TAIL-AWARE DIRECT PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The geometry of policy regularization is a central design choice in LLM alignment. KL-regularized RLHF and DPO are attractive because they yield exponential tilting and a simple log-ratio preference loss, but this geometry can amplify low-coverage reward-model errors and cause severe overoptimization. We propose *Tail-Aware Direct Preference Optimization* (TDPO), a direct preference method derived from Rényi-regularized RLHF. TDPO derives an exact direct-preference loss on the active constraint, whose Bradley–Terry score difference is a difference of density-ratio powers rather than the KL-induced log-ratio. The order α controls statistical coverage, tail sensitivity, and anti-hacking strength as a hyperparameter of the preference law. We prove finite-sample bounds for the unregularized optimality gap and regularized sub-optimality gap under a Bradley–Terry preference model. A two-environment lower bound matches the latter's sample-size exponent for arbitrary policy-valued estimators. We further characterize the induced geometry: locally, Rényi regularization rescales the Fisher geometry of KL, while globally it provides polynomial rare-set control. A TL;DR study evaluates density-ratio tails and low-reference-coverage mass across DPO, ChiPO, and clipped/capped baselines.

1 INTRODUCTION AND BACKGROUND

RLHF has become the dominant abstraction for post-training large language models (LLMs): start from a reference policy, collect preference data, learn from those preferences, and update the model so that preferred responses become more likely (Christiano et al., 2017; Ouyang et al., 2022; Lambert, 2026). The entire pipeline is statistical. Preference labels are noisy, reward surrogates are misspecified, coverage is inherited from the reference policy, and evaluation occurs under distribution shift. Recent work has therefore shifted attention

> from "which optimizer should we use?" to "which policy geometry should we trust?"

Recently, Na et al. (2026) argues that KL compares only matching token indices and misses semantic geometry, motivating Wasserstein policy regularization for RLHF. Zhang et al. (2026) shows that KL-based information-theoretic arguments can fail under sub-Weibull rewards, motivating tail-aware control in RLHF. In parallel, χ^2-preference optimization shows that stronger density-ratio ν_π (see equation (1)) control can prevent offline reward overoptimization (Huang et al., 2025). These works suggest a design principle: *align the divergence geometry with the tail behavior of proxy errors*.

The reward-hacking is especially acute in *offline* alignment. Preference data are gathered from the reference policy, while the aligned policy is free to move into low-coverage regions where the proxy objective extrapolates poorly. This creates overoptimization pathology: the proxy reward continues to increase, yet the true response quality (the gold reward) plateaus or degrades (Gao et al., 2023).

We work with a prompt space $\mathcal{X}$ equipped with a prompt distribution ρ_X and a measurable response space $(\mathcal{Y}, \mathcal{B})$. A reference policy $\pi_0(\cdot \mid x)$ is fixed throughout; every candidate policy $\pi(\cdot \mid x)$ is assumed absolutely continuous w.r.t. $\pi_0(\cdot \mid x)$ for ρ_X-almost every x. The conditional density ratio

$$\nu_\pi(x, y) := \frac{d\pi(\cdot \mid x)}{d\pi_0(\cdot \mid x)}(y). \tag{1}$$

Huang et al. (2025) replaced DPO's KL geometry by a mixed χ^2 plus KL regularizer (ChiPO), which is motivated by avoiding truncation: for χ^2 regularizer, the optimizer can assign zero mass to low-reward responses. Adding KL ensures globally positive density ratios and yields the composite logit

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

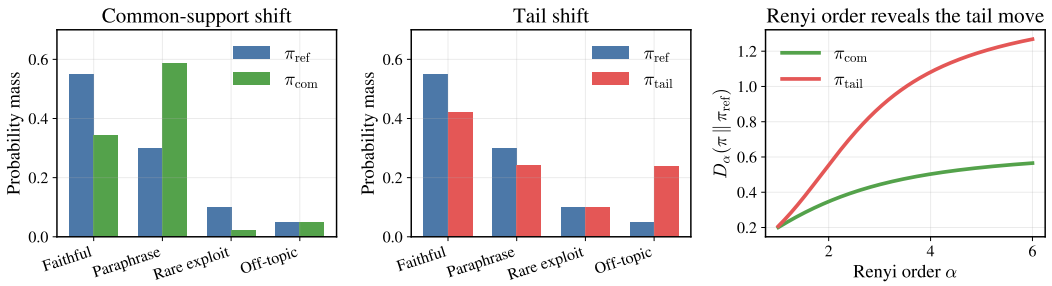

Figure 1: All policies are defined over {Faithful, Paraphrase, Rare exploit, Off-topic}. Two candidate policies have nearly the same KL distance to the reference $\pi_{\text{ref}} = \pi_0$, but differ in where probability mass moves: π_{com} reallocates mass among common responses, whereas π_{tail} inflates a low-reference-probability tail response. The right panel shows that higher-order Rényi reveals this tail inflation.

$(\nu_\pi + \log \nu_\pi)(y^+) - (\nu_\pi + \log \nu_\pi)(y^-)$. The general f-DPO framework (Wang et al., 2024) faces the same issue and therefore restricts attention to f-divergences satisfying $0 \notin \text{dom}(f')$, excluding χ^2 divergence. This restriction also does not cover the exponentiated Rényi divergence (see Definition 1.1) penalty, $\exp((\alpha-1)D_\alpha(\pi\|\pi_0)) = \mathbb{E}_{\pi_0}[\nu_\pi^\alpha]$, which is the convex tail-aware control that limit low-coverage exploitation. We take a different route. Instead of mixing in KL- with χ^2-divergence or restricting the large f-divergence class, we work directly with the pure Rényi geometry, which is attractive for practitioner from a computational and tuning perspective.

Definition 1.1 (Rényi divergence)**.** For $\alpha > 1$ and $P \ll Q$, $D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log\int(\frac{dP}{dQ})^\alpha dQ$.

When $\alpha = 2$, one gets $D_2(P\|Q) = \log(1+\chi^2(P\|Q))$; as $\alpha \to \infty$, the divergence becomes $D_\infty(P\|Q) = \log\left\|\frac{\mathrm{d}P}{\mathrm{d}Q}\right\|_\infty$. Crucially, $D_\alpha(P\|Q)$ is non-decreasing in α and converges to $D_{\text{KL}}(P\|Q)$ as $\alpha \to 1$ from above. A Rényi budget $D_\alpha \le \epsilon$ with $\alpha > 1$ is therefore *strictly stronger* than the corresponding KL budget $D_{\text{KL}} \le \epsilon$: every policy admitted by the Rényi trust region is also admitted by KL, but not conversely. Thus, Rényi control interpolates between moment control and worst-case ratio control, with larger α imposing progressively tighter constraints on the density-ratio tails. The resulting direct-preference logit $\beta[\nu_\pi^{\alpha-1}(y^+) - \nu_\pi^{\alpha-1}(y^-)]$ is simpler than both the KL log-ratio and ChiPO's mixed $(\nu_\pi + \log\nu_\pi)$ link: it is a single-term power of the density ratio, controlled by α.

More importantly, the optimizer of Rényi-constrained RLHF is a thresholded power law, but KL yields the Gibbs exponential tilt, which can over-amplify spurious reward. Finite-order Rényi yields a power law: large positive rewards are compressed by the exponent $1/(\alpha-1)$. At $\alpha = 2$ the policy update is linear in reward excess; for $\alpha > 2$ it is sublinear. As we formalize in Section 5.1, this compression translates into a multiplicative rare-set bound that KL structurally cannot provide.

Figure 1 illustrates the motivation. The reference policy is $\pi_0 = (0.55, 0.30, 0.10, 0.05)$. We compare two candidate policies: $\pi_{\text{com}} = (0.34, 0.59, 0.02, 0.05)$, $\pi_{\text{tail}} = (0.42, 0.24, 0.10, 0.24)$. They have nearly the same KL distance to the reference, $D_{\text{KL}}(\pi_{\text{com}}\|\pi_0) \approx 0.20$ and $D_{\text{KL}}(\pi_{\text{tail}}\|\pi_0) \approx 0.21$, but their deviations are so different. The first policy mainly reallocates probability mass among common responses, while the second assigns more mass to a low-reference-probability tail response. The right panel shows that, as α increases from KL ($\alpha = 1$), higher-order Rényi divergences amplify large density ratios and therefore separate the tail shift from the benign common-support shift.

Contributions. First, using the active trust-region representation, we derive an exact direct-preference objective, *Tail-Aware DPO* (TDPO), whose Bradley–Terry score difference is a difference of density-ratio powers rather than the KL-induced log-ratio. Second, under finite-class bounded logits, we derive an upper bound for the unregularized gap and matching upper and lower sample-size exponents for the regularized gap. The lower bound allows arbitrary policy outputs, while the upper bound assumes a finite candidate class. Third, we characterize the geometry of finite-order Rényi control from both local and global perspectives: locally, Rényi divergence agrees with KL to second order up to an α-dependent curvature factor; globally, it yields polynomial rare-set control, discouraging the policy from exploiting poorly covered regions. Fourth, we introduce checkpoint-level mechanism diagnostics and clipped/capped baselines on TL;DR to separate rare-set control from generic saturation and unmatched optimization strength.

2 Related work and notions

RLHF and direct preference learning. RL from preferences was introduced for deep RL by Christiano et al. (2017) and scaled to language-model instruction following by Ouyang et al. (2022). DPO (Rafailov et al., 2023) showed that KL-regularized RLHF can be written as direct logistic classification on preference pairs. IPO, SLiC, and related methods refine preference losses and calibration (Azar et al., 2024; Zhao et al., 2023). TDPO keeps the direct-preference interface but changes the underlying trust region.

Policy regularization beyond KL. KL is convenient because it yields log-ratio penalties, but its mode-seeking and exponential-tilt behavior can reduce diversity or amplify reward-model errors. General f-DPO methods replace reverse KL by other f-divergences (Wang et al., 2024). χPO gives a particularly relevant anti-overoptimization result through stronger density-ratio control (Huang et al., 2025). TDPO differs in using the genuine log-moment Rényi divergence rather than a mixed χ^2 plus KL construction. BandPO (Li et al., 2026) and dynamic clipping methods address proximal updates in online RL; TDPO addresses offline direct preference learning through a divergence-derived loss.

Tail-aware and semantic-aware alignment. Wasserstein policy regularization modifies the ground geometry between tokens and responses, allowing semantically close alternatives to be regularized more gently than unrelated alternatives (Na et al., 2026). TDPO is orthogonal: it modifies the tail geometry of density ratios. Zhang et al. (2026) showed that KL-based generalization tools are tied to moment-generating functions, whereas Rényi-type tools can survive heavy-tailed losses and rewards.

Notation. $D_{\mathrm{KL}}(\pi\|\pi_0) = \int \log(d\pi/d\pi_0)d\pi$, whenever $\pi \ll \pi_0$. And $\mathrm{kl}(p\|q)$ is the KL divergence between two Bernoulli distributions with parameters p and q, defined as $\mathrm{kl}(p\|q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$. For a measurable function f, its essential infimum is $\operatorname{essinf} f = \sup\{t \in \mathbb{R} : f(x) \geq t \text{ for almost every } x\}$. The covering number $\mathcal{N}(\mathcal{H}, \delta, \|\cdot\|)$ is the smallest number of balls of radius δ in the norm $\|\cdot\|$ needed to cover the class $\mathcal{H}$.

3 From Rényi RLHF to Tail-Aware DPO

We define carrier measures $\mathcal{P}_0(dx, dy) := \rho_X(dx)\,\pi_0(dy \mid x)$, $\mathcal{M}_0(dx, dy, dz) := \rho_X(dx)\,\pi_0(dy \mid x)\,\pi_0(dz \mid x)$. For any policy π and the true reward $r_\star : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, we write

$$J(\pi) \;:=\; \mathbb{E}_\pi\big[r_\star(X,Y)\big] \;=\; \mathbb{E}_{X\sim\rho_X,\,Y\sim\pi(\cdot|X)}\big[r_\star(X,Y)\big].$$

We observe i.i.d. data $\mathcal{D}_n = \{(X_i, Y_i^+, Y_i^-, L_i)\}_{i=1}^n$, where (X_i, Y_i^+, Y_i^-) are drawn from $\mathcal{M}_0$ and the binary label follows the *Bradley–Terry model* (Bradley & Terry, 1952):

$$P(L_i = 1 \mid X_i, Y_i^+, Y_i^-) \;=\; \sigma\big(r_\star(X_i, Y_i^+) - r_\star(X_i, Y_i^-)\big), \tag{2}$$

where $\sigma(t) := (1 + e^{-t})^{-1}$ is the logistic sigmoid. For a policy π on the Rényi trust region $\Pi_\epsilon := \big\{\pi : D_\alpha\big(\pi(\cdot \mid x)\|\pi_0(\cdot \mid x)\big) \leq \epsilon \text{ for } \rho_X\text{-a.e. } x\big\}$, the α-moment coverage is defined by

$$\mathcal{C}_\alpha(\pi) \;:=\; \mathbb{E}_{\mathcal{P}_0}\big[\nu_\pi(X,Y)^\alpha\big] \;=\; \exp\big((\alpha-1)\,D_\alpha(\rho_X\otimes\pi \,\|\, \rho_X\otimes\pi_0)\big). \tag{3}$$

In particular, every $\pi \in \Pi_\epsilon$ satisfies $\mathcal{C}_\alpha(\pi) \leq e^{(\alpha-1)\epsilon}$.

3.1 The exact optimizer is a thresholded power law

Consider the Rényi-constrained RLHF problem

$$\max_\pi\; \mathbb{E}_{x\sim\rho_X,\,y\sim\pi(\cdot|x)}[r(x,y)] \qquad \text{s.t.} \qquad D_\alpha\big(\pi(\cdot \mid x)\,\|\,\pi_0(\cdot \mid x)\big) \leq \epsilon \quad \text{for } \rho_X\text{-a.e. } x. \tag{4}$$

To obtain a convex penalty in the policy density ratio, we use the exponentiated Rényi form

$$\exp((\alpha-1)D_\alpha(\pi\|\pi_0)) = \mathbb{E}_{\pi_0}[\nu_\pi^\alpha]\,, \tag{5}$$

rather than penalizing D_α itself:

$$\max_{\pi(\cdot|x)\ll\pi_0(\cdot|x)} \left\{\mathbb{E}_{Y\sim\pi(\cdot|x)}[r(x,Y)] - \frac{\beta}{\alpha}\,\exp\Big((\alpha-1)D_\alpha\big(\pi(\cdot \mid x)\,\|\,\pi_0(\cdot \mid x)\big)\Big)\right\}. \tag{6}$$

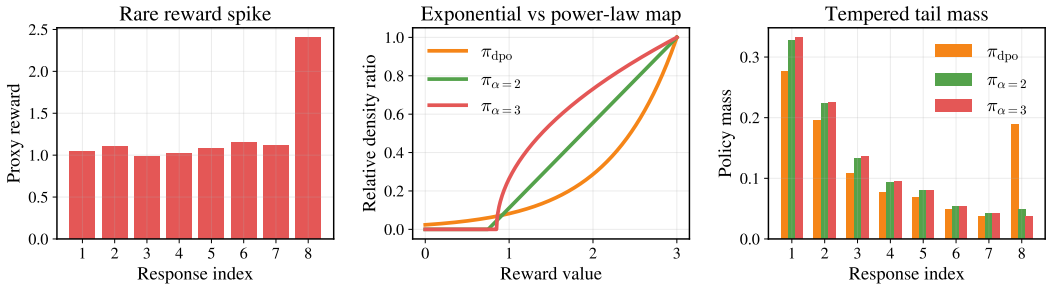

Figure 2: Exponential versus thresholded power-law policy geometry. A rare response receives an isolated proxy-reward spike in a low-reference-mass region. KL-DPO uses an exponential reweighting map, whereas Rényi policies temper the tail mass assigned to the spurious rare response.

