Adaptive Rates for Interactive Decision Making

Atul Ganju[1] and Karthik Sridharan[1]

[1]Cornell University

Working Paper, January 8, 2025

1 Introduction

Recently, progress has been made towards characterizing the worst-case learnability of interactive decision making. In particular, a recent statistical complexity measure which balances between exploration and exploitation has been shown to be both necessary and sufficient for sample-efficient interactive learning. However, it remains to obtain model-dependent bounds for interactive decision making problems. In this paper, we address this limitation by developing a framework for constructing algorithms that can achieve regret bounds.

Model-dependent bounds for interactive decision making have been a topic of interest with recent results obtaining small loss and variance-aware bounds for contextual bandits [2, 4].

NOTE: Need to complete this

1.1 Learning Framework

We consider the Contextual Decision Making with Structured Observations (C-DMSO) setting by [1]. The learning protocol proceeds in T rounds, where for each round $t = 1, \dots, T$:

1. Nature provides the learner with a context $x_t \in \mathcal{X}$, where $\mathcal{X}$ is the context space.
2. The learner selects a decision $\pi_t \in \Pi$, where Π is the decision space.
3. Nature selects a loss $l_t \in \mathcal{L}$ and observation $o_t \in \mathcal{O}$ based on the learner's decision, where $\mathcal{L} \subseteq \mathbb{R}$ is the loss space and $\mathcal{O}$ is the observation space.
4. The loss and observation are observed by the learner.

We focus on the following case of the C-DMSO framework with adversarially generated contexts and stochastically generated outcomes: On each round t, nature arbitrarily generates the context x_t and, after the learner selects decision π_t, the outcome (l_t, o_t) they observe is sampled independently from an unknown distribution $M^*(x_t, \pi_t)$, where $M^* : \mathcal{X} \times \Pi \to \Delta(\mathcal{L} \times \mathcal{O})$ is a mapping from context decision pairs to distributions over outcomes. To allow for learning through function approximation we assume the learner has access to a parametric model class which aims to capture the underlying distribution of outcomes. Specifically, we assume there is a parameter space Θ known to the learner which parameterizes a model class $\mathcal{M} = \mathcal{M}_\Theta = \{\mathcal{M}_\theta : \theta \in \Theta\}$ that aims to model M^*. Our work makes the following standard realizability assumption:

Assumption 1.1 (Realizability)**.** *There exists a $\theta^* \in \Theta$ such that $M_{\theta^*} = M^*$.*

Furthermore, for each model $M_\theta \in \mathcal{M}$, we denote $f_\theta(x, \pi) = \mathbb{E}_{(l,\cdot)\sim M_\theta(x,\pi)}[l]$ to be the mean loss observed under true model M_θ when nature selects context x and the learner responds with action π, and denote $\pi_\theta(x) = \operatorname{argmin}_{\pi\in\Pi} f_\theta(x, \pi)$ to be the decision-making policy that attains the lowest loss when outcomes are generated according to the model M_θ. Allowing the learner to pick a distribution $p_t \in \Delta(\Pi)$ over decisions, we measure the performance of the learner via the following regret formulation:

$$\mathbf{Reg} := \sum_{t=1}^{T} \mathbb{E}_{\pi_t \sim p_t} \left[f^*(x_t, \pi^*(x_t)) - f_{\theta^*}(x_t, \pi_t(x_t)) \right],$$

where we abbreviate f_{θ^*} as f^* and π_{θ^*} as π^*.

This general framework is able to capture many interactive learning problems from contextual bandits to episodic reinforcement learning NOTE: show this.

A uniform regret bound in this setting is achievable if there exists a randomized algorithm selecting p_t such that:

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{\pi_t \sim p_t} \left[f^*(x_t, \pi^*(x_t)) - f_{\theta^*}(x_t, \pi_t(x_t)) \right]\right] \le \mathcal{B}_T \qquad \forall\, x_{1:T},\ \forall\, M_\theta \in \mathcal{M},$$

where $a_{1:T} = \{a_1, \ldots, a_T\}$. Such bounds only depend on the complexity of the model class $\mathcal{M}$ and the time horizon T. The seminal work that introduced the C-DMSO learning framework also establishes the decision estimation coefficient (DEC), a statistical complexity measure that is both necessary and sufficient for the learnability of interactive learning problems within this framework [1].

