Open Problem: Model Selection for Contextual Bandits

Dylan J. Foster DYLANF@MIT.EDU
Massachusetts Institute of Technology

Akshay Krishnamurthy AKSHAY@CS.UMASS.EDU
Microsoft Research NYC

Haipeng Luo HAIPENGL@USC.EDU
University of Southern California

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

In statistical learning, algorithms for model selection allow the learner to adapt to the complexity of the best hypothesis class in a sequence. We ask whether similar guarantees are possible for contextual bandit learning.

1. Introduction

Model selection is the fundamental statistical task of choosing a hypothesis class using data, with statistical guarantees dating back to Vapnik's structural risk minimization principle. Despite decades of research on model selection for supervised learning and the ubiquity of model selection procedures such as cross-validation in practice, very little is known about model selection in interactive learning and reinforcement learning settings where exploration is required. Focusing on contextual bandits, a simple reinforcement learning setting, we ask: *Can model selection guarantees be achieved in contextual bandit learning, where a learner must balance exploration and exploitation to make decisions online?*

2. Problem Formulation

We consider the adversarial contextual bandit setting (Auer et al., 2002). The setting is defined by a context space $\mathcal{X}$ and a finite action space $\mathcal{A} := \{1, \dots, K\}$. The learner interacts with nature for T rounds, where in round t: (1) nature selects a context $x_t \in \mathcal{X}$ and loss $\ell_t \in [0,1]^{\mathcal{A}}$, (2) the learner observes x_t and chooses action a_t, and (3) the learner observes $\ell_t(a_t)$. We allow for an adaptive adversary, so that x_t and ℓ_t may depend on $a_1, \dots, a_{t-1}$. In the usual problem formulation, the learner is given a policy class $\Pi \subset (\mathcal{X} \to \mathcal{A})$, and the goal is to minimize regret to Π:

$$\mathrm{Reg}(\Pi) := \max_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(\pi(x_t))\right].$$

When Π is finite, the well-known Exp4 algorithm (Auer et al., 2002) achieves the optimal regret bound of $O(\sqrt{KT\log|\Pi|})$.

The Model Selection Problem. In the contextual bandit model selection problem, we assume that the policy class under consideration decomposes as a nested sequence:[1]

$$\Pi_1 \subset \Pi_2 \subset \cdots \subset \Pi_M = \Pi.$$

The goal of the learner is to achieve low regret to all classes in the sequence simultaneously, with the regret to policy class Π_m scaling only with $\log|\Pi_m|$. Intuitively, this provides a luckiness guarantee: if a good policy lies in a small policy class, the algorithm discovers this quickly.

To motivate the precise guarantee we ask for, let us recall what is known in the simpler *full-information* online learning setting, where the learner gets to see the entire loss vector ℓ_t at the end of each round. Here, the minimax rate is $O(\sqrt{T\log|\Pi|})$, and it can be shown (Foster et al. (2015); see also Orabona and Pál (2016)) that a variant of the exponential weights algorithm guarantees

$$\mathrm{Reg}(\Pi_m) \le O\Big(\sqrt{T(\log|\Pi_m| + \log m)}\Big), \quad \text{for all } m \in [M]. \tag{1}$$

In other words, by paying a modest additive overhead of $\log m$, we can compete with all M policy classes simultaneously. The most basic variant of our open problem asks whether the natural analogue of Eq. (1) can be attained for contextual bandits.

Open Problem 1a *Design a contextual bandit algorithm that for any sequence $\Pi_1 \subset \Pi_2 \subset \cdots \Pi_M$ ensures*

$$\max_{\pi\in\Pi_m} \mathbb{E}\left[\sum_{t=1}^{T}\ell_t(a_t) - \sum_{t=1}^{T}\ell_t(\pi(x_t))\right] \le O\Big(\sqrt{KT(\log|\Pi_m| + \log m)}\Big), \quad \textit{for all } m \in [M]. \tag{2}$$

We also welcome the following weaker guarantees.

Open Problem 1b *Design a contextual bandit algorithm that for any sequence $\Pi_1 \subset \Pi_2 \subset \cdots \Pi_M$ ensures either:*

1. $\mathrm{Reg}(\Pi_m) \le O\Big(\mathrm{poly}(K, M, \log\log|\Pi|) \cdot \sqrt{T\log|\Pi_m|}\Big)$ *for all* $m \in [M]$.
2. $\mathrm{Reg}(\Pi_m) \le O\big(\mathrm{poly}(K, M, \log\log|\Pi|) \cdot T^{\alpha} \log^{1-\alpha}|\Pi_m|\big)$ *for all* $m \in [M]$, *where* $\alpha \in [1/2, 1)$.