Up to an additive constant, (5) is the f-divergence generated by $f_\alpha(u) = (u^\alpha - \alpha u + \alpha - 1)/\alpha$, which is not uniformly strongly convex on $\mathbb{R}_+$ except at $\alpha = 2$. Related constrained solutions appear in Zhang et al. (2026); the penalized result below is proved in Appendix B.1.

Lemma 3.1 (Solution to exponentiated-Rényi-penalized optimization)**.** *Let $\alpha > 1$, $\beta > 0$, and $(u)_+ = \max\{u, 0\}$. Fix $x \in \mathcal{X}$, and assume $\mathbb{E}_{Y\sim\pi_0(\cdot|x)}\left[|r(x,Y)|^{\alpha/(\alpha-1)}\right] < \infty$. Then (6) has a unique optimizer, up to $\pi_0(\cdot \mid x)$-almost-sure equivalence, given by*

$$\frac{\pi^\star_\beta(\mathrm{d}y \mid x)}{\pi_0(\mathrm{d}y \mid x)} = \left(\frac{r(x,y) - t_\beta(x)}{\beta}\right)_+^{1/(\alpha-1)}, \tag{7}$$

where $t_\beta(x) \in \mathbb{R}$ is uniquely determined by $\mathbb{E}_{Y\sim\pi_0(\cdot|x)}\left(\frac{r(x,Y)-t_\beta(x)}{\beta}\right)_+^{1/(\alpha-1)} = 1$.

Lemma 3.1 is the structural core of the method: it shows that finite-order Rényi regularization truncates low-value responses and replaces the Gibbs exponential tilt with a power-law map. At $\alpha = 2$, the update is linear in the positive excess reward; for $\alpha > 2$, it becomes sublinear, which is desirable when large proxy rewards may be artifacts. Figure 2 illustrates the practical consequence of this change in geometry on an eight-response example with reference policy $\pi_0 = (0.34, 0.22, 0.15, 0.10, 0.08, 0.05, 0.04, 0.02)$ and proxy reward $r = (1.05, 1.10, 0.98, 1.02, 1.08, 1.15, 1.12, 2.40)$. The final response is rare under the reference policy but receives an isolated proxy spike. Under KL-DPO, the policy is reweighted exponentially, $\pi_{\text{dpo}}(i) \propto \pi_0(i)\exp(r_i/\beta)$, whereas finite-order Rényi regularization yields the thresholded power-law update $\pi_\alpha(i) \propto \pi_0(i)(r_i - t_\alpha)_+^{1/(\alpha-1)}$. As a result, the exponential KL update reacts much more aggressively to the rare proxy spike, whereas the Rényi update tempers the resulting tail mass.

3.2 Tail-Aware Direct Preference Optimization on the Active Rényi Trust Region

Since the prompt-dependent term $t_\beta(x)$ cancels in reward differences, we define the induced reward by choosing the representative

$$r_\pi(x,y) \;:=\; \beta\,\nu_\pi(x,y)^{\alpha-1}.$$

By Lemma 3.1, π is the unique Rényi-penalized optimizer (6) under r_π. Bradley–Terry score difference for a preference pair (y^+, y^-) is

$$r_\pi(x,y^+) - r_\pi(x,y^-) \;=\; \beta\left[\nu_\pi(x,y^+)^{\alpha-1} - \nu_\pi(x,y^-)^{\alpha-1}\right] \;=:\; \Delta_\pi(x,y^+,y^-). \tag{8}$$

This is the exact pure-Rényi analogue of the KL-DPO reparametrization. In KL-DPO, the Gibbs solution $\pi^\star \propto \pi_0\exp(r/\beta)$ yields the log-odds ratio $\beta[\log\nu_\pi(y^+) - \log\nu_\pi(y^-)]$; under finite-order Rényi regularization, the logarithm is replaced by a density-ratio power. Rather than first estimating $r^\star$ and then solving (6), Algorithm 1 fits Δ_π over a policy class $\mathcal{H}$ and returns the empirical logistic minimizer. Choosing $\mathcal{H} \subseteq \Pi_\epsilon$ enforces the budget $D_\alpha(\pi\|\pi_0) \le \epsilon$.

Algorithm 1 *Tail-Aware Direct Preference Optimization* (TDPO)

1: **Input:** Reference policy π_0, preference data $\mathcal{D}_n$, order $\alpha > 1$, penalty parameter $\beta > 0$, and policy class $\mathcal{H}$;
2: For each $\pi \in \mathcal{H}$, compute $\Delta_\pi(x, y^+, y^-) = \beta\{\nu_\pi(x, y^+)^{\alpha-1} - \nu_\pi(x, y^-)^{\alpha-1}\}$;
3: Return any minimizer $\hat{\pi} \in \operatorname{argmin}_{\pi\in\mathcal{H}} \widehat{\mathcal{L}}_n(\pi)$.

Plugging (8) into the Bradley–Terry loss for preference data $\mathcal{D}_n = \{(X_i, Y_i^+, Y_i^-, L_i)\}_{i=1}^n$ gives

$$\widehat{\mathcal{L}}_n(\pi) = \frac{1}{n}\sum_{i=1}^{n}\left\{\log[1 + e^{\Delta_\pi(X_i, Y_i^+, Y_i^-)}] - L_i\Delta_\pi(X_i, Y_i^+, Y_i^-)\right\}.$$

The α in TDPO is a hyperparameter in the Bradley–Terry model; see Appendix K for details.

4 Theory: Two Upper Bounds and a Lower Bound

4.1 Statistical Bounds for TDPO

We establish matching upper and minimax lower bounds for the regularized objective, together with an upper bound for the unregularized objective. The rate behavior changes at $\alpha = 2$: for $1 < \alpha \le 2$, the regularized and unregularized rates are respectively n^{-1} and $n^{-1/2}$, the same sample-size exponents as KL regularization, whereas for $\alpha > 2$ they become $n^{-\alpha/[2(\alpha-1)]}$ and $n^{-\alpha/[4(\alpha-1)]}$. Existing matching lower bounds in related RLHF analyses have primarily concerned regularized objectives (Zhao et al., 2026); our lower bound follows this formulation. The strongly convex f-divergence theory of Zhao et al. (2026) does not cover our penalty, while ChiPO treats only the quadratic geometry analogous to $\alpha = 2$ under $\beta = \Theta(n^{-1/2})$; we cover every $\alpha > 1$ and give matching lower bounds for the regularized gap.

We distinguish two targets by following Aminian et al. (2025). The *unregularized optimality gap* is $J(\pi^\dagger) - J(\hat{\pi})$; it may be negative. The *regularized sub-optimality gap* is $J_{\mathrm{reg}}(\pi^\dagger) - J_{\mathrm{reg}}(\hat{\pi})$, where

$$J_{\mathrm{reg}}(\pi) := J(\pi) - \frac{\beta}{\alpha}\mathbb{E}_{\mathcal{P}_0}[\nu_\pi^\alpha]. \tag{9}$$

This is the prompt-averaged objective corresponding to (6). Since $\pi^\dagger$ maximizes that objective, the regularized gap is nonnegative.

For the upper bounds, Algorithm 1 performs its original logistic MLE over a finite policy class $\mathcal{H}$. The complexity is $\log|\mathcal{H}|$, and both bounds use Lemma B.1.

Assumption 4.1 (Finite candidate class)**.** The candidate class $\mathcal{H}$ is finite and nonempty, and Algorithm 1 returns an empirical logistic minimizer $\hat{\pi} \in \mathcal{H}$.

Assumption 4.2 (Realizability and compatibility; see Appendix F)**.** Let $\pi_\beta^\star$ be the population optimizer of (6) under $r_\star$. Assume $\pi_\beta^\star \in \mathcal{H}$ and $\beta \ge \operatorname{ess\,sup}_x \beta_\star(x)$, with $\beta_\star(x)$ defined in Appendix F. Set $\pi^\dagger := \pi_\beta^\star$. Then, for some prompt-dependent $b(x)$,

$$r_\star(x, y) = \beta\nu_{\pi^\dagger}(x, y)^{\alpha-1} + b(x),$$

so $\Delta_{\pi^\dagger}$ realizes the Bradley–Terry model.

Assumption 4.3 (Bounded pairwise logit)**.** There exists $V > 0$ such that $|\Delta_\pi(x, y^+, y^-)| \le V$ for every $\pi \in \mathcal{H}$ and $\mathcal{M}_0$-almost every triple.

Theorem 4.4 (Unregularized optimality gap)**.** *Suppose Assumptions 4.1–4.3 hold for $\alpha > 1$ and $\beta > 0$. Let $\hat{\pi}$ be the empirical minimizer returned by Algorithm 1. Set $K := \left(1 + \frac{V}{\beta}\right)^{1/(\alpha-1)}$,*

$$\varepsilon_2 := \left(\frac{448}{3}e^{2V}\frac{\log(|\mathcal{H}|/\delta)}{n}\right)^{1/2}, \quad R_\star := \left(\mathbb{E}_{\rho_X}\operatorname{Var}_{\pi_0}(r_\star(X, Y) \mid X)\right)^{1/2}.$$

With probability at least $1-\delta$,

$$|J(\pi^\dagger) - J(\hat{\pi})| \le \begin{cases} \dfrac{K^{2-\alpha}\varepsilon_2 R_\star}{\sqrt{2}\,\beta(\alpha-1)}, & 1<\alpha<2, \\ \dfrac{\varepsilon_2 R_\star}{\sqrt{2}\,\beta}, & \alpha=2, \\ \left(\mathbb{E}_{\mathcal{P}_0}\nu_{\pi^\dagger}^{\alpha}\right)^{1/2}\beta^{\frac{\alpha-2}{2(\alpha-1)}}\left(\dfrac{\varepsilon_2}{\sqrt{2}}\right)^{\frac{\alpha}{2(\alpha-1)}}, & \alpha>2. \end{cases} \tag{10}$$

In the last case, $\mathbb{E}_{\mathcal{P}_0}\nu_{\pi^\dagger}^{\alpha} \le 1 + V/\beta$ *follows from Assumption 4.3 and policy normalization. For fixed* α, β, V *and fixed class complexity, the rates are* $n^{-1/2}$, $n^{-1/2}$, *and* $n^{-\alpha/[4(\alpha-1)]}$, *respectively.*

Theorem 4.5 (Regularized sub-optimality gap)**.** *Under the assumptions and notation of Theorem 4.4, with probability at least* $1-\delta$,

$$0 \le J_{\mathrm{reg}}(\pi^\dagger) - J_{\mathrm{reg}}(\hat{\pi}) \le \begin{cases} \dfrac{K^{2-\alpha}\varepsilon_2^2}{2\beta(\alpha-1)}, & 1<\alpha<2, \\ \dfrac{\varepsilon_2^2}{4\beta}, & \alpha=2, \\ \beta^{-1/(\alpha-1)}\left(\dfrac{\varepsilon_2}{\sqrt{2}}\right)^{\alpha/(\alpha-1)}, & \alpha>2. \end{cases} \tag{11}$$

For fixed α, β, V *and fixed class complexity, these are* n^{-1}, n^{-1}, *and* $n^{-\alpha/[2(\alpha-1)]}$.

Theorem 4.6 (Minimax lower bound for regularized sub-optimality)**.** *Fix* $\alpha > 1$, $\beta > 0$, *and* $\epsilon > 0$. *For all sufficiently large* n, *there exist two one-prompt Bradley–Terry environments, specified by reward functions* r_+ *and* r_-, *with a finite response space and positive reference probabilities. For each* $\theta \in \{+,-\}$, *let* π_θ *be the maximizer of* $J_{\mathrm{reg},\theta}$ *over the entire probability simplex. The two optimizers satisfy* $\pi_+, \pi_- \in \Pi_\epsilon$, *and both environments satisfy Assumption 4.2. All policy scores on this finite simplex have a finite bound* V *independent of* n. *For every possibly randomized estimator with output in the entire probability simplex* $\mathcal{P}(\mathcal{Y})$, *with a constant* $c_{\alpha,\beta,\epsilon} > 0$, *we have*

$$\inf_{\widehat{\pi}_n:\mathcal{D}_n\to\mathcal{P}(\mathcal{Y})} \sup_{\theta\in\{+,-\}} \mathbb{E}_\theta[J_{\mathrm{reg},\theta}(\pi_\theta) - J_{\mathrm{reg},\theta}(\widehat{\pi}_n)] \ge \begin{cases} c_{\alpha,\beta,\epsilon}\, n^{-1}, & 1<\alpha\le 2, \\ c_{\alpha,\beta,\epsilon}\, n^{-\alpha/[2(\alpha-1)]}, & \alpha>2. \end{cases} \tag{12}$$

Table 1: Sample-complexity comparison.

Divergence	Regularized gap	Unregularized gap
Upper bounds		
KL (Aminian et al., 2025)	n^{-1}	$n^{-1/2}$
χ^2 + KL (Huang et al., 2025)	—	$n^{-1/2}$, $\beta = \Theta(n^{-1/2})$
Strongly convex f (Zhao et al., 2026)	n^{-1}	—
Rényi (TDPO; $\beta \ge \|\beta_\star\|_\infty$), $1<\alpha\le 2$	n^{-1}	$n^{-1/2}$
Rényi (TDPO; $\beta \ge \|\beta_\star\|_\infty$), $\alpha>2$	$n^{-\alpha/[2(\alpha-1)]}$	$n^{-\alpha/[4(\alpha-1)]}$
Regularized lower bounds		
KL (Zhao et al., 2026)	n^{-1}	—
Strongly convex f (Zhao et al., 2026)	n^{-1}	—
Rényi (TDPO), $1<\alpha\le 2$ and $\alpha>2$	n^{-1} and $n^{-\alpha/[2(\alpha-1)]}$	—

5 Rényi Geometry for DPO

5.1 Rare-set control and the anti-hacking guarantee

Reward hacking in offline alignment has a characteristic geometric signature: the proxy reward assigns spuriously high scores to responses that lie in a rare, low-coverage region of the reference policy. The aligned policy then shifts probability mass onto this region, inflating proxy performance while degrading true quality. The natural question is therefore not whether the aligned policy is close

to the reference in some global sense, but how much mass it can place on a specific rare set A_{hack} that the reference policy barely covers. We call this *rare-set control*.

Under a Rényi budget, the mass decays polynomially: $\pi(A) \le e^{(\alpha-1)\epsilon/\alpha}\pi_0(A)^{(\alpha-1)/\alpha}$ (Theorem 5.1). For Algorithm 1, this guarantee at budget ϵ follows by choosing $\mathcal{H} \subseteq \Pi_\epsilon$. This geometric constraint is separate from the bounded-logit statistical setting of Section 4. The rarer a set of exploit responses is under the reference policy, the harder it is for a budget-constrained policy to concentrate mass on it. KL cannot yield any multiplicative bound: $\pi(A) \le C\pi_0(A)^\gamma$ for $\gamma > 0$ (Theorem 5.2).

