_b \left(\mu(x)^2 - \bar{\mu}(b)^2\right)\left(\frac{1}{\lambda(x)} - \frac{1}{\bar{\lambda}(b)}\right)\,\mathrm{d}x \\
&= \int_b \left(\mu(x) + \bar{\mu}(b)\right)\left(\mu(x) - \bar{\mu}(b)\right)\left(\frac{1}{\lambda(x)} - \frac{1}{\bar{\lambda}(b)}\right)\,\mathrm{d}x \\
&\leq \int_b 2|\mu(x) - \bar{\mu}(b)|\left|\frac{1}{\lambda(x)} - \frac{1}{\bar{\lambda}(b)}\right|\,\mathrm{d}x \\
&\leq 2L_\beta\left(\frac{\sqrt{d}}{B}\right)^\beta \int_b \left|\frac{1}{\lambda(x)} - \frac{1}{\bar{\lambda}(b)}\right|\,\mathrm{d}x.
\end{aligned}
$$

Since $1/\lambda$ is of class $\mathcal{C}^1$, Taylor-Lagrange inequality yields, using the fact that there exists $x_0 \in b$ such that $\bar{\lambda}(b) = \lambda(x_0)$:

$$
\left|\frac{1}{\lambda(x)} - \frac{1}{\bar{\lambda}(b)}\right| \leq \left\|\left(\frac{1}{\lambda}\right)'\right\|_\infty |x - x_0| \leq \frac{\|\lambda'\|_\infty}{\lambda_{\min}^2}\frac{\sqrt{d}}{B}.
$$

We obtain therefore

$$I_1 \leq 2L_\beta \left\|\lambda'\right\|_\infty \sqrt{d}^{\beta+1} \frac{1}{\lambda_{\min}^2} B^{-(1+\beta+d)} = \mathcal{O}\left(\frac{B^{-(1+\beta+d)}}{\lambda_{\min}^2}\right).$$

Bounding I_2:

We have

$$I_2 = \bar{\mu}(b)^2 \int_b \left(\frac{1}{\lambda(x)} - \frac{1}{\bar{\lambda}(b)}\right) \mathrm{d}x \leq \int_b \left(\frac{1}{\lambda(x)} - \frac{1}{\bar{\lambda}(b)}\right) \mathrm{d}x$$

because $\displaystyle\int_b \left(\frac{1}{\lambda(x)} - \frac{1}{\bar{\lambda}(b)}\right) \mathrm{d}x \geq 0$ from Jensen's inequality.

Without loss of generality we can assume that the bin b is the closed cuboid $[0, 1/B]^d$. We suppose that for all $x \in b$, $\lambda(x) > 0$.

Since λ is of class $\mathcal{C}^\infty$, we have the following Taylor series expansion:

$$\lambda(x) = \lambda(0) + \sum_{i=1}^d \frac{\partial \lambda(0)}{\partial x_i} x_i + \frac{1}{2} \sum_{i,j} \frac{\partial^2 \lambda(0)}{\partial x_i \partial x_j} x_i x_j + o(\|x\|^2).$$

Integrating over the bin b we obtain

$$\bar{\lambda}(b) = \lambda(0) + \frac{1}{2}\frac{1}{B} \sum_{i=1}^d \frac{\partial \lambda(0)}{\partial x_i} + \frac{1}{8}\frac{1}{B^2} \sum_{i \neq j} \frac{\partial^2 \lambda(0)}{\partial x_i \partial x_j} + \frac{1}{6}\frac{1}{B^2} \sum_{i=1}^d \frac{\partial^2 \lambda(0)}{\partial x_i^2} + o\left(\frac{1}{B^2}\right).$$

