The Pareto Frontier of model selection for general Contextual Bandits

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent progress in model selection raises the question of the fundamental limits of these techniques. Under specific scrutiny has been model selection for general contextual bandits with nested policy classes, resulting in a COLT2020 open problem. It asks whether it is possible to obtain simultaneously the optimal single algorithm guarantees over all policies in a nested sequence of policy classes, or if otherwise this is possible for a trade-off $\alpha \in [\frac{1}{2}, 1)$ between complexity term and time: $\ln(|\Pi_m|)^{1-\alpha}T^{\alpha}$. We give a disappointing answer to this question. Even in the purely stochastic regime, the desired results are unobtainable. We present a Pareto frontier of up to logarithmic factors matching upper and lower bounds, thereby proving that an increase in the complexity term $\ln(|\Pi_m|)$ independent of T is unavoidable for general policy classes. As a side result, we also resolve a COLT2016 open problem concerning second-order bounds in full-information games.

1 Introduction

Contextual multi-armed bandits are a fundamental problem in online learning [Auer et al., 2002, Langford and Zhang, 2007, Chu et al., 2011, Abbasi-Yadkori et al., 2011]. The contextual bandit problem proceeds as a repeated game between a learner and an adversary. At every round of the game the adversary prepares a pair of a context and a loss over an action space, the learner observes the context and selects an action from the action space and then observes only the loss of the selected action. The goal of the learner is to minimize their cumulative loss. The performance measure, known as *regret*, is the difference between the learner's cumulative loss and the smallest loss of a fixed policy, belonging to an apriori determined policy class, mapping contexts to actions. Given a single contextual bandit instance with finite sized policy class, the well-known Exp4 algorithm [Auer et al., 2002] achieves the optimal regret bound of $\mathcal{O}(\sqrt{KT\ln(|\Pi|)}$. Regret guarantees degrade with the complexity of the policy class, therefore a a learner might want to leverage "guesses" about the optimal policy. Given policy classes $\Pi_1 \subset \cdots \subset \Pi_M$, a learner would ideally suffer regret scaling only with the complexity of Π_{m^*}, the smallest policy class containing the optimal policy π^*. While these kind of results are obtainable in full-information games, in which the learner gets to observe the loss for all actions, they are impossible for multi-armed bandits [Lattimore, 2015]. In some aspects, contextual bandits are an intermediate setting between full-information and multi-armed bandits and it is unknown if model selection is possible. Foster et al. [2020b] stated model selection in contextual bandits as a relevant open problem in COLT2020. Any positive result for model selection in contextual bandits would imply a general way to treat multi-armed bandits with a switching baseline. Furthermore any negative result is conjectured to implicate negative results on another unresolved open problem on second order bounds for full-information games [Freund, 2016].

In this paper, we give a fairly complete answer to the questions above.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

1. We provide a Pareto frontier of upper bounds for model selection in contextual bandits with finite sized policy classes.
2. We present matching lower bounds that shows that our upper bounds are tight, thereby resolve the motivating open problems [Foster et al., 2020b].
3. We present a novel impossibility result for adapting to the number of switch points under adaptive adversaries [Besbes et al., 2014].
4. We negatively resolve an open problem on second order bounds for full-information [Freund, 2016].

Related work. A problem closely related to contextual bandits with finite policy classes are linear contextual bandits. Model selection in linear contextual bandit problems has recently received significant attention, however none of these resuls transfer to the finite policy case. In the linear bandits problem the m-th policy class is a subset of $\mathbb{R}^{d_m}$ and the losses $\ell_{t,\pi(x)}, \pi \in \Pi_m$ are linear, that is $\ell_{t,\pi(x)} = \langle \theta_m, \phi_m(x, \pi(x)) \rangle + \xi$. Here $\phi_m : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^{d_m}$ is a feature embedding mapping from context-action pairs into $\mathbb{R}^{d_m}$, ξ is mean-zero sub-Gaussian noise with variance proxy equal to one and $\theta_m \in \mathbb{R}^{d_m}$ is an unknown parameter.

Foster et al. [2019] assume the contexts are also drawn from an unknown distribution $x \sim \mathcal{D}$ and propose an algorithm which does not incur more than $\tilde{O}\left(\frac{1}{\gamma^3}(i^*T)^{2/3}(Md_{i^*})^{1/3}\right)$, where γ^3 is the smallest eigenvalue of the covariance matrix of feature embeddings $\Sigma = \mathbb{E}_{x\sim\mathcal{D}}\left[\frac{1}{M}\sum_{a\in\mathcal{A}} \phi_M(x,a)\phi_M(x,a)^\top\right]$. Pacchiano et al. [2020b] propose a different approach based on the corralling algorithm of Agarwal et al. [2017] which enjoys a $\tilde{O}(d_{i^*}\sqrt{T})$ regret bound for finite action sets and $\tilde{O}(d_{i^*}^2\sqrt{T})$ bound for arbitrary action sets $\mathcal{A}$. Later, Pacchiano et al. [2020a] design an algorithm which enjoys a gap-dependent guarantee under the assumption that all of the miss-specified models have regret $R_i(t) \geq \Delta t, \forall t \in [T]$. Under such an assumption, the authors recover a regret bounds of the order $\tilde{O}(d_{i^*}\sqrt{T} + d_{i^*}^4/\Delta)$ for arbitrary action sets. Cutkosky et al. [2020] also manage to recover the $O(d_{i^*}\sqrt{T})$ and $O(d_{i^*}^2\sqrt{T})$ bounds for the model selection problems through their corralling algorithm. Ghosh et al. [2021] propose an algorithm which enjoys $\tilde{O}\left(\frac{d_M^2}{\gamma^{4.65}} + \sqrt{d_{m^*}T}\right)$ in the finite arm setting, where $\gamma = \min\{|\theta_{m^*,i}| : |\theta_{m^*,i}| > 0\}$ is the smallest, in absolute value, entry of θ_{m^*}. Their algorithm also enjoys a similar guarantee for arbitrary action sets with $\sqrt{d_{m^*}T}$ replaced by $d_{m^*}\sqrt{T}$. Zhu and Nowak [2021] show that it is impossible to achieve the desired regret guarantees of $\sqrt{d_{m^*}T}$ without additional assumptions by showing a result similar to the one of Lattimore [2015]. The work of Lattimore [2015] states that in the stochastic multi-armed bandit problem it is impossible to achieve $\sqrt{T}$ regret to a fixed arm, without suffering at least $K\sqrt{T}$ regret to a different arm.