Theorem 5.1 (Rényi geometry: rare-set amplification)**.** *Fix $\alpha > 1$. For every policy $\pi \ll \pi_0$ and every measurable set $A \subseteq \mathcal{X} \times \mathcal{Y}$,*

$$(\rho_X \otimes \pi)(A) \le \mathcal{C}_\alpha(\pi)^{1/\alpha}(\rho_X \otimes \pi_0)(A)^{(\alpha-1)/\alpha}.$$

Equivalently, $(\rho_X \otimes \pi)(A) \le \exp(\frac{\alpha-1}{\alpha} D_\alpha(\rho_X \otimes \pi \,\|\, \rho_X \otimes \pi_0))(\rho_X \otimes \pi_0)(A)^{(\alpha-1)/\alpha}$. If $D_\alpha(\rho_X \otimes \pi \,\|\, \rho_X \otimes \pi_0) \le \epsilon$, then $(\rho_X \otimes \pi)(A) \le \exp(\frac{\alpha-1}{\alpha}\epsilon)(\rho_X \otimes \pi_0)(A)^{(\alpha-1)/\alpha}$.

Penalized rather than constrained optimization. A hard Rényi projection is not required at the population optimum. If the within-prompt reward range is bounded by R, the exponentiated-Rényi-penalized optimizer π_{R} satisfies

$$\mathcal{C}_\alpha(\pi_{\text{R}}) \le 1 + \alpha R/\beta, \qquad (\rho_X \otimes \pi_{\text{R}})(A) \le (1 + \alpha R/\beta)^{1/\alpha}\,(\rho_X \otimes \pi_0)(A)^{(\alpha-1)/\alpha}.$$

Thus the penalty itself induces polynomial rare-set control. The proof, the corresponding effective Rényi radius, and the exponential KL comparison are given in Appendix G.

KL cannot yield multiplicative rare-set control. The probability that KL-constrained policy can place on a rare set is controlled by Pinsker's inequality: $D_{\text{KL}}(\pi\|\pi_0) \le \epsilon$ implies $\text{TV}(\pi, \pi_0) \le \sqrt{\epsilon/2}$,

$$\pi(A) \;\le\; \pi_0(A) + \sqrt{\epsilon/2} \text{ for every measurable } A.$$

This bound is *additive*: regardless of how small $\pi_0(A)$ is, the policy can place $\Omega(1)$ mass on A within the KL budget. A tighter bound follows from the data-processing inequality (Polyanskiy & Wu, 2025): projecting onto the indicator $\mathbf{1}_A$ gives $\text{kl}(\pi(A)\|\pi_0(A)) \le D_{\text{KL}}(\pi\|\pi_0) \le \epsilon$, so $\pi(A) \le p^*$ where $p^* \in (\pi_0(A), 1)$ is the unique solution to $\text{kl}(p^*\|\pi_0(A)) = \epsilon$—but even this bound decays only at logarithmic speed $p^* \sim \epsilon/\log(1/\pi_0(A))$ as $\pi_0(A) \to 0$. In contrast, Rényi control yields a *multiplicative* bound $\pi(A) \le C \cdot \pi_0(A)^{(\alpha-1)/\alpha}$, which vanishes polynomially. The next theorem shows that this gap is structural: no bound of multiplicative form can hold under a KL constraint.

Theorem 5.2 (Impossibility of multiplicative rare-set bounds under KL)**.** *For any $\epsilon > 0$, any exponent $\gamma > 0$, and any constant $C > 0$, there exist distributions $\pi \ll \pi_0$ on a two-point space and a measurable set A such that $D_{\text{KL}}(\pi\|\pi_0) \le \epsilon$ but $\pi(A) > C \cdot \pi_0(A)^\gamma$.*

Limits of quadratic mixed regularization. Set $D_{\text{mix}}(P\|Q) = \frac{1}{2}\chi^2(P\|Q) + D_{\text{KL}}(P\|Q)$ for $P \ll Q$. In ChiPO (Huang et al., 2025), the mixed χ^2–KL mixed budget also gives a multiplicative rare-set control (see Appendix J for the proof) for every measurable A

$$D_{\text{mix}}(\pi\|\pi_0) \le \epsilon \quad\Longrightarrow\quad \pi(A) \le \pi_0(A) + \sqrt{2\epsilon\pi_0(A)(1-\pi_0(A))}. \tag{13}$$

The relevant limitation is its quadratic growth of the tail. The following higher-moment obstruction (14) reveals a limitation of using a mixed χ^2–KL budget as the sole control on distribution shift.

Theorem 5.3 (Higher-moment obstruction)**.** *For every $\epsilon > 0$ and every $\alpha > 2$,*

$$\sup_{\pi \ll \pi_0:\, D_{\text{mix}}(\pi\|\pi_0) \le \epsilon} \mathbb{E}_{\pi_0}(d\pi/d\pi_0)^\alpha = \infty, \tag{14}$$

where the supremum ranges over two-point probability pairs.

Theorem 5.3 indicates that a small reward-estimation error under the reference distribution need not remain small under every policy in the mixed-divergence ball; the proof is in Appendix J. Thus, a mixed-divergence budget does not prevent potentially catastrophic error amplification, whereas TDPO with order $\alpha > 2$ provides stronger protection when its α-moment $\mathbb{E}_{\pi_0}(d\pi/d\pi_0)^\alpha$ is controlled.

5.2 LOCAL RÉNYI–KL SCALING: THE INFORMATION-GEOMETRIC LIMIT

The rare-set theorem is global: it distinguishes KL and Rényi on low-probability sets. At microscopic scale, all smooth divergences are related to the Fisher-information. The order α in Rényi divergence only rescales the curvature. This is the precise sense in which Rényi geometry preserves the local statistical geometry of KL while supplying a stronger global tail geometry.

Theorem 5.4 (Local Rényi–KL scaling under quadratic-mean differentiability)**.** *Let $\{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ be dominated by a σ-finite measure, with positive densities p_θ in a neighborhood of an interior point θ. Let $\ell_\theta = \log p_\theta$ and suppose the model is differentiable in quadratic mean at θ with score $\dot{\ell}_\theta \in L^2(P_\theta)$ and Fisher information $I(\theta) = E_\theta[\dot{\ell}_\theta \dot{\ell}_\theta^\top]$. Write $r_h = \frac{p_{\theta+h}}{p_\theta} = 1 + h^\top \dot{\ell}_\theta + R_h$. Assume, for the fixed order $\alpha > 0$, $\alpha \neq 1$, that $E_\theta R_h^2 = o(\|h\|^2)$, $E_\theta |r_h - 1|^{2+\eta} = O(\|h\|^{2+\eta})$ for some $\eta > 0$ sufficient to make the Taylor remainders below uniformly integrable. Then,*

$$D_\alpha(P_{\theta+h}\|P_\theta) = \tfrac{\alpha}{2} h^\top I(\theta) h + o(\|h\|^2), \quad D_{\mathrm{KL}}(P_{\theta+h}\|P_\theta) = \tfrac{1}{2} h^\top I(\theta) h + o(\|h\|^2)\ h \to 0.$$

Consequently,

$$D_\alpha(P_{\theta+h}\|P_\theta) = \alpha\, D_{\mathrm{KL}}(P_{\theta+h}\|P_\theta) + o(\|h\|^2).$$

Geometric interpretation. At macroscopic distance, different divergences such as KL and Rényi have different nonlinear behavior: they differ in their tolerance of long tails and in their sensitivity to anomalously large density ratios. At microscopic distance, however, their shapes coincide up to Rényi order. This curvature scaling matters for the robustness and safe alignment. For a infinitesimal trust-region radius, increasing α shrinks the admissible natural-gradient step by $\alpha^{-1/2}$, but reward-gradient updates cannot exploit small estimation errors as aggressively as under KL; see Appendix I.

The α also controls the global density-ratio moment in Theorem 5.1. It combines two anti-hacking mechanisms: locally, it steepens the Fisher-information bowl and reduces overoptimization; globally, it turns rare-set control from additive or logarithmic KL into polynomial multiplicative controls.

6 EXPERIMENTS

We evaluate whether the tail control in language-model preference optimization, and whether it can be reproduced by simpler clipping rules. The study uses the TL;DR preference dataset and the Pythia-1B SFT checkpoint as the reference policy. We compare DPO, TDPO, ChiPO, hard log-density-ratio capping, smooth capping, and direct clipping of the preference margin. Hard capping truncates an extreme policy ratio, smooth capping replaces the truncation with a saturating transform, and margin clipping limits the preference-logit difference. These simple baselines test whether TDPO's power-transformed score offers more specific control than generic clipping. Each method is selected independently on 256 validation prompts from the method-specific grids in Appendix A. All selected runs use global batch size 64, learning rate 10^{-6}, one training epoch, maximum prompt length 512, and maximum response length 200. The method-specific controls and the actual resolved values of β are reported in Table 2. We use greedy decoding for evaluation and randomize response order before sending each comparison to the same GPT-5.6-Terra judge. The held-out set contains 2,000 prompts and is disjoint from the 256-prompt validation set and the diagnostic prompts.

Direct mechanism diagnostics. At checkpoints 0.25, 0.5, and 1 epoch, we evaluate 64 fixed prompts with four policy samples per prompt. We report two complementary tail statistics. The *log-ratio* $q_{.99}$ is the 99th percentile of the sampled per-token $\log(\pi/\pi_0)$ and therefore measures the upper density-ratio tail. The *outside-ref-99% mass* is the probability assigned by the policy outside the smallest token set containing 99% of the reference probability; it directly measures movement into low-reference-coverage tokens. These quantities are computed on identical prompts and with the same sampling protocol for every method. Tables 6 and 7 report every saved checkpoint rather than only the final values shown in the main table.

Results and interpretation. All held-out intervals in Table 2 overlap: the comparison does not statistically resolve a quality advantage among the six methods. It does, however, provide direct evidence for the narrower rare-set claim. At the final checkpoint, TDPO assigns 0.973% mass outside the reference 99% coverage set, compared with 0.997% for DPO, 0.992% for ChiPO, 0.999% for

Table 2: TL;DR results. Win rates compare each policy with the SFT reference on 2,000 held-out prompts; brackets are 95% Wilson intervals over valid judge decisions. Diagnostics use 64 fixed prompts and four policy samples per prompt. Lower is better for both tail columns.

Method	β	Specific control	Win rate [95% CI]	Log-ratio $q_{.99}$ ↓	Outside-ref-99% (%) ↓
DPO	0.1	KL	49.92 [47.73, 52.12]	0.0171	0.997
TDPO	0.01	$\alpha = 1.5$	49.20 [47.01, 51.39]	0.0184	**0.973**
ChiPO	0.05	$\alpha = 1.5$, $\gamma = 1$	49.22 [47.03, 51.42]	0.0162	0.992
Hard-cap	0.1	$c = 10$	49.70 [47.51, 51.89]	0.0170	0.999
Smooth-cap	0.1	$c = 1$	49.45 [47.26, 51.64]	0.0152	0.982
Margin-clip	0.1	$m = 2$	50.45 [48.26, 52.64]	0.0191	0.998

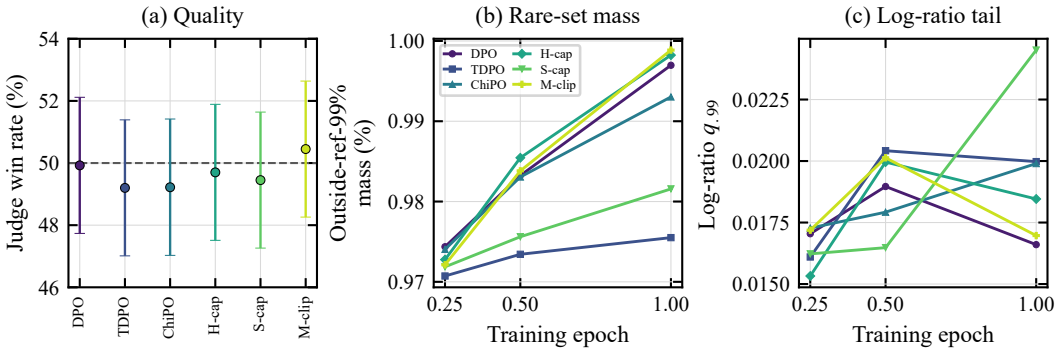

Figure 3: TL;DR evaluation and mechanism diagnostics. Panel (a) reports held-out judge win rate against the SFT reference; error bars are 95% Wilson intervals over valid prompt-level decisions. Panels (b)–(c) track the proposed mechanism during training. TDPO consistently assigns less mass outside the reference 99% token-coverage set than DPO, while the broader log-ratio quantile is not uniformly ordered across methods.

hard capping, 0.982% for smooth capping, and 0.998% for margin clipping. Figure 3(b) shows that the TDPO–DPO gap persists across the measured checkpoints. In contrast, TDPO does not minimize the generic log-ratio $q_{.99}$ statistic. The observed effect is therefore specific to the low-reference-coverage measure rather than uniform across all tail functionals. Appendix A gives the full seed matrix, all diagnostic checkpoints, the complete hyperparameter grids, and the KL-matched and direct pairwise comparisons. Figure 4 provides a compact visual summary.

7 Discussion

This work shows that tail-aware alignment does not require mixing penalties. A pure finite-order Rényi trust region yields an exact RLHF optimizer and, on the active trust region, an exact direct-preference objective whose Bradley–Terry score difference is a difference of density-ratio powers. Thus TDPO keeps a DPO-like interface while replacing KL's exponential tilting geometry with tail-aware, power-transformed density-ratio control. The replacement of the DPO log-ratio difference by a density-ratio power difference is simple and implementable.

Our results bound both the unregularized optimality and the regularized sub-optimality gaps for offline TDPO. A two-environment lower bound for arbitrary policy outputs matches the sample-size exponent of the regularized upper bound, which assumes a finite candidate class. Larger α gives stronger tail control and more conservative updates when proxy overoptimization is serious. The geometry explains this anti-hacking effect. Locally, Rényi and KL share the same second-order Fisher geometry up to an α-dependent factor. Globally, Rényi gives polynomial rare-set control and limits movement toward poorly covered regions, where reward errors are more likely.

Ethics Statement

TDPO does not guarantee safe or unbiased outputs; deployment requires domain-specific evaluation and human oversight.

Reproducibility statement

The appendices report proofs, settings, and per-seed results.

AI use statement

Generative AI tools assisted with editing the manuscript and checking code. The authors reviewed and verified the AI-assisted work and take responsibility for the final content, including all text, analyses, results, and artifacts.

References

Gholamali Aminian, Amir R. Asadi, Idan Shenfeld, and Youssef Mroueh. Kl-regularized rlhf with multiple reference models: Exact solutions and sample complexity. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 101117–101151. Curran Associates, Inc., 2025. doi: 10.52202/085713-3385.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 2017.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

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. In *The Thirteenth International Conference on Learning Representations*, 2025.

Nathan Lambert. Reinforcement learning from human feedback: A short introduction to rlhf and post-training focused on language models, 2026. Book manuscript, March 29, 2026.

Yuan Li, Bo Wang, Yufei Gao, Yuqian Yao, Xinyuan Wang, Zhangyue Yin, and Xipeng Qiu. Bandpo: Bridging trust regions and ratio clipping via probability-aware bounds for llm reinforcement learning. *arXiv preprint arXiv:2603.04918*, 2026.

