Open Problem: Tight Online Confidence Intervals for RKHS Elements

Sattar Vakili SATTAR.VAKILI@MTKRESEARCH.COM
MediaTek Research, UK

Jonathan Scarlett SCARLETT@COMP.NUS.EDU.SG
National University of Singapore

Tara Javidi TJAVIDI@ENG.UCSD.EDU
University of California, San Diego

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

Confidence intervals are a crucial building block in the analysis of various online learning problems. The analysis of kernel-based bandit and reinforcement learning problems utilize confidence intervals applicable to the elements of a reproducing kernel Hilbert space (RKHS). However, the existing confidence bounds do not appear to be tight, resulting in suboptimal regret bounds. In fact, the existing regret bounds for several kernelized bandit algorithms (e.g., GP-UCB, GP-TS, and their variants) may fail to even be sublinear. It is unclear whether the suboptimal regret bound is a fundamental shortcoming of these algorithms or an artifact of the proof, and the main challenge seems to stem from the online (sequential) nature of the observation points. We formalize the question of online confidence intervals in the RKHS setting and overview the existing results.

Keywords: RKHS, Gaussian Processes, Confidence Intervals, Bayesian Optimization, Bandits, Reinforcement Learning.

1. Introduction

The kernel trick provides an elegant and natural technique to extend linear models to non-linear models with a great representation power. In the past decade, numerous works have studied bandit and reinforcement learning problems under the assumption that the reward function conforms to a kernel-based model (Srinivas et al., 2010; Krause and Ong, 2011; Wang and de Freitas, 2014; Nguyen et al., 2017; Scarlett et al., 2017; Chowdhury and Gopalan, 2017; Wang et al., 2018; Kandasamy et al., 2018; Javidi and Shekhar, 2018; Yang et al., 2020; Shekhar and Javidi, 2020; Bogunovic et al., 2020; Zhou et al., 2020; Vakili et al., 2020, 2021; Cai and Scarlett, 2021; Zhang et al., 2021).

The analysis of online learning problems with a kernel-based model typically utilizes confidence intervals applicable to the elements of a reproducing kernel Hilbert space (RKHS). However, the state-of-the-art confidence intervals in this setting (Chowdhury and Gopalan, 2017) do not appear to be tight, resulting in suboptimal regret bounds. The main challenge seems to stem from the online (sequential) nature of the observation points, in contrast to an offline (fixed in advance) design. We first overview the existing results, and then formalize the open problem of tight confidence intervals for the RKHS elements under the online setting. We also discuss the consequences of these bounds

on the regret performance. For clarity of exposition, we focus on bandit problems and the GP-UCB algorithm (Srinivas et al., 2010; Chowdhury and Gopalan, 2017), but the problem is equally relevant to reinforcement learning problems and other algorithms such as GP-TS.

2. Problem Setup

Consider a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with respect to a finite Borel measure, where $\mathcal{X} \subset \mathbb{R}^d$ is a compact set. Let $\mathcal{H}_k$ denote the RKHS corresponding to k, defined as a Hilbert space equipped with an inner product $\langle ., . \rangle_{\mathcal{H}_k}$ satisfying the following: $k(., x) \in \mathcal{H}_k, \forall x \in \mathcal{X}$, and $\langle f, k(., x) \rangle_{\mathcal{H}_k} = f(x), \forall x \in \mathcal{X}, \forall f \in \mathcal{H}_k$ (reproducing property). The typical assumption in kernel-based models is that the *objective function* f satisfies $f \in \mathcal{H}_k$ for a known kernel k. Let $\{\lambda_m\}_{m=1}^{\infty}$ and $\{\phi_m\}_{m=1}^{\infty}$ denote the Mercer eigenvalues and eigenfeatures of k, respectively (see, e.g., Kanagawa et al., 2018, Theorem 4.1). Using Mercer's representation theorem (see, e.g., Kanagawa et al., 2018, Theorem 4.2), an alternative representation for $f \in \mathcal{H}_k$ is given by

$$f(x) = \mathbf{w}^\top \mathbf{\Lambda}^{\frac{1}{2}} \boldsymbol{\phi}(x), \tag{1}$$

where $\mathbf{w} = [w_1, w_2, ...]^\top$ and $\boldsymbol{\phi}(x) = [\phi_2(x), \phi_2(x), ...]^\top$ are the (possibly infinite-dimensional) *weight* and feature vectors, and $\mathbf{\Lambda}$ is a (possibly infinite dimensional) diagonal matrix with $\mathbf{\Lambda}_{i,j} = \lambda_i$, if $i = j$. The RKHS norm of f satisfies $\|f\|_{\mathcal{H}_k} = \|\mathbf{w}\|_{\ell^2}$.

