Open Problem: Tight Bounds for Kernelized Multi-Armed Bandits with Bernoulli Rewards

Marco Mussi MARCO.MUSSI@POLIMI.IT
Simone Drago SIMONE.DRAGO@POLIMI.IT
Alberto Maria Metelli ALBERTOMARIA.METELLI@POLIMI.IT
Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan, Italy

Abstract

We consider Kernelized Bandits (KBs) to optimize a function $f : \mathcal{X} \to [0, 1]$ belonging to the Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$. Mainstream works on kernelized bandits focus on a subgaussian noise model in which observations of the form $f(\mathbf{x}_t) + \epsilon_t$, being ϵ_t a subgaussian noise, are available (Chowdhury and Gopalan, 2017). Differently, we focus on the case in which we observe realizations $y_t \sim \text{Ber}(f(\mathbf{x}_t))$ sampled from a Bernoulli distribution with parameter $f(\mathbf{x}_t)$. While the Bernoulli model has been investigated successfully in multi-armed bandits (Garivier and Cappé, 2011), logistic bandits (Faury et al., 2022), bandits in metric spaces (Magureanu et al., 2014), it remains an open question whether tight results can be obtained for KBs. This paper aims to draw the attention of the online learning community to this open problem.

Keywords: Concentration, Regret, Bernoulli Rewards, Kernelized Bandits, MAB, RKHS

1 Introduction

In this work, we present three open problems related to Kernelized Bandits (KBs, Chowdhury and Gopalan 2017) for optimizing a function $f : \mathcal{X} \to [0, 1]$ belonging in the Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$. We assume to observe samples $y_t \sim \text{Ber}(f(\mathbf{x}_t))$ from a Bernoulli distribution with parameter $f(\mathbf{x}_t)$. In the following, we revise the literature about Bernoulli observations coupled with different bandit structures and the subgaussian noise model for KBs.

Bernoulli Samples. Garivier and Cappé (2011) developed the first optimal algorithm (`KL-UCB`) for regret minimization in Multi-Armed Bandits (MABs) with Bernoulli rewards (and no structure on the arms). `KL-UCB` leverages *optimism* and a concentration bound based on the Kullback-Leibler divergence (KL, Kullback and Leibler, 1951) succeeding in asymptotically matching the lower bound (Lai and Robbins, 1985). Several works extended MABs with Bernoulli rewards to account for structure, including metric spaces (Magureanu et al., 2014) and linear (logistic) models (Faury et al., 2022).

Kernelized Bandits. The seminal work (Srinivas et al., 2010) introduce `GP-UCB`, the first no-regret solution based on Gaussian Processes (GPs, Rasmussen and Williams, 2006). `GP-UCB` enjoys regret guarantees both in the cases when f is indeed sampled from a GP and when f belongs to a suitable RKHS (*agnostic* case). (Chowdhury and Gopalan, 2017) derive an improved version of `GP-UCB`, called `IGP-UCB`, working under subgaussian noise model. The analysis is based on a novel *self-normalized concentration inequality* for subgaussian samples $f(\mathbf{x}_t) + \epsilon_t$ that extends and generalizes that of (Abbasi-Yadkori et al., 2011) for linear models. These solutions can be adapted to learn also in the presence of Bernoulli rewards[1] at the price of (possibly) looser guarantees.

1. A Bernoulli random variable is λ-subgaussian, with $\lambda = 1/2$.

	Subgaussian	*Bernoulli*
No Structure	Lattimore and Szepesvári 2020 (Corollary 5.5)	Garivier and Cappé 2011 (Theorem 10)
Linear	Abbasi-Yadkori et al. 2011 (Theorem 2)	Faury et al. 2022 (Proposition 3)
Metric Space	Kleinberg et al. 2008 (Theorem 4.2)	Magureanu et al. 2014 (Theorem 2)
RKHS	Chowdhury and Gopalan 2017 (Theorem 2)	**Open Problem**

Table 1: Summary of the state-of-the-art in concentration bounds.