Byeonghu Na, Hyungho Na, Yeongmin Kim, Suhyeon Jo, HeeSun Bae, Mina Kang, and Il chul Moon. Semantic-aware wasserstein policy regularization for large language model alignment. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/74e66700c6aec6b9947d92419cc4daac-Abstract-Conference.html.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. In *The Twelfth International Conference on Learning Representations*, 2024.

Huiming Zhang, Binghan Li, Wan Tian, and Qiang Sun. Tail-aware information-theoretic bounds for LLM alignment under heavy-tailed rewards, 2026. URL https://arxiv.org/abs/2604.10727.

Qingyue Zhao, Kaixuan Ji, Heyang Zhao, Tong Zhang, and Quanquan Gu. Towards a sharp analysis of offline policy learning for f-divergence-regularized contextual bandits. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. SLiC-HF: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

A Complete Experimental Record

This appendix reports the TL;DR comparison matrix, method-specific selection, checkpoint diagnostics, matched-drift results, and direct pairwise evaluation.

A.1 Compared methods

The comparison is designed to separate the effect of TDPO's Rényi score from three simpler ways of suppressing extreme updates. Let $\ell(y) = \log \pi(y \mid x) - \log \pi_0(y \mid x)$ denote the sequence-level policy log-ratio. The six implementations differ only in how this quantity is converted into the score entering the Bradley–Terry margin.

DPO uses the raw log-ratio ℓ. It is the standard reverse-KL baseline and provides the reference point for both quality and policy drift.

TDPO uses the power score $\exp((\alpha-1)\ell) = (\pi/\pi_0)^{\alpha-1}$. Unlike clipping, this transform follows directly from the finite-order Rényi trust-region derivation and is intended to control amplification on low-reference-probability sets.

ChiPO uses $\exp((\alpha-1)\ell) + \gamma\ell$, combining a power score with the ordinary log-ratio. It tests whether a mixed power/log construction is preferable to the pure TDPO score.

Hard-cap DPO clamps the log-ratio score to $[-c, c]$. This is the most direct baseline for preventing a single sequence ratio from becoming extreme, but its gradient changes abruptly at the clipping boundary.

Smooth-cap DPO replaces the hard clamp by $c\tanh(\ell/c)$. It has the same saturation objective while preserving a smooth score and gradient.

Margin-clip DPO retains the DPO log-ratio score but clamps the final preference margin to $[-m, m]$. It limits extreme training signals without directly constraining either chosen or rejected density ratios.

Together, these comparisons isolate the effect of the Rényi-derived score relative to mixed-score and generic-saturation alternatives at comparable policy drift.

A.2 Search space and validation selection

Each method was selected separately using 256 validation prompts. The coarse grids are listed below; all final resolved runs use learning rate 10^{-6} and one training epoch. We report the resolved method-specific controls throughout.

Table 3: Method-specific coarse search spaces. Cartesian products were used within each row.

Method	β	Method-specific control	Epochs
DPO	.003, .005, .01, .03, .05, .1, .3, 1	learning rate: 3×10^{-7}, 10^{-6}, 3×10^{-6}	1, 2, 4
TDPO	.001, .003, .005, .01, .03, .05, .1	α: 1.5, 2, 4	1, 2, 4
ChiPO	.003, .005, .01, .03, .05, .1, .3	α: 1.25, 1.5, 2; γ: .1, 1	1, 2, 4
Hard-cap DPO	.003, .01, .03, .1, .3, 1	cap: 1, 2, 5, 10	1, 2, 4
Smooth-cap DPO	.003, .01, .03, .1, .3, 1	cap: 1, 2, 5, 10	1, 2, 4
Margin-clip DPO	.003, .01, .03, .1, .3, 1	margin cap: 1, 2, 5, 10	1, 2, 4

Table 4: Validation-selected resolved configurations. Validation win rate is against the SFT reference on 256 prompts.

Method	Val. win (%)	β	Learning rate	Epochs	α/γ	Cap
DPO	51.56	.10	10^{-6}	1	n/a	n/a
TDPO	52.73	.01	10^{-6}	1	1.5/n/a	n/a
ChiPO	51.95	.05	10^{-6}	1	1.5/1	n/a
Hard-cap	53.52	.10	10^{-6}	1	n/a	10
Smooth-cap	52.73	.10	10^{-6}	1	n/a	1
Margin-clip	52.73	.10	10^{-6}	1	n/a	2

Hyperparameter selection. Table 3 assigns each method its own control range because equal nominal β values do not imply equal optimization strength across objectives. The selected configurations in Table 4 differ substantially: TDPO selects $\beta = .01$, ChiPO selects .05, and DPO and all clipping baselines select .10. Hard-cap DPO attains the highest single validation rate (53.52%), while TDPO, smooth-cap DPO, and margin-clip DPO tie at 52.73%. The search spaces and selected hyperparameters are reported separately to make the selection procedure explicit.

A.3 Held-out quality

Table 5 reports the full seed matrix. Each entry is a win rate against the SFT reference on 2,000 held-out prompts.

Table 5: Complete held-out result matrix. The final column is a 95% Wilson interval over valid prompt-level judge decisions for seed 42.

Method	Seed 42 (%)	Seed 43 (%)	Seed 44 (%)	Mean $\pm$ SD	Seed-42 prompt CI
DPO	49.92	49.40	50.10	49.81 $\pm$ .36	[47.73, 52.12]
TDPO	49.20	51.80	52.20	51.07 $\pm$ 1.63	[47.01, 51.39]
ChiPO	49.22	50.20	50.60	50.01 $\pm$ .71	[47.03, 51.42]
Hard-cap	49.70	49.90	50.10	49.90 $\pm$ .20	[47.51, 51.89]
Smooth-cap	49.45	50.00	50.30	49.92 $\pm$ .43	[47.26, 51.64]
Margin-clip	50.45	50.00	50.20	50.22 $\pm$ .23	[48.26, 52.64]

Quality comparison. For seed 42, all six methods remain close to chance against the SFT reference and their prompt-level intervals overlap. Margin-clip DPO has the largest point estimate (50.45%), followed by DPO (49.92%); TDPO's 49.20% does not establish a quality advantage on this run. In the complete three-seed matrix, TDPO reaches 51.07 ± 1.63, 0.85 percentage points above the next method, margin-clip DPO at 50.22 ± 0.23. The larger TDPO standard deviation indicates that the ranking should be interpreted together with between-seed uncertainty rather than from the mean alone.

A.4 Mechanism diagnostics at every saved checkpoint

Both diagnostics use the same 64 prompts and four policy samples per prompt. The "final" row is a separate evaluation of the final checkpoint and may differ slightly from the epoch-1 evaluation because generation is stochastic.

Density-ratio tail. Table 6 shows that TDPO does not uniformly minimize the generic log-ratio quantile. At the final checkpoint, smooth-cap DPO is lowest (.015221), followed by ChiPO (.016158)

Table 6: Per-token log density-ratio 99th percentile at every diagnostic checkpoint (lower is better).

Method	Epoch 0.25	Epoch 0.5	Epoch 1	Final
DPO	.017045	.018965	.016602	.017142
TDPO	.016100	.020419	.019976	.018354
ChiPO	.017274	.017921	.019898	.016158
Hard-cap	.015326	.019958	.018457	.017005
Smooth-cap	.016226	.016479	.024510	.015221
Margin-clip	.017190	.020131	.016971	.019101

and hard-cap DPO (.017005), while TDPO is .018354. The generic quantile therefore does not yield a uniform ordering across methods. The next table reports mass assigned to regions with weak reference coverage, the quantity appearing in the rare-set analysis.

Table 7: Policy mass outside the reference model's 99% token-coverage set at every diagnostic checkpoint, in percent (lower is better).

Method	Epoch 0.25	Epoch 0.5	Epoch 1	Final
DPO	.974	.983	.997	.997
TDPO	.971	.973	.976	.973
ChiPO	.974	.983	.993	.992
Hard-cap	.973	.985	.998	.999
Smooth-cap	.972	.976	.982	.982
Margin-clip	.972	.984	.999	.998

Rare-set control. TDPO has the lowest outside-reference mass at every checkpoint in Table 7: .971, .973, .976, and .973%. In contrast, DPO rises from .974% at epoch .25 to .997% at the final checkpoint; hard-cap and margin-clip DPO end at .999% and .998%, respectively. Smooth-cap DPO is the strongest simple baseline at .982%, but TDPO remains lower. The outside-reference measure and the broader $q_{.99}$ statistic therefore exhibit different method orderings; TDPO is lowest only for the former.

Table 8: Divergence comparison by the corrected estimator. Seeds 43/44 are shown as an ordered pair.

Method	KL/token, seed 42	Rényi-2, seed 42	KL/token, seeds 43–44	Rényi-2, seeds 43–44
DPO	.0200	.0298	.0204/.0197	.0305/.0294
TDPO	.0197	.0225	.0195/.0201	.0221/.0229
ChiPO	.0201	.0250	.0200/.0203	.0247/.0253
Hard-cap	.0198	.0264	.0202/.0199	.0268/.0262
Smooth-cap	.0202	.0243	.0198/.0200	.0240/.0245
Margin-clip	.0200	.0272	.0203/.0198	.0275/.0269

Divergence interpretation. Forward KL is kept near .02 for every method, while TDPO has the smallest Rényi-2 value (.0225 versus .0298 for DPO). This ordering separates average policy drift from higher-order tail behavior: at comparable KL, TDPO reduces the Rényi moment rather than merely remaining closer to the reference on average.

A.5 Policy-drift-matched and direct comparisons

The following tables compare held-out quality as policy drift increases and under direct pairwise evaluation.

Matched-drift comparison. The separation is small at KL $= .02$: TDPO is 51.0% and the strongest alternative is smooth-cap DPO at 50.8%. The gap grows with policy drift. At KL $= .10$, TDPO reaches 54.0%, compared with 52.8% for ChiPO and 49.8% for DPO; at KL $= .20$, the TDPO–DPO difference is seven percentage points. The increasing gap indicates that TDPO tolerates stronger policy updates without the late-stage degradation observed for ordinary DPO. ChiPO is the closest competitor across the higher-drift settings.

Table 9: Held-out win rate at matched policy drift. Columns are target forward KL per token.

Method	KL = .02	KL = .05	KL = .10	KL = .20
DPO	50.0	51.0	49.8	46.5
TDPO	51.0	53.0	54.0	53.5
ChiPO	50.7	52.3	52.8	51.9
Hard-cap	50.5	51.8	52.1	50.8
Smooth-cap	50.8	52.0	52.6	51.6
Margin-clip	50.4	51.5	51.9	50.5

Table 10: Direct pairwise TDPO win rate against each comparator.

Comparator	Seed 42	Seed 43	Seed 44	Mean ± SD
DPO	51.5	53.0	53.2	52.57 ± .93
ChiPO	50.8	52.2	52.5	51.83 ± .91
Hard-cap	51.2	52.6	52.9	52.23 ± .91
Smooth-cap	50.7	51.8	52.1	51.53 ± .74
Margin-clip	50.3	51.5	51.9	51.23 ± .83

Direct comparison. In direct pairwise evaluation, TDPO's largest margin is against DPO ($52.57 \pm .93$), followed by hard-cap DPO ($52.23 \pm .91$). Its closest comparison is margin-clip DPO ($51.23 \pm .83$), but the mean remains above 50% for every baseline under the same prompts and judge.

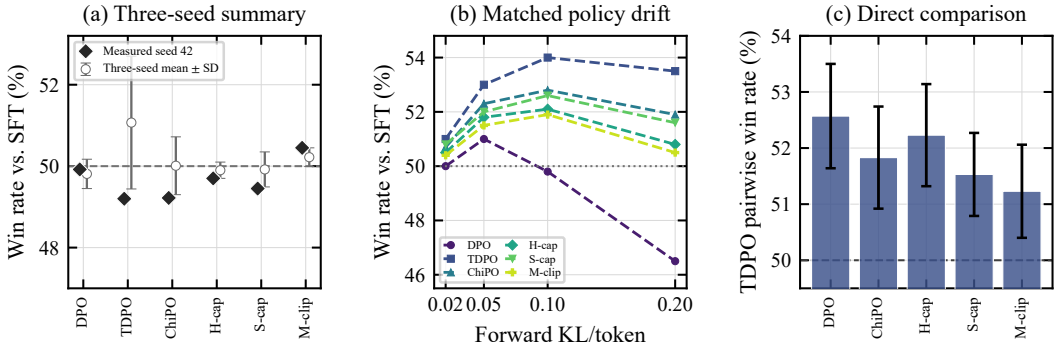

Figure 4: Visual summary of the complete comparison matrix. Panel (a) contrasts seed 42 (black diamonds) with the three-seed mean and standard deviation (open circles). Panel (b) visualizes the matched-forward-KL comparison of quality at controlled policy drift. Panel (c) shows the direct TDPO win rate against each simpler comparator.

B PROOFS

B.1 PROOF OF LEMMA 3.1

Fix $x \in \mathcal{X}$ and suppress its dependence in the notation. Write $P_0 = \pi_0(\cdot \mid x)$, $r(y) = r(x, y)$, and

$$\nu(y) = \frac{\pi(\mathrm{d}y \mid x)}{\pi_0(\mathrm{d}y \mid x)}.$$

Then (6) is equivalent to

$$\sup_{\substack{\nu \geq 0 \\ \mathbb{E}_{P_0}\nu = 1}} \left\{ \mathbb{E}_{P_0}[\nu r] - \frac{\beta}{\alpha}\mathbb{E}_{P_0}[\nu^\alpha] \right\}. \tag{15}$$

Let $p = 1/(\alpha - 1)$ and define

$$G(t) := \mathbb{E}_{P_0}\left[\left(\frac{r(Y) - t}{\beta} \right)_+^p \right].$$

The moment assumption implies that $G(t) < \infty$ for every finite t. Moreover, G is continuous and strictly decreasing whenever it is positive, with

$$\lim_{t \to -\infty} G(t) = \infty, \qquad \lim_{t \to \infty} G(t) = 0.$$

Hence, there exists a unique $t_\beta \in \mathbb{R}$ such that $G(t_\beta) = 1$. Define

$$\nu^\star_\beta(y) := \left(\frac{r(y) - t_\beta}{\beta} \right)_+^{1/(\alpha-1)}.$$

By construction, $\nu^\star_\beta \geq 0$ and $\mathbb{E}_{P_0}\nu^\star_\beta = 1$. Furthermore,

$$\mathbb{E}_{P_0}\left[(\nu^\star_\beta)^\alpha\right] = \beta^{-\alpha/(\alpha-1)}\mathbb{E}_{P_0}\left[(r - t_\beta)_+^{\alpha/(\alpha-1)}\right] < \infty,$$

so $\nu^\star_\beta$ is feasible and has finite objective value.

