Regularized Contextual Bandits

Xavier Fontaine
CMLA, ENS Cachan
CNRS, Université Paris-Saclay

Quentin Berthet
Statistical Laboratory
DPMMS, University of Cambridge

Vianney Perchet
CMLA, ENS Cachan
CNRS, Université Paris-Saclay
& Criteo Research, Paris

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 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, 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, yielding intermediate 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") 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).

Contextual bandits have been extensively studied in the last decades and several improvements upon multi-armed bandits algorithms have been applied to contextual bandits (Agrawal and Goyal, 2013; Perchet and Rigollet, 2013; Dudik et al., 2011). They 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. 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

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

rewards. This setting has been studied in Srinivas et al. (2010), Perchet and Rigollet (2013) and Slivkins (2014). 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.

We introduce regularization to the problem of stochastic multi-armed bandits. It is a widely-used technique in machine learning to avoid overfitting or to solve ill-posed problems. Here, the regularization forces 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 well. This has already motivated previous work such as Conservative Bandits (Wu et al., 2016), where an additional arm corresponding to the handcrafted policy is added. By adding regularization, the agent can be sure to end up close to the chosen policy. Within this setting, the form of the objective function is not a classical bandit loss anymore, but contains a regularization term on the global policy. Regularized bandit problems, with no context, have been studied in (Berthet and Perchet, 2017), with applications in online experiment design (Berthet and Chandraskeran, 2016), motivated by computational-statistical tradeoffs (Berthet and Rigollet, 2013; Wang et al., 2016a; Berthet, 2014; Wang et al., 2016b; Baldin and Berthet, 2018; Berthet and Ellenberg, 2019).

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. Based on nonparametric statistics we obtain parameter-independent intermediate convergence rates when the regularization function depends on the context value.

The remaining of this paper is organized as follows. We present the setting and problem in Section 2. Our algorithm is described in Section 3. Sections 4 and 5 are devoted to deriving the convergence rates. Lower bounds are detailed in Section 6 and experiments are presented 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 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 $\mu_k(X) \doteq \mathbb{E}(Y^{(k)}|X)$ 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 (β-Hölder)**.** *For all $k \in [K]$*[1]*,*

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

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 add a 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 $\partial\Delta^K$. 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 known and given to the agent, while the μ_k functions are unknown and must be learned. We want to minimize 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 prove some propositions, the convexity of ρ will not be enough and we will need strong convexity. We will also be led to consider S-smooth functions:

Definition 1. *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.*

[1] $[K] = \{1, \cdots, K\}$

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 2. *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. Note that $R(T)$ is actually a cumulative regret, since p_T is the vector of the empirical frequency of each arm, *i.e.* the normalized total number of pulls of each arm. Earlier choices affect this variable unalterably so that we face a trade-off between exploration and exploitation.

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 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 have an explicit dependency on x since q depends on the context values, which was not the case of the entropy (which only depends on x through p). 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 for the minimization. Let us now note $\tilde{\mu} = \mu + \lambda k$. We are now 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.

2.3 The Upper-Confidence Frank-Wolfe Algorithm

We now briefly present the Upper-Confidence Frank-Wolfe algorithm (UC-FW) from Berthet and Perchet (2017), that will be an important tool of our own algorithm. This algorithm is designed to optimize an unknown convex function $L : \Delta^K \to \mathbb{R}$. At each time step $t \geq 1$ the feedback available is a noisy estimate of $\nabla L(p_t)$, where p_t is the vector of proportions of each action. The algorithm chooses the arm k minimizing a lower confidence estimate of the gradient value (similarly as in the UCB algorithm (Auer et al., 2002)) and updates the proportions vector accordingly. Slow and fast rates for this algorithm are derived by the authors.

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 $1/B$. Inspired by UCBo-grams (Rigollet and Zeevi, 2010) 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 $\operatorname{diam}(b) = \sqrt{d}/B$ 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 functions L_b for each bin $b \in \mathcal{B}$. We fall precisely in the setting of Subsection 2.3 and we propose consequently the following algorithm: for each time step $t \geq 1$, given the context value X_t, we run one iteration of the UC-FW 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 $\alpha_k^{(b)}(T/B^d)$ times arm k for all $k \in [K]$
3: **end for**
4: **for** $t \geq 1$ **do**
5: Receive context X_t from the environment
6: $b_t \leftarrow$ bin of X_t
7: Perform one iteration of the UC-FW 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 Algorithm 1 consists in a pre-sampling stage where all arms are sampled a certain amount of time. It guarantees that $p_T(k)$ is bounded away from 0 so that p_T is bounded away from the boundary of Δ^K, which will be required when L_b is not smooth on $\partial\Delta^K$.

In the remaining of this paper, we derive slow and fast rates of convergence for this algorithm.

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 $b \in \mathcal{B}$, $p_b^\star = \operatorname{arginf}_{p\in\Delta^K} L_b(p)$ 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 3. *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 4. *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 and fast rates of convergence. The proofs are deferred to Appendix A and Appendix B.

4.1 Slow Rates

The analysis of the UC-FW algorithm gives the following bound.

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[2]

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

Some regularization functions are not S-smooth on Δ^K, for example the entropy whose Hessian is not bounded on Δ^K. The following proposition shows that the previous result 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,

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

We want to bound

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

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

With Equation (2) and convex analysis tools 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*[3]

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

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

The detailed proof of the theorem (see Appendix A) consists in choosing a value of B balancing between the estimation and the approximation errors. Since $\beta \in (0, 1]$, we 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*.

