Fast Rates for Offline Contextual Bandits with Forward-KL Regularization under Single-Policy Concentrability

Qingyue Zhao* Kaixuan Ji* Heyang Zhao Quanquan Gu

Department of Computer Science, University of California, Los Angeles

NeurIPS 2026 Spotlight Paper: arXiv:2605.09214

Poster Session: Fri, Dec 11, 2026 • 8:30 AM – 11:30 AM AEDT • **Hall C1**

Motivation: Reverse KL vs. Forward KL

Offline Policy Learning with KL Regularization: $J(\pi) := \mathbb{E}[r] - \eta^{-1}\mathrm{KL}$

- **Reverse KL** $\mathrm{KL}\left(\pi \| \pi^{\mathsf{ref}}\right)$ (RLHF): globally strongly convex, giving **fast** $\widetilde{\Theta}(1/\epsilon)$ **rates** when data covers **only** π^* (Tiapkin et al., 2023; Zhao et al., 2026; Foster et al., 2025).
- **Forward KL** $\mathrm{KL}\left(\pi^{\mathsf{ref}} \| \pi\right)$: mass-covering alternative widely used in f-DPO, diffusion alignment, and RLHF (Wang et al., 2023; Shan et al., 2025; Tang & Munos, 2025).
- **Apparent barrier**: $-\log x$ loses global strong convexity, and prior offline analysis gives only **slow** $\widetilde{O}(1/\epsilon^2)$ **rates** under local-neighborhood coverage (Aminian et al., 2025).

Central Question

Does flipping Reverse KL to **Forward KL** inherently slow down offline learning—or can we still achieve **fast** $\widetilde{O}(1/\epsilon)$ **sample efficiency** when logged data covers **only the optimal policy**? **Yes!**

Problem Setup: Forward-KL Offline Contextual Bandits

Model, Rational Optimal Policy, and Coverage

- **Offline** i.i.d. **data**: $\mathcal{D} = \{(s_i, a_i, r_i)\}_{i=1}^n$ with $s_i \sim \rho$, $a_i \sim \pi^{\mathsf{ref}}(\cdot|s_i)$, and $r_i = r(s_i, a_i) + \varepsilon_i$.
- **Objective ($\eta > 0$) & unique rational optimizer**:

$$J_{\mathrm{FKL}}(\pi; r) := \langle r, \pi \rangle - \eta^{-1} \mathsf{KL}\left(\pi^{\mathsf{ref}} \| \pi\right), \quad \pi^*_{\mathrm{FKL}}(a|s) = \frac{\eta^{-1}\pi^{\mathsf{ref}}(a|s)}{\lambda_s - r(s,a)}.$$

- **Target suboptimality**:

$$\mathrm{SubOpt}_{\mathrm{FKL}}(\widehat{\pi}) := J_{\mathrm{FKL}}(\pi^*_{\mathrm{FKL}}; r) - J_{\mathrm{FKL}}(\widehat{\pi}; r) \le \epsilon.$$

- **Single-policy concentrability**:

$$C^{\pi^*} := \sup_{s \in \mathcal{S},\, a \in \mathcal{A}} \frac{\pi^*_{\mathrm{FKL}}(a|s)}{\pi^{\mathsf{ref}}(a|s)}.$$

Main Results & Pessimistic Algorithm (FKL-PCB)

Type	Reference	Setting	Sample Complexity	Coverage
Upper Bound	Aminian et al. (2025)	Function Approx.	$\widetilde{O}(C_{\varepsilon_{\mathrm{FKL}}} \log N_{\mathcal{G}} \,/\, \epsilon^2)$	Local FKL-ball ($C_{\varepsilon_{\mathrm{FKL}}}$)
Upper Bound	**FKL-PCB (This Work)**	**Tabular**	$\widetilde{O}(\eta (C^{\pi^*})^2 \lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert \,/\, \epsilon)$	**Single-policy (C^{π^*})**
Upper Bound	**FKL-PCB (This Work)**	**Function Approx.**	$\widetilde{O}(\eta C^{\pi^*} D^2_{\pi^*} \log N_{\mathcal{G}}(\epsilon) \,/\, \epsilon)$	**Single-policy ($C^{\pi^*}, D^2_{\pi^*}$)**
Lower Bound	**This Work**	**Tabular**	$\Omega(\eta \lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert \,/\, \epsilon)$	Uniform π^{ref} ($C^{\pi^*} = \Theta(1)$)

Unified Algorithm: FKL-PCB (Tabular & Function Approx. $\mathcal{G}$)

1. Build a clipped pessimistic reward $\widehat{g} \in [0, 1]$ from $\bar{g}$ satisfying $0 \le r - \widehat{g} \le 2\,\mathrm{Bonus}$ w.h.p.
2. Output the exact regularized maximizer under $\widehat{g}$:

$$\widehat{\pi} = \operatorname*{argmax}_{\pi \in \Delta(\mathcal{A}|\mathcal{S})} J_{\mathrm{FKL}}(\pi; \widehat{g}) \iff \widehat{\pi}(a|s) = \frac{\eta^{-1} \pi^{\mathsf{ref}}(a|s)}{\widehat{\lambda}_s - \widehat{g}(s,a)}.$$