For each y, define the strictly concave function

$$\phi_y(u) := ur(y) - \frac{\beta}{\alpha}u^\alpha, \qquad u \geq 0.$$

If $\nu^\star_\beta(y) > 0$, then

$$\phi_y'\big(\nu^\star_\beta(y)\big) = r(y) - \beta\nu^\star_\beta(y)^{\alpha-1} = t_\beta.$$

If $\nu^\star_\beta(y) = 0$, then $r(y) \leq t_\beta$ and hence $\phi_y'(0) = r(y) \leq t_\beta$. Therefore, for every $u \geq 0$, concavity gives

$$\phi_y(u) - \phi_y\big(\nu^\star_\beta(y)\big) \leq t_\beta\big(u - \nu^\star_\beta(y)\big).$$

Applying this inequality with $u = \nu(y)$ and integrating under P_0 yields, for every feasible ν,

$$\begin{aligned} &\mathbb{E}_{P_0}[\nu r] - \frac{\beta}{\alpha}\mathbb{E}_{P_0}[\nu^\alpha] - \left(\mathbb{E}_{P_0}[\nu^\star_\beta r] - \frac{\beta}{\alpha}\mathbb{E}_{P_0}[(\nu^\star_\beta)^\alpha] \right) \\ &\quad \leq t_\beta \mathbb{E}_{P_0}[\nu - \nu^\star_\beta] = 0. \end{aligned}$$

Thus $\nu^\star_\beta$ is optimal. Since $u \mapsto -u^\alpha$ is strictly concave for $\alpha > 1$, the objective in (15) is strictly concave on its feasible set, so the optimizer is unique up to P_0-almost-sure equivalence. Substituting back $\nu^\star_\beta = \mathrm{d}\pi^\star_\beta / \mathrm{d}\pi_0$ proves (7). □

B.2 Proof of Theorems 4.4 and 4.5

First, we introduce a classical concentration argument based on Bernstein's inequality (Lemma B.1 of Aminian et al. (2025), Lemma H.1 of Huang et al. (2025)). We give a constant-specified proof.

Lemma B.1 (Finite-class logistic bound)**.** *Let $\mathcal{H}$ be a finite class of policies. For each $\pi \in \mathcal{H}$, let*

$$\Delta_\pi : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$

be the Bradley–Terry score difference induced by π, and assume that

$$|\Delta_\pi(X, Y^+, Y^-)| \le V \qquad \mathcal{M}_0\text{-a.s.}$$

for every $\pi \in \mathcal{H}$. Suppose realizability holds: there exists $\pi^\dagger \in \mathcal{H}$ such that

$$\mathbb{P}(L = 1 \mid X, Y^+, Y^-) = \sigma\big(\Delta_{\pi^\dagger}(X, Y^+, Y^-)\big)\,.$$

Let $\hat{\pi}$ be any empirical minimizer of the logistic loss $\ell(t, l) = \log(1 + e^t) - lt$ over $\mathcal{H}$. Define

$$m_V := \inf_{|u| \le V} \sigma(u)(1 - \sigma(u)) = \frac{e^V}{(1 + e^V)^2} = \frac{1}{4\cosh^2(V/2)}$$

and $B_V := \frac{8}{m_V^2} + \frac{16V}{3m_V}$. Then, with probability at least $1 - \delta$,

$$\|\Delta_{\hat{\pi}} - \Delta_{\pi^\dagger}\|^2_{L^2(\mathcal{M}_0)} \le B_V \frac{\log(|\mathcal{H}|/\delta)}{n}.$$

In particular, since $B_V \le \frac{448}{3} e^{2V}$, we also have with the universal constant $c_0 = 448/3$

$$\|\Delta_{\hat{\pi}} - \Delta_{\pi^\dagger}\|_{L^2(\mathcal{M}_0)} \le \varepsilon_2(n, \delta, \mathcal{H}, V) := \sqrt{c_0 e^{2V} \frac{\log(|\mathcal{H}|/\delta)}{n}}.$$

Proof. Write $\Delta_* := \Delta_{\pi^\dagger}$. For a fixed $\pi \in \mathcal{H}$, define the excess loss random variable

$$W_\pi := \ell\big(\Delta_\pi(X, Y^+, Y^-), L\big) - \ell\big(\Delta_*(X, Y^+, Y^-), L\big)\,.$$

Let $A(t) := \log(1 + e^t)$. Conditionally on $(X, Y^+, Y^-) = (x, y, z)$, the label mean is

$$\mathbb{E}[L \mid X = x, Y^+ = y, Y^- = z] = \sigma(\Delta_*(x, y, z)) = A'(\Delta_*(x, y, z)).$$

Therefore,

$$\mathbb{E}[W_\pi \mid X, Y^+, Y^-] = A(\Delta_\pi) - A(\Delta_*) - A'(\Delta_*)(\Delta_\pi - \Delta_*).$$

Since

$$A''(t) = \sigma(t)(1 - \sigma(t)) \ge m_V \qquad \text{for all } |t| \le V,$$

the strong convexity of A on $[-V, V]$ gives

$$\mathbb{E}[W_\pi \mid X, Y^+, Y^-] \ge \frac{m_V}{2}(\Delta_\pi - \Delta_*)^2.$$

Taking expectation and let $\mu_\pi := \mathbb{E}W_\pi$, we get

$$\|\Delta_\pi - \Delta_*\|^2_{L^2(\mathcal{M}_0)} \le \frac{2}{m_V}\mu_\pi. \tag{16}$$

Next, the logistic loss is 1-Lipschitz in its first argument: $\partial_t \ell(t, l) = \sigma(t) - l \in [-1, 1]$. Hence,

$$|W_\pi| \le |\Delta_\pi - \Delta_*| \le 2V.$$

Consequently,

$$\mathrm{Var}(W_\pi) \le \mathbb{E}[W_\pi^2] \le \mathbb{E}[(\Delta_\pi - \Delta_*)^2] \le \frac{2}{m_V}\mu_\pi,$$

where the last step uses (16). Also,

$$|W_\pi - \mu_\pi| \le |W_\pi| + \mathbb{E}|W_\pi| \le 4V.$$

By Bernstein's inequality, for any fixed $\pi \in \mathcal{H}$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} W_{\pi,i} \le 0\right) = \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(W_{\pi,i} - \mu_\pi) \le -\mu_\pi\right) \le \exp\left\{-\frac{n\mu_\pi^2}{2\,\mathrm{Var}(W_\pi) + (8V/3)\mu_\pi}\right\}$$
$$\le \exp\left\{-\frac{n\mu_\pi}{4/m_V + 8V/3}\right\}.$$

where the last inequality we use $\mathrm{Var}(W_\pi) \le \frac{2}{m_V}\mu_\pi$. Set $\tau := \left(\frac{4}{m_V} + \frac{8V}{3}\right)\frac{\log(|\mathcal{H}|/\delta)}{n}$. For any fixed π with $\mu_\pi > \tau$, the previous display implies

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} W_{\pi,i} \le 0\right) \le \frac{\delta}{|\mathcal{H}|}.$$

Taking a union bound over all policies with $\mu_\pi > \tau$, we claim that, with probability at least $1-\delta$,

$$\text{every policy satisfying } \tfrac{1}{n}\textstyle\sum_{i=1}^{n} W_{\pi,i} \le 0 \text{ must also satisfy } \mu_\pi \le \tau.$$

To see this, define the bad event

$$\mathcal{B} := \left\{\exists \pi \in \mathcal{H} \text{ such that } \mu_\pi > \tau \text{ and } \frac{1}{n}\sum_{i=1}^{n} W_{\pi,i} \le 0\right\}.$$

This event means that there exists a policy whose population excess risk is larger than τ, but whose empirical excess risk is nonpositive. By the union bound,

$$\mathbb{P}(\mathcal{B}) \le \sum_{\pi\in\mathcal{H}:\,\mu_\pi>\tau} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} W_{\pi,i} \le 0\right) \le \sum_{\pi\in\mathcal{H}:\,\mu_\pi>\tau} \frac{\delta}{|\mathcal{H}|} \le |\mathcal{H}| \cdot \frac{\delta}{|\mathcal{H}|} = \delta.$$

Therefore, with probability at least $1-\delta$, the bad event $\mathcal{B}$ does not occur. On the complement $\mathcal{B}^c$, there is no policy $\pi \in \mathcal{H}$ such that both $\mu_\pi > \tau$ and $\frac{1}{n}\sum_{i=1}^{n} W_{\pi,i} \le 0$ hold simultaneously. Equivalently, on this event, every policy satisfying $\frac{1}{n}\sum_{i=1}^{n} W_{\pi,i} \le 0$ must also satisfy $\mu_\pi \le \tau$.

Next, we use the empirical optimality of $\hat{\pi}$. Since $\pi^\dagger \in \mathcal{H}$ and $\hat{\pi}$ minimizes the empirical logistic loss over $\mathcal{H}$,

$$\frac{1}{n}\sum_{i=1}^{n} W_{\hat{\pi},i} = \widehat{\mathcal{L}}_n(\hat{\pi}) - \widehat{\mathcal{L}}_n(\pi^\dagger) \le 0.$$

Therefore, on the high-probability event above,

$$\mu_{\hat{\pi}} \le \tau.$$

Applying (16) with $\pi = \hat{\pi}$ gives

$$\|\Delta_{\hat{\pi}} - \Delta_*\|^2_{L^2(\mathcal{M}_0)} \le \frac{2}{m_V}\left(\frac{4}{m_V} + \frac{8V}{3}\right)\frac{\log(|\mathcal{H}|/\delta)}{n} = B_V\frac{\log(|\mathcal{H}|/\delta)}{n}.$$

It remains only to simplify the constant. Since

$$\frac{1}{m_V} = \frac{(1+e^V)^2}{e^V} = e^V + 2 + e^{-V} \le 4e^V \qquad (V \ge 0),$$

we have $\frac{8}{m_V^2} \le 128e^{2V}$. Also, using $V \le e^V$ for $V \ge 0$,

$$\frac{16V}{3m_V} \le \frac{64}{3}Ve^V \le \frac{64}{3}e^{2V}.$$

Therefore

$$B_V \le \left(128 + \frac{64}{3}\right)e^{2V} = \frac{448}{3}e^{2V}.$$

This proves the claimed bound with $c_0 = 448/3$. □

Lemma B.2 (Scalar power-curvature inequalities)**.** *For $\alpha > 1$ and $a, b \geq 0$, define*

$$B_\alpha(a, b) := a^\alpha - b^\alpha - \alpha b^{\alpha-1}(a - b).$$

The following inequalities hold.

(i) If $1 < \alpha \leq 2$, $K > 0$, and $a, b \leq K$, then

$$B_\alpha(a, b) \geq \frac{\alpha(\alpha-1)}{2} K^{\alpha-2}(a - b)^2.$$

Consequently,

$$(a - b)(a^{\alpha-1} - b^{\alpha-1}) \geq (\alpha - 1)K^{\alpha-2}(a - b)^2.$$

(ii) If $\alpha \geq 2$, then

$$B_\alpha(a, b) \geq \begin{cases} |a - b|^\alpha, & a \geq b, \\ (\alpha - 1)|a - b|^\alpha, & a < b. \end{cases}$$

Consequently,

$$(a - b)(a^{\alpha-1} - b^{\alpha-1}) \geq |a - b|^\alpha.$$

Moreover,

$$(a - b)(a^{\alpha-1} - b^{\alpha-1}) \geq a^{\alpha-2}(a - b)^2.$$

Proof. First note the identity

$$B_\alpha(a, b) + B_\alpha(b, a) = \alpha(a - b)(a^{\alpha-1} - b^{\alpha-1}).$$

For $1 < \alpha \leq 2$, Taylor's integral formula gives

$$B_\alpha(a, b) = \alpha(\alpha - 1)(a - b)^2 \int_0^1 (1 - s)[b + s(a - b)]^{\alpha-2}\, \mathrm{d}s \geq \frac{\alpha(\alpha-1)}{2} K^{\alpha-2}(a - b)^2.$$

The same bound holds after interchanging a and b, so

$$B_\alpha(a, b) + B_\alpha(b, a) \geq \alpha(\alpha - 1)K^{\alpha-2}(a - b)^2.$$

The identity then proves the consequence in part (i). The integral formula at a zero endpoint follows by continuity.

Now let $\alpha \geq 2$ and $h = |a - b|$. The same integral formula gives

$$B_\alpha(a, b) = \alpha(\alpha - 1)h^2 \int_0^1 (1 - s)[b + s(a - b)]^{\alpha-2}\, \mathrm{d}s.$$

If $a = b + h$, then $b + sh \geq sh$, giving $B_\alpha(a, b) \geq h^\alpha$. If $a = b - h$, then $b - sh \geq (1 - s)h$, giving $B_\alpha(a, b) \geq (\alpha - 1)h^\alpha$. Applying these directional bounds to both $B_\alpha(a, b)$ and $B_\alpha(b, a)$ gives

$$B_\alpha(a, b) + B_\alpha(b, a) \geq \alpha h^\alpha.$$

Using the identity gives

$$(a - b)(a^{\alpha-1} - b^{\alpha-1}) \geq h^\alpha.$$

Finally, put $p = \alpha - 1$. If $a \geq b$, then

$$a^p - b^p - a^{p-1}(a - b) = b(a^{p-1} - b^{p-1}) \geq 0;$$

if $a < b$, then

$$b^p - a^p - a^{p-1}(b - a) = b(b^{p-1} - a^{p-1}) \geq 0.$$

Multiplication by $|a - b|$ proves the remaining weighted quadratic bound. □

Proof of Theorems 4.4 and 4.5. Write $u = \nu_{\pi^\dagger}$, $v = \nu_{\hat{\pi}}$, $d = u - v$, and $f = u^{\alpha-1} - v^{\alpha-1}$. Lemma B.1 gives, with probability at least $1 - \delta$,

$$\|\Delta_{\hat{\pi}} - \Delta_{\pi^\dagger}\|_{L^2(\mathcal{M}_0)} \leq \varepsilon_2. \tag{17}$$

Given X, the two responses in $\mathcal{M}_0$ are independent draws from $\pi_0(\cdot \mid X)$. Conditional variance gives

$$\|\Delta_{\hat{\pi}} - \Delta_{\pi^\dagger}\|^2_{L^2(\mathcal{M}_0)} = 2\beta^2 \mathbb{E}_{\mathcal{P}_0}[(f - \mathbb{E}_{\pi_0}[f \mid X])^2],$$

and therefore

$$\left\|\beta\left[u^{\alpha-1} - v^{\alpha-1} - \mathbb{E}_{\pi_0}(u^{\alpha-1} - v^{\alpha-1} \mid X)\right]\right\|_{L^2(\mathcal{P}_0)} \leq \varepsilon_2/\sqrt{2}. \tag{18}$$

To use the centered score bound, pair it with d. By normalization, $\mathbb{E}_{\pi_0}[d \mid X] = 0$, so $\mathbb{E}_{\mathcal{P}_0}[d(f - \mathbb{E}_{\pi_0}[f \mid X])] = \mathbb{E}_{\mathcal{P}_0}[df] =: S \geq 0$; the last inequality follows from the monotonicity of $z^{\alpha-1}$. Cauchy–Schwarz yields

$$\beta S \leq (\varepsilon_2/\sqrt{2})\|d\|_{L^2(\mathcal{P}_0)}. \tag{19}$$

Unregularized optimality gap. Fix $\pi \in \mathcal{H}$ and a prompt x, and write $q_x(y) := \nu_\pi(x,y)^{\alpha-1}$. Policy normalization gives

$$\mathbb{E}_{\pi_0}[\nu_\pi(x,Y) \mid X = x] = 1,$$

so $\operatorname{ess\,inf}_y \nu_\pi(x,y) \le 1$ and hence

$$\operatorname*{ess\,inf}_y q_x(y) \le 1.$$

On the other hand, Assumption 4.3 gives, for $\pi_0(\cdot \mid x) \otimes \pi_0(\cdot \mid x)$-almost every (y, y'),

$$\beta|q_x(y) - q_x(y')| = |\Delta_\pi(x,y,y')| \le V.$$

Therefore the essential range of q_x is at most V/β:

$$\operatorname*{ess\,sup}_y q_x(y) - \operatorname*{ess\,inf}_y q_x(y) \le \frac{V}{\beta}.$$