These results are possible thanks to specifically designed concentration bounds, which are (almost) optimal for their specific settings. As we can notice from Table 1, the only missing solution is the one to learn with kernelized structure in the presence of Bernoulli rewards. The goal of this work is to raise the attention of the online learning community on this gap in the literature.

2 Problem Formulation

In this section, we describe the setting, the learning problem, and the considered assumptions.

Setting. We consider the problem of sequentially maximizing a fixed and unknown function $f : \mathcal{X} \to [0,1]$ over a decision set $\mathcal{X} \subseteq \mathbb{R}^d$ (also called action set). At every round $t \in [\![T]\!] := \{1, \ldots, T\}$, being $T \in \mathbb{N}$ the learning horizon, the algorithm $\mathfrak{A}$ chooses an action $\mathbf{x}_t \in \mathcal{X}$ based on the history of past observations $\mathcal{G}_t := \{(\mathbf{x}_s, y_s)\}_{s \in [\![t-1]\!]}$ and observes a random variable $y_t \sim \text{Ber}(f(\mathbf{x}_t))$, where $\text{Ber}(p)$ denotes a Bernoulli distribution with parameter $p \in [0,1]$.

Learning Problem. The goal of the learning algorithm $\mathfrak{A}$ is to minimize the *regret*:

$$R_T(\mathfrak{A}) := T\, f(\mathbf{x}^\star) - \sum_{t \in [\![T]\!]} f(\mathbf{x}_t) \qquad \text{where} \qquad \mathbf{x}^\star \in \arg\max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}). \tag{1}$$

Regularity Conditions. We consider the frequentist-type regularity assumption that is usually employed in KBs (Srinivas et al., 2010; Chowdhury and Gopalan, 2017). Let $\mathcal{H}_k$ be a RKHS induced by the kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ so that every function $h \in \mathcal{H}_k$ satisfies the *reproducing property* $h(x) = \langle h, k(\cdot, x) \rangle_{\mathcal{H}_k}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ is the inner product defined on the space $\mathcal{H}_k$. We denote with $\|h\|_{\mathcal{H}_k} = \sqrt{\langle h, h \rangle_{\mathcal{H}_k}}$ the RKHS norm. We enforce the following standard assumption prescribing that f belongs to the RKHS with bounded norm.

Assumption 2.1 (Regularity Conditions) *f belongs to the RKHS, i.e., $f \in \mathcal{H}_k$, and:*
- *the function f has a bounded RKHS norm, i.e., $\|f\|_{\mathcal{H}_k} \leqslant B < +\infty$ (B is known);*
- *the kernel function k is bounded, i.e., $k(\mathbf{x}, \mathbf{x}) \leqslant 1$ for every $\mathbf{x} \in \mathcal{X}$.*

3 Open Problems

3.1 Open Problem 1: Estimation

Can we effectively estimate $f(\mathbf{x})$ in a new point $\mathbf{x} \in \mathcal{X}$ based on the history of past observations $\mathcal{G}_t := \{(\mathbf{x}_s, y_s)\}_{s \in [\![t-1]\!]}$ where y_s are Bernoulli samples?

When the observations are of the form $y_t = f(\mathbf{x}_t) + \epsilon_t$ with ϵ_t being a λ-subgaussian noise, the standard approach consists in resorting to GPs. We consider a prior GP model $\mathcal{GP}(0, k(\cdot, \cdot))$ for function f and a Gaussian likelihood model $\mathcal{N}(0,\nu^2)$ for the noise ϵ_t.[2] Given the history $\mathcal{G}_t := \{(\mathbf{x}_s, y_s)\}_{s\in[\![t-1]\!]}$, the posterior distribution of f is $\mathcal{GP}(\mu_t(\cdot), k_t(\cdot,\cdot))$, where, for every $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$:

$$\mu_t(\mathbf{x}) := \mathbf{k}_t(\mathbf{x})^\top(\mathbf{K}_t + \nu^2\mathbf{I})^{-1}\mathbf{y}_t, \qquad k_t(\mathbf{x},\mathbf{x}') := k(\mathbf{x},\mathbf{x}) - \mathbf{k}_t(\mathbf{x})^\top(\mathbf{K}_t + \nu^2\mathbf{I})^{-1}\mathbf{k}_t(\mathbf{x}'),$$