Kernelized Bandits: Consider an online learning setting where a learning algorithm is allowed to collect a sequence of noisy observations $\{(x_i, y_i)\}_{i=1}^{\infty}$, where $y_i = f(x_i) + \epsilon_i$ with ϵ_i being well-behaved noise terms. The objective is to get as close as possible to the maximum of f. The performance of the algorithm is measured in terms of regret, defined as the cumulative loss in the values of the objective function at observation points, compared to a global maximum:

$$\mathcal{R}(N) = \sum_{i=1}^{N} \left(f(x^*) - f(x_i) \right), \tag{2}$$

where $x^* \in \text{argmax}_{x \in \mathcal{X}} f(x)$ is a global maximum. Under the assumption $f \in \mathcal{H}_k$, this setting is often referred to as that of kernelized bandits, Gaussian process (GP) bandits, or Bayesian optimization. The latter two terms are motivated by the algorithm design which often employs a GP surrogate model. Throughout this paper, we make the following assumptions.

Assumption 1 *The RKHS norm of f is bounded as $\|f\|_{\mathcal{H}_k} \leq B$, for some $B > 0$. Moreover, the noise terms are i.i.d. sub-Gaussian random variables, i.e., for some $R > 0$, $\mathbb{E}[\exp(\eta \epsilon_i)] \leq \exp(\frac{\eta^2 R^2}{2})$, $\forall \eta \in \mathbb{R}, \forall i \in \mathbb{N}$.*

In online learning problems, the observation points are collected sequentially. In particular, the observation point x_{i+1} is determined after all the values $\{(x_j, y_j)\}_{j=1}^{i}$ are revealed. This is in contrast to an offline design, where the data points are fixed in advance. We next formalize this distinction.

Definition 1 *i) In the **online setting**, for the sigma algebras $\mathcal{F}_i = \sigma(x_1, x_2, \ldots, x_{i+1}, \epsilon_1, \epsilon_2, \ldots, \epsilon_i)$, $i \geq 1$, it holds that x_i and ϵ_i are $\mathcal{F}_{i-1}$ and $\mathcal{F}_i$ measurable, respectively. ii) In the **offline setting**, for all $i \geq 1$, it holds that x_i is independent of all ϵ_j, $j \geq 1$.*

Surrogate GP Model: It is useful for algorithm design to employ a zero-mean surrogate GP model $\hat{f}$ with kernel k which provides a surrogate posterior mean (regressor) and a surrogate posterior variance (uncertainty estimate) for the kernel-based model. Defining $\mu_n(x) = \mathbb{E}\big[\hat{f}(x)|\{(x_i, y_i)\}_{i=1}^n\big]$ and $\sigma_n^2(x) = \mathbb{E}\big[(\hat{f}(x) - \mu_n(x))^2|\{(x_i, y_i)\}_{i=1}^n\big]$, it is well known that $\mu_n(x) = \mathbf{z}_n^\top(x)\mathbf{y}_n$ and $\sigma_n^2(x) = k(x,x) - \mathbf{k}_n^\top(x)(\lambda^2\mathbf{I}_n + \mathbf{K}_n)^{-1}\mathbf{k}_n(x)$, where $\mathbf{k}_n(x) = [k(x,x_1), k(x,x_2), \ldots, k(x,x_n)]^\top$, $\mathbf{K}_n$ is the positive definite kernel matrix $[\mathbf{K}_n]_{i,j} = k(x_i, x_j)$, $\mathbf{z}_n(x) = (\lambda^2\mathbf{I}_n + \mathbf{K}_n)^{-1}\mathbf{k}_n(x)$, $\mathbf{I}_n$ is the identity matrix of dimension n, and $\lambda > 0$ is a regularization parameter.

3. Confidence Intervals Applicable to RKHS Elements

Deriving confidence intervals applicable to RKHS elements is significantly more challenging in the online setting compared to the offline setting. In the latter case, for any fixed $x \in \mathcal{X}$, we have with probability at least $1 - \delta$ that $|f(x) - \mu_n(x)| \leq \rho_0(\delta)\sigma_n(x)$, where $\rho_0(\delta) = B + \frac{R}{\lambda}\sqrt{2\log(\frac{2}{\delta})}$, B and R are the parameters specified in Assumption 1, and λ is the regularization parameter of the surrogate GP model. Moreover, when f is Lipschitz (or Hölder) continuous (that is true with typical kernels; see, Shekhar and Javidi, 2020), this easily extends to a uniform guarantee: With probability at least $1-\delta$, we have uniformly in x that $|f(x) - \mu_n(x)| = \mathcal{O}\big(\big(B + \frac{R}{\lambda}\sqrt{d\log(n) + \log(\frac{1}{\delta})}\big)\sigma_n(x)\big)$, where the implied constants in $\mathcal{O}(.)$ depend on the Lipschitz (or Hölder) continuity parameters.