Alternatively, prove that no algorithm can achieve item 2 above for any value $\alpha \in [1/2, 1)$.

The first item here differs from Open Problem 1a only in the dependence on K, M, and $\log\log|\Pi|$ factors, which we do not believe represent the most challenging aspect of the problem. The second item is a further relaxation of the original guarantee. Here we simply ask that the model selection algorithm has regret sublinear in T whenever $\sqrt{T\log|\Pi_m|}$ is sublinear. In other words, if policy class Π_m is learnable on its own, the model selection algorithm should have sublinear regret to it. To attain this behavior it is essential that the exponents α and $1-\alpha$ sum to one. Indeed, it is relatively easy to design algorithms with exponents that do not sum to one,[2] but we do not know of an algorithm satisfying item 2 above for any $\alpha \in [1/2, 1)$. This stands in contrast to other problems involving adaptivity and data-dependence in contextual bandits (Agarwal et al., 2017a), where attaining adaptive guarantees with suboptimal dependence on T is straightforward, and the primary challenge is to attain $\sqrt{T}$-type regret bounds. We also welcome a lower bound showing that this type of model selection guarantees is not possible for contextual bandits.

[1] It is also natural to consider infinite sequences of policy classes, but we restrict to finite sequences for simplicity.

[2] For example we can attain regret $\sqrt{T}\cdot\log|\Pi_m|$ for all m by running Exp4 with a particular prior over policies.

Stochastic Setting. The model selection problem for contextual bandits has yet to be solved even for the stochastic setting, and even when the model is well-specified. Here, we assume: (1) $\{(x_t, \ell_t)\}_{t=1}^T$ are drawn i.i.d. from a fixed distribution $\mathcal{D}$; (2) Each class Π_m is induced by a class of regression functions $\mathcal{F}_m \subset (\mathcal{X} \times \mathcal{A} \to [0,1])$, in the sense that $\Pi_m = \{\pi_f \mid f \in \mathcal{F}_m\}$, where $\pi_f(x) := \operatorname{argmin}_{a \in \mathcal{A}} f(x,a)$; (3) The problem is realizable/well-specified in the sense that there exists index $m_\star$ and regression function $f_\star \in \mathcal{F}_{m_\star}$ such that $\mathbb{E}[\ell(a) \mid x] = f_\star(x,a)$, for all x, a.

The final version of our open problem asks for a model selection guarantee when the problem is stochastic and well-specified. Note that these assumptions imply that the optimal unconstrained policy (in terms of expected loss) is $\pi_{f_\star}$. As such, here we only ask for a regret bound against class $\Pi_{m_\star}$. From an algorithmic perspective, this is the easiest version of the problem.

Open Problem 2 *For some value $\alpha \in [1/2, 1)$, design an algorithm for contextual bandits that for any sequence $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \mathcal{F}_M = \mathcal{F}$, whenever data is stochastic and realizable, ensures*

$$\mathbb{E}\left[\sum_{t=1}^T \ell_t(a_t) - \sum_{t=1}^T \ell_t(\pi_{f_\star}(x_t))\right] \leq O\big(\operatorname{poly}(K, M, \log\log|\mathcal{F}|) \cdot T^\alpha \log^{1-\alpha}|\mathcal{F}_{m_\star}|\big). \qquad (3)$$

Alternatively, prove that no algorithm can achieve this guarantee for any value of $\alpha \in [1/2, 1)$.

We offer \$300 for the first solution to either Open Problem 1 or Open Problem 2.