Technical Obstacles: Why Existing Fast-Rate Proofs Fail

1. Lack of Uniform Strong Convexity on $\Delta(\mathcal{A})$

$\nabla^2_\pi \mathsf{KL}\left(\pi^{\mathsf{ref}} \| \pi\right) = \mathrm{diag}(\pi^{\mathsf{ref}}/\pi^2) \implies \Theta\big(\min_a \pi^{\mathsf{ref}}(a)\big)$-strongly convex,

so L^1 strong-convexity analyses pay $(\min_a \pi^{\mathsf{ref}}(a))^{-1}$ (**all-policy coverage**).

2. Failure of Reverse-KL Midpoint Monotonicity

Interpolating $r_u = (1-u)r + u\widehat{r}$ bounds suboptimality by a midpoint term $F(\bar{u})$ at $\pi_{\bar{u}}$. Under pessimism ($g := r - \widehat{r} \geq 0$), reverse KL satisfies $F'(u) \leq 0$ ($\pi_{\bar{u}} \to \pi^*$), whereas forward KL yields an **indefinite tilted skewness**:

$$F'(u) = -\eta^2 Z_u \cdot \mathbb{E}_{a\sim\mu_u}\Big[w_u(a)\big(g(a) - \mathbb{E}_{\mu_u} g\big)^3\Big] \quad \text{(indefinite for } g \geq 0\text{)}.$$

Key Technique: Simplex Conjugacy & Self-Bounding

Bypassing Midpoints on $\Delta(\mathcal{A})$ ($h(\pi) := \eta^{-1}\mathsf{KL}\left(\pi^{\mathsf{ref}}\|\pi\right)$, $g := r - \widehat{r} \geq 0$)

1. **Simplex Conjugate-Bregman Duality** ($y_a := 1 - \widehat{\pi}(a)/\pi^*(a) < 1$, $\phi(y) := -y - \log(1-y)$):

   $$\mathrm{SubOpt}_{\mathrm{FKL}}(\widehat{\pi}) = \mathsf{B}_{h^*}\left(r, \widehat{r}\right) = \mathsf{B}_h\left(\widehat{\pi}, \pi^*\right) = \eta^{-1}\mathbb{E}_{a\sim\pi^{\mathsf{ref}}}\left[\phi(y_a)\right].$$

2. **Fenchel–Young** ($\mathrm{SubOpt}_{\mathrm{FKL}}(\widehat{\pi}) \leq \langle g, \pi^* - \widehat{\pi}\rangle$, $z_a := 2\eta\frac{\pi^*(a)}{\pi^{\mathsf{ref}}(a)}g(a)$):

   $$\langle g, \pi^* - \widehat{\pi}\rangle = \tfrac{1}{2}\eta^{-1}\mathbb{E}_{\pi^{\mathsf{ref}}}[y_a z_a] \leq \tfrac{1}{2}\eta^{-1}\mathbb{E}_{\pi^{\mathsf{ref}}}\left[\phi(y_a) + \phi^*(z_a)\right].$$

3. **Self-Bounding + Pessimism**
   ($z_a \geq 0 \implies \phi^*(z_a) = z_a - \log(1+z_a) \leq \frac{1}{2}z_a^2$):

$$\mathrm{SubOpt}_{\mathrm{FKL}}(\widehat{\pi}) \leq \eta^{-1}\mathbb{E}_{\pi^{\mathsf{ref}}}\left[\phi^*(z_a)\right] \leq 2\eta\sum_{a\in\mathcal{A}}\frac{\pi^*_{\mathrm{FKL}}(a)^2}{\pi^{\mathsf{ref}}(a)}\left(r(a) - \widehat{r}(a)\right)^2.$$

Takeaways & Open Questions

Summary

- **First $\widetilde{O}(\epsilon^{-1})$ fast rates** for forward-KL offline CBs under **single-policy concentrability** (tabular & $\mathcal{G}$).
- **Simplex conjugate-Bregman + Fenchel–Young self-bounding**: exploits pessimism ($z_a \geq 0$) without midpoint monotonicity.
- **Rate-optimal hardness**: $\Omega(\eta|\mathcal{S}||\mathcal{A}|\epsilon^{-1})$ under strong regularization, transitioning back to $\Omega(\epsilon^{-2})$ as $\eta \to \infty$.

Open Directions

- Sharpening coverage exponents ($(C^{\pi^*})^2$ vs. C^{π^*}) and proving function-approximation fast-rate lower bounds.
- Extending to multi-step MDPs, misspecified rewards, and online forward-KL regret ($z = -1$ pole).

Thank You! **NeurIPS 2026 Spotlight** Paper: arXiv:2605.09214

Poster Session: Fri, Dec 11, 2026 • 8:30 AM – 11:30 AM AEDT • **Hall C1**