Consequently

$$\begin{aligned}
\int_b \frac{\mathrm{d}x}{\bar{\lambda}(b)} &= \frac{1}{B^d \bar{\lambda}(b)} \\
&= \frac{1}{B^d \lambda(0)} \frac{1}{1 + \frac{1}{2\lambda(0)}\frac{1}{B} \sum_{i=1}^d \frac{\partial \lambda(0)}{\partial x_i} + \frac{1}{\lambda(0)}\frac{1}{B^2} \left(\frac{1}{8} \sum_{i \neq j} \frac{\partial^2 \lambda(0)}{\partial x_i \partial x_j} + \frac{1}{6} \sum_{i=1}^d \frac{\partial^2 \lambda(0)}{\partial x_i^2}\right) + o\left(\frac{1}{B^2}\right)} \\
&= \frac{1}{B^d \lambda(0)} \left(1 - \frac{1}{2\lambda(0)}\frac{1}{B} \sum_{i=1}^d \frac{\partial \lambda(0)}{\partial x_i} - \frac{1}{\lambda(0)}\frac{1}{B^2} \left(\frac{1}{8} \sum_{i \neq j} \frac{\partial^2 \lambda(0)}{\partial x_i \partial x_j} + \frac{1}{6} \sum_{i=1}^d \frac{\partial^2 \lambda(0)}{\partial x_i^2}\right)\right. \\
&\qquad \left. + \frac{1}{4\lambda(0)^2}\frac{1}{B^2} \left(\sum_{i=1}^d \frac{\partial \lambda(0)}{\partial x_i}\right)^2 + o\left(\frac{1}{B^2}\right)\right) \\
&= \frac{1}{B^d \lambda(0)} - \frac{1}{2\lambda(0)^2}\frac{1}{B^{d+1}} \sum_{i=1}^d \frac{\partial \lambda(0)}{\partial x_i} - \frac{1}{\lambda(0)^2}\frac{1}{B^{d+2}} \left(\frac{1}{8} \sum_{i \neq j} \frac{\partial^2 \lambda(0)}{\partial x_i \partial x_j} + \frac{1}{6} \sum_{i=1}^d \frac{\partial^2 \lambda(0)}{\partial x_i^2}\right) \\
&\qquad + \frac{1}{4\lambda(0)^3}\frac{1}{B^{d+2}} \left(\sum_{i=1}^d \frac{\partial \lambda(0)}{\partial x_i}\right)^2 + o\left(\frac{1}{B^2}\right).
\end{aligned}$$

Let us now compute the Taylor series development of $1/\lambda$. We have:

$$\frac{\partial}{\partial x_i}\frac{1}{\lambda(x)} = -\frac{1}{\lambda(x)^2}\frac{\partial \lambda(x)}{\partial x_i} \quad \text{and} \quad \frac{\partial^2}{\partial x_i \partial x_j}\frac{1}{\lambda(x)} = -\frac{1}{\lambda(x)^2}\frac{\partial^2 \lambda(x)}{\partial x_i \partial x_j} + \frac{2}{\lambda(x)^3}\frac{\partial \lambda(x)}{\partial x_i}\frac{\partial \lambda(x)}{\partial x_j}.$$

This lets us write

$$\frac{1}{\lambda(x)} = \frac{1}{\lambda(0)} - \frac{1}{\lambda(0)^2}\sum_{i=1}^{d}\frac{\partial\lambda(0)}{\partial x_i}x_i - \frac{1}{2}\frac{1}{\lambda(0)^2}\sum_{i,j}\frac{\partial^2\lambda(0)}{\partial x_i\partial x_j}x_ix_j + \frac{1}{\lambda(0)^3}\sum_{i,j}\frac{\partial\lambda(0)}{\partial x_i}\frac{\partial\lambda(0)}{\partial x_j}x_ix_j + o(\|x\|^2)$$

$$\begin{aligned}\int_b \frac{\mathrm{d}x}{\lambda(x)} = \frac{1}{\lambda(0)}\frac{1}{B^d} - \frac{1}{2\lambda(0)^2}\frac{1}{B^{d+1}}\sum_{i=1}^{d}\frac{\partial\lambda(0)}{\partial x_i} - \frac{1}{\lambda(0)^2}\frac{1}{B^{d+2}}\left(\frac{1}{8}\sum_{i\neq j}\frac{\partial^2\lambda(0)}{\partial x_i\partial x_j} + \frac{1}{6}\sum_{i=1}^{d}\frac{\partial^2\lambda(0)}{\partial x_i^2}\right)\\ + \frac{1}{\lambda(0)^3}\frac{1}{B^{d+2}}\left(\frac{1}{4}\sum_{i\neq j}\frac{\partial\lambda(0)}{\partial x_i}\frac{\partial\lambda(0)}{\partial x_j} + \frac{1}{3}\sum_{i=1}^{d}\left(\frac{\partial\lambda(0)}{\partial x_i}\right)^2\right) + o\left(\frac{1}{B^{d+2}}\right).\end{aligned}$$

