Instance-Optimal PAC Algorithms for Contextual bandits

Zhaoqi Li*, Lillian Ratliff*, Houssam Nassif**, Kevin Jamieson*, Lalit Jain*

* University of Washington, ** Amazon

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

What is the best policy that gives personalized recommendations to different users in an experiment?

Problem Statement

- At each time $t = 1,2,\cdots$:
  - $c_t \sim \nu \in \triangle_C$ arrives, action $a_t \in A$ from $p_{c_t} \in \triangle_A$
  - Receive reward r_t, $\mathbb{E}[r_t | c_t, a_t] = r(c_t, a_t) \in \mathbb{R}$
- Learn $\pi_* := \arg\max_{\pi\in\Pi} V(\pi) := \arg\max_{\pi\in\Pi} \mathbb{E}_{c\sim\nu}[r(c,\pi(c))]$
- Allow context space C and policy class Π to be infinite

Goal: an *instance-optimal* and *computationally efficient* algorithm for (ϵ,δ)—PAC learning that hits the *lower bound*

Related Work

Method	Sample Complexity	Policy Classes
EXP4/ILTCB	$\frac{\lvert\mathcal{A}\rvert\log(\lvert\Pi\rvert/\delta)}{\epsilon^2}$	Agnostic
AdaCB [1]	$\frac{\lvert\mathcal{A}\rvert\log(\lvert\Pi\rvert)}{\epsilon\ \Delta_{\min}}\mathfrak{C}_\Pi^{\text{pol}}$	Agnostic
LinUCB/LinTS	$\frac{d^2}{\epsilon\ \Delta_{\min}}$	Linear Realizable
Reward-free LinUCB [2]	$\frac{d^2}{\epsilon^2}\log(1/\delta)$	Linear Realizable
This work	$\rho_{\Pi,0}\log(\lvert\Pi\rvert/\delta)$	Linear Realizable

Table: Known sample complexity results

low-regret algorithms are inefficient!

Reduction to Linear Realizability

- $\exists \phi : C \times A \to \mathbb{R}^d$ such that $r(c,a) = \langle \phi(c,a), \theta^* \rangle$ for $\theta^* \in \Theta \subset \mathbb{R}^d$
- Let $\phi_\pi := \mathbb{E}_{c\sim\nu}[\phi(c,\pi(c))] \Rightarrow V(\pi) = \langle \phi_\pi, \theta_* \rangle$
- Agnostic: $\theta^* \in \mathbb{R}^{|C|\times|A|}$, $[\theta^*]_{c,a} = r(c,a) \Rightarrow r(c,a) = \langle \mathrm{vec}(\mathbf{e}_c\mathbf{e}_a^\top), \theta^* \rangle$ $\quad \phi(c,a)$

Sample Complexity Lower Bound

Theorem [Li et al. 2022] Let τ be the stopping time of the algorithm. Any $(0,\delta)$-PAC algorithm satisfies $\mathbb{E}[\tau] \geq \rho_{\Pi,0}\log(1/2.4\delta)$ where

$$\rho_{\Pi,\epsilon} := \min_{p_c\in\triangle_A,\ \forall c\in C}\ \max_{\pi\in\Pi\setminus\pi_*} \frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{\boxed{\mathbb{E}_{c\sim\nu}[\sum_{a\in A} p_{c,a}\phi(c,a)\phi(c,a)^\top]}^{-1}}}{(\langle \phi_{\pi_*} - \phi_\pi, \theta_* \rangle \vee \epsilon)^2}.$$

