Instance-Optimal Pure Exploration for Linear Bandits on Continuous Arms

Sho Takemori[1], Yuhei Umeda[1], Aditya Gopalan[2] 1: Fujitsu Limited, 2: Indian Institue of Science

Overview

- **Background:** The stochastic (linear) bandit problem for *continuous* arm sets $\mathcal{X}$ is well-studied on the cumulative regret minimization setting. However, existing research in the *pure exploration* setting is sparse.
- **Objective:** Efficiently compute an asymptotically optimal arm sampling distribution $(\pi_t)_{t\geq 1}$ on the arm set $\mathcal{X}$.
- **Challenge 1:** Such a sampling distribution involves optimization over the space $\mathcal{P}(\mathcal{X})$ of probability measures on $\mathcal{X}$, which can be *infinite dimensional*
- **Challenge 2:** And the objective function is *non-smooth*. Simply applying existing methods for the finite-armed setting via discretization would be computationally expensive.
- **Contribution:** Assuming computation oracles for quadratic and fractional quadratic objectives on the arm set, we propose a tractable algorithm (in terms of the number of oracle calls) that achieves an asymptotically optimal sampling distribution.

Problem Formulation

- We consider the ϵ-BAI (best arm identification) problem with Bayesian reward setting on a compact arm set $\mathcal{X} \subset \mathbb{R}^d$.
- **Reward function:** Reward function $f : \mathcal{X} \to \mathbb{R}$ with $f(x) = \theta_f \cdot x, \theta_f \sim \mathcal{N}(0_d, 1_d)$.
- **Sampling Rule:** For each round $t = 1, \dots,$ a learner selects an arm $x_t \sim \pi_t$, and observes a random reward $y_t = f(x_t) + \omega_t$, where $\omega_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \lambda^2)$. We call $(\pi_t)_{t\geq 1}$ a *sampling rule*.
- **Posterior probability:** Based on observations up to t, the posterior mean $\mu_t : \mathcal{X} \to \mathbb{R}$ and covariance matrix Σ_t are defined (here, $V_t := \lambda 1_d + \sum_{s=1}^t x_s x_s^\top$). P_t: the posterior probability measure conditioned on $\mathcal{F}_t$ (conditioned on $\mathcal{F}_t$, $f_t(x) = \theta_t \cdot x$ with $\theta_t \sim \mathcal{N}(\mu_t, \Sigma_t^{-1})$).
- **Recommendation Rule:** ζ_t: an estimation of an ϵ-optimal arm at round t. Formally, $(\zeta_t)_{t\geq 1}$: a sequence of $\mathcal{F}_t$-meas. $\mathcal{X}$-valued R.V.
- **Objective:** THe objective of the learner to miniize the posterior probability $P_t(\zeta_t \notin \mathcal{X}^*(\epsilon))$ of misidentification, where $\mathcal{X}^*(\epsilon) := \{x \in \mathcal{X} : f(x) > \sup_{\xi \in \mathcal{X}} f(\xi) - \epsilon\}$.

Asymptotic analysis of posterior probability

***Lemma* 1.** *Assume* $\lim_{t\to\infty} \zeta_t$ *converges to* $\zeta_\infty \in \mathcal{X}^*(\epsilon)$ *a.s., and* $\lim_{t\to} \mu_t(x) = f(x)$ *a.s. for any* $x \in \mathcal{X}$*. Suppose* $\inf_{t\geq 1} \lambda_{\min}(V(\overline{\pi}_t)) > 0$, *where* $\overline{\pi}_t := \frac{1}{t}\sum_{s=1}^t \pi_s$*. Then,*

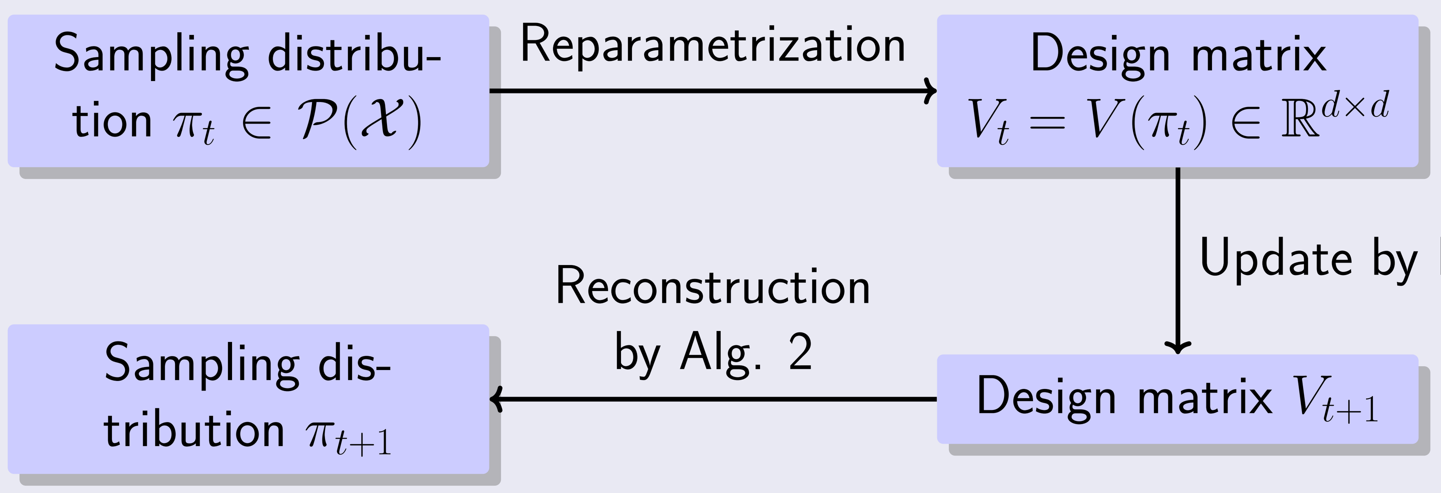
$$-\frac{1}{2}\limsup_{t\to\infty} \left(\Gamma^*(V(\overline{\pi}_t); \zeta_\infty, f)\right)^{-1} \leq \liminf_{t\to\infty} \frac{1}{t} \log P_t\left(\zeta_t \notin \mathcal{X}^*(\epsilon)\right)$$
$$\leq \limsup_{t\to\infty} \frac{1}{t}\log P_t\left(\zeta_t \notin \mathcal{X}^*(\epsilon)\right) \leq -\frac{1}{2}\liminf_{t\to\infty} \left(\Gamma^*(V(\overline{\pi}_t); \zeta_\infty, f)\right)^{-1}.$$