3. Challenges and Partial Progress

Many natural algorithmic strategies for model selection fail under bandit feedback. These include (a) running Exp4 over all policies with a non-uniform prior adapted to the nested policy class structure, (b) the Corral aggregation strategy (Agarwal et al., 2017b), and (c) an adaptive version of the classical ϵ-greedy strategy (Langford and Zhang, 2008). These strategies all require tuning parameters (e.g., the learning rate) in terms of the class index m of interest, and naive tuning gives guarantees of the form $\tilde{O}(T^\alpha \log^\beta |\Pi_m|)$ for $\alpha + \beta > 1$. Adaptive online learning algorithms like AdaNormalHedge (Luo and Schapire, 2015) and Squint (Koolen and Van Erven, 2015) also fail because they do not adequately handle bandit feedback.[3] We refer the reader to Foster et al. (2019) for more details on these strategies in the context of model selection. The main point here is that model selection for contextual bandits appears to require new algorithmic ideas, even when we are satisfied with weak $O(T^\alpha \log^{1-\alpha}|\Pi_m|)$-type rates where $\alpha > 1/2$.

In a recent paper (Foster et al., 2019), we showed that a guarantee of the form Eq. (3) *is* achievable when $\mathcal{F}_m$ consists of linear functions in d_m dimensions, under distributional assumptions on $\mathcal{D}$. Our strategy was inspired by the fact that if the optimal loss $L^\star = \mathbb{E}\left[\ell(\pi_{f_\star}(x))\right]$ is known, one can test if a given class $\mathcal{F}_m$ contains the optimal policy by running a standard contextual bandit algorithm and checking whether it substantially underperforms relative to $L^\star$. In our linear setup, we showed that one can estimate a surrogate for the optimal loss $L^\star$ at a “sublinear” rate, which allowed us to run this testing strategy and achieve a guarantee akin to Eq. (3) with no prior information. However, we do not know if this strategy can succeed beyond specialized settings where sublinear loss estimation is possible. Along these lines, Locatelli and Carpentier (2018) also observe that knowledge of $L^\star$ can enable adaptive guarantees in Lipschitz bandits, where adaptivity is not possible in the absence of such information (such lower bounds do not appear to carry over to the contextual case).

[3] Their regret bounds do not contain the usual “local norm” term used in the analysis of Exp4 and other bandit algorithms.

For (non-contextual) multi-armed bandits, several lower bounds demonstrate that model selection is *not* possible. Lattimore (2015) shows that for multi-armed bandits, if we want to ensure $O(\sqrt{T})$ regret against a single fixed arm instead of the usual $O(\sqrt{KT})$ rate, we must incur $\Omega(K\sqrt{T})$ regret to one of the remaining arms in the worst case. This precludes a model selection guarantee of the form $\sqrt{T|\mathcal{A}_m|}$ for nested action sets $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \ldots$, which is a natural analogue of Eq. (2) for bandits.[4] Related lower bounds are also known for Lipschitz bandits (Locatelli and Carpentier, 2018; Krishnamurthy et al., 2019). On the positive side, Chatterji et al. (2019) show that, with distributional assumptions, it is possible to adapt between multi-armed bandits and linear contextual bandits.

4. Consequences and Connections to Other Problems

Switching Regret for Bandits. In full-information online learning, algorithms for *switching regret* (Herbster and Warmuth, 1998) simultaneously ensure that for all sequences of actions $a_1^\star, \ldots, a_T^\star$, $\mathbb{E}\left[\sum_{t=1}^T \ell_t(a_t) - \sum_{t=1}^T \ell_t(a_t^\star)\right] \le O\left(\sqrt{S(a_{1:T}^\star)\cdot T}\right)$, where $S(a_{1:T}^\star)$ denotes the number of switches in the sequence. In the (non-contextual) multi-armed bandit setting, with no prior knowledge of the number of switches S, the best guarantee we are aware of is $\mathbb{E}\left[\sum_{t=1}^T \ell_t(a_t) - \sum_{t=1}^T \ell_t(a_t^\star)\right] \le O\left(\sqrt{S(a_{1:T}^\star)\cdot KT} + T^{3/4}\right)$ which can be attained by combining the Bandits-over-Bandits strategy from Cheung et al. (2019) with Exp3.[5] A solution to Open Problem 1a would immediately yield a nearly-optimal switching regret bound of $\widetilde{O}(\sqrt{S(a_{1:T}^\star)\cdot KT})$ for bandits by choosing the mth policy class Π_m to be the set of all sequences $a_1^\star, \ldots, a_T^\star$ with at most m switches.[6] Solving Open Problem 1a would also lead to similar improvements in switching regret for contextual bandits.