$A(p)$ variance

gap

Algorithm

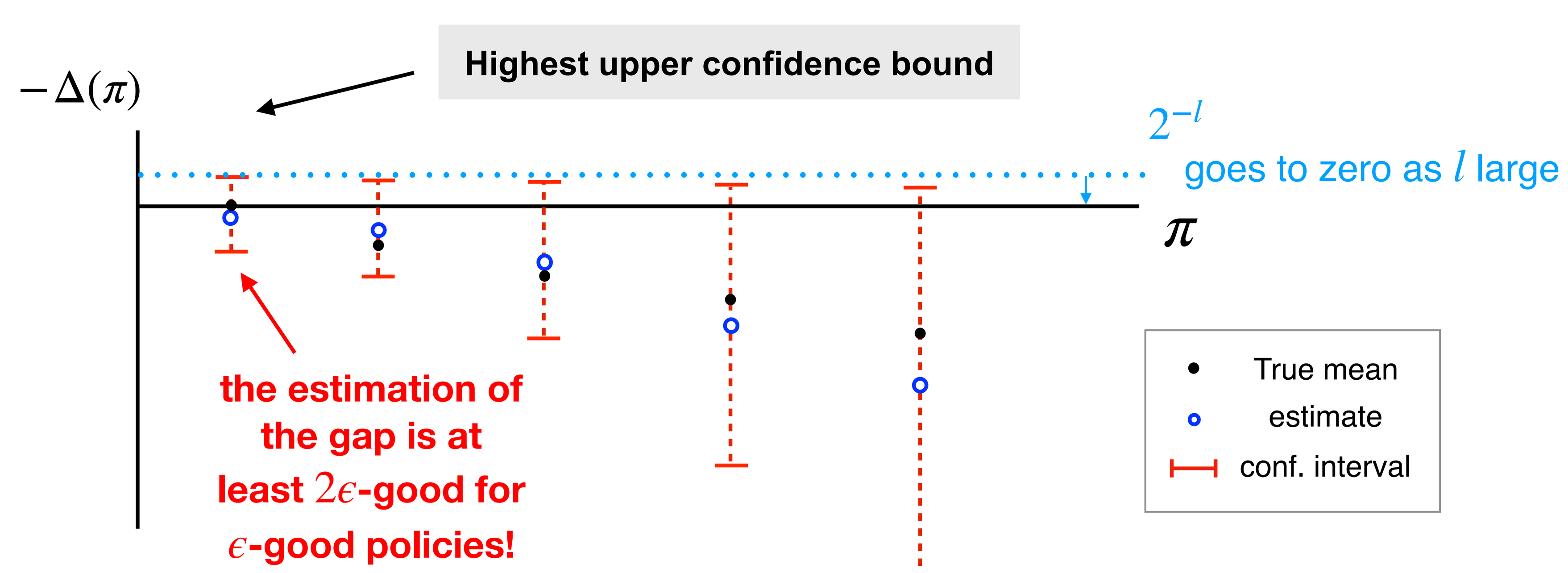
Define the gap $\Delta(\pi,\pi') := \langle \phi_{\pi'} - \phi_\pi, \theta_* \rangle$. In round l, given $\{(c_s,a_s,r_s)\}_{s=1}^n$, $\hat{\Delta}_l^{IPW}(\pi,\pi') := \frac{1}{n}A(p^{(l)})^{-1}\sum_{s=1}^n \phi(c_s,a_s)r_s$

Input: Π

Initialize $\Pi_1 = \Pi$, estimate $\hat{\pi}_0$

for $l = 1,2,\cdots$

1. Choose $p_c^{(l)} \in \triangle_A, \forall c \in C$ and n_l such that
$$\min_{p_c\in\triangle_A,\forall c\in C}\ \max_{\pi\in\Pi}\left(-\hat{\Delta}_l(\pi,\hat{\pi}_{l-1}) + \sqrt{\frac{\|\phi_{\hat{\pi}_{l-1}} - \phi_\pi\|^2_{A(p)^{-1}}\log(2l^2|\Pi|/\delta)}{n_l}}\right) \leq 2^{-l}$$
2. For $t \in [n_l]$, for each context c_t, sampling $a_t \sim p_{c_t}^{(l)}$ and compute IPW estimate $\hat{\Delta}_l(\pi,\hat{\pi}_{l-1})$ for each $\pi \in \Pi$
3. Update
$$\hat{\pi}_l = \arg\min_{\pi\in\Pi} \hat{\Delta}_l(\pi,\hat{\pi}_{l-1})$$



Computationally Efficient Algorithm and Upper Bound

Definition (argmax oracle). Given contexts and cost vectors $(c_1,v_1),\cdots,(c_n,v_n) \in C \times \mathbb{R}^{|A|}$, it returns $\arg\max_{\pi\in\Pi}\sum_{t=1}^n v_t(\pi(c_t))$.