Chatterji et al. [2020] study the problem of selecting between an algorithm for the linear contextual bandit problem and the simple stochastic multi-armed bandit problem, that is they aim to achieve simultaneously a regret guarantee which is instance-dependent optimal for the stochastic multi-armed bandit problem and optimal for the finite arm stochastic linear bandit problem. The proposed results only hold under additional assumptions. More generally, the study of the corralling problem, in which we are presented with multiple bandit algorithms and would like to perform as well as the best one, was initiated by Agarwal et al. [2017]. Other works which fall into the corralling framework are that of Foster et al. [2020a] who study the miss-specified linear contextual bandit problem, that is the observed losses are linear up to some unknown ϵ miss-specification, and the work of Arora et al. [2021] who study the corralling problem for multi-armed stochastic bandit algorithms.

Our work also shows an impossibility result for the stochastic bandit problem with non-stationary rewards. Auer [2002] first investigates the problem under the assumption that there are L distributional changes throughout the game and gives an algorithm with a $\tilde{O}(\sqrt{KLT})$ *dynamic regret*[1] bound, under the assumption that L is known. Auer et al. [2019] achieves similar regret guarantees without assuming that the number if switches (or changes) of the distribution is known. A different measurement of switches is the total variation of changes in distribution $V_T = \sum_{t=2}^{T} \|\mathbb{E}[\ell_t] - \mathbb{E}[\ell_{t-1}]\|_\infty$. Multiple works give dynamic regret bounds of the order $\tilde{O}(V_T^{1/3}T^{2/3})$ (hiding dependence on the size of the

[1] In dynamic regret the comparator is the best action for the current distribution.

General CB	Upper bound	Lower bound
adaptive adversary	$\mathcal{O}(\max\{\mathfrak{C}, \frac{\ln\lvert\Pi_m\rvert}{\mathfrak{C}}\}\sqrt{MKT})$	$\Omega(\max\{\mathfrak{C}, \frac{\ln\lvert\Pi_m\rvert}{\mathfrak{C}}\}\sqrt{\frac{KT}{\ln(K)}})$
oblivious adversary / stochastic	$\mathcal{O}(\max\{\mathfrak{C}, \frac{\ln\lvert\Pi_m\rvert}{\mathfrak{C}}\}\sqrt{MKT})$	$\Omega(\max\{\mathfrak{C}, \frac{\ln\lvert\Pi_m\rvert}{\mathfrak{C}}\}\sqrt{T})$
S-switch	Upper bound	Lower bound
adaptive adversary	$\tilde{\mathcal{O}}(\max\{\mathfrak{C}, \frac{S}{\mathfrak{C}}\}\sqrt{KT})$	$\Omega(\max\{\mathfrak{C}, \frac{S}{\mathfrak{C}}\}\sqrt{KT})$
oblivious adversary	$\tilde{\mathcal{O}}(\sqrt{SKT} + T^{3/4})$	$\Omega(\sqrt{SKT})$
stochastic	$\tilde{\mathcal{O}}(\sqrt{SKT})$	$\Omega(\sqrt{SKT})$

Table 1: Overview of our results. Our novel contributions are in bold; lower bounds only hold if the expressions are not exceeding $\Theta(T)$. The stochastic/oblivious lower bounds hold only for proper algorithms.

policy class) when V_T is known, including for extensions of the multi-armed bandit problem like contextual bandits and linear contextual bandits [Besbes et al., 2014, Luo et al., 2018, Besbes et al., 2015, Wei et al., 2017]. Cheung et al. [2019], Zhao et al. [2020] further show algorithms which enjoy a parameter free regret bound of the order $\tilde{O}(V_T^{1/4}T^{3/4})$ (hiding dependence on dimensionality) for the linear bandits problem. The lower bound in Table 1 might seem to contradict such results. In Section 5.1 we carefully explain why this is not the case.

Finally, our lower bounds apply to the problem of devising an algorithm which simultaneously enjoys a second order bound over any fraction of experts. Cesa-Bianchi et al. [2007] first investigate the problem of second order bounds for the experts problem, in which the proposed algorithm maintains a distribution p_t over the set of K experts, during every round of the game. The experts are assumed to have stochastic losses ℓ_t and the work shows an algorithm with $\tilde{O}(\sqrt{\sum_{t=1}^{T}\mathbb{V}_{i\sim p_t}[\ell_{t,i}]\log K})$ regret guarantee. Chaudhuri et al. [2009], Chernov and Vovk [2010], Luo and Schapire [2015], Koolen and Van Erven [2015] study a different experts problem in which the comparator class for the regret changes from the best expert in hindsight to the uniform distribution over the best $\lfloor \epsilon K \rfloor$ experts for an arbitrary positive ϵ. The above works propose algorithms which achieve a $\tilde{O}(\sqrt{T\log(1/\epsilon)})$ regret bound for all ϵ simultaneously. Freund [2016] asks if there exists an algorithm which enjoys both guarantees at the same time, that is, does there exist an algorithm with regret bound $\tilde{O}(\sqrt{\sum_{t=1}^{T}\mathbb{V}_{i\sim p_t}[\ell_{t,i}]\log(1/\epsilon)})$ which holds simultaneously for all positive ϵ.