- Here, for $V \in \mathbb{R}^{d\times d}$, $\zeta \in \mathcal{X}$, and a function $\mu : \mathcal{X} \to \mathbb{R}$, we define
$$\Gamma^*(V; \zeta, \mu) := \sup_{\xi \in \mathcal{X}} \frac{\|\zeta - \xi\|^2_{V^{-1}}}{(\epsilon + \mu(\zeta) - \mu(\xi))^2}.$$
- Intuitively, this lemma implies that the posterior probability $P_t\left(\zeta_t \notin \mathcal{X}^*(\epsilon)\right)$ exponentially decays as t increases, and its decay rate is given as $\lim_{t\to\infty} \left(\Gamma^*(V(\overline{\pi}_t); \zeta_\infty, f)\right)^{-1}$.

Optimization Objective

- Lemma 1 indicates an asymptotically optimal sampling policy gives a solution to the following optimization objective:
$$\tau^*_{\mathcal{X}}(f; \zeta_\infty) := \inf_{\pi \in \mathcal{P}(\mathcal{X})} \sup_{\xi \in \mathcal{X}} \frac{\|\zeta_\infty - \xi\|^2_{V(\pi)^{-1}}}{(\epsilon + f(\zeta_\infty) - f(\xi))^2}.$$
- This is an optimization problem over the space of probability measure $\mathcal{P}(\mathcal{X})$, which can be infinite-dimensional in our setting.
- Due to the inner supremum, the objective can be non-smooth.

Proposed Method



Pseudo Code (simplified)

Algorithm 1 Main Algorithm

1: Initialize: $\pi_1 = \pi_{\text{exp}}$.
2: **for** $t = 1, 2, \dots,$ **do**
3: Play $x_t \sim \pi_t$ and observe a noisy reward y_t
4: $V_t = V(\pi_t)$ {Reparametrization $\pi_t \mapsto V_t$.}
5: **// Computation of a subgradient $g_t \in \mathbb{R}^{d\times d}$ of the objective function at V_t.**
6: **// Update in the matrix space.**
7: $W_{t+1} = V_t - \eta_t g_t$.
8: **// Approx. projection and distribution-reconstruction.**
9: $\pi_{t+1} \leftarrow$ Algorithm 2 with $W = W_{t+1}, n = n_t$
10: **end for**

Algorithm 2 Approximate Projection by the Frank-Wolfe Algorithm

Input: $W \in \mathbb{R}^{d\times d}$, $n \geq 1$, $\widetilde{\pi}^{(0)} \in \mathcal{P}(\mathcal{X})$, $(\gamma_i)_{i\geq 1}$
for $i = 1, 2, \dots, n$ **do**
$a_i = \operatorname{argmax}_{x\in\mathcal{X}} x^\top (W - V(\widetilde{\pi}^{(i-1)}))x$
$\widetilde{\pi}^{(i)} = (1 - \gamma_i)\widetilde{\pi}^{(i-1)} + \gamma_i \delta(a_i)$
end for
Output $\widetilde{\pi}^{(n)}$

Main Theoretical Result

***Theorem* 1** (informal)**.** $(\pi_t)_{t\geq 1}$ *be the sampling rule of 1, and* $(\zeta_t)_{t\geq 1}$ *be a recommendation rule with* $\zeta_\infty = \lim_{t\to\infty} \zeta_t$ *a.s. Under some assumptions (e.g.,* $\|\zeta_t - \zeta_\infty\| = O(t^{-\nu})$ *with* $\nu > 0$*), the following holds:*

$$\lim_{t\to\infty} \frac{1}{t} \log P_t(\zeta_t \notin \mathcal{X}^*(\epsilon)) = -\frac{1}{2\tau^*_{\mathcal{X}}(f; \zeta_\infty)}.$$

Experiments

- **Setting:** $\mathcal{X} = \{(\cos(\theta), \sin(\theta)) : \theta \in [0, \theta_1]\} \subset \mathbb{R}^2$, and $f(x) = (\cos(\theta_f), \sin(\theta_f)) \cdot x$ with $\theta_f = a\pi, \theta_0 = 0, \theta_1 = b\pi$.
- **Evaluation metric:** (an upper bound of) $P_t(\zeta_t \notin \mathcal{X}^*(\epsilon))$ denoted by p_t
- **Baselines:** Uniformly random (Uniform) and MVR [Vakili et al., 2021]