Towards a Sharp Analysis of Offline Policy Learning for f-Divergence-Regularized Contextual Bandits

Qingyue Zhao[*1], **Kaixuan Ji**[*1], **Heyang Zhao**[*1], **Tong Zhang**[2] and **Quanquan Gu**[1]

[1]University of California, Los Angeles [2]University of Illinois Urbana-Champaign

Data Model: Offline Contextual Bandits

- $(s, a) \in \mathcal{S} \times \mathcal{A}$: input-output (e.g., prompt-response)
- $\rho \in \Delta(\mathcal{S})$: input distribution
- $\pi^{\mathrm{ref}} \in \Delta(\mathcal{A}|\mathcal{S})$: reference model
- $\mathcal{G} \subset (\mathcal{S} \times \mathcal{A} \to [0, 1])$: known function class
  - $\forall \epsilon_{\mathrm{c}} > 0, \mathcal{N}_{\mathcal{G}}(\epsilon_{\mathrm{c}}) = \mathrm{poly}(\epsilon_{\mathrm{c}}^{-1})$: low complexity (i.e., nice covering number)
- $r \in \mathcal{G}$: **realizable** mean reward function
- $\mathcal{D} \coloneqq \{(s_i, a_i, r_i)\}_{i=1}^n$: i.i.d. prompt-response-reward dataset
  - $s_i \sim \rho$, $a_i \sim \pi^{\mathrm{ref}}(\cdot|s_i)$
  - $r_i = r(s_i, a_i) + \varepsilon_i$: reward with subgaussian noise
    - *See also the pairwise-comparison (dueling) counterpart in our paper!*

Targets of Interest

- **Objective** $J(\pi) \coloneqq \mathbb{E}_{s\sim\rho}\Big[\langle r(s, \cdot), \pi(\cdot|s)\rangle - \eta^{-1} D_f\big(\pi(\cdot|s) \| \pi^{\mathrm{ref}}(\cdot|s)\big)\Big]$
  - D_f: f-divergence
    - Different f's induce different realizations, including TV, KL, ...
  - η^{-1}: regularization intensity (also known as the " temperature")
- **Optimal Policy** $\pi^* \leftarrow \mathrm{argmax}_\pi J(\pi)$
  - e.g., $\pi^*(\cdot|s) \propto \pi^{\mathrm{ref}}(\cdot|s) \exp\big(\eta \cdot r(s, a)\big)$ for $D_f = \mathsf{KL}$
- **Learners** We consider learners with general function approximation. Formally, the agent knows $\mathcal{D}$, $\mathcal{G}$, and π^{ref}
  - The agent may optionally know other hyperparamters like η
  - The agent is not allowed to further sample from ρ, π^{ref}, and $\rho \times \pi^{\mathrm{ref}}$
- **Performance Metric** For any learner $\widehat{\pi}$, we are interested in the sample complexity n for it to achieve $J(\pi^*) - J(\widehat{\pi}) \le \epsilon$
  - Results regarding $\mathrm{argmax}_{a\in\mathcal{A}} r(s, a) - \langle r(s, \cdot), \widehat{\pi}(\cdot|s)\rangle \le \epsilon$ are well-established, all of which are $\asymp \epsilon^{-2}$, intuitively because the statistical rate for reward estimation is already needs at least $\gtrsim n^{-1/2}$
  - The convergence against regularized objectives themselves are less-explored even in the i.i.d. setting

Motivation

What is the weakest data **coverage** *condition for post-training to be* **statistically** *optimal against f-divergence-***regularized objectives***?*

- The answer depends on the "curvature" of the regularizing divergence
- Scope: offline setting with absolute or dueling feedback

Key Results

Regularizer		2312.11456	2411.04625	Ours
KL	Upper	$d\epsilon^{-2}$	$D_{\mathrm{sup}}^2 \eta\epsilon^{-1} \log \mathcal{N}$	$D_{\pi^*}^2 \eta\epsilon^{-1} \log \mathcal{N}$
	Lower	-	$\eta\epsilon^{-1} \log \mathcal{N}$	$C^{\pi^*} \eta\epsilon^{-1} \log \mathcal{N}$
f-divergence w/ α-strongly convex f	Upper	-	-	$\alpha^{-1}\eta\epsilon^{-1} \log \mathcal{N}$
	Lower	-	-	$\alpha^{-1}\eta\epsilon^{-1} \log \mathcal{N}$

$$D_{\mathcal{G}}^2((s, a); \pi) = \sup_{g,h\in\mathcal{G}} \frac{\big(g(s, a) - h(s, a)\big)^2}{\mathbb{E}_{(s',a')\sim\rho\times\pi}[(g(s', a') - h(s', a'))^2]}$$