Notation. For any $N \in \mathbb{N}$, $[N]$ denotes the set $\{1, \ldots, N\}$. $\tilde{O}$ notation hides poly-logarithmic factors in the horizon T and the number of arms K but not in the size of the policy classes $|\Pi_m|$.

2 Problem setting

We consider the contextual bandit problem with general policy classes of finite size. There are K arms and nested policy classes $(\Pi_m)_{m=1}^{M}$, where a policy $\pi \in \Pi_m, \pi : \mathcal{X} \to [K]$ is a mapping from an arbitrary context space $\mathcal{X}$ into the set of K arms. The game is played for T rounds and at any time t, the agent observes a context $x_t \in \mathcal{X}$, selects arm $A_t \in [K]$ and observes the loss ℓ_{t,A_t} from an otherwise unobserved loss vector $\ell_t \in [K]$. We measure an algorithm's performance in terms of pseudo-regret, which is the expected cumulative regret of the player against following a fixed policy in hindsight

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

Environments. We distinguish between *stochastic* environments and *oblivious* or *adaptive* adversaries. In stochastic environments, there are unknown distribution $P_{\mathcal{X}}, Q$ such that $x_t \sim P_{\mathcal{X}}$ and $\ell_t \sim Q(\cdot|x_t)$ are i.i.d. samples. In the adversarial regime, the distributions can change over time, i.e. $x_t \sim P_{\mathcal{X},t}, \ell_t \sim Q_t(\cdot|x_t)$. When the choices are fixed at the beginning of the game, the adversary is called oblivious, while an adaptive adversary can chose $P_{\mathcal{X},t}, Q_t$ based on all observations up to time $t-1$.

Often the stochastic-adversarial hybrid problem has been studied with adversarially chosen context but stochastic losses. In our work, all upper bounds hold in the stronger notion where both the losses and the contexts are adaptive, while the lower bounds hold for the weaker notion where only the contexts are adaptive.

Definition 1. *We call an algorithm* proper*, if at any time t, the algorithm follows the recommendation of a policy π_t, and if the choice of π_t by the algorithm, is independent of the context x_t.*

Example. EXP4 is proper.

We introduce the notion of *proper* algorithms, because it greatly reduces the technicalities for lower bound proofs in the non-adaptive regimes. We give compelling arguments of why the lower bound holds for general algorithms as well, but do not provide a formal proof for that conjecture.

Open problem [Foster et al., 2020b]. The regret upper bounds for all regimes introduced above for a fixed policy class Π of finite size are of the order $\tilde{O}(\sqrt{\ln(|\Pi|)KT})$ and can be achieved by the Exp4 algorithm [Auer et al., 2002]. The question asked by Foster et al. [2020b]: For a nested sequence of policies $\Pi_1 \subset \Pi_2 \subset \cdots \subset \Pi_M$, is there a universal $\alpha \in [\frac{1}{2}, 1)$ such that a regret bound of

$$\mathrm{Reg}(T, \Pi_m) = \mathrm{PolyLog}(K, M)\tilde{\mathcal{O}}\left(\ln(|\Pi_m|)^{1-\alpha}T^{\alpha}\right) \tag{1}$$

is obtainable for all $m \in [M]$ simultaneously?

W.l.o.g. we can assume that $M = \mathcal{O}(\ln\ln(|\Pi_M|)) = \mathcal{O}(\ln(T))$. Otherwise we take a subset of policy classes that includes Π_M and where two consequent policy classes at least square in size. Due to nestedness, any guarantees on this subset of models imply up to constants the same bounds on the full set.

S-switch A motivating example for studying nested policy classes is the S-switch problem. The context is simply $x_t = t \in [T]$ and the set of policies is given by

$$\Pi_S = \left\{ \pi \middle| \sum_{t=1}^{T-1} \mathbb{I}\{\pi_t \neq \pi_{t+1}\} \leq S \right\},$$

the set of policies that changes its action not more than S many times. Any positive result for contextual bandits with finite sized policy classes would provide algorithms that adapt to the number of switch points, since $\ln|\Pi_S| = \tilde{\mathcal{O}}(S)$. To make clear what problem we are considering, we are using $\mathrm{Reg}_{SW}(T, S)$ to denote the regret in the switching problem.

3 Upper bounds

In this section, we generalize the Hedged-FTRL algorithm [Foster et al., 2020a] to obtain an upper bound for model selection over a large collection of $\sqrt{T}$ regret algorithms.