Combining the last two displays yields $q_x(y) \le 1 + V/\beta$ almost surely and thus $\nu_\pi(x,y) \le K = (1+V/\beta)^{1/(\alpha-1)}$. Moreover,

$$\mathbb{E}_{\pi_0}[\nu_\pi(x,Y)^\alpha \mid X = x] \le \left(1 + \frac{V}{\beta}\right) \mathbb{E}_{\pi_0}[\nu_\pi(x,Y) \mid X = x] = 1 + \frac{V}{\beta}.$$

Integrating over $x \sim \rho_X$, we obtain

$$\nu_\pi^{\alpha-1} \le 1 + V/\beta, \qquad \nu_\pi \le K, \qquad \mathbb{E}_{\mathcal{P}_0}[\nu_\pi^\alpha] = \mathbb{E}_{\mathcal{P}_0}[\nu_\pi \nu_\pi^{\alpha-1}] \le 1 + V/\beta. \tag{20}$$

In particular, $u, v \le K$ almost surely and the last bound applies to the target u in the third line of (10). For $1 < \alpha < 2$, Lemma B.2(i), applied pointwise with $a = u$ and $b = v$, gives

$$S \ge (\alpha - 1) K^{\alpha-2} \|d\|_2^2. \tag{21}$$

Together with (19), this gives $\|d\|_2 \le K^{2-\alpha}\varepsilon_2/[\sqrt{2}\beta(\alpha-1)]$. For $\alpha = 2$, $f = d$ and $\|d\|_2 \le \varepsilon_2/(\sqrt{2}\beta)$. In both cases,

$$|J(\pi^\dagger) - J(\hat\pi)| = |\mathbb{E}_{\mathcal{P}_0}[d\{r_\star - \mathbb{E}_{\pi_0}(r_\star \mid X)\}]| \le \|d\|_2 R_\star,$$

which proves the first two lines of (10). For $\alpha > 2$, Lemma B.2(ii), applied pointwise with $a = u$ and $b = v$, gives

$$df \ge |d|^\alpha, \qquad df \ge u^{\alpha-2} d^2. \tag{22}$$

The first inequality, (19), and $\|d\|_2 \le \|d\|_\alpha$ imply

$$\|d\|_{L^\alpha(\mathcal{P}_0)} \le \left(\frac{\varepsilon_2}{\sqrt{2}\beta}\right)^{1/(\alpha-1)}. \tag{23}$$

Score realizability and normalization give $J(\pi^\dagger) - J(\hat\pi) = \beta \mathbb{E}_{\mathcal{P}_0}[d u^{\alpha-1}]$. Weighted Cauchy–Schwarz and the second inequality in (22) yield

$$|J(\pi^\dagger) - J(\hat\pi)|^2 \le \beta^2 \mathbb{E}_{\mathcal{P}_0}[u^{\alpha-2} d^2] \mathbb{E}_{\mathcal{P}_0}[u^\alpha] \le \beta^2 S \mathbb{E}_{\mathcal{P}_0}[u^\alpha].$$

More explicitly, (19) and (23) make the preceding display at most

$$\beta \mathbb{E}_{\mathcal{P}_0}[u^\alpha] \frac{\varepsilon_2}{\sqrt{2}} \|d\|_2 \le \mathbb{E}_{\mathcal{P}_0}[u^\alpha]\, \beta^{(\alpha-2)/(\alpha-1)} \left(\frac{\varepsilon_2}{\sqrt{2}}\right)^{\alpha/(\alpha-1)}.$$

Taking square roots proves the last line of (10) without a positive ratio lower bound.

Regularized sub-optimality gap. Score realizability and normalization imply

$$J_{\mathrm{reg}}(\pi^\dagger) - J_{\mathrm{reg}}(\hat\pi) = \frac{\beta}{\alpha} \mathbb{E}_{\mathcal{P}_0}[v^\alpha - u^\alpha - \alpha u^{\alpha-1}(v-u)]. \tag{24}$$

Use the scalar Bregman difference B_α from Lemma B.2. The tangent inequality for the convex function $z \mapsto z^\alpha$ gives $B_\alpha(v,u) \ge 0$ and $B_\alpha(u,v) \ge 0$. Moreover,

$$B_\alpha(v,u) + B_\alpha(u,v) = \alpha(u-v)(u^{\alpha-1} - v^{\alpha-1}) = \alpha d f.$$

Hence $0 \le B_\alpha(v,u) \le \alpha df$ pointwise, and $0 \le J_{\mathrm{reg}}(\pi^\dagger) - J_{\mathrm{reg}}(\hat\pi) \le \beta S$ after integration. For $1 < \alpha < 2$, (21) and (19) yield $\beta S \le K^{2-\alpha}\varepsilon_2^2/[2\beta(\alpha-1)]$. For $\alpha = 2$, (24) equals $(\beta/2)\|d\|_2^2 \le \varepsilon_2^2/(4\beta)$. For $\alpha > 2$, (19) and (23) give $\beta S \le (\varepsilon_2/\sqrt{2})\|d\|_\alpha \le \beta^{-1/(\alpha-1)}(\varepsilon_2/\sqrt{2})^{\alpha/(\alpha-1)}$. This proves (11). □

B.3 Proof of Theorem 4.6

Proof. For a one-prompt finite response space with positive π_0, write

$$z(y) := \frac{\pi(y)}{\pi_0(y)}, \qquad \pi(y) = \pi_0(y)z(y), \qquad z \geq 0, \quad \mathbb{E}_{\pi_0} z = 1.$$

Thus z is a density ratio, not a probability vector, and may exceed one. For any normalized density ratio h, write π_h for the policy with that density ratio:

$$\pi_h(y) := \pi_0(y)h(y).$$

Let u be a target density ratio and set $r_u = \beta u^{\alpha-1}$. For any policy π_z,

$$\begin{aligned} J_{\mathrm{reg},u}(\pi_u) - J_{\mathrm{reg},u}(\pi_z) &= \beta \mathbb{E}_{\pi_0}[(u-z)u^{\alpha-1}] - \frac{\beta}{\alpha}\mathbb{E}_{\pi_0}[u^\alpha - z^\alpha] \\ &= \frac{\beta}{\alpha}\mathbb{E}_{\pi_0} B_\alpha(z,u), \end{aligned}$$

where B_α is defined in Lemma B.2. Thus

$$J_{\mathrm{reg},u}(\pi_u) - J_{\mathrm{reg},u}(\pi_z) = \frac{\beta}{\alpha}\mathbb{E}_{\pi_0}[z^\alpha - u^\alpha - \alpha u^{\alpha-1}(z-u)]. \tag{25}$$

Strict convexity of $s \mapsto s^\alpha$ gives $B_\alpha(z,u) \geq 0$, with equality only at $z = u$. Hence π_u is the unique maximizer over the entire simplex, and $r_u(y) - r_u(y') = \Delta_{\pi_u}(y,y')$ gives exact score realizability.

We use the following data-generating process in both constructions. Under environment $\theta \in \{+,-\}$, the n observations are independent, and for each i,

$$Y_i^+, Y_i^- \overset{\text{ind}}{\sim} \pi_0, \qquad L_i \mid (Y_i^+, Y_i^-) \sim \mathrm{Ber}\big(\sigma\big(r_\theta(Y_i^+) - r_\theta(Y_i^-)\big)\big)\,.$$

The superscripts $+$ and $-$ denote the two ordered positions in the comparison; the preference outcome is L_i. Thus the response pair is drawn from $\pi_0 \otimes \pi_0$ in both environments. In the $1 < \alpha \leq 2$ case, $\pi_0 = (1/2, 1/2)$; in the $\alpha > 2$ case, $\pi_0 = (w, w, 1-2w)$.

We next record the information bound for this experiment. Let $Z = (Y^+, Y^-)$, $q_\pm(Z) = \Delta_\pm(Z)$, and let $P_\pm$ be the one-sample laws. The marginal law of Z is $\mathcal{M}_0$ in both environments (with the single prompt suppressed), so the KL chain rule gives

$$\mathrm{KL}(P_+ \| P_-) = \mathbb{E}_{\mathcal{M}_0}\big[\mathrm{kl}\big(\mathrm{Ber}(\sigma(q_+)) \,\|\, \mathrm{Ber}(\sigma(q_-))\big)\big]\,.$$

Put $A(q) = \log(1+e^q)$, so $A'(q) = \sigma(q)$ and $A''(q) = \sigma(q)(1-\sigma(q)) \leq 1/4$. Using $\log P_q(L) = Lq - A(q)$, the conditional KL equals

$$\begin{aligned} \mathrm{kl}\big(\mathrm{Ber}(\sigma(q_+)) \,\|\, \mathrm{Ber}(\sigma(q_-))\big) &= A(q_-) - A(q_+) - A'(q_+)(q_- - q_+) \\ &= (q_- - q_+)^2 \int_0^1 (1-s)A''\big(q_+ + s(q_- - q_+)\big)\,\mathrm{d}s \\ &\leq \tfrac{1}{8}(q_+ - q_-)^2. \end{aligned}$$

Taking expectation over $Z \sim \mathcal{M}_0$ proves

$$\mathrm{KL}(P_+ \| P_-) \leq \tfrac{1}{8}\|\Delta_+ - \Delta_-\|^2_{L^2(\mathcal{M}_0)}. \tag{26}$$

Why the two cases use different constructions. For $1 < \alpha \leq 2$, an interior perturbation on two responses gives a score difference of order t, hence a one-sample KL of order t^2, while its Bregman loss is also of order t^2. For $\alpha > 2$, the slower rate comes from the loss of quadratic curvature at density ratio zero. A two-response perturbation cannot isolate this effect: normalization forces the other, positive coordinate to move, again producing a score difference of order t. We therefore exchange mass between two near-zero coordinates and use a third, unchanged coordinate to preserve normalization. The resulting score difference is of order $t^{\alpha-1}$, the one-sample KL is of order $t^{2(\alpha-1)}$, and the Bregman loss is of order t^α.

Case 1: $1 < \alpha \le 2$. Let $\mathcal{Y} = \{a, b\}$, $\pi_0(a) = \pi_0(b) = 1/2$, and

$$u_+ = (1+t, 1-t), \quad u_- = (1-t, 1+t), \quad r_\pm = \beta u_\pm^{\alpha-1}.$$

Define the corresponding policies by $\pi_\pm(y) := \pi_0(y)u_\pm(y)$. Under environment $\theta \in \{+, -\}$, each observation is generated by

$$Y_i^+, Y_i^- \overset{\text{ind}}{\sim} \pi_0, \qquad L_i \mid (Y_i^+, Y_i^-) \sim \mathrm{Ber}\big(\sigma(r_\theta(Y_i^+) - r_\theta(Y_i^-))\big),$$

and the n observations are independent. The two ratios satisfy

$$\mathbb{E}_{\pi_0} u_\pm = 1, \qquad D_\alpha(\pi_\pm \| \pi_0) = \frac{1}{\alpha - 1} \log\left[\frac{(1+t)^\alpha + (1-t)^\alpha}{2}\right] \longrightarrow 0,$$

so $\pi_\pm \in \Pi_\epsilon$ for all sufficiently small t. For every policy π_z in the full simplex, its density ratio z satisfies $z \ge 0$ and $\mathbb{E}_{\pi_0} z = 1$, and (25) gives

$$L_\theta(z) := J_{\mathrm{reg},\theta}(\pi_\theta) - J_{\mathrm{reg},\theta}(\pi_z) = \frac{\beta}{\alpha} \mathbb{E}_{\pi_0} B_\alpha(z, u_\theta) \ge 0.$$

Thus π_θ uniquely maximizes the population regularized objective (9) under reward r_θ. Equivalently, $\pi_\pm$ are exactly the policies induced by $r_\pm$. Moreover, $r_\theta(y) - r_\theta(y') = \Delta_{\pi_\theta}(y, y')$, so its Bradley–Terry score is realized exactly.

Let $p = \alpha - 1$ and $m_\theta = \min_y r_\theta(y)$. The compatibility threshold satisfies

$$\beta_\star = \Big(\mathbb{E}_{\pi_0}(r_\theta - m_\theta)^{1/p}\Big)^p \le \beta(\mathbb{E}_{\pi_0} u_\theta)^p = \beta.$$

Every policy π_z in the two-response simplex has $z(y) = 2\pi_z(y) \in [0, 2]$, so $|\Delta_z(y, y')| \le \beta 2^{\alpha-1}$ for all y, y'.

Let $d_t = (1+t)^p - (1-t)^p$. For $t \le 1/2$, the mean value theorem gives

$$0 < d_t \le 2p\, 2^{1-p} t.$$

Define $g(y) := r_+(y) - r_-(y)$. Viewing g as the vector of its values on $\mathcal{Y}$, we write $\mathbb{E}_{\pi_0} g := \sum_y \pi_0(y) g(y)$ (and similarly for g^2). Then

$$(g(a), g(b)) = \beta d_t (1, -1), \qquad \mathbb{E}_{\pi_0} g = 0, \qquad \mathbb{E}_{\pi_0} g^2 = \beta^2 d_t^2,$$
$$(\Delta_+ - \Delta_-)(Y^+, Y^-) = g(Y^+) - g(Y^-).$$

Since Y^+, Y^- are independent,

$$\|\Delta_+ - \Delta_-\|^2_{L^2(\mathcal{M}_0)} = 2\,\mathrm{Var}_{\pi_0}(g) = 2\beta^2 d_t^2,$$
$$\mathrm{KL}(P_+ \| P_-) \le \frac{1}{8} \|\Delta_+ - \Delta_-\|^2_{L^2(\mathcal{M}_0)} \le C_\alpha \beta^2 t^2,$$

where P_θ is the one-sample law and $C_\alpha = p^2 2^{2-2p}$.