And then

$$I_2 \leq \frac{1}{12}\frac{1}{\lambda(0)^3}\frac{1}{B^{d+2}}\sum_{i=1}^{d}\left(\frac{\partial\lambda(0)}{\partial x_i}\right)^2 + o\left(\frac{1}{B^{d+2}}\right).$$

Since the derivatives of λ are bounded we obtain that

$$I_2 = \mathcal{O}\left(\frac{B^{-2-d}}{\lambda_{\min}^3}\right).$$

Bounding I_3:

$$\begin{aligned}I_3 &= \int_b \frac{1}{\bar{\lambda}(b)}\left(\mu(x)^2 - \bar{\mu}(b)^2\right)\mathrm{d}x\\ &= \frac{1}{\bar{\lambda}(b)}\int_b \left(\mu(x) - \bar{\mu}(b)\right)^2\mathrm{d}x\\ &\leq \frac{1}{\lambda_{\min}}L_\beta^2 d^\beta B^{-(2\beta+d)} = \mathcal{O}\left(\frac{B^{-(2\beta+d)}}{\lambda_{\min}}\right).\end{aligned}$$

Putting this together we have $I = \mathcal{O}\left((dL_\beta^2\|\nabla\lambda\|_\infty^2)\frac{B^{-(2\beta+d)}}{\lambda_{\min}^3}\right)$. And finally

$$L(\tilde{p}^\star) - L(p^\star) \leq \mathcal{O}\left(KdL_\beta^2\|\nabla\lambda\|_\infty^2\frac{B^{-2\beta}}{\zeta\lambda_{\min}^3}\right).$$

□

Lemma 6 (Regularity of η)**.** *If η is the distance of the optimum $p^\star$ to the boundary of Δ^K as defined in Definition 6, and if the μ_k functions are all β-Hölder and λ of class $\mathcal{C}^1$, then η is β-Hölder. More precisely we have*

$$\forall x, y \in b, |\eta(x) - \eta(y)| \leq \sqrt{\frac{K}{K-1}}\frac{\|\lambda\|_\infty + \|\lambda'\|_\infty}{\zeta\lambda_{\min}(b)^2}|x-y|^\beta = \frac{C_L}{\lambda_{\min}(b)^2}|x-y|^\beta.$$

Proof. Let $x \in \mathcal{X}$. Since $\eta(x) = \mathrm{dist}(p_b^\star, \partial\Delta^K)$ we obtain

$$\eta(x) = \sqrt{\frac{K}{K-1}}\min_i p_i^\star(x).$$

And

$$\begin{aligned}p^\star(x) &= \mathrm{argmin}\langle\mu(x), p(x)\rangle + \lambda(x)\rho(p(x))\\ &= \nabla(\lambda(x)\rho)^*(-\mu(x))\\ &= \nabla\rho^*\left(-\frac{\mu(x)}{\lambda(x)}\right).\end{aligned}$$

Since ρ is ζ-strongly convex, $\nabla\rho^*$ is $1/\zeta$-Lipschitz continuous.
Therefore, for $x, y \in b$,

$$\begin{aligned}
|p^\star(x) - p^\star(y)| &\leq \frac{1}{\zeta}\left|\frac{\mu(x)}{\lambda(x)} - \frac{\mu(y)}{\lambda(y)}\right| \\
&\leq \frac{1}{\zeta}\left|\frac{\mu(x)-\mu(y)}{\lambda(x)}\right| + \frac{1}{\zeta}|\mu(y)|\left|\frac{1}{\lambda(x)} - \frac{1}{\lambda(y)}\right| \\
&\leq \frac{1}{\zeta\lambda_{\min}(b)}|x-y|^\beta + \frac{1}{\zeta}\frac{\|\lambda'\|_\infty}{\lambda_{\min}(b)^2}|x-y|
\end{aligned}$$

since all μ_k are bounded by 1 (the losses are bounded by 1). □

Proof of Lemma 2. We consider a well-behaved bin b. There exists $x_1 \in b$ such that $\lambda(x_1) \geq c_1 B^{-\beta/3}$. Since λ is $\mathcal{C}^\infty$ on $[0,1]^d$, it is in particular Lipschitz-continuous on b. And therefore

$$\forall x \in b,\ \lambda(x) \geq c_1 B^{-\beta/3} - \|\lambda'\|_\infty \operatorname{diam}(b) \geq c_1 B^{-\beta/3} - \|\lambda'\|_\infty \operatorname{diam}(b)^{\beta/3} = B^{-\beta/3}.$$