Theorem 1. *For any $\mathfrak{C} > 0$, we can tune Hedged-FTRL over a selection of M instances of EXP4 operating on policy classes $\Pi_1, \ldots \Pi_M$, such that the following regret bound holds uniformly over all $m \in [M]$*

$$\mathrm{Reg}(T, \Pi_m) = \tilde{\mathcal{O}}\left(\max\left\{\mathfrak{C}, \frac{\ln|\Pi_m|}{\mathfrak{C}}\right\}\sqrt{MKT}\right).$$

Hedged-FTRL. (α, R)-hedged FTRL, introduced in Foster et al. [2020a], works as follows. Define $\rho_{t,m}^{-1} = \min_{s\in[t]} p_{s,m}$, and initialize $B_{0,m} = \rho_{1m}^{\alpha} R$. The algorithm runs FTRL as a top algorithm over a collection of base algorithms. At any time t, the top selects a base algorithm M_t to play the current action, observes its loss and gives as feedback an important weighted loss to the selected base. The Top algorithm is updated based on the observed loss. Whenever the base played at round t would satisfy $\rho_{t+1,M_t} > \rho_{t,M_t}$, the loss fed to the top algorithm is adjusted with a bias b_{t,M_t}, such that the cumulative biases track the quantity $\rho_{t+1,m}^{\alpha} R$. This has been shown to be always possible. We extend this algorithm in two ways. We make R dependent on the base algorithm, $R \in \mathbb{R}_+^M$, and we change the definition of ρ by selecting a threshold $\beta \le p_1$, such that $\rho_{t,m}^{-1} = \min\{\beta_m, \min_{s\in[t]} p_{s,m}\}$. The condition for adjusting the biases reads

Input: $\alpha, R, \beta, Top, (Base_i)_{i=1}^M$
for $t = 1, \dots, T$ **do**
  Get M_t, q_{t,M_t} from Top
  Let $Base_{M_t}$ play the next round and receive A_t
  Play A_t and observe ℓ_{t,A_t}
  Update $Base_{M_t}$ with $\ell_{t,A_t}/q_{t,M_t}$
  Update Top with (M_t, ℓ_t)
  if q_{t+1} *would violate Eq.* (2) **then**
    Bias losses by b_t to ensure Eq. (2).
  end
end

Algorithm 1: Hedged FTRL

$$\forall m \in [M] : \; B_{0,m} + \sum_{s=1}^{t} b_{s,m} = \rho_{t+1,m}^{\alpha} R_m \,. \tag{2}$$

The condition in Equation 2 is motivated in a similar way to the stability condition in the work of Agarwal et al. [2017]. Algorithm 1 constructs an unbiased estimator for the loss vector, $\ell_{t,A_t}/q_{t,M_t}\mathbf{e}_{M_t}$, and updates each base algorithm accordingly. A similar update is present in the CORRAL algorithm Agarwal et al. [2017] in which each of the base learners also receives an importance weighted loss. The regret of the base learners is assumed to scale with the variance of the importance weighted losses. This assumption is natural and in practice holds for all bandit or expert algorithms. The scaling of the regret, however, must be appropriately bounded as Agarwal et al. [2017] show, otherwise no corralling or model selection guarantees are possible. Formally, the following stability property is required. If an algorithm $\mathcal{B}$ enjoys a regret bound R under environment $\mathcal{V}$ with loss sequence $(\ell_t)_{t=1}^T$, then the algorithm is (α, R)-*stable* if it enjoys a regret bound of the order $\mathbb{E}[\rho_{\max}^{\alpha}]R$ under the environment $\mathcal{V}'$ of importance weighted losses $(\hat{\ell}_t)_{t=1}^T$, where $\rho_{\max}$ is the maximum variance of the T losses and the expectation is taken with respect to any randomness in $\mathcal{B}$. Essentially all bandit and expert algorithms used in practice are (α, R)-stable with $\alpha \le 1/2$, e.g., Exp4 is $(1/2, \sqrt{KT\ln(|\Pi|)})$-stable. The bias terms in Algorithm 1 intuitively cancel the additional variance introduced by the importance weighted losses and this is why we require the biases to satisfy Equation 2.

Theorem 2. *Given a collection of base algorithms $(\mathcal{B}_m)_{m=1}^M$ which are $(1/2, \sqrt{\mathfrak{C}_m T})$-stable, that is*

$$\forall m \in [M] : \; \mathrm{Reg}_{\mathrm{Imp}}(T, \mathcal{B}_m) \le \mathbb{E}[\sqrt{\rho_{Tm}}]\sqrt{\mathfrak{C}_m T} \,,$$

and any $\mathfrak{C} \ge 0$, then the regret of $(1/2, R, \beta)$-hedged Tsallis-Inf with $R_m = \sqrt{\mathfrak{C}_m T}$, $\beta_m = \frac{1}{M}\max\{1, \frac{\mathfrak{C}^2}{\mathfrak{C}_m}\}$ satisfies a simultaneous regret of

$$\forall m \in [M] : \; \mathrm{Reg}(T, \mathcal{B}_m) \le 2\max\left\{\mathfrak{C}, \frac{\mathfrak{C}_m}{\mathfrak{C}}\right\}\sqrt{MT} + \sqrt{2MT} \,.$$

The analysis follows closely the proof of Foster et al. [2020a] and is postponed to Appendix A.

Theorem 2 recovers the bounds of Pacchiano et al. [2020b] for model selection in linear bandits, but holds in more general settings including adaptive adversaries in both contexts and losses. It neither requires nestedness of the policies nor that the policies operate on the same action or context space.

Proof of Theorem 1. The EXP4 algorithm initialized with policy class Π_m satisfies the condition of Theorem 2 with $\mathfrak{C}_m = \mathcal{O}(\ln|\Pi_m|)$, as shown in Agarwal et al. [2017]. Hence Theorem 1 is a direct corollary of Theorem 2. □

4 Lower bounds

We present lower bounds that match the upper bounds from Section 3 up to logarithmic factors, thereby proving a tight Pareto frontier of worst-case regret guarantees in model selection for contextual bandits.

In the first part of this section, we consider a special instance of S-switch with *adaptive* adversary. The proof technique based on Pinsker's inequality is folklore and leads to the following theorem.