Second-Order Regret Bounds for Online Learning. Consider full-information online learning, and let P_t denote the algorithm's distribution over policies at time t. An unresolved COLT 2016 open problem of Freund (2016) asks whether there exists an algorithm for this setting with regret at most $O\left(\sqrt{\sum_{t=1}^T \mathrm{Var}_{\pi\sim P_t}(\ell_t(\pi(x_t)))\cdot \log(1/\varepsilon)}\right)$ against the top ε-quantile of policies for all $\varepsilon > 0$ simultaneously. A slight strengthening of Freund's open problem asks for the following bound:

$$\sum_{t=1}^T \mathbb{E}_{\pi\sim P_t}\,\ell(\pi(x_t)) - \mathbb{E}_{\pi\sim Q}\,\ell_t(\pi(x_t)) \le O\left(\sqrt{\sum_{t=1}^T \mathrm{Var}_{\pi\sim P_t}(\ell_t(\pi(x_t)))\cdot \mathrm{KL}(Q\|P_1)}\right), \quad \forall Q \in \Delta_\Pi. \tag{4}$$

Eq. (4) implies the weaker quantile bound by choosing Q to be uniform over the top ε-fraction of policies and P_1 to be the uniform distribution over all policies. While the $\log(1/\varepsilon)$-type quantile bound does not seem to imply Eq. (4) directly, historically KL-based bounds have quickly followed quantile bounds (Chaudhuri et al., 2009; Luo and Schapire, 2015; Koolen and Van Erven, 2015).

The guarantee in Eq. (4) would immediately yield a positive resolution to Open Problem 1a via the following reduction: (1) Choose $P_1(\pi) \propto \frac{1}{|\Pi_m|m^2}$ for all $\pi \in \Pi_m$; (2) To handle bandit feedback, draw $a_t \sim p_t$ and feed importance weighted losses $\hat{\ell}_t(a) := \frac{\ell_t(a)}{p_t(a)}\mathbb{I}\{a_t = a\}$ into the full-information algorithm at each round, where $p_t(a) := \sum_{\pi\in\Pi} P_t(\pi)\,\mathbb{I}\{\pi(x_t) = a\}$. Conversely, a lower bound showing that Eq. (2) is not attainable would imply that no full-information algorithm can achieve Eq. (4), which strongly suggests that the quantile bound in Freund's open problem is also not attainable.

[4] This does not preclude a guarantee of the form Eq. (2), however, since we pay for the maximum number of actions.

[5] Auer et al. (2002) achieves regret $\widetilde{O}(\sqrt{S\cdot KT})$, but only when a bound S on the switches is known a-priori.

[6] Formally, this is accomplished by setting $\mathcal{X} = [T]$ and $\pi(t) = a_t^\star$.

References

Alekh Agarwal, Akshay Krishnamurthy, John Langford, Haipeng Luo, and Robert E. Schapire. Open problem: First-order regret bounds for contextual bandits. In *Conference on Learning Theory*, 2017a.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. *Conference on Learning Theory*, 2017b.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 2002.

Niladri S Chatterji, Vidya Muthukumar, and Peter L Bartlett. OSOM: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Kamalika Chaudhuri, Yoav Freund, and Daniel J Hsu. A parameter-free hedging algorithm. In *Advances in neural information processing systems*, 2009.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Dylan J Foster, Alexander Rakhlin, and Karthik Sridharan. Adaptive online learning. In *Advances in Neural Information Processing Systems*, pages 3375–3383, 2015.

Dylan J Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. In *Advances in Neural Information Processing Systems*, 2019.

Yoav Freund. Open problem: Second order regret bounds based on scaling time. In *Conference on Learning Theory*, 2016.

Mark Herbster and Manfred K Warmuth. Tracking the best expert. *Machine learning*, 1998.

Wouter M Koolen and Tim Van Erven. Second-order quantile methods for experts and combinatorial games. In *Conference on Learning Theory*, 2015.

Akshay Krishnamurthy, John Langford, Aleksandrs Slivkins, and Chicheng Zhang. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. *Conference on Learning Theory*, 2019.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, 2008.

Tor Lattimore. The pareto regret frontier for bandits. In *Advances in Neural Information Processing Systems*, 2015.

Andrea Locatelli and Alexandra Carpentier. Adaptivity to smoothness in X-armed bandits. In *Conference on Learning Theory*, 2018.

Haipeng Luo and Robert E Schapire. Achieving all with no parameters: Adanormalhedge. In *Conference on Learning Theory*, 2015.

Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. In *Advances in Neural Information Processing Systems*, 2016.