Lemma 6 shows that η is β-Hölder continuous (with constant denoted by $C_L/\lambda_{\min}^2$) and therefore we have

$$\forall x \in b,\ \eta(x) \geq c_2 B^{-\beta/3} - \frac{C_L}{\lambda_{\min}(b)^2}\operatorname{diam}(b)^\beta = B^{-\beta/3}.$$

□

Lemma 7. *If ρ is convex, η is an increasing function of λ.*

Proof. As in the proof of Proposition 2 we use the KKT conditions to find that on a bin b (without the index k for the arm):

$$\bar{\mu}(b) + \bar{\lambda}(b)\nabla\rho(p_b^\star) + \xi = 0.$$

Therefore

$$p_b^\star = (\nabla\rho)^{-1}\left(-\frac{\xi + \bar{\mu}(b)}{\bar{\lambda}(b)}\right).$$

Since ρ is convex, $\nabla\rho$ is an increasing function and its inverse as well. Consequently $p_b^\star$ is an increasing function of $\bar{\lambda}(b)$, and since $\eta(b) = \sqrt{K/(K-1)}\min_i p_{b,i}^\star$, η is also an increasing function of $\bar{\lambda}(b)$. □

Proof of Theorem 3. Since B will be chosen as an increasing function of T we only consider T sufficiently large in order to have $c_1 B^{-\beta/3} < \delta_1$ and $c_2 B^{-\beta/3} < \delta_2$. To ensure this we can also take smaller δ_1 and δ_2. Moreover we lower the value of δ_2 or δ_1 to be sure that $\frac{\delta_2}{c_2} = \eta(\frac{\delta_1}{c_1})$. These are technicalities needed to simplify the proof.

The proof will be divided into several steps. We will first obtain lower bounds on λ and η for the "well-behaved bins". Then we will derive bounds for the approximation error and the estimation error. And finally we will put that together to obtain the intermediate convergence rates.

As in the proofs on previous theorems we will denote the constants C_k with increasing values of k.

- **Lower bounds on η and λ:**

  Using a technique from Rigollet and Zeevi (2010) we notice that without loss of generality we can index the B^d bins with increasing values of $\bar{\lambda}(b)$. Let us note $\mathcal{IB} = \{1, \dots, j_1\}$ and $\mathcal{WB} = \{j_1 + 1, \dots, B^d\}$. Since η is an increasing function of λ (cf Lemma 7), the $\eta(b_j)$ are also increasingly ordered.

  Let $j_2 \geq j_1$ be the largest integer such that $\bar{\lambda}(b_j) \leq \dfrac{\delta_1}{c_1}$. Consequently we also have that j_2 is the largest integer such that $\eta(b_j) \leq \dfrac{\delta_2}{c_2}$.

Let $j \in \{j_1+1, \dots, j_2\}$. The bin b_j is a well-behaved bin and Lemma 2 shows that $\bar{\lambda}(b_j) \geq B^{-\beta/3}$. Then $\bar{\lambda}(b_j) + (c_1-1)B^{-\beta/3} \leq c_1\bar{\lambda}(b_j) \leq \delta_1$ and we can apply the margin condition (cf Assumption 3) which gives

$$\mathbb{P}_X(\lambda(x) \leq \bar{\lambda}(b_j) + (c_1-1)B^{-\beta/3}) \leq C_m(c_1\bar{\lambda}(b_j))^{6\alpha}.$$

But since the context are uniformly distributed and since the $\bar{\lambda}(b_j)$ are increasingly ordered we also have that

$$\mathbb{P}_X(\lambda(x) \leq \bar{\lambda}(b_j) + (c_1-1)B^{-\beta/3}) \geq \mathbb{P}_X(\lambda(x) \leq \bar{\lambda}(b_j)) \geq \frac{j}{B^d}.$$

This gives $\bar{\lambda}(b_j) \geq \frac{1}{c_1 C_m^{1/6\alpha}} \left(\frac{j}{B^d}\right)^{1/6\alpha}$. The same computations give $\eta(b_j) \geq \frac{1}{c_2 C_m^{1/6\alpha}} \left(\frac{j}{B^d}\right)^{1/6\alpha}$. We note $C_\gamma \doteq \min((c_1 C_m^{1/6\alpha})^{-1}, (c_2 C_m^{1/6\alpha})^{-1}))$ and $\gamma_j \doteq C_\gamma \left(\frac{j}{B^d}\right)^{1/\alpha}$. Consequently $\bar{\lambda}(b_j) \geq \gamma_j$ and $\eta(b_j) \geq \gamma_j$.