Theorem 3. *For any $K \geq 3$, sufficiently large T, and any algorithm with regret guarantee*

$$\mathrm{Reg}_{SW}(T, 1) = \mathcal{O}(\mathfrak{C}\sqrt{KT}),$$

there exists for any number of switches $S = \Omega(\mathfrak{C}^2)$ a stochastic bandit problem such that

$$\mathrm{Reg}_{SW}(T, S) = \Omega\left(\min\left\{\frac{S}{\mathfrak{C}}\sqrt{KT}, T\right\}\right).$$

This bound holds even when the agent is informed about the number of switches up to time t.

Since this bound holds even when the agent is informed about when a switch occurs, we can restrict the policy class to policies that only switch arms whenever the agent is informed about a switch in the environment. This as a contextual bandit problem with context $\mathcal{X} = [S+1]$ and $|\Pi_S| = \Theta(K^S)$ policies. Hence Theorem 3 implies a lower bound of $\mathrm{Reg}(T, \Pi_S) = \Omega\left(\min\left\{\frac{\ln|\Pi_S|}{\mathfrak{C}\ln(K)}\sqrt{KT}, T\right\}\right)$.

In the second part of the section, we consider the *stochastic* regime. Our lower bound construction is non-standard and relies on bounding the total variation between problem instances directly without the use of Pinsker's inequality.

Theorem 4. *There exists policy classes $\Pi_1 \subset \Pi_2$ [2] with $|\Pi_2| = \Omega(\mathfrak{C}^2)$, such that if the regret of a proper algorithm is upper bounded in any environment by*

$$\mathrm{Reg}(T, \Pi_1) = \mathcal{O}(\mathfrak{C}\sqrt{T})$$

then there exists an environment such that

$$\mathrm{Reg}(T, \Pi_2) = \Omega\left(\max\left\{\mathfrak{C}, \frac{\ln|\Pi_2|}{\mathfrak{C}}\right\}\sqrt{T}\right).$$

These theorems directly provide negative answers to [Foster et al., 2020b].

Corollary 1. *There is no $\alpha \in [\frac{1}{2}, 1)$ that satisfies the regret guarantee of open problem (1) for any algorithm in the adaptive adversarial regime or any proper algorithm in the stochastic case.*

Proof. By Theorems 3 and 4, for any $\alpha > 0$ there exists $K = 3$, $M = 2$, $|\Pi_1| = 1, |\Pi_2| = \Theta(\exp(T^\alpha))$. Assume that

$$\mathrm{Reg}(T, \Pi_1) \leq C_T T^\alpha = C_T T^{\alpha - \frac{1}{2}}\sqrt{T},$$

where $C_T = \mathrm{PolyLog}(T)$. Hence by Theorem 3 and Theorem 4 there exist environments where

$$\mathrm{Reg}(T, \Pi_2) = \Omega\left(\frac{T^\alpha}{C_T T^{\alpha - \frac{1}{2}}}\sqrt{T}\right) = \tilde{\Omega}(T).$$

□

4.1 Adaptive adversary: S-switch(Δ) Problem

We present the adaptive environment in which model selection fails and the proof of Theorem 3.

The adversary switches the reward distribution up to S many times, thereby segmenting the time into $S+1$ phases $(1, \ldots, \tau_1, \tau_1 + 1, \ldots, \tau_2, \ldots, \tau_S, \ldots T)$. We denote $x_t \in [S+1]$ as the counter of phases and assume the agent is given this information. For each phase $x_t \in [S]$, the adversary selects

[2] In Foster et al. [2020b] open problem 2, they ask about model based contextual bandit with realizability. Our lower bound is providing an instance of that.

an optimal arm $(a_s^*)_{s=1}^S$ uniformly at random among the first $K-1$ arms. If $x_t \le S$, the losses are i.i.d. Bernoulli random variables with means

$$\mathbb{E}[\ell_{t,i}] = \frac{1}{2} - \begin{cases} \frac{7}{8}\Delta & \text{for } i = K \\ \Delta & \text{for } i = a^*_{x_t} \\ 0 & \text{otherwise.} \end{cases}$$

In phase $S+1$, all losses are 0 until the end of the game. The adversary decides on the switching points based on an adaptive strategy. A switch from phase $i < S+1$ to $i+1$ occurs when the player has played $\mathrm{N}_{\max} = \lceil \frac{K-1}{30\Delta^2} \rceil$ times an arm in $[K-1]$ in phase i. We can see this problem either as a special case of S-switch problem, or alternatively as a contextual bandit problem with $|\Pi_S| = K^{S+1}$ policies.

The lower bound proof for S-switch(Δ) relies on the following Lemma, which is proven in Appendix B.

Lemma 1. *Let an agent interact with a $K-1 \ge 2$ armed bandit problem with centered Bernoulli losses and randomized best arm of gap $\Delta \le \frac{1}{8\sqrt{3}}$ for an adaptive number of time steps N. If the probability of $N \ge \mathrm{N}_{\max} = \lceil \frac{K-1}{192\Delta^2} \rceil$ is at least $\frac{1}{2}$, then the regret after $\mathrm{N}_{\max}$ time-steps conditioned on the event $N \ge \mathrm{N}_{\max}$ is lower bounded by*

$$\mathrm{Reg} \ge \frac{\Delta}{4} \mathrm{N}_{\max} \ .$$

Proof of Theorem 3. Let $\mathcal{T} = x_T - 1$ be the number of switches the agent triggers from the adversary in the game. We set $\Delta = \min\left\{ \frac{S\sqrt{K-1}}{3072\mathfrak{C}\sqrt{T}}, \frac{1}{8\sqrt{3}} \right\}$, which guarantees

$$\mathrm{N}_{\max} = \left\lceil \max\left\{ \frac{49152\mathfrak{C}^2 T}{S^2}, K-1 \right\} \right\rceil \le \frac{T}{2S} \ .$$