where $\mathbf{k}_t(\mathbf{x}) := (k(\mathbf{x}_1,\mathbf{x}),\dots,k(\mathbf{x}_t,\mathbf{x}))^\top$, $\mathbf{K}_t := (k(\mathbf{x}_i,\mathbf{x}_j))_{i,j\in[\![t]\!]}$, and $\mathbf{y}_t := (y_1,\dots,y_t)^\top$. This allows to have the estimate $\mu_t(\mathbf{x})$ and an index of uncertainty $\sigma_t^2(\mathbf{x}) := k_t(\mathbf{x},\mathbf{x})$ of the estimate.

This approach can be employed when $y_t \sim \mathrm{Ber}(f(\mathbf{x}_t))$, since a Bernoulli variable is $1/2$-subgaussian. However, the drawback is that $\mu_t(\mathbf{x})$ is not guaranteed to lie in $[0,1]$, although the true $f(\mathbf{x}) \in [0,1]$, being the parameter of a Bernoulli distribution (Figure 1).

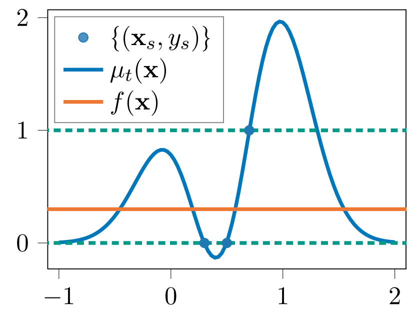

Figure 1: GP estimate example of $f \in [0,1]$.

A first attempt to fix this consists of keeping the prior $\mathcal{GP}(0, k(\cdot,\cdot))$ for the unknown function f and *change the likelihood model* to a Bernoulli one. However, this attempt is unsuccessful since the posterior computation would require evaluating the conditional distribution $\Pr(y_t|f(\mathbf{x}_t))$ that is not well defined when $f \sim \mathcal{GP}(0,k(\cdot,\cdot))$ since $f(\mathbf{x}_t)$ may take values outside $[0,1]$.

A second attempt consists in *changing both the prior and the likelihood model* to overcome the "incompatibility" between the GP and the Bernoulli likelihood model. Aiming for a conjugate prior update, we should use a *Beta prior and a Bernoulli likelihood* model. However, as noted in (Rolland et al., 2019), enforcing correlation with Beta distributions is challenging differently from the Gaussian case. The notion of "Beta process" was introduced in survival analysis but displays a too-limited modeling power (Hjort, 1990; Paisley and Carin, 2009). Furthermore, there is no consensus on one definition of *multivariate Beta* distribution. A common approach (Westphal, 2019) bases on a Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\zeta})$ defined over the support $\{0,1\}^{2^t}$ (with parameter $\boldsymbol{\zeta} = (\zeta_1,\dots,\zeta_{2^t})^\top \in \mathbb{R}^{2^t}_{\geqslant 0}$) from which to sample a probability vector $\mathbf{p}_t = (p_1,\dots,p_{2^t})^\top \sim \mathrm{Dir}(\boldsymbol{\zeta})$ needed to define the multivariate Beta variable as $\boldsymbol{\theta} = \mathbf{H}_t\mathbf{p}_t$ where $\mathbf{H}_t = (\mathrm{bin}(0)|\dots|\mathrm{bin}(2^t))$ and $\mathrm{bin}(n)$ is the binary encoding of number n. Although this allows for a simple posterior calculation, it is completely unstructured and does not allow embedding the structure enforced by the kernel k.