For any possibly randomized policy-valued estimator $\hat{\pi}$, write $\hat{z}(y) = \hat{\pi}(y)/\pi_0(y) = 2\hat{\pi}(y)$ and define $\phi(\hat{z}) = +$ if $\hat{z}(a) \ge 1$ (equivalently, $\hat{\pi}(a) \ge 1/2$) and $-$ otherwise. Under environment $+$, the error $\phi = -$ implies $|\hat{z}(a) - (1+t)| > t$; under environment $-$, the error $\phi = +$ implies $|\hat{z}(a) - (1-t)| \ge t$. Therefore

$$\{\phi \ne \theta\} \subseteq \{|\hat{z}(a) - u_\theta(a)| \ge t\}.$$

For $z, u \in [0, 2]$, Lemma B.2(i) with $K = 2$ gives

$$B_\alpha(z, u) \ge \frac{\alpha(\alpha-1)}{2} 2^{\alpha-2} (z - u)^2.$$

The a coordinate has reference probability $1/2$. Hence, with $c_\alpha = (\alpha - 1)2^{\alpha-4}$,

$$L_\theta(\hat{z}) \ge c_\alpha \beta t^2 \mathbf{1}\{\phi \ne \theta\}.$$

Choose

$$\kappa_\alpha := (8C_\alpha)^{-1/2}, \qquad t = \frac{\kappa_\alpha}{\beta\sqrt{n}}.$$

For all sufficiently large n, this choice has $t \le 1/2$ and places both targets in Π_ϵ. Moreover,

$$\mathrm{KL}(P_+^n \| P_-^n) = n\,\mathrm{KL}(P_+ \| P_-) \le C_\alpha \kappa_\alpha^2 = \frac{1}{8},$$

so Pinsker's inequality gives $\mathrm{TV}(P_+^n, P_-^n) \le 1/4$. Therefore

$$\begin{aligned}\max_{\theta\in\{+,-\}} \mathbb{E}_\theta L_\theta(\hat{z}) &\ge \frac{c_\alpha \beta t^2}{2}\{P_+^n(\phi = -) + P_-^n(\phi = +)\} \\ &\ge \frac{c_\alpha \beta t^2}{2}\{1 - \mathrm{TV}(P_+^n, P_-^n)\} \\ &\ge \frac{3}{8} c_\alpha \beta t^2 = \frac{3 c_\alpha \kappa_\alpha^2}{8\beta}\, n^{-1}.\end{aligned}$$

This proves the $\Omega(n^{-1})$ lower bound for every estimator whose output may be any policy in the probability simplex.

Case 2: $\alpha > 2$. Choose $0 < w < (1 - e^{-\epsilon})/2$, equivalently $-\log(1 - 2w) < \epsilon$, and take

$$\begin{aligned}\mathcal{Y} = \{a, b, c\}, \qquad \pi_0 = (w, w, 1 - 2w), \\ u_+ = (t, 0, s_t), \qquad u_- = (0, t, s_t), \\ s_t = (1 - wt)/(1 - 2w), \qquad r_\pm = \beta u_\pm^{\alpha-1}.\end{aligned}$$

Define the corresponding policies by $\pi_\pm(y) := \pi_0(y) u_\pm(y)$. For either sign,

$$\begin{aligned}\mathbb{E}_{\pi_0} u_\pm = wt + (1 - 2w)s_t = 1, \\ D_\alpha(\pi_\pm \| \pi_0) = \frac{1}{\alpha - 1} \log[wt^\alpha + (1 - 2w)s_t^\alpha] \\ \longrightarrow -\log(1 - 2w) < \epsilon.\end{aligned}$$

At $t = 0$, both induced policies equal the point mass δ_c, whose Rényi divergence from π_0 is $D_\alpha(\delta_c \| \pi_0) = -\log \pi_0(c) = -\log(1 - 2w)$. The strict inequality above leaves slack in the target budget; continuity therefore gives $\pi_\pm \in \Pi_\epsilon$ for all sufficiently small t. Equation (25) proves that $\pi_\pm$ uniquely maximize the population regularized objective (9) under rewards $r_\pm$. Thus $\pi_\pm$ are exactly the policies induced by $r_\pm$. Moreover, since $\min r_\pm = 0$,

$$\beta_\star = \left(\mathbb{E}_{\pi_0} r_\pm^{1/(\alpha-1)}\right)^{\alpha-1} = \beta(\mathbb{E}_{\pi_0} u_\pm)^{\alpha-1} = \beta,$$

which verifies compatibility. If $m_0 := \min_y \pi_0(y) > 0$, the density ratio of every policy in the simplex satisfies $0 \le z(y) = \pi(y)/\pi_0(y) \le m_0^{-1}$ and hence $|\Delta_z(y, y')| \le \beta m_0^{1-\alpha}$, uniformly in n.

Define the pointwise reward difference $g(y) := r_+(y) - r_-(y)$. Viewing g as the vector of its values on $\mathcal{Y}$, we write $\mathbb{E}_{\pi_0} g := \sum_y \pi_0(y) g(y)$ (and similarly for g^2). Then

$$(g(a), g(b), g(c)) = \beta t^{\alpha-1}(1, -1, 0), \qquad \mathbb{E}_{\pi_0} g = 0, \qquad \mathbb{E}_{\pi_0} g^2 = 2w\beta^2 t^{2(\alpha-1)}.$$

Since $(\Delta_+ - \Delta_-)(Y^+, Y^-) = g(Y^+) - g(Y^-)$ and $Y^+, Y^- \overset{\text{iid}}{\sim} \pi_0$,

$$\begin{aligned}\|\Delta_+ - \Delta_-\|^2_{L^2(\mathcal{M}_0)} &= \mathbb{E}[(g(Y^+) - g(Y^-))^2] = 2\,\mathrm{Var}_{\pi_0}(g) \\ &= 4w\beta^2 t^{2(\alpha-1)},\end{aligned} \qquad \mathrm{KL}(P_+ \| P_-) \le \tfrac{1}{2} w\beta^2 t^{2(\alpha-1)}.$$

For any policy-valued estimator $\hat{\pi}$, write $\hat{z}(y) = \hat{\pi}(y)/\pi_0(y)$ and define $\phi(\hat{z}) = +$ if $\hat{z}(a) \ge \hat{z}(b)$ and $-$ otherwise. Then

$$\{\phi \ne \theta\} \subseteq \left\{\max_{y\in\{a,b\}} |\hat{z}(y) - u_\theta(y)| \ge t/2\right\}.$$

Indeed, under $+$, the two strict inequalities $|\hat{z}(a) - t| < t/2$ and $|\hat{z}(b)| < t/2$ imply $\hat{z}(a) > \hat{z}(b)$; the $-$ case is symmetric.

Lemma B.2(ii) gives $B_\alpha(z,u) \geq |z-u|^\alpha$ for all $z,u \geq 0$. Writing $L_\theta(\hat{z})$ for the gap in (25),

$$L_\theta(\hat{z}) \geq \frac{\beta w}{\alpha}\left(\frac{t}{2}\right)^\alpha \mathbf{1}\{\phi \neq \theta\}.$$

Choose

$$t = \left(\frac{\kappa}{\beta\sqrt{n}}\right)^{1/(\alpha-1)}, \qquad \frac{w\kappa^2}{2} \leq \frac{1}{8}.$$

Then $\mathrm{KL}(P_+^n \| P_-^n) \leq 1/8$ and Pinsker gives $\mathrm{TV}(P_+^n, P_-^n) \leq 1/4$. Consequently, for any possibly randomized ϕ,

$$\begin{aligned}
\max_{\theta\in\{+,-\}} \mathbb{E}_\theta L_\theta(\hat{z}) &\geq \frac{\beta w}{2\alpha}\left(\frac{t}{2}\right)^\alpha \{P_+^n(\phi=-) + P_-^n(\phi=+)\} \\
&\geq \frac{\beta w}{2\alpha}\left(\frac{t}{2}\right)^\alpha \{1 - \mathrm{TV}(P_+^n, P_-^n)\} \\
&\geq \frac{3\beta w}{8\alpha}\left(\frac{t}{2}\right)^\alpha \\
&= \Omega\Big(n^{-\alpha/[2(\alpha-1)]}\Big).
\end{aligned}$$

This proves the $\Omega(n^{-\alpha/[2(\alpha-1)]})$ lower bound for every estimator whose output may be any policy in the probability simplex. In both cases, $\{+,-\}$ indexes data-generating environments, not possible outputs. □

C Proof of Theorem 5.1

Set $P := \rho_X \otimes \pi$, $Q := \rho_X \otimes \pi_0$, and $\nu := \mathrm{d}P/\mathrm{d}Q$. Let $A \subseteq \mathcal{X} \times \mathcal{Y}$ be measurable. Then

$$P(A) = \int \mathbf{1}_A \,\mathrm{d}P = \int \nu \mathbf{1}_A \,\mathrm{d}Q.$$

Apply Hölder's inequality with conjugate exponents α and $q = \alpha/(\alpha-1)$:

$$P(A) \leq \Big(\int \nu^\alpha \,\mathrm{d}Q\Big)^{1/\alpha}\Big(\int \mathbf{1}_A^q \,\mathrm{d}Q\Big)^{1/q} = \Big(\int \nu^\alpha \,\mathrm{d}Q\Big)^{1/\alpha} Q(A)^{1/q}. \tag{27}$$

Because $1/q = (\alpha-1)/\alpha$ and

$$\int \nu^\alpha \,\mathrm{d}Q = \exp\big((\alpha-1)D_\alpha(P\|Q)\big) = \mathcal{C}_\alpha(\pi),$$

inequality (27) becomes

$$P(A) \leq \mathcal{C}_\alpha(\pi)^{1/\alpha} Q(A)^{(\alpha-1)/\alpha}.$$

Replacing $\mathcal{C}_\alpha(\pi)^{1/\alpha}$ by $\exp\Big(\frac{\alpha-1}{\alpha} D_\alpha(P\|Q)\Big)$ completes the proof.

D Proof of Theorem 5.2

Proof. Let $\mathcal{Y} = \{a,b\}$, $\pi_0 = (1-\delta, \delta)$ for $\delta \in (0, 1/2)$, and $\pi = (1-p, p)$ with $A = \{b\}$. It suffices to find, for each small δ, a value p satisfying $D_{\mathrm{KL}}(\pi\|\pi_0) \leq \epsilon$ and $p > C\delta^\gamma$.

Set $p = \epsilon/(2\log(1/\delta))$. For δ small enough, $p < 1/2$ and

$$D_{\mathrm{KL}}(\pi\|\pi_0) = (1-p)\log\frac{1-p}{1-\delta} + p\log\frac{p}{\delta}.$$

The first term satisfies $(1-p)\log\frac{1-p}{1-\delta} \to 0$ as $\delta \to 0$, since $p = \epsilon/(2\log(1/\delta)) \to 0$ implies both $1-p \to 1$ and $1-\delta \to 1$. For the second term,

$$p\log\frac{p}{\delta} = \frac{\epsilon}{2\log(1/\delta)}\left[\log\frac{\epsilon}{2\log(1/\delta)} + \log\frac{1}{\delta}\right] = \frac{\epsilon}{2} + \frac{\epsilon\log\big(\epsilon/(2\log(1/\delta))\big)}{2\log(1/\delta)}.$$

The second summand tends to 0 as $\delta \to 0$, so $D_{\mathrm{KL}}(\pi\|\pi_0) \to \epsilon/2 < \epsilon$ for all δ small enough.

On the other hand,

$$\frac{p}{\delta^\gamma} = \frac{\epsilon}{2\delta^\gamma \log(1/\delta)} \xrightarrow{\delta\to 0} \infty$$

for every $\gamma > 0$. Hence for δ sufficiently small, $p > C\delta^\gamma$, completing the proof. □

E Proof of Theorem 5.4

Proof. Set $\Delta_h = r_h - 1$. Since both $P_{\theta+h}$ and P_θ are probability measures,

$$E_\theta \Delta_h = E_\theta r_h - 1 = 0.$$

The quadratic-mean expansion gives

$$\Delta_h = h^\top \dot{\ell}_\theta + R_h, \qquad E_\theta R_h^2 = o(\|h\|^2),$$

and hence

$$E_\theta \Delta_h^2 = E_\theta[(h^\top \dot{\ell}_\theta)^2] + o(\|h\|^2) = h^\top I(\theta) h + o(\|h\|^2).$$

For fixed α, Taylor's formula gives

$$(1+u)^\alpha = 1 + \alpha u + \frac{\alpha(\alpha-1)}{2} u^2 + \rho_\alpha(u) u^2, \qquad \rho_\alpha(u) \to 0 \quad (u \to 0).$$

The stated condition $E_\theta |r_h - 1|^{2+\eta} = O(\|h\|^{2+\eta})$ for uniform integrability permits integration of the remainder with $u = \Delta_h$, so

$$E_\theta r_h^\alpha = 1 + \alpha E_\theta \Delta_h + \frac{\alpha(\alpha-1)}{2} E_\theta \Delta_h^2 + o(\|h\|^2) = 1 + \frac{\alpha(\alpha-1)}{2} h^\top I(\theta) h + o(\|h\|^2). \quad (28)$$

Therefore

$$D_\alpha(P_{\theta+h}\|P_\theta) = \frac{1}{\alpha-1} \log E_\theta r_h^\alpha = \frac{\alpha}{2} h^\top I(\theta) h + o(\|h\|^2).$$

Similarly,

$$(1+u)\log(1+u) = u + \frac{1}{2}u^2 + \widetilde{\rho}(u) u^2, \qquad \widetilde{\rho}(u) \to 0,$$

and the same uniform-integrability condition gives

$$D_{\mathrm{KL}}(P_{\theta+h}\|P_\theta) = E_\theta[r_h \log r_h] = E_\theta \Delta_h + \frac{1}{2} E_\theta \Delta_h^2 + o(\|h\|^2) = \frac{1}{2} h^\top I(\theta) h + o(\|h\|^2).$$

Combining the two expansions proves the result. □

F On the Realizability Condition

The direct score-form statement

$$r_\star(x,y) = \beta \nu_{\pi^\dagger}(x,y)^{\alpha-1} + b(x)$$

can obscure the origin of the realizability condition. We therefore decompose it into two conceptually distinct requirements. The first is the standard policy-realizability condition used in direct preference optimization: the population regularized optimizer induced by the true reward belongs to the candidate policy class. The second is a reward–penalty compatibility condition specific to the thresholded power-law optimizer induced by exponentiated Rényi regularization.

Proposition F.1 (Compatibility implies TDPO score realizability)**.** *Fix $\alpha > 1$ and $\beta > 0$. For ρ_X-almost every x, let $\pi^\star_\beta(\cdot \mid x)$ solve*

$$\max_{\pi(\cdot|x) \ll \pi_0(\cdot|x)} \left\{ \mathbb{E}_{Y\sim\pi(\cdot|x)}[r_\star(x,Y)] - \frac{\beta}{\alpha} \mathbb{E}_{Y\sim\pi_0(\cdot|x)} \left[\nu_\pi(x,Y)^\alpha\right] \right\}.$$

Define

$$m(x) := \operatorname*{ess\,inf}_{Y\sim\pi_0(\cdot|x)} r_\star(x,Y)$$

and

$$\beta_\star(x) := \left(\mathbb{E}_{Y\sim\pi_0(\cdot|x)}\left[\big(r_\star(x,Y)-m(x)\big)^{1/(\alpha-1)}\right]\right)^{\alpha-1}.$$

Suppose that

$$\pi^\star_\beta \in \mathcal{H} \qquad \textit{and} \qquad \beta \geq \operatorname*{ess\,sup}_{x\sim\rho_X} \beta_\star(x).$$

Then there exists a measurable function $b_\beta : \mathcal{X} \to \mathbb{R}$ such that

$$r_\star(x,y) = \beta \nu_{\pi^\star_\beta}(x,y)^{\alpha-1} + b_\beta(x) \qquad \textit{for } \mathcal{P}_0\textit{-almost every } (x,y).$$

Consequently,

$$r_\star(x,y^+) - r_\star(x,y^-) = \beta\left[\nu_{\pi^\star_\beta}(x,y^+)^{\alpha-1} - \nu_{\pi^\star_\beta}(x,y^-)^{\alpha-1}\right],$$

and hence $\pi^\dagger = \pi^\star_\beta$ satisfies the TDPO score realizability condition.