Theorem [Li et al. 2022] The algorithm returns an (ϵ,δ)-PAC policy with at most $O(\rho_{\Pi,\epsilon}\log(|\Pi|/\delta)\log_2(1/\epsilon))$ samples and $\mathrm{poly}(|A|,\epsilon^{-1},\log(1/\delta),\log(|\Pi|))$ calls to argmax oracle.

We start with the primal problem, which is the design itself:

$$\min_{p_c\in\triangle_A,\forall c\in C}\ \max_{\pi\in\Pi} -\hat{\Delta}_l(\pi,\hat{\pi}_{l-1}) + \boxed{\sqrt{\frac{\|\phi_{\hat{\pi}_{l-1}} - \phi_\pi\|^2_{A(p)^{-1}}\log(1/\delta)}{n}}}$$

convex in $p_c, \forall c \in C$ $\Downarrow$ strong duality holds!

$$= \min_{p_c\in\triangle_A,\forall c\in C}\ \max_{\pi\in\Pi}\ \min_{\gamma_\pi\geq 0} -\hat{\Delta}_l(\pi,\hat{\pi}_{l-1}) + \gamma_\pi\|\phi_{\hat{\pi}_{l-1}} - \phi_\pi\|^2_{A(p)^{-1}} + \frac{\log(1/\delta)}{\gamma_\pi n}$$

The dual problem is:

agnostic setting $\Rightarrow$ analytical solution!

$$\max_{\lambda\in\triangle_\Pi}\ \min_{\gamma_\pi\geq 0}\ \boxed{\min_{p_c\in\triangle_A,\forall c\in C}\sum_{\pi\in\Pi}\lambda_\pi\Big(-\hat{\Delta}_l(\pi,\hat{\pi}_{l-1}) + \gamma_\pi\|\phi_{\hat{\pi}_{l-1}} - \phi_\pi\|^2_{A(p)^{-1}}} + \frac{\log(1/\delta)}{\gamma_\pi n}\Big)$$

$$= \max_{\lambda\in\triangle_\Pi}\ \min_{\gamma}\sum_{\pi\in\Pi}\lambda_\pi\Big(-\hat{\Delta}_l(\pi,\hat{\pi}_{l-1}) + \frac{\log(1/\delta)}{\gamma_\pi n}\Big) + \mathbb{E}_{c\sim\nu}\Big[\Big(\sum_{a\in A}\sqrt{(\lambda\odot\gamma)^\top t_a^{(c)}}\Big)^2\Big] =: \max_{\lambda\in\triangle_\Pi}\ \min_{\gamma} h_l(\lambda,\gamma)$$

where $t_a^{(c)} = \mathbf{1}\{\pi(c) = a\} + \mathbf{1}\{\hat{\pi}_{l-1}(c) = a\} - 2\mathbf{1}\{\pi(c) = \hat{\pi}_{l-1}(c)\}$.

concave in λ and locally strongly convex in γ $\Rightarrow$ can solve the saddle point problem!

To get a sparse solution of λ, we use the **Frank-Wolfe** subroutine. In each step t of Frank-Wolfe, we compute

$$\pi_t = \arg\max_{s\in\triangle_\Pi} s^\top\nabla_\lambda h_l(\lambda^t,\gamma^t) = \arg\max_{\pi\in\Pi}\left[\nabla_\lambda h_l(\lambda^t,\gamma^t)\right]_\pi$$

which could be computed using an **argmax** oracle.

Reference

[1] Dylan J. Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104* (2020).

[2] Andrea Zanette, Kefan Dong, Jonathan Lee, and Emma Brunskill. Design of experiments for stochastic contextual linear bandits. *Advances in Neural Information Processing Systems*, 34, 2021.