Let us now compute the number of ill-behaved bins:

$$\begin{aligned}
\#\{b \in \mathcal{B}, b \notin \mathcal{WB}\} &= B^d \ \mathbb{P}(b \notin \mathcal{WB}) \\
&= B^d \ \mathbb{P}(\forall x \in \mathcal{B},\ \eta(x) \leq c_2 B^{-\beta/3} \text{ or } \forall x \in \mathcal{B},\ \lambda(x) \leq c_1 B^{-\beta/3}) \\
&\leq B^d \ \mathbb{P}(\eta(\bar{x}) \leq c_2 B^{-\beta/3} \text{ or } \lambda(\bar{x}) \leq c_1 B^{-\beta/3}) \\
&\leq C_m(c_1^{6\alpha} + c_2^{6\alpha}) B^d B^{-2\alpha\beta} \doteq C_I B^d B^{-2\alpha\beta}
\end{aligned}$$

where $\bar{x}$ is the mean context value in the bin b. Consequently if $j \geq j^\star \doteq C_I B^d B^{-2\alpha\beta}$, then $b_j \in \mathcal{WB}$. Let $\hat{j} \doteq C_I B^d B^{-\alpha\beta} \geq j^\star$. Consequently for all $j \geq j^\star$, $b_j \in \mathcal{WB}$.

We want to obtain an upper-bound on the constant $S\bar{\lambda}(b_j) + \frac{K}{\eta(b_j)^4 \bar{\lambda}(b_j)^2}$ that arises in the fast rate for the estimation error. For the sake of clarity we will remove the dependency in b_j and denote this constant $C = S\lambda + \frac{K}{\lambda^2 \eta^4}$.

In the case of the entropy regularization $S = 1/\min_i p_i^\star$. Since $\eta = \sqrt{K/(K-1)}\min_i p_i^\star$, we have that $\min_i p_i^\star = \sqrt{(K-1)/K}\eta \geq \eta/2$. Consequently $S \leq 2/\gamma_j$ and, on a well-behaved bin b_j, for $j \leq j_2$,

$$C \leq \frac{K + 2\|\lambda\|_\infty}{\gamma_j^6} \doteq \frac{C_F}{\gamma_j^6}, \tag{5}$$

where the subscript F stands for "Fast". When $j \geq j_2$, we have $\bar{\lambda}(b_j) \geq \delta_1/c_1$ and $\eta(b_j) \geq \delta_2/c_2$ and consequently

$$C \leq \frac{K}{(\delta_1/c_1)^2(\delta_2/c_2)^4} + \frac{2\|\lambda\|_\infty}{\delta_2/c_2} \doteq C_{\max}.$$

Let us notice than λ being known by the agent, the agent knows the value of $\bar{\lambda}(b)$ on each bin b and can therefore order the bins. Consequently the agent can sample, on every well-behaved bin, each arm $T\gamma_j/2$ times and be sure that $\min_i p_i \geq \gamma_j/2$. On the first $\lfloor \hat{j} \rfloor$ bins the agent will sample each arm $\bar{\lambda}(b)\sqrt{T/B^d}$ times as in the proof of Proposition 2.

- **Approximation Error:**

  We now bound the approximation error. We separate the bins into two sets: $\{1, \dots, \lfloor j^\star \rfloor\}$ and $\{\lceil j^\star \rceil, \dots, B^d\}$. On the first set we use the slow rates of Proposition 7 and on the second set we use the fast rates of Proposition 8.