- $D_{\mathcal{G}}^2$-related notions are offline counterparts of **Eluder** dimension
- $D_{\mathrm{sup}}^2 = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} D_{\mathcal{G}}^2((s, a); \pi^{\mathrm{ref}})$: D^2-based **all**-policy concentrability
- $D_{\pi^*}^2 = \mathbb{E}_{(s,a)\sim\rho\times\pi^*}[\, D_{\mathcal{G}}^2((s, a); \pi^{\mathrm{ref}})\,]$: D^2-based **single**-policy concentrability
- $D_{\pi^*}^2 \le D_{\mathrm{sup}}^2$ for sure
  - D_{sup}^2 can be **much larger** than $D_{\pi^*}^2$ in non-trivial cases
- $C^{\pi^*} = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\pi^*(a|s)}{\pi^{\mathrm{ref}}(a|s)}$: density-ratio-based **single**-policy concentrability
  - **Limitation**: the worst-case notion C^{π^*} may not match the instance-dependent notion $D_{\pi^*}^2$ in general (sometimes $D_{\pi^*}^2/C^{\pi^*} \approx SA$)
  - Resolution to this limitation? Check our follow-up refinement on online MABs at `https://arxiv.org/pdf/2603.02155`
- Subtlety for low temperature: as $\eta \to \infty$, regularization **vanishes**
  - In the rigorous hardness results for KL, $\eta\epsilon^{-1}$ is replaced by $\eta\epsilon^{-1} \wedge \epsilon^{-2}$

Reminder: Definition and Realizations of f-divergence

- $D_f(P\|Q) \coloneqq \sum_x Q(x) f\big(P(x)/Q(x)\big)$
  - Requirement: convex f with $f(1) = 0$
  - $f(0) \coloneqq f(0+)$, $0f(0/0) \coloneqq 0$, and $0f(1/0) = f'(+\infty)$

Realizations	$\mathsf{TV}(P\|Q)$	$\mathsf{KL}(P\|Q)$	$\chi^2(P\|Q)$
f	$\lvert x-1\rvert/2$	$x \log x$	$(x-1)^2/2$
Convexity of f	convex	**strictly** convex	**strongly** convex

- However, $\pi \mapsto \mathsf{KL}(\pi\|\pi^{\mathrm{ref}})$ is $\Theta(1)$-**strongly** convex w.r.t. the TV distance

Algorithm Template for $D_f = \mathsf{KL}$: No tricky algorithmic tweaks

Require: regularization η, reference policy π^{ref}, function class $\mathcal{G}$, dataset $\mathcal{D}$

1: Reward learning via ordinary least square

$$\bar{g} \in \mathop{\mathrm{argmin}}_{g\in\mathcal{G}} \sum_{(s_i,a_i,r_i)\in\mathcal{D}} \big(g(s_i, a_i) - r_i\big)^2$$

2: Let $\widehat{g} \leftarrow \bar{g} - \Gamma_n$, where $\Gamma_n : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ is the pessimism term

Ensure: $\widehat{\pi}(a|s) \propto \pi^{\mathrm{ref}}(a|s) \exp\big(\eta \cdot \widehat{g}(s, a)\big)$

Algorithm Template for D_f w/ Strongly Convex f: No pessimism!

1: Reward learning via ordinary least square

$$\bar{g} \in \mathop{\mathrm{argmin}}_{g\in\mathcal{G}} \sum_{(s_i,a_i,r_i)\in\mathcal{D}} \big(g(s_i, a_i) - r_i\big)^2$$

2: Compute the optimal policy under $\bar{g}$ for $s \in \mathcal{S}$ as

$$\widehat{\pi}(\cdot|s) \leftarrow \mathop{\mathrm{argmax}}_{\pi(\cdot|s)\in\Delta(\mathcal{A})} \langle \pi(\cdot|s), \bar{g}(s, \cdot)\rangle + \eta^{-1} D_f\big(\pi(\cdot|s) \| \pi^{\mathrm{ref}}(\cdot|s)\big)$$

Ensure: $\widehat{\pi}$

D_f w/ s.c. f provides stronger regularization than KL, $\widehat{\pi}$ won't deviate from π^{ref} too much $\Longrightarrow$ no pess. needed

(Comments are welcome!) See also our worse-case refinement for the tabular setting *Near-Optimal Regret for KL-Regularized Multi-Armed Bandits* (arXiv:2603.02155) @ the SPOT workshop on Monday