The regret is bounded by $\mathrm{Reg}_{SW}(T,S) \ge \mathbb{P}(\mathcal{T} \ne S)\frac{\Delta}{8}(T - S\,\mathrm{N}_{\max}) = (1 - \mathbb{P}(\mathcal{T} = S))\frac{\Delta T}{16}$, since the agent cannot have played the optimal action more than $S\,\mathrm{N}_{\max}$ times without triggering S switches. If the probability of triggering the S's switch is below $\frac{1}{2}$, we are done. Otherwise by Lemma 1, the regret against a non-switching baseline on arm K is bounded by $\mathrm{Reg}_{SW}(T,1) \ge \mathbb{P}(\mathcal{T} = S)\frac{\Delta}{8}\,\mathrm{N}_{\max}\,S$. By assumption, we have $\mathrm{Reg}_{SW}(T,1) \le \mathfrak{C}\sqrt{(K-1)T}$, hence $\mathbb{P}(\mathcal{T} = S) \le \frac{8\mathfrak{C}\sqrt{(K-1)T}}{\Delta\,\mathrm{N}_{\max}\,S}$. Plugging this into the bound above, yields

$$\mathrm{Reg}_{SW}(T,S) \ge \left(1 - \frac{8\mathfrak{C}\sqrt{(K-1)T}}{\Delta\,\mathrm{N}_{\max}\,S}\right)\frac{\Delta T}{16} \ge \frac{\Delta T}{32} = \Omega\left(\min\left\{\frac{S}{\mathfrak{C}}\sqrt{KT}, T\right\}\right) \ .$$

□

4.2 Stochastic lower bound

We present the stochastic environment in which proper algorithms fail and the proof of Theorem 4.

We are defining environments $(\mathcal{E}_i)_{i=0}^k$ with $\Pi_2 = \{\pi_i | i \in [k] \cup \{0\}\}$ policies and $\Pi_1 = \{\pi_0\}$. In all environments, we have $K = 3$ and π_0 always chooses action 3, while $(\pi_i)_{i=1}^k$ are playing an action from $\{1,2\}$ uniformly at random. (In other words, the context is $\mathcal{X} = \{-1,1\}^k$ with x_t sampled uniformly at random.)

In each environment, the losses of actions $\{1,2\}$ at any time step satisfy $\ell_{t,1} = 1 - \ell_{t,2}$, which are conditioned on x_t independent Bernoulli random variables, with mean

$$\mathbb{E}_{\mathcal{E}_0}[\ell_{t,1}] = \mathbb{E}_{\mathcal{E}_0}[\ell_{t,2}] = \frac{1}{2} \qquad \text{and } \forall i \in [k] : \mathbb{E}_{\mathcal{E}_i}[\ell_{t,\pi_i(x_t)}] = \frac{1}{2}(1-\Delta) \ .$$

Action 3 gives a constant loss of $\frac{1}{2} - \frac{1}{4}\Delta$ in all environments. This construction ensures the following properties:

1) Observing x_t without $\ell_{t,1}$ or $\ell_{t,2}$ provides no information to distinguish between the environments.
2) The losses for all policies $j \in [k] \setminus \{i\}$ are i.i.d. Bernoulli random variables with mean $\frac{1}{2}$ under $\mathcal{E}_i$.

Additionally, we require the following Lemma whose proof is deferred to Appendix B.

Lemma 2. *Let $\Delta \leq \frac{1}{4}$ and $k \geq e^{20} + 1$. If the number of times $\ell_{t,1}$ or $\ell_{t,2}$ is observed for a proper algorithm is bounded by $N = \lfloor \frac{\ln(k-1)}{20\Delta^2} \rfloor$ almost surely and $\Delta^2 N \geq \frac{1}{2}$, then for any measurable event E under environments $(\mathcal{E}_i)_{i=0}^k$, it holds that*

$$\min_{i \in [k]} \mathbb{P}_{\mathcal{E}_i}(E) - \mathbb{P}_{\mathcal{E}_0}(E) \leq \frac{17}{\sqrt[4]{k-1}} \leq \frac{1}{4}\,.$$

Proof of Theorem 4. Let $c_1 = \frac{1}{160}, c_2 = \frac{1}{10} c_1^{-2}$. We assume w.l.o.g. that $c_2 \mathfrak{C}^2 \leq \ln(k) \leq \frac{1}{2}T$. If $\ln(k) = \mathcal{O}(\mathfrak{C}^2)$, then the regret is lower bounded by the regret of Π_1, if $\ln(k) = \Omega(T)$, the optimal regret is linear in T anyway. Pick $\Delta = \min\{\frac{c_1 \ln(k)}{\mathfrak{C}\sqrt{T}}, \frac{1}{4}\}$. This choice implies $N = \lfloor \frac{\ln(k)}{20\Delta^2} \rfloor \leq \frac{T}{2}$. Denote $\mathcal{N}$ the random number of plays in action $\{1, 2\}$, i.e. the number of observations of the full information game. The regret in environment $\mathcal{E}_0$ is given by

$$\mathrm{Reg}_{\mathcal{E}_0}(T, \Pi_1) = \mathbb{E}_{\mathcal{E}_0}[\mathcal{N}] \frac{\Delta}{4} \leq \mathfrak{C}\sqrt{T} \leq \frac{c_1 \ln(k)}{\Delta}\,.$$

