Proposal: KL-Regularized RLHF (and χ^2 Preference Optimization) Sharp Rates and Coverage Effects in Contextual Bandits

October 27, 2025

Background and Motivation. Reinforcement learning from human feedback (RLHF) frequently applies a divergence penalty that keeps a learned policy close to a reference policy (e.g., the pre-trained LM). A canonical objective is a regularized maximum-return criterion, often with (reverse) Kullback–Leibler (KL) divergence; recent theory shows this regularization can *sharply* reduce sample complexity in contextual bandits (a standard lens for per-prompt response selection in LLMs). Parallel work argues that replacing KL with the χ^2 divergence (χ^2 *Preference Optimization,* χPO) yields robustness aligned with single-policy concentrability—a gold-standard coverage condition in offline RLHF. This proposal targets a clean, theory-first synthesis of these guarantees in a compact setting suitable for a course project.[1]

Setting (Contextual Bandit RLHF Model). Let $\mathcal{X}$ denote prompts (contexts), $\mathcal{Y}$ candidate responses (actions). A policy $\pi(\cdot \mid x)$ maps $x \in \mathcal{X}$ to a distribution over $\mathcal{Y}$. A reference policy π_0 (the base LM) provides coverage. For a bounded reward $r : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ and divergence weight $\beta > 0$, consider the f-divergence regularized objective

$$J_f(\pi) \;:=\; \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi(\cdot|x)}[\, r(x,y)\,] \;-\; \beta\, \mathbb{E}_{x\sim\mathcal{D}}\Big[D_f\big(\pi(\cdot \mid x) \,\|\, \pi_0(\cdot \mid x)\big)\Big], \tag{1}$$

where D_f is KL ($f(t) = t \log t$) or χ^2 ($f(t) = \frac{1}{2}(t-1)^2$). We let $\pi_f^\star \in \arg\max_\pi J_f(\pi)$ and call $\widehat{\pi}$ *ε-optimal* if $J_f(\pi_f^\star) - J_f(\widehat{\pi}) \le \varepsilon$. Throughout, we measure data requirements by the minimal $n(\varepsilon)$ such that an estimator using n i.i.d. rounds achieves ε-optimality with high probability.

Coverage. Following recent alignment theory, we quantify how well π_0 "covers" near-optimal actions via a single-policy concentrability-style coefficient, e.g.

$$\mathrm{Cov} \;:=\; \sup_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \frac{\pi_f^\star(y \mid x)^2}{\pi_0(y \mid x)} \quad \in [1, \infty], \tag{2}$$

or, in sparse cases, a minimum-mass proxy $\kappa := \inf_{x,y:\, \pi_f^\star(y|x)>0} \pi_0(y \mid x) > 0$. These are standard in contextual bandit/RLHF analyses and govern the stability of importance-weighted estimates.

[1] This proposal follows the *CS 542 Stat RL: Project Guideline* (sections on proposal scope and avoiding plagiarism). *Source available from course materials.* :contentReference[oaicite:1]index=1

Papers to Study (Seed Papers).

- **Sharp Analysis for KL-Regularized Contextual Bandits and RLHF** [1]. Establishes *linear* sample complexity in $1/\varepsilon$ (vs. $1/\varepsilon^2$) for KL-regularized contextual bandits, and clarifies when coverage enters *additively* rather than multiplicatively.
- **Correcting the Mythos of KL-Regularization: Direct Alignment without Overoptimization via χ^2 Preference Optimization** [2]. Shows that a χ^2-based variant of DPO achieves guarantees under *single-policy concentrability*, providing pessimism and robustness to reward-model misspecification in offline RLHF.

Main Theoretical Results to Reproduce (Tentative).

R1. Sharp KL Rate. For the KL-regularized objective J_{KL} in (1), reproduce the result that there exists an algorithm achieving

$$n(\varepsilon) \;=\; \tilde{\mathcal{O}}(1/\varepsilon), \tag{3}$$

under standard boundedness/realizability assumptions, and that with sufficient coverage the dependence on Cov (or κ^{-1}) is *additive* rather than multiplicative [1].

R2. χ^2 Preference Optimization Guarantee. For J_{χ^2} in (1), reproduce the generalization and robustness guarantees of χPO under *single-policy concentrability*, highlighting how the curvature of D_{χ^2} controls estimation error and mitigates overoptimization in offline settings [2].

R3. Unified Statement (optional if space is tight). Present a succinct comparison theorem indicating when J_{KL} and J_{χ^2} yield comparable $n(\varepsilon) = \tilde{\mathcal{O}}(1/\varepsilon)$ rates, emphasizing the role of local Bregman curvature at π_0 and the coverage parameters in (2).

Scope. The project will focus on contextual bandits (no horizon dependence), reproduce a carefully chosen subset of the above results with full mathematical detail, and present a unified notation and proof roadmap tailored to RLHF-style learning. No new algorithms are proposed here; the emphasis is on crisp reproduction, clear assumptions, and tight dependence on $(\varepsilon, \beta, \mathrm{Cov})$.

References

References

[1] Heyang Zhao, Chenlu Ye, Quanquan Gu, and Tong Zhang. *Sharp Analysis for KL-Regularized Contextual Bandits and RLHF.* arXiv:2411.04625, 2024. URL: arxiv.org/abs/2411.04625.

[2] Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D. Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J. Foster. *Correcting the Mythos of KL-Regularization: Direct Alignment without Overoptimization via Chi-Squared Preference Optimization.* arXiv:2407.13399, 2024. URL: arxiv.org/abs/2407.13399.