These attempts focus on deriving a "proper" Bayesian update. Since even for the subgaussian KBs, GPs are just an estimation tool, we may consider *non-Bayesian* updates. (Goetschalckx et al., 2011) proposes a sample-sharing method in which samples contribute weighted by the kernel k:

$$\alpha_t(\mathbf{x}) := \alpha_0 + \sum_{s\in[\![t]\!]} y_s k(\mathbf{x},\mathbf{x}_s), \qquad \beta_t(\mathbf{x}) := \beta_0 + \sum_{s\in[\![t]\!]} (1-y_s) k(\mathbf{x},\mathbf{x}_s).$$

This approach has convergence guarantees when f is Lipschitz continuous. Other approaches leverage on Logistic Gaussian Processes (Leonard, 1978) or Gaussian Process Copulas (Wilson and Ghahramani, 2010), and they all involve non-Bayesian updates due to the analytical intractability.

3.2 Open Problem 2: Concentration

Can we derive concentration guarantees for the deviation $|f(\mathbf{x}) - \mu_t(\mathbf{x})|$ (being $\mu_t(\mathbf{x})$ a suitable estimator of $f(\mathbf{x})$) which is tight for the Bernoulli observations?

2. The GP model is used for estimation and the true f may not be sampled from the GP (Chowdhury and Gopalan, 2017).

For the λ-subgaussian case, by resorting to the GP-based estimator presented in Section 3.1 (Srinivas et al., 2010), it is possible, under Assumption 2.1, to achieve the following concentration bound for the deviation that holds w.p. $1-\delta$ simultaneously for every $t \geqslant 1$ and $\mathbf{x} \in \mathcal{X}$:

$$|\mu_{t-1}(\mathbf{x}) - f(\mathbf{x})| \leqslant \left(B + \lambda\sqrt{2\left(\gamma_{t-1} + 1 + \log(1/\delta)\right)}\right)\sigma_{t-1}(\mathbf{x}), \tag{2}$$

where $\gamma_t = \max_{\mathcal{A}\subset\mathcal{X}:|\mathcal{A}|=t} I(\mathbf{y}_{\mathcal{A}};\mathbf{f}_{\mathcal{A}})$ is the *maximum information gain* at time t. This result is obtained by deriving a *self-normalized concentration inequality* that bounds a suitable weighted norm of the noise process $(\epsilon_1, \dots, \epsilon_{t-1})^\top$ (Chowdhury and Gopalan, 2017, Theorem 1). Clearly, Equation (2) holds for Bernoulli distributions too, being them subgaussian with $\lambda = 1/2$.

While Equation (2) is likely tight for subgaussian observations, it fails to capture the stronger concentration rate that characterizes Bernoulli random variables. Indeed, in the unstructured case (i.e., MABs with no correlation between the arms), Garivier and Cappé (2011) obtain the stronger concentration bound, holding w.p. $1-\delta$ for a fixed $\mathbf{x} \in \mathcal{X}$ and simultaneously for every $t \geqslant 0$:

$$\begin{array}{ll} \text{Let:} & u_t(\mathbf{x}) := \max\{q \geqslant \mu_t(\mathbf{x}) : N_t(\mathbf{x}) d(\mu_t(\mathbf{x}), q) \leqslant c_1 \log(t/\delta) + c_2 \log\log(t/\delta)\}, \\ \text{then:} & f(\mathbf{x}) \leqslant u_t(\mathbf{x}), \end{array} \tag{3}$$

where $d(a,b) = a\log(a/b) + (1-a)\log((1-a)/(1-b))$ for $a, b \in [0,1]$ is the Bernoulli KL-divergence, $\mu_t(\mathbf{x}) = \sum_{s\in[\![t]\!]} y_s \mathbf{1}\{\mathbf{x}_s = \mathbf{x}\}/N_t(\mathbf{x})$, $N_t(\mathbf{x}) = \sum_{s\in[\![t]\!]} \mathbf{1}\{\mathbf{x}_s = \mathbf{x}\}$, and $c_1, c_2 > 0$ are universal constants. Equation (3) evaluates the distance using the KL-divergence $d(\cdot,\cdot)$ between Bernoulli parameters and, therefore, delivers a stronger concentration bound compared to that of Equation (2). The derivation of this result (Garivier and Cappé, 2011) makes use of a *martingale*-based argument deeply depending on the moment-generating function of the Bernoulli distribution that seems not to be easily extensible to the KB setting in which correlation among arms is present.