We obtain that, for $\alpha < 1/2$,

$$
\begin{aligned}
L(\tilde{p}^\star) - L(p^\star) &\leq L_\beta d^{\beta/2} \sum_{j=1}^{\lfloor j^\star \rfloor} B^{-\beta-d} + \|\rho\|_\infty \|\nabla\lambda\|_\infty \sqrt{d} \sum_{j=1}^{\lfloor j^\star \rfloor} B^{-1-d} + (KdL_\beta^2 \|\nabla\lambda\|_\infty^2) \sum_{j=\lceil j^\star \rceil}^{B^d} \frac{B^{-2\beta-d}}{\bar{\lambda}(b_j)^3} \\
&\leq C_I L_\beta d^{\beta/2} B^{-\beta} B^{-2\alpha\beta} + (KdL_\beta^2 \|\nabla\lambda\|_\infty^2) \left(\sum_{j=\lceil j^\star \rceil}^{j_2} \frac{B^{-2\beta-d}}{\gamma_j^3} + \sum_{j=j_2+1}^{B^d} \frac{B^{-2\beta-d}}{(c_1/\delta_1)^3} \right) + o(B^{-2\alpha\beta-\beta}) \\
&\leq C_I L_\beta d^{\beta/2} B^{-2\alpha\beta-\beta} + (KdL_\beta^2 \|\nabla\lambda\|_\infty^2) \left(\frac{B^{-2\beta-d}}{C_\gamma^3} \sum_{j=\lceil j^\star \rceil}^{j_2} \left(\frac{j}{B^d} \right)^{-1/2\alpha} + B^{-2\beta} \left(\frac{\delta_1}{c_1} \right)^3 \right) + o(B^{-2\alpha\beta-\beta}) \\
&\leq C_I L_\beta d^{\beta/2} B^{-2\alpha\beta-\beta} + (KdL_\beta^2 \|\nabla\lambda\|_\infty^2) \frac{1}{C_\gamma^3} B^{-2\beta} \int_{C_I B^{-2\alpha\beta}}^1 x^{-1/2\alpha} \, \mathrm{d}x + o(B^{-2\alpha\beta-\beta}) \\
&\leq \left(C_I L_\beta d^{\beta/2} + KdL_\beta^2 \|\nabla\lambda\|_\infty^2 \frac{2\alpha}{1-2\alpha} \frac{C_I^{(2\alpha-1)/2\alpha}}{C_\gamma^3} \right) B^{-\beta-2\alpha\beta} + o(B^{-2\alpha\beta-\beta}) = \mathcal{O}\left(B^{-\beta-2\alpha\beta}\right)
\end{aligned}
$$

since $\alpha < 1/2$. We step from line 3 to 4 thanks to a series-integral comparison.

For $\alpha = 1/2$ we get

$$
L(\tilde{p}^\star) - L(p^\star) \leq \left(C_I L_\beta d^{\beta/2} + \left(KdL_\beta^2 \|\nabla\lambda\|_\infty^2 \right) (\delta_1^3 c_1^{-3} + 2\beta C_\gamma^{-3} \log(B)) \right) B^{-2\beta} + o(B^{-2\beta}) = \mathcal{O}\left(B^{-2\beta} \log(B)\right).
$$

And for $\alpha > 1/2$ we have

$$
L(\tilde{p}^\star) - L(p^\star) \leq \left(KdL_\beta^2 \|\nabla\lambda\|_\infty^2 \right) \left(\frac{1}{C_\gamma^3} \frac{2\alpha}{2\alpha-1} + \left(\frac{\delta_1}{c_1} \right)^3 \right) B^{-2\beta} + o(B^{-2\beta}) = \mathcal{O}\left(B^{-2\beta}\right)
$$

because $\beta + 2\alpha\beta > 2\beta$.

Let us note

$$
\begin{aligned}
\xi_1 &\doteq \left(C_I L_\beta d^{\beta/2} + KdL_\beta^2 \|\nabla\lambda\|_\infty^2 \frac{2\alpha}{1-2\alpha} \frac{C_I^{(2\alpha-1)/2\alpha}}{C_\gamma^3} \right); \\
\xi_2 &\doteq \left(C_I L_\beta d^{\beta/2} + \left(KdL_\beta^2 \|\nabla\lambda\|_\infty^2 \right) (\delta_1^3 c_1^{-3} + 2\beta C_\gamma^{-3} \log(B)) \right); \\
\xi_3 &\doteq \left(KdL_\beta^2 \|\nabla\lambda\|_\infty^2 \right) \left(\frac{1}{C_\gamma^3} \frac{2\alpha}{2\alpha-1} + \left(\frac{\delta_1}{c_1} \right)^3 \right); \\
\xi_{app} &\doteq \max(\xi_1, \xi_2, \xi_3).
\end{aligned}
$$

Finally we obtain that the approximation error is bounded by $\xi_{app} B^{-\min(\beta+2\alpha\beta, 2\beta)} \log(B)$ with $\alpha > 0$.