In the online setting, strong uniform bounds are also well-known in the case of a linear model $f(x) = \mathbf{w}^\top x$: Abbasi-Yadkori et al. (2011) proved that, with probability $1 - \delta$, uniformly over x,

$$|f(x) - \mu_n(x)| \leq \rho_n(\delta)\sigma_n(x), \tag{3}$$

where $\rho_n(\delta) = B + \frac{R}{\lambda}\sqrt{d\log(\frac{1+n\bar{x}^2/\lambda^2}{\delta})}$ and $\bar{x} = \max_{x\in\mathcal{X}} \|x\|_{\ell^2}$. The crux of the proof is a *self-normalized bound for vector valued martingales* $S_n = \sum_{i=1}^n \epsilon_i x_i$ (Abbasi-Yadkori et al., 2011, Theorem 1), which yields the following *confidence ellipsoid* for $\mathbf{w}$ (Abbasi-Yadkori et al., 2011, Theorem 2): $\|\mathbf{w} - \hat{\mathbf{w}}_n\|_{V_n} \leq \lambda\rho_n(\delta)$, with probability at least $1-\delta$, where $V_n = \lambda^2\mathbf{I}_d + \sum_{i=1}^n x_i x_i^\top$. This confidence ellipsoid for $\mathbf{w}$ can then be represented in terms of the confidence interval for $f(x)$ given in (3). Notice that the linear model is a special case of (1) with $\mathbf{w} = [w_1, w_2, \ldots, w_d]^\top$ and $\boldsymbol{\phi}(x) = x$ being d dimensional weight and feature vectors respectively, and $\Lambda = \mathbf{I}_d$ being the square identity matrix of dimension d.

Chowdhury and Gopalan (2017) built on the self-normalized bound for the vector valued martingales to prove the following theorem for the kernel-based models.

Theorem 2 *Under Assumption 1, in the online setting, with probability at least $1 - \delta$, we have for all $x \in \mathcal{X}$ that*

$$|f(x) - \mu_n(x)| \leq \rho_n(\delta)\sigma_n(x), \tag{4}$$

where $\rho_n(\delta) = B + R\sqrt{2\left(\gamma_{n-1} + 1 + \log(\frac{1}{\delta})\right)}$, *and* $\gamma_n = \sup_{\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}} \log\det(\lambda^2 \mathbf{I}_n + \mathbf{K}_n)$ *is the maximal information gain at time* n*, which is closely related to the effective dimension associated with the kernel (e.g., see Srinivas et al. (2010); Valko et al. (2013)).*

Our open problem is concerned with improving this confidence interval.

Open Problem. Under Assumption 1, in the online setting, consider the general problem of proving a confidence interval of the following form uniformly in $x \in \mathcal{X}$:

$$|f(x) - \mu_n(x)| \leq \rho_n(\delta)\sigma_n(x), \text{ with probability at least } 1 - \delta. \tag{5}$$

What is the lowest growth rate of $\rho_n(\delta)$ with n? In particular, is it possible to reduce the confidence interval width in Theorem 2 by an $\tilde{\mathcal{O}}(\sqrt{\gamma_n})$ factor?

4. Discussion

Following standard UCB-based bandit algorithm techniques, it can be shown that the GP-UCB algorithm (namely, repeatedly choosing x to maximize the current upper confidence bound) attains

$$\mathcal{R}(N) = \tilde{\mathcal{O}}(\rho_N(\delta)\sqrt{N\gamma_N}), \text{ with probability at least } 1 - \delta. \tag{6}$$

Substituting $\rho_N(\delta)$ from Theorem 2, we have $\mathcal{R}(N) = \tilde{\mathcal{O}}(\gamma_N\sqrt{N})$. Unfortunately, this is not always sublinear in N, since γ_N can grow faster than $\sqrt{N}$, e.g., in the case of the Matérn family of kernels. Hence, the regret bound can be trivial in many cases of interest. It is unknown whether this suboptimal regret bound is a fundamental shortcoming of GP-UCB or a result of suboptimal confidence intervals, but the latter appears likely to be the most significant factor. The same question can be asked about the analysis of many other bandit algorithms including GP-TS (Chowdhury and Gopalan, 2017) and GP-EI (Nguyen et al., 2017), as well as KOVI in the reinforcement learning setting (Yang et al., 2020).