Hence

$$\mathbb{E}_{\mathcal{E}_0}[\mathcal{N}] \leq 4 \frac{c_1 \ln(k)}{\Delta^2} \leq 80 c_1 N\,.$$

Given any algorithm $\mathcal{A}$, we create a modified algorithm $\underline{\mathcal{A}}$ that uses a stopping time to commit to action 3 after it played N times action $\{1, 2\}$. By Markov inequality, the probability of $\mathcal{A}$ hitting the stopping time on $\mathcal{E}_0$ is bounded by $80c_1 \leq \frac{1}{2}$. Denote that event by E. By choice of our constants, the conditions for Lemma 2 are met, which implies

$$\mathbb{P}_{\mathcal{E}_{i^*}}(E) := \min_{i \in [k]} \mathbb{P}_{\mathcal{E}_i}(E) \leq \frac{1}{2} + \frac{1}{4} = \frac{3}{4}\,.$$

Hence

$$\mathrm{Reg}_{\mathcal{E}_{i^*}}(T, \Pi_2, \mathcal{A}) \geq \frac{1}{4} \mathrm{Reg}_{\mathcal{E}_{i^*}}(T, \Pi_2, \underline{\mathcal{A}}) \geq \frac{1}{4}(T - N)\frac{\Delta}{4} \geq \frac{\Delta T}{32} = \Omega\left(\min\left\{\frac{\ln(k)\sqrt{T}}{\mathfrak{C}}, T\right\}\right)\,.$$

□

Removing the properness requirement. It is not clear if non-proper algorithms might be able to circumvent the lower bound. We believe this is not the case for the following reason. In order to improve on the lower bound, an agent needs to run a hypothesis test between $\mathcal{E}_0$ and $(\mathcal{E}_i)_{i=1}^k$ without incurring too much regret in the case $\mathcal{E}_0$.Being improper means the agent can "cherry-pick" the content x_t for which to observe the losses of either arm 1 or 2. In an extreme case, the agent could commit to a single context $x \in \mathcal{X}$ for which to reveal the losses. It can be easily seen, that the sample complexity to rule out environment $\mathcal{E}_0$ reduces from $\Omega(\frac{\ln k}{\Delta^2})$ to $\Omega(\frac{1}{\Delta^2})$. Generally speaking, the hypothesis test becomes easier when the agent observes a significant amount of correlated contexts, i.e. $|\frac{\langle x_t, x_s \rangle}{k}| \gg 0$. We argue that this does not pose a problem in practice, since the contexts are drawn i.i.d. and we are interested in the regime where $k = \exp(T^\beta)$, for some $\beta > 0$. Hence $k \gg T$ and with high probability we have $\forall s, t \in [T] : |\frac{\langle x_t, x_s \rangle}{k}| \leq \epsilon_T$, where $\lim_{T \to \infty} \epsilon_T = 0$.

We therefore **conjecture** that the lower bound statement of Theorem 4 holds even for algorithms that are not *proper*. We do not pursue the proof of that conjecture since we believe the formal argument to be highly technical and the weaker version is sufficient to resolve the motivating open problem regarding second order bounds for full-information.

5 Implications

The relevance of open problem Eq. (1) has been motivated by its potential implications for other problems such as the S-switch bandit problem and an unresolved COLT2016 open problem on improved second order bounds for full-information. Our negative result for Eq. (1) indeed lead to the expected insights.

5.1 S-switch

Our lower bound in the *adaptive* regime shows that adapting to the number of switches is hopeless if the timing of the switch points is not independent of the players actions. Any algorithm adaptive to the number of switches in the regime with *oblivious* adversary must break in the *adaptive* case, which rules out bandit over bandit approaches based on importance sampling [Agarwal et al., 2017]. The successful algorithm proposed in Cheung et al. [2019] is using a bandit over bandit approach without importance sampling. Nonetheless, all components have adaptive adversarial guarantees. The algorithm splits the time horizon into equal intervals of length L. It initializes EXP3 with $\ln(T)$ arms, corresponding to a grid of learning rates. For each epoch, the EXP3 top algorithm samples an arm and starts a freshly initialized instance of EXP3.S using the learning rate corresponding to the selected arm. This instance is run over the full epoch of length L. It collects the accumulated losses $L_{sum} = \sum_{t=1}^{L} \ell_t$ of the algorithm and feeds the loss L_{sum}/L to the EXP3 top algorithm.

If all algorithms in the protocol enjoy guarantees against adaptive adversaries, why do bandit over bandit break against adaptive adversaries? Adaptive adversaries are assumed to pick the losses ℓ_t independent of the choice of arm A_t of the agent at round t. In the bandit over bandit protocol, the loss of the arm of the top algorithm dependents on the losses that the selected base suffers in the epoch. An adaptive adversary can adapt the losses in the epoch based on the actions of the base algorithm, that means the loss ℓ_t is not chosen independent of the action A_t. Hence the protocol is broken and the adaptive adversarial regret bounds do not hold.