An adaptive regret bound in this setting is achievable if there exists a randomized algorithm selecting p_t such that:

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{\pi_t \sim p_t} \left[f^*(x_t, \pi^*(x_t)) - f_{\theta^*}(x_t, \pi_t(x_t)) \right]\right] \le \mathcal{B}_T(M_\theta) \qquad \forall\, x_{1:T},\ \forall\, M_\theta \in \mathcal{M},$$

These bounds not only depend on the complexity of the model class $\mathcal{M}$ and the time horizon T, but also on the specific model M_θ which governs the observed outcomes.

2 Small Loss Bound:

In this section we show that our analysis recovers the small-loss bound for contextual bandits. We take inspiration from recent work [3]. First, we decompose regret in the same way as they do with the averaged DEC, an information game term, an underestimation error term, and an optimality gap term. Then we bound the underestimation error in terms of information gain and loss of our strategy in the same way as they do. What remains is the averaged DEC term and two terms that can be upper bounded efficiently by a mixable cost function which we have efficient algorithms for bounding.

$$\mathbb{E}[\mathbf{Reg}] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{\pi_t \sim p_t} \left[f_{\theta^*}(x_t, \pi_t(x_t)) - f^*(x_t, \pi^*(x_t)) \right]\right]$$

= NOTE: Convert to new notation, elaborate on mixability part of proof, try to bound NOTE: underestimation term in terms of adaptive rate to get proof that generalizes easier

$$\begin{aligned}
&= \mathbb{E}\left[\sum_{t=1}^{T} \overline{\mathrm{DEC}}_{\mu,t}(\pi_t) + \mu \cdot \overline{\mathrm{IG}}_t(\pi_t) + \mathrm{UE}_t + \mathrm{OG}_t\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T} \overline{\mathrm{DEC}}_{\mu,t}(\pi_t) + 4\mu \cdot \mathrm{IG}_t(\pi_t) + \mathrm{UE}_t + \mathrm{OG}_t\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T} \overline{\mathrm{DEC}}_{\mu,t}(\pi_t) + 4\mu \cdot \mathrm{IG}_t(\pi_t) + \left(2\gamma \cdot \mathbb{E}_{A_t \sim \pi_t}[\ell(\theta_0, A_t) | X_t, \mathcal{H}_{t-1}] + \frac{1}{\gamma} \cdot \mathrm{IG}_t(\pi_t)\right) + \mathrm{OG}_t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \overline{\mathrm{DEC}}_{\mu,t}(\pi_t) + \left(4\mu + \frac{1}{\gamma}\right) \cdot \mathrm{IG}_t(\pi_t) + \mathrm{OG}_t + 2\gamma \cdot \mathbb{E}_{A_t \sim \pi_t}[\ell(\theta_0, A_t) | X_t, \mathcal{H}_{t-1}]\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \frac{5}{4} \cdot \left(\overline{\mathrm{DEC}}_{\mu,t}(\pi_t) + \left(4\mu + \frac{1}{\gamma}\right) \cdot \mathrm{IG}_t(\pi_t) + \left(1 - \frac{\lambda\beta}{2}\right) \cdot \mathrm{OG}_t + 2\gamma \cdot \mathbb{E}_{A_t \sim \pi_t}[\ell(\theta_0, A_t)|X_t, \mathcal{H}_{t-1}]\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \frac{5}{4} \cdot \left(\overline{\mathrm{DEC}}_{1/10\lambda,t}(\pi_t) + \frac{1}{2\lambda} \cdot \mathrm{IG}_t(\pi_t) + (1 - \lambda) \cdot \mathrm{OG}_t + 20\lambda \cdot \mathbb{E}_{A_t \sim \pi_t}[\ell(\theta_0, A_t)|X_t, \mathcal{H}_{t-1}]\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \frac{5}{4} \cdot \left(\overline{\mathrm{DEC}}_{1/10\lambda,t}(\pi_t) + \frac{1}{2\lambda} \cdot \mathrm{IG}_t(\pi_t) + (1 - \lambda) \cdot \mathrm{OG}_t\right) + 25\lambda \cdot \mathbb{E}_{A_t \sim \pi_t}[\ell(\theta_0, A_t)|X_t, \mathcal{H}_{t-1}]\right]
\end{aligned}
$$