When $\lambda = 0$ we are in the usual contextual bandit setting. The propositions of this section hold and we recover the slow rates from Perchet and Rigollet (2013).

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 the Upper-Confidence Frank-Wolfe algorithm.

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

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

This bound depends on several parameters of the problem: λ, distance η of the optimum to the boundary

[3]The notation $\mathcal{O}_{L_\beta,\beta,K,d}$ means that there is a hidden constant depending on L_β, β, K and d. The constant can be found in the proof in Appendix A.

of the simplex, strong convexity and smoothness constants. Since λ can be arbitrarily small, η can be small as well and S large. Therefore the "constant" factor can explode despite the convergence rate being "fast": these terms describe only the dependency in T.

As in the previous section we want to consider regularization functions ρ that are not smooth on $\partial\Delta^K$. To do so we 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 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.

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 :

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

From this expression we obtain the following slow and fast rates of convergence. These rates are the same as in Section 4 in term of the powers of B but have worse dependency in λ.

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:*

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

The important point 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:

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:*

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

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 5. *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*

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

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

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

The fast rates obtained in Subsection 4.2 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, 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*

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

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 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 λ and η can be arbitrarily small. The idea is to use the fast rates on the "well-behaved bins" and the slow rates (independent of λ and η) on the "ill-behaved bins". This is 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

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

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*

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

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	$\left(\frac{T}{\log(T)}\right)^{\frac{1}{2\beta+d}}$	$\left(\frac{T}{\log^2(T)}\right)^{\frac{1}{2\beta+d}}$
$R(T)$	$\left(\frac{T}{\log(T)}\right)^{\frac{-\beta}{2\beta+d}}$	$\left(\frac{T}{\log^2(T)}\right)^{\frac{-2\beta}{2\beta+d}}$

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

Table 1 shows that the slow rates do not depend on the constants, so that we can use them on the "ill-behaved bins".

Theorem 3 (Intermediate rates)**.** *Applying Algorithm 1 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 presampling 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 obtained is $T^{-\frac{\beta}{2\beta+d}}$ 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*

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

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 arms' rewards follow 3 different probability distributions (Poisson, exponential and Bernoulli), with β-Hölder mean functions.

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.

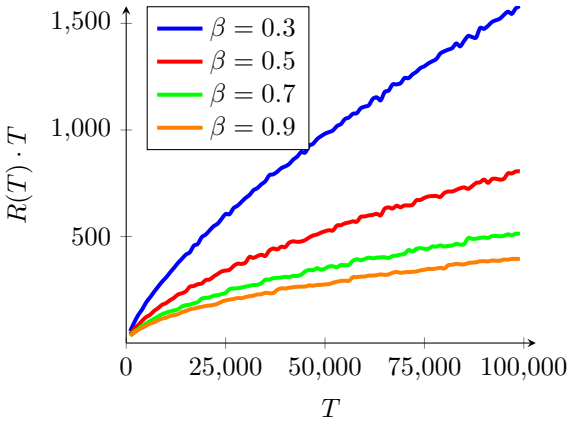

Figure 1: Regret as a Function of T

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.

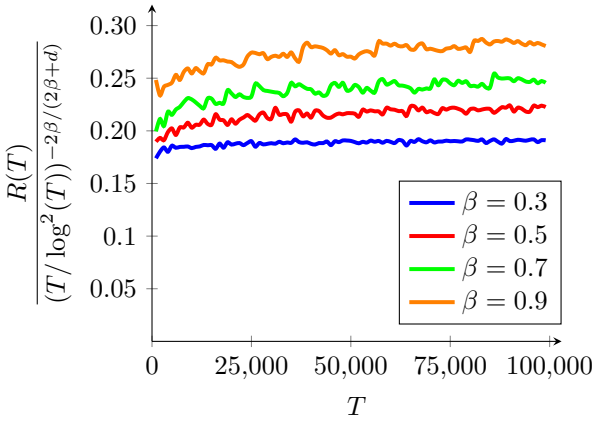

Figure 2: Normalized Regret as a Function of T

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, and validated by our experiments. 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.

Acknowledgments

Xavier Fontaine was supported by grants from Région Ile-de-France. Quentin Berthet was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. Vianney Perchet has benefited from the support of the FMJH Program Gaspard Monge in optimization and operations research (supported in part by EDF), from the Labex LMH and from the CNRS through the PEPS program.

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.

Baldin, N. and Berthet, Q. (2018). Optimal link prediction with matrix logistic regression. *Preprint*.

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

Berthet, Q. (2014). Optimal testing for planted satisfiability problems. *Electron. J. Stat.*

Berthet, Q. and Chandraskeran, V. (2016). Resource allocation for statistical estimation. *Proceedings of the IEEE*.

Berthet, Q. and Ellenberg, J. (2019). Detection of planted solutions for flat satisfiability problems. *AIStats 2019*.

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.

Berthet, Q. and Rigollet, P. (2013). Complexity theoretic lower bounds for sparse principal component detection. *J. Mach. Learn. Res. (COLT)*, 30.

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.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 1015–1022, USA. Omnipress.

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.

Wang, T., Berthet, Q., and Samworth, R. J. (2016a). Statistical and computational trade-offs in estimation of sparse principal components. *Ann. Statist.*

Wang, T., Berthet, Q., and Y.Plan (2016b). Average-case hardness of rip certification. *NIPS*.

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.