3.3 Open Problem 3: Regret Minimization

Can we design regret minimization algorithms which achieve a $\log T$ regret guarantee, highlighting the dependence on $d(f(\mathbf{x}), f(\mathbf{x}^\star))$ when $\mathcal{X}$ is finite?

Under Assumption 2.1, by applying the improved `GP-UCB` presented in `IGP-UCB` (Chowdhury and Gopalan, 2017, Theorem 2), we obtain a *worst-case* $\widetilde{\mathcal{O}}(\sqrt{T})$ regret guarantee w.p. $1-\delta$:

$$R_T(\texttt{IGP-UCB}) \leqslant \mathcal{O}\left(B\sqrt{\gamma_T} + \sqrt{T\gamma_T\left(\gamma_T + \log(1/\delta)\right)}\right).$$

The study of *instance-dependent* regret bounds for KBs is introduced in (Shekhar and Javidi, 2022), focusing on the *packing* properties of the RKHS and still achieving regret bounds of order $\widetilde{\mathcal{O}}(T^\alpha)$ for some $\alpha \in (0,1)$. Here, instead, when $\mathcal{X}$ is finite, we are interested in understanding if achieving $\log T$ regret is possible for KBs with Bernoulli observations. Indeed, in the unstructured case (and $|\mathcal{X}| < +\infty$), the `KL-UCB` (Garivier and Cappé, 2011) achieves the tight *instance-dependent* bound:

$$R_T(\texttt{KL-UCB}) \leqslant \mathcal{O}\left(\sum_{\mathbf{x}\in\mathcal{X}} \frac{\log T}{d(f(\mathbf{x}), f(\mathbf{x}^\star))}\right).$$

We perceive that this should be possible since, when $|\mathcal{X}| < +\infty$, using the trivial kernel $k(\mathbf{x}, \mathbf{x}') = \mathbf{1}\{\mathbf{x} = \mathbf{x}'\}$ for every $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, the KB reduces to an unstructured MAB. Furthermore, we conjecture that this possibility (at least for optimistic algorithms) is strictly related to the open problem of Section 3.2. Indeed, the bound of Equation (3) is specifically designed for the `KL-UCB` algorithm (Garivier and Cappé, 2011).

References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2312–2320, 2011.

S. R. Chowdhury and A. Gopalan. On kernelized multi-armed bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 844–853, 2017.

L. Faury, M. Abeille, K. Jun, and C. Calauzènes. Jointly efficient and optimal algorithms for logistic bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*, pages 546–580. PMLR, 2022.

A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, volume 19 of *JMLR Proceedings*, pages 359–376, 2011.

R. Goetschalckx, P. Poupart, and J. Hoey. Continuous correlated beta processes. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1269–1274, 2011.

N. L. Hjort. Nonparametric bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, pages 1259–1294, 1990.

R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. *CoRR*, abs/0809.4882, 2008.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

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

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

T. Leonard. Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(2):113–132, 1978.

S. Magureanu, R. Combes, and A. Proutière. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 975–999, 2014.

J. W. Paisley and L. Carin. Nonparametric factor analysis with beta process priors. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 382 of *ACM International Conference Proceeding Series*, pages 777–784. ACM, 2009.

C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. Springer, 2006.

P. Rolland, A. Kavis, A. Immer, A. Singla, and V. Cevher. Efficient learning of smooth probability functions from bernoulli tests with guarantees. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 5459–5467. PMLR, 2019.

S. Shekhar and T. Javidi. Instance dependent regret analysis of kernelized bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 19747–19772. PMLR, 2022.

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

M. Westphal. Simultaneous inference for multiple proportions: a multivariate beta-binomial model. *arXiv preprint arXiv:1911.00098*, 2019.

A. G. Wilson and Z. Ghahramani. Copula processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2460–2468. Curran Associates, Inc., 2010.