- **Estimation Error:**

  We proceed in a similar manner as for the approximation error, except that we do not split the bins around $j^\star$ but around $\hat{j}$.

  In a similar manner to the proofs of Theorems 1 and 2 we only need to consider the terms of dominating order from Propositions 1 and 4. As before we consider the same event A (cf the proof of Proposition 1)

and we note $C_A \doteq 4B^d(1+\|\lambda\rho\|_\infty)$. We obtain, for $\alpha<1$, using (5):

$$\begin{aligned}
\mathbb{E}L(\tilde{p}_T) - L(\tilde{p}^\star) &= \frac{1}{B^d}\sum_{b\in\mathcal{B}} \mathbb{E}L_b(\tilde{p}_T) - L(p_b^\star) \\
&= \frac{1}{B^d}\sum_{j=\lceil\hat{j}\rceil}^{B^d} \mathbb{E}L_b(\tilde{p}_T) - L(p_b^\star) + \frac{1}{B^d}\sum_{j=1}^{\lfloor\hat{j}\rfloor} \mathbb{E}L_b(\tilde{p}_T) - L(p_b^\star) \\
&\leq \frac{1}{B^d}\sum_{j=\lceil\hat{j}\rceil}^{B^d} 2C\frac{\log^2(T)}{T/B^d} + \frac{1}{B^d}\sum_{j=1}^{\lfloor\hat{j}\rfloor} 4\sqrt{12K}\sqrt{\frac{\log(T)}{T/B^d}} + C_A e^{-\frac{T}{12B^d}} \\
&\leq 2C_F\sum_{j=\lceil\hat{j}\rceil}^{j_2}\frac{\log^2(T)}{T}\gamma_j^{-6} + \sum_{j=j_2+1}^{B^d} 2C_{\max}\frac{\log^2(T)}{T} + 6\sqrt{3K}\sqrt{\frac{\log(T)}{T}}B^{d/2}B^{-\alpha\beta} + C_A e^{-\frac{T}{12B^d}} \\
&\leq \frac{2C_F}{C_\gamma^6}\frac{\log^2(T)}{T}\sum_{j=\lceil\hat{j}\rceil}^{j_2}\left(\frac{j}{B^d}\right)^{-1/\alpha} + 2C_{\max}\frac{\log^2(T)}{T}B^d + 6\sqrt{3K}\sqrt{\frac{\log(T)}{T}}B^{d/2-\alpha\beta} + C_A e^{-\frac{T}{12B^d}} \\
&\leq \frac{2C_F}{C_\gamma^6}\frac{\log^2(T)}{T}B^d\int_{C_I B^{-\alpha\beta}}^1 x^{-1/\alpha}\,\mathrm{d}x + 2C_{\max}\frac{\log^2(T)}{T}B^d + 6\sqrt{3K}\sqrt{\frac{\log(T)}{T}}B^{d/2-\alpha\beta} + C_A e^{-\frac{T}{12B^d}} \\
&\leq \frac{2C_F}{C_\gamma^6}\frac{\log^2(T)}{T}B^d\frac{\alpha}{1-\alpha}B^{\beta(1-\alpha)} + 2C_{\max}\frac{\log^2(T)}{T}B^d + 6\sqrt{3K}\sqrt{\frac{\log(T)}{T}}B^{d/2-\alpha\beta} + C_A e^{-\frac{T}{12B^d}} \\
&\leq \frac{2C_F}{C_\gamma^6}\frac{\log^2(T)}{T}\frac{\alpha}{1-\alpha}B^{d+\beta-\alpha\beta} + 6\sqrt{3K}\sqrt{\frac{\log(T)}{T}}B^{d/2-\alpha\beta} + 2C_{\max}\frac{\log^2(T)}{T}B^d + C_A e^{-\frac{T}{12B^d}}.
\end{aligned}$$