which implies,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} (1 - 25\lambda) \cdot \mathbb{E}_{A_t \sim \pi_t}[\ell_t(\theta_0, A_t)|X_t, \mathcal{H}_{t-1}] - \ell_t^*(\theta_0)\right] \\
&\quad \leq \mathbb{E}\left[\sum_{t=1}^{T} \frac{5}{4} \cdot \left(\overline{\mathrm{DEC}}_{1/10\lambda,t}(\pi_t) + \frac{1}{2\lambda} \cdot \mathrm{IG}_t(\pi_t) + (1 - \lambda) \cdot \mathrm{OG}_t\right)\right]
\end{aligned}
$$

Assuming we use the inverse gap weighting strategy, we focus in on the latter two terms in the summation.

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{2\lambda} \cdot \mathrm{IG}_t(\pi_t) + (1 - \lambda) \cdot \mathrm{OG}_t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{A_t \sim \pi_t}\left[\mathbb{E}_{\theta_t \sim Q_t}\left[\frac{1}{2\lambda} \cdot \mathcal{D}_{\mathrm{H}}^2(p_t(L_t|\theta_t, A_t), p_t(L_t|\theta_0, A_t)) + (1 - \lambda) \cdot (\ell_t^*(\theta_t) - \ell_t^*(\theta_0))\right]\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{A_t \sim \pi_t}\left[-\frac{1}{2\lambda} \cdot \log\left(\mathbb{E}_{\theta_t \sim Q_t}\left[\left(\frac{p_t(L_t|\theta_t, A_t)}{p_t(L_t|\theta_0, A_t)}\right)^{1/2}\right]\right) + (1 - \lambda) \cdot \mathbb{E}_{\theta_t \sim Q_t}\left[(\ell_t^*(\theta_t) - \ell_t^*(\theta_0))\right]\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{A_t \sim \pi_t}\left[- \frac{1}{2\lambda} \cdot \left(\log\left(\mathbb{E}_{\theta_t \sim Q_t}\left[(p_t(L_t|\theta_t, A_t))^{1/2}\right]\right) - \log\left((p_t(L_t|\theta_0, A_t))^{1/2}\right)\right)\right.\right. \\
&\qquad \left.\left. + (1 - \lambda) \cdot \mathbb{E}_{\theta_t \sim Q_t}\left[(\ell_t^*(\theta_t) - \ell_t^*(\theta_0))\right]\right]\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{A_t \sim \pi_t}\left[- \frac{1}{2\lambda} \cdot \left(\log\left(\mathbb{E}_{\theta_t \sim Q_t}\left[(p_t(L_t|\theta_t, A_t))^{1/2}\right]\right) - \log\left((p_t(L_t|\theta_0, A_t))^{1/2}\right)\right)\right.\right. \\
&\qquad \left.\left. + \frac{(1 - \lambda)}{\beta} \cdot \log\left(\mathbb{E}_{\theta_t \sim Q_t}\left[\exp\left\{\beta\left(\ell_t^*(\theta_t) - \ell_t^*(\theta_0)\right)\right\}\right]\right)\right]\right].
\end{aligned}
$$

Taking the cost function $c_t(Q) = \log((\mathbb{E}_{\theta_t \sim Q_t}[(p_t(L_t|\theta_t, A_t))^{1/2})^{-1/2\lambda} \cdot (\mathbb{E}_{\theta_t \sim Q_t}[\exp\{\beta l_t^*(\theta_t)\}])^{(1-\lambda)/\beta}])$. Since this loss function is λ-mixable when we set $\beta = -2\lambda(1 - \lambda)$ which we can do since the last equality is true for any β, we get the small loss bound.

References

[1] D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin. The statistical complexity of interactive decision making, 2023.

[2] D. J. Foster and A. Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination, 2021.

[3] G. Neu, M. Papini, and L. Schwartz. Optimistic information directed sampling, 2024.

[4] A. Pacchiano. Second order bounds for contextual bandits with function approximation, 2024.