5.2 Second order bounds for full information.

In an unresolved COLT2016 open problem, Freund [2016] asks if it is possible to ensure a regret bound of order

$$\mathrm{Reg}_\varepsilon = \tilde{\mathcal{O}}\left(\sqrt{\sum_{t=1}^{T} \mathbb{V}_{i\sim p_t}[\ell_{t,i}] \ln(\frac{1}{\varepsilon})}\right) , \tag{3}$$

against the best ε proportion of policies simultaneously for all ε. We go even a step further and show that the lower bound construction from Section 4.2 directly provides a negative answer for any $\alpha < 1$ to the looser bound

$$\mathrm{Reg}_\varepsilon = \tilde{\mathcal{O}}\left(\sqrt{\sum_{t=1}^{T} (\mathbb{V}_{i\sim p_t}[\ell_{t,i}] + T^\alpha) \ln(\frac{1}{\varepsilon})}\right) . \tag{4}$$

Theorem 5. *An algorithm satisfying Eq. (4) for $\alpha < 1$ implies the existence of a proper algorithm that violates the lower bound for the counter example in the proof of Theorem 4.*

Theorem 5 has the following interpretation. For any fixed $\alpha \in [0, 1)$, there is no algorithm which enjoys a regret upper bound as in Equation 3 for all problem instances s.t. $\sum_{t=1}^{T} (\mathbb{V}_{i\sim p_t}[\ell_{t,i}]) = \Theta(T^\alpha)$. This implies we can not hope for a polynomial improvement, in terms of time horizon, over the existing bound of $\tilde{O}(\sqrt{T \ln(1/\epsilon)})$. The detailed proof is found in Appendix B. The high level idea is to initialize the full-information algorithm satisfying Eq. (4) with a sufficient number of copies of the baseline policy π_0 and to feed importance weighted losses of the experts (i.e. policies) to that algorithm.

As we mention in Section 1, the case $\alpha = 1$ is obtainable. Our reduction relates the adaptation to variance to the model selection problem. As in Eq. (1), α is the trade-off between time and complexity. An algorithm satisfying Eq. (4) with $\alpha = 1$ merely allows to recover the trivial $\mathcal{O}(T)$ bound for model selection, and hence does not lead to a contradiction.

6 Conclusion

We derived the Pareto Frontier of minimax regret for model selection in Contextual bandits. Our results have resolved several open problems [Foster et al., 2020b, Freund, 2016].

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Our main regret upper bounds are found in Theorem 1 and Theorem 2. Our main lower bounds are found in Theorem 3 and Theorem 4 with Corollary 1 solving the COLT2020 open problem and Theorem 5 solving the COLT2016 open problem.
   (b) Did you describe the limitations of your work? [Yes] See paragraph regarding removing the properness requirement at the end of page 8.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This paper is theoretical in nature and we do not foresee any immediate societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] For complete proofs we refer the reader to the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, volume 11, pages 2312–2320, 2011.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. 2017.

Raman Arora, Teodor Vanislavov Marinov, and Mehryar Mohri. Corralling stochastic bandit algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 2116–2124. PMLR, 2021.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1), 2002.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, pages 138–158. PMLR, 2019.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27:199–207, 2014.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Esseen Carl-Gustav. On the liapunoff limit of error in the theory of probability. *Arkiv for matematik, astronomi och fysik, A: 1–19*, 1942.

Nicolo Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2):321–352, 2007.

Niladri Chatterji, Vidya Muthukumar, and Peter Bartlett. Osom: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 1844–1854. PMLR, 2020.

Kamalika Chaudhuri, Yoav Freund, and Daniel Hsu. A parameter-free hedging algorithm. *arXiv preprint arXiv:0903.2851*, 2009.

Alexey Chernov and Vladimir Vovk. Prediction with advice of unknown number of experts. *arXiv preprint arXiv:1006.0475*, 2010.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1079–1087. PMLR, 2019.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Ashok Cutkosky, Abhimanyu Das, and Manish Purohit. Upper confidence bounds for combining stochastic bandits. *arXiv preprint arXiv:2012.13115*, 2020.

Dylan J Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. *arXiv preprint arXiv:1906.00531*, 2019.

Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33, 2020a.

Dylan J Foster, Akshay Krishnamurthy, and Haipeng Luo. Open problem: Model selection for contextual bandits. In *Conference on Learning Theory*, pages 3842–3846. PMLR, 2020b.

Yoav Freund. Open problem: Second order regret bounds based on scaling time. In *Conference on Learning Theory*, pages 1651–1654. PMLR, 2016.

Avishek Ghosh, Abishek Sankararaman, and Ramchandran Kannan. Problem-complexity adaptive model selection for stochastic linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 1396–1404. PMLR, 2021.

Wouter M Koolen and Tim Van Erven. Second-order quantile methods for experts and combinatorial games. In *Conference on Learning Theory*, pages 1155–1175. PMLR, 2015.

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007.

Tor Lattimore. The pareto regret frontier for bandits. *arXiv preprint arXiv:1511.00048*, 2015.

Haipeng Luo and Robert E Schapire. Achieving all with no parameters: Adaptive normalhedge. *arXiv preprint arXiv:1502.05934*, 2015.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *Conference On Learning Theory*, pages 1739–1776. PMLR, 2018.

Aldo Pacchiano, Christoph Dann, Claudio Gentile, and Peter Bartlett. Regret bound balancing and elimination for model selection in bandits and rl. *arXiv preprint arXiv:2012.13045*, 2020a.

Aldo Pacchiano, My Phan, Yasin Abbasi-Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvari. Model selection in contextual stochastic bandit problems. *arXiv preprint arXiv:2003.01704*, 2020b.

Ilya Tyurin. New estimates of the convergence rate in the lyapunov theorem. *arXiv preprint arXiv:0912.0726*, 2009.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Tracking the best expert in non-stationary stochastic environments. *arXiv preprint arXiv:1712.00578*, 2017.

Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 746–755. PMLR, 2020.

Yinglun Zhu and Robert Nowak. Pareto optimal model selection in linear bandits. *arXiv preprint arXiv:2102.06593*, 2021.

Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021.