Comparing the results under the online and offline settings, we see a stark contrast of an $\mathcal{O}(\sqrt{\gamma_n})$ factor in the width of confidence intervals. We expect that the $\mathcal{O}(\sqrt{\gamma_n})$ factor in the confidence interval width in the online setting can be replaced by an $\tilde{\mathcal{O}}(d\log(n))$ term, resulting in an $\tilde{\mathcal{O}}(\sqrt{dN\gamma_N})$ regret bound. Roughly speaking, we are suggesting that a square root of the effective dimension of the kernel in the regret bound can be traded off for a square root of the input dimension.

Of significant theoretical importance is a less practical algorithm *SupKernelUCB* (Valko et al., 2013), which achieves an $\tilde{\mathcal{O}}(\sqrt{N\gamma_N})$ regret bound for the kernelized bandit problem with a finite action set ($|\mathcal{X}| < \infty$). The finite action set assumption can be relaxed to compact domains using a discretization argument contributing only an $\tilde{\mathcal{O}}(\sqrt{d\log(N)})$ factor to the regret bound (see, Cai and Scarlett, 2021, Appendix A.4). This bound is tight for the cases where a lower bound on regret is known, namely for commonly used squared exponential and Matérn kernels (Scarlett et al., 2017; Vakili et al., 2021). In view of this discussion, the above-mentioned improvement is information-theoretically feasible, but it remains to determine whether GP-UCB can achieve it.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Ilija Bogunovic, Andreas Krause, and Jonathan Scarlett. Corruption-tolerant Gaussian process bandit optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1071–1081, 2020.

Xu Cai and Jonathan Scarlett. On lower bounds for standard and robust Gaussian process bandit optimization. In *Proceedings of International Conference on Machine Learning*, 2021.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *Proceedings of International Conference on Machine Learning*, pages 844–853, 2017.

Tara Javidi and Shekhar Shekhar. Gaussian process bandits with adaptive discretization. *Electronic Journal of Statistics*, 12(2):3829–3874, 2018.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv:1805.08845v1 [stat.ML]*, 2018.

Kirthevasan Kandasamy, Akshay Krishnamurthy, Jeff Schneider, and Barnabás Póczos. Parallelised Bayesian optimisation via Thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 133–142, 2018.

Andreas Krause and Cheng S. Ong. Contextual Gaussian process bandit optimization. In *Advances in Neural Information Processing Systems 24*, pages 2447–2455, 2011.

Vu Nguyen, Sunil Gupta, Santu Rana, Cheng Li, and Svetha Venkatesh. Regret for expected improvement over the best-observed value and stopping condition. In *Asian Conference on Machine Learning*, pages 279–294, 2017.

Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. Lower bounds on regret for noisy Gaussian process bandit optimization. In *Conference on Learning Theory*, pages 1723–1742, 2017.

Shubhanshu Shekhar and Tara Javidi. Multi-scale zero-order optimization of smooth functions in an RKHS. *arXiv:2005.04832*, 2020.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *International Conference on Machine Learning*, pages 1015–1022, 2010.

Sattar Vakili, Henry Moss, Artem Artemev, Vincent Dutordoir, and Victor Picheny. Scalable Thompson sampling using sparse Gaussian process models. *arXiv:2006.05356v3 [stat.ML]*, 2020.

Sattar Vakili, Kia Khezeli, and Victor Picheny. On information gain and regret bounds in Gaussian process bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 82–90, 2021.

Michal Valko, Nathan Korda, Rémi Munos, Ilias Flaounas, and Nello Cristianini. Finite-time analysis of kernelised contextual bandits. In *Conference on Uncertainty in Artificial Intelligence*, pages 654–663, 2013.

Zi Wang, Beomjoon Kim, and Leslie Pack Kaelbling. Regret bounds for meta Bayesian optimization with an unknown Gaussian process prior. In *Advances in Neural Information Processing Systems*, pages 10477–10488, 2018.

Ziyu Wang and Nando de Freitas. Theoretical analysis of Bayesian optimisation with unknown Gaussian process hyper-parameters. *arXiv:1406.7758*, 2014.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael Jordan. Provably efficient reinforcement learning with kernel and neural function approximations. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Weitong Zhang, Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural Thompson sampling. In *International Conference on Learning Representations*, 2021.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with UCB-based exploration. In *International Conference on Machine Learning*, pages 11492–11502, 2020.