- **Putting things together:**

  We note $C_\alpha \doteq \frac{2C_F}{C_\gamma^6}\frac{\alpha}{1-\alpha}$. This leads to the following bound on the regret:

  $$R(T) \leq C_\alpha\frac{\log^2(T)}{T}B^{d+\beta-\alpha\beta} + 6\sqrt{3K}\sqrt{\frac{\log(T)}{T}}B^{d/2-\alpha\beta} + 2C_{\max}\frac{\log^2(T)}{T}B^d + C_A e^{-\frac{T}{12B^d}} + \xi_{app}B^{-\min(2\beta,\beta+2\alpha\beta)}\log(B).$$

  Choosing $B = \left(\frac{T}{\log^2(T)}\right)^{1/(2\beta+d)}$ we get

  $$R(T) \leq (C_\alpha + 6\sqrt{3}K)\left(\frac{T}{\log^2(T)}\right)^{-\beta(1+\alpha)/(2\beta+d)} + o\left(\frac{T}{\log^2(T)}\right)^{-\beta(1+\alpha)/(2\beta+d)}$$

  which is valid for $\alpha\in(0,1)$.

  Finally we have

  $$R(T) = \mathcal{O}\left(\left(\frac{T}{\log^2(T)}\right)^{-\beta(1+\alpha)/(2\beta+d)}\right).$$

□

D Proofs of Lower Bounds

Proof of Theorem 4. We consider the model with $K=2$ where $\mu(x) = (-\eta(x), \eta(x))^\top$, where η is a β-Hölder function on $\mathcal{X} = [0,1]^d$. We note that η is uniformly bounded over $\mathcal{X}$ as a consequence of smoothness, so

one can take λ such that $|\eta(x)| < \lambda$. We denote by $e = (1/2, 1/2)$ the center of the simplex, and we consider the loss

$$L(p) = \int_{\mathcal{X}} \big(\langle \mu(x), p(x) \rangle + \lambda \|p(x) - e\|^2 \big) \, \mathrm{d}x.$$

Denoting by $p_0(x)$ the vector $e + \mu(x)/(2\lambda)$, we have that $p_0(x) \in \Delta^2$ for all $x \in \mathcal{X}$. Further, we have that

$$\langle \mu(x), p(x) \rangle + \lambda \|p(x) - e\|^2 = \lambda \|p(x) - p_0(x)\|^2 + 1/(4\lambda) \|\mu(x)\|^2 \,,$$

since $\langle \mu(x), e \rangle = 0$. As a consequence, L is minimized at p_0 and

$$L(p) - L(p_0) = \int_{\mathcal{X}} \lambda \|p(x) - p_0(x)\|^2 \, \mathrm{d}x = 1/(2\lambda) \int_{\mathcal{X}} |\eta(x) - \eta_0(x)|^2 \, \mathrm{d}x \,.$$

where η is such that $p(x) = \big(1/2 - \eta(x)/(2\lambda), 1/2 + \eta(x)/(2\lambda)\big)$. As a consequence, for any algorithm with final variable $\hat{p}_T$, we can construct an estimator $\hat{\eta}_T$ such that

$$\mathbb{E}[L(\hat{p}_T)] - L(p_0) = 1/(2\lambda) \mathbb{E} \int_{\mathcal{X}} |\hat{\eta}_T(x) - \eta_0(x)|^2 \, \mathrm{d}x \,,$$

where the expectation is taken over the randomness of the observations Y_t, with expectation $\pm\eta(X_t)$, with sign depending on the known choice $\pi_t = 1$ or 2. As a consequence, any upper bound on the regret for a policy implies an upper bound on regression over β-Hölder functions in dimension d, with T observations. This yields that, in the special case where ρ is the 1-strongly convex function equal to the squared ℓ_2 norm

$$\inf_{\hat{p}} \sup_{\substack{\mu \in \mathcal{H}_\beta \\ \rho = \ell_2^2}} \mathbb{E}[L(\hat{p}_T)] - L(p_0) \geq \inf_{\hat{\eta}} \sup_{\eta \in \mathcal{H}_\beta} 1/(2\lambda) \mathbb{E} \int_{\mathcal{X}} |\hat{\eta}_T(x) - \eta_0(x)|^2 \, \mathrm{d}x \; \geq C T^{-\frac{2\beta}{2\beta + d}} \,.$$

The final bound is a direct application of Theorem 3.2 in Györfi et al. (2006). □