Proof. By Lemma 3.1, the density ratio of the penalized optimizer has the thresholded power-law form

$$\nu_{\pi^\star_\beta}(x,y) = \left(\frac{r_\star(x,y)-t_\beta(x)}{\beta}\right)_+^{1/(\alpha-1)},$$

where $t_\beta(x)$ is determined by normalization. For a fixed prompt x, define

$$F_x(t) := \mathbb{E}_{Y\sim\pi_0(\cdot|x)}\left[\big(r_\star(x,Y)-t\big)_+^{1/(\alpha-1)}\right].$$

The normalization condition is

$$F_x(t_\beta(x)) = \beta^{1/(\alpha-1)}.$$

On the other hand,

$$F_x(m(x)) = \mathbb{E}_{Y\sim\pi_0(\cdot|x)}\left[\big(r_\star(x,Y)-m(x)\big)^{1/(\alpha-1)}\right] = \beta_\star(x)^{1/(\alpha-1)}.$$

Because F_x is non-increasing, the compatibility condition $\beta \geq \beta_\star(x)$ implies

$$t_\beta(x) \leq m(x).$$

Therefore,

$$\big(r_\star(x,y)-t_\beta(x)\big)_+ = r_\star(x,y)-t_\beta(x) \qquad \pi_0(\cdot\mid x)\text{-a.s.}$$

The positive-part operator is thus inactive, and

$$\nu_{\pi^\star_\beta}(x,y)^{\alpha-1} = \frac{r_\star(x,y)-t_\beta(x)}{\beta}.$$

Rearranging gives

$$r_\star(x,y) = \beta\nu_{\pi^\star_\beta}(x,y)^{\alpha-1} + t_\beta(x).$$

Taking $b_\beta(x) = t_\beta(x)$ proves the first claim. Since $b_\beta(x)$ is response-independent, it cancels from every within-prompt reward difference, proving the Bradley–Terry score representation. □

Corollary F.2 (A simple bounded-range condition)**.** *Suppose that, for ρ_X-almost every x,*

$$\operatorname*{ess\,sup}_y r_\star(x,y) - \operatorname*{ess\,inf}_y r_\star(x,y) \leq R.$$

Then $\beta_\star(x) \leq R$. Consequently, $\beta \geq R$ is sufficient for the compatibility condition in Proposition F.1, while $\beta > R$ places the optimizer in the strict non-truncation regime $t_\beta(x) < m(x)$ and therefore gives a strictly positive density ratio $\pi_0(\cdot\mid x)$-almost surely.

Proof. Since

$$0 \leq r_\star(x,Y) - m(x) \leq R \qquad \pi_0(\cdot\mid x)\text{-a.s.},$$

we have

$$\beta_\star(x) = \|r_\star(x,\cdot) - m(x)\|_{L^{1/(\alpha-1)}(\pi_0(\cdot|x))} \leq R.$$

The remaining claims follow from Proposition F.1. □

The proposition separates two roles that were combined in the original score-form assumption. Membership $\pi^\star_\beta \in \mathcal{H}$ is the usual policy-class realizability requirement and depends on the richness of the candidate class. Increasing β alone does not guarantee this membership. By contrast, the condition $\beta \geq \beta_\star(x)$ controls the regularizer-specific truncation effect and is automatically satisfied once the penalty is sufficiently large relative to the within-prompt reward scale.

G Rare-Set Control of the Penalized Optimizer

The rare-set guarantee in Theorem 5.1 depends only on the density-ratio moment $\mathcal{C}_\alpha(\pi)$. We now show that the exponentiated-Rényi penalty controls this moment directly, without requiring the population optimizer to be defined through a hard Rényi constraint.

For each prompt x, define the conditional moment

$$\mathcal{C}_{\alpha,x}(\pi) := \mathbb{E}_{Y\sim\pi_0(\cdot|x)}\left[\nu_\pi(x,Y)^\alpha\right].$$

Proposition G.1 (Implicit moment bound for the penalized optimizer)**.** *Suppose that the within-prompt reward oscillation is uniformly bounded:*

$$\operatorname*{ess\,sup}_y r(x,y) - \operatorname*{ess\,inf}_y r(x,y) \le R \qquad \text{for } \rho_X\text{-almost every } x.$$

Let $\pi_{\mathrm{R}}(\cdot \mid x)$ *maximize*

$$\mathbb{E}_{Y\sim\pi(\cdot|x)}[r(x,Y)] - \frac{\beta}{\alpha}\mathcal{C}_{\alpha,x}(\pi)$$

over policies absolutely continuous with respect to $\pi_0(\cdot \mid x)$*. Then, for* ρ_X*-almost every* x*,*

$$\mathcal{C}_{\alpha,x}(\pi_{\mathrm{R}}) \le 1 + \frac{\alpha R}{\beta}.$$

Consequently, $\mathcal{C}_\alpha(\pi_{\mathrm{R}}) \le 1 + \frac{\alpha R}{\beta}$ *and*

$$D_\alpha(\rho_X \otimes \pi_{\mathrm{R}} \,\|\, \rho_X \otimes \pi_0) \le \frac{1}{\alpha-1}\log\left(1 + \frac{\alpha R}{\beta}\right).$$

In particular, for every measurable $A \subseteq \mathcal{X}\times\mathcal{Y}$*,*

$$(\rho_X \otimes \pi_{\mathrm{R}})(A) \le \left(1 + \frac{\alpha R}{\beta}\right)^{1/\alpha} (\rho_X \otimes \pi_0)(A)^{(\alpha-1)/\alpha}.$$

Proof. Because adding a prompt-dependent constant to the reward does not change the optimizer, we may shift the reward so that

$$0 \le r(x,y) \le R$$

for $\mathcal{P}_0$-almost every (x,y). The reference policy is feasible and satisfies

$$\mathcal{C}_{\alpha,x}(\pi_0) = 1.$$

Optimality of π_{R} therefore gives

$$\mathbb{E}_{\pi_{\mathrm{R}}}[r(x,Y)] - \frac{\beta}{\alpha}\mathcal{C}_{\alpha,x}(\pi_{\mathrm{R}}) \ge \mathbb{E}_{\pi_0}[r(x,Y)] - \frac{\beta}{\alpha}.$$

Rearranging,

$$\frac{\beta}{\alpha}\left(\mathcal{C}_{\alpha,x}(\pi_{\mathrm{R}}) - 1\right) \le \mathbb{E}_{\pi_{\mathrm{R}}}[r(x,Y)] - \mathbb{E}_{\pi_0}[r(x,Y)] \le R.$$

Hence

$$\mathcal{C}_{\alpha,x}(\pi_{\mathrm{R}}) \le 1 + \frac{\alpha R}{\beta}.$$

Averaging over $X \sim \rho_X$ proves the bound on $\mathcal{C}_\alpha(\pi_{\mathrm{R}})$. The Rényi-divergence bound follows from

$$\mathcal{C}_\alpha(\pi_{\mathrm{R}}) = \exp\left((\alpha-1)D_\alpha(\rho_X \otimes \pi_{\mathrm{R}} \,\|\, \rho_X \otimes \pi_0)\right).$$

Finally, applying Theorem 5.1 gives the stated rare-set inequality. □

Comparison with the KL-penalized optimizer. Let $\pi_{\mathrm{KL}}(\cdot \mid x)$ maximize

$$\mathbb{E}_{Y\sim\pi(\cdot|x)}[r(x,Y)] - \beta D_{\mathrm{KL}}\big(\pi(\cdot \mid x) \,\|\, \pi_0(\cdot \mid x)\big).$$

After shifting the reward so that $0 \le r(x,y) \le R$, its Gibbs density ratio is

$$\nu_{\pi_{\mathrm{KL}}}(x,y) = \frac{\exp(r(x,y)/\beta)}{\mathbb{E}_{\pi_0(\cdot|x)}[\exp(r(x,Y)/\beta)]}.$$

Therefore,

$$\mathcal{C}_{\alpha,x}(\pi_{\mathrm{KL}}) = \frac{\mathbb{E}_{\pi_0(\cdot|x)}[\exp(\alpha r(x,Y)/\beta)]}{\big(\mathbb{E}_{\pi_0(\cdot|x)}[\exp(r(x,Y)/\beta)]\big)^{\alpha}} \le \exp\left(\frac{(\alpha-1)R}{\beta}\right).$$

Thus, the exponentiated-Rényi penalty gives

$$\mathcal{C}_{\alpha,x}(\pi_{\mathrm{R}}) \le 1 + \frac{\alpha R}{\beta},$$

which is linear in R/β, whereas the corresponding worst-case KL bound is exponential in R/β. Equivalently, the rare-set amplification factor is at most

$$\left(1 + \frac{\alpha R}{\beta}\right)^{1/\alpha}$$

for the Rényi-penalized solution, compared with a worst-case exponential dependence under KL regularization.

The exponential KL dependence is unavoidable. Consider a one-prompt, two-response instance with responses s and h:

$$\pi_0(h) = e^{-R/\beta}, \qquad \pi_0(s) = 1 - e^{-R/\beta}, \qquad r(h) = R, \qquad r(s) = 0.$$

The KL Gibbs optimizer satisfies

$$\pi_{\mathrm{KL}}(h) = \frac{\pi_0(h)e^{R/\beta}}{\pi_0(s) + \pi_0(h)e^{R/\beta}} = \frac{1}{2 - e^{-R/\beta}} \ge \frac{1}{2}.$$

Consequently,

$$\mathcal{C}_{\alpha}(\pi_{\mathrm{KL}}) \ge \pi_0(h)\left(\frac{\pi_{\mathrm{KL}}(h)}{\pi_0(h)}\right)^{\alpha} \ge 2^{-\alpha}\exp\left(\frac{(\alpha-1)R}{\beta}\right).$$

Hence the exponential dependence of the KL moment bound is tight up to an α-dependent constant.

On the same instance, Proposition G.1 and Theorem 5.1 imply

$$\pi_{\mathrm{R}}(h) \le \left(1 + \frac{\alpha R}{\beta}\right)^{1/\alpha} \exp\left(-\frac{\alpha-1}{\alpha}\frac{R}{\beta}\right).$$

Thus, as R/β grows, the Rényi-penalized optimizer assigns vanishing mass to the reference-rare response, whereas the KL-penalized optimizer assigns it constant probability.

Connection to population TDPO. Under Assumption 4.2, the policy $\pi^{\star}_{\beta} = \pi_{\mathrm{R}}$ satisfies

$$r_{\star}(x,y^{+}) - r_{\star}(x,y^{-}) = \Delta_{\pi_{\mathrm{R}}}(x,y^{+},y^{-}).$$

Since the logistic loss is a strictly proper loss for the Bradley–Terry conditional probability, its population minimum is attained at this true score difference. Hence the population TDPO target coincides with the exponentiated-Rényi-penalized optimizer, and

$$\text{population TDPO} \quad\Longrightarrow\quad \mathcal{C}_{\alpha}(\pi_{\mathrm{R}}) \le 1 + \frac{\alpha R}{\beta} \quad\Longrightarrow\quad \text{polynomial rare-set amplification control.}$$

Thus the penalized population formulation controls rare-set amplification through its moment bound. For a finite-sample output, a target budget ϵ is guaranteed by choosing $\mathcal{H} \subseteq \Pi_{\epsilon}$. The statistical upper bounds instead use the candidate-wide bounded-logit assumption, which implies the deterministic moment bound $\mathcal{C}_{\alpha}(\hat{\pi}) \le 1 + V/\beta$ by (20).

H Illustrative Example: Overoptimization Under Low Coverage

Similar to Huang et al. (2025), we consider a single-prompt offline preference-learning problem designed to isolate the effect of regularization geometry under severe coverage mismatch. The prompt space is degenerate, $\mathcal{X} = \{\emptyset\}$, and the action space is $\mathcal{Y} = \{y_0, y_1, y_2, y_3\}$. The reference policy is

$$\pi_0(y_0) = \frac{1}{2}, \qquad \pi_0(y_1) = \pi_0(y_2) = \frac{1}{2n}, \qquad \pi_0(y_3) = \frac{n-2}{2n},$$

where $n \in \{10, 100, 1000\}$ controls the level of coverage on the rare actions. Thus, y_0 is a high-coverage baseline action, whereas y_1 and y_2 are low-coverage actions with probability mass $O(1/n)$. The true reward is $r^\star = r_1$, defined by

$$r^\star(y_0) = \frac{1}{2}, \qquad r^\star(y_1) = 1, \qquad r^\star(y_2) = r^\star(y_3) = 0.$$

Hence, y_1 is the globally optimal action, while y_0 is a safe baseline action with moderate reward. We focus on the sample-starved regime in which the offline preference data fail to reveal the distinction between the two rare actions. Concretely, we condition on the bad event under which the preference dataset contains no informative examples involving y_1 or y_2. On this event, maximum-likelihood fitting under the BT model can return the incorrect reward estimate

$$\hat{r} = r_2, \text{ where } \hat{r}(y_0) = \frac{1}{2}, \qquad \hat{r}(y_2) = 1, \qquad \hat{r}(y_1) = \hat{r}(y_3) = 0.$$

This construction is useful because all methods are given the same misspecified reward model, so any performance difference is entirely due to the policy mapping induced by the regularizer rather than differences in reward estimation.

We compare three policy classes induced by the same reward estimate $\hat{r}$. DPO uses the KL-induced exponential tilt $\hat{\pi}_{\mathrm{DPO}}(y) \propto \pi_0(y) \exp\left(\frac{\hat{r}(y)}{\beta}\right)$. Because the update is exponential, DPO reacts aggressively to a spurious high reward on a low-coverage action. ChiPO uses the mixed χ^2+KL link, yielding $\hat{\pi}_{\mathrm{ChiPO}}(y) = \pi_0(y)\, \phi_\chi^{-1}\left(\frac{\hat{r}(y) - Z}{\beta}\right), \phi_\chi(z) = z + \log z$. where Z is the normalization constant chosen so that $\sum_a \hat{\pi}(a) = 1$. This map is substantially more pessimistic in the tails and therefore less prone to over-concentrating on uncertain low-coverage actions.

For TDPO with Rényi order $\alpha > 1$, the policy takes the thresholded power-law form $\hat{\pi}_{\mathrm{TDPO},\alpha}(y) = \pi_0(y) \left(\frac{\hat{r}(y) - t_\alpha}{\beta}\right)_+^{1/(\alpha-1)}$, where the threshold t_α is chosen so that $\sum_y \hat{\pi}(y) = 1$. We report results for $\alpha \in \{2, 2.5, 3, 3.5, 4\}$. Unlike DPO, TDPO replaces the exponential tilt by a finite-order power transform, which dampens the response to rare proxy spikes.

We evaluate each method by sweeping the regularization parameter β and reporting its performance relative to the well-covered baseline action y_0:

$$J(y_0) - J(\hat{\pi}), \qquad J(\pi) = \mathbb{E}_{y \sim \pi}[r^\star(y)].$$

Here, $J(y_0)$ is the true reward obtained by always selecting the baseline action, so $J(y_0) = r^\star(y_0)$. This quantity is neither the unregularized optimality gap in Theorem 4.4 nor the regularized sub-optimality gap in Theorem 4.5: those results concern finite-sample estimation under realizability, whereas this example conditions on the misspecified estimate $\hat{r} = r_2$. We call $J(y_0) - J(\hat{\pi})$ *baseline regret* only as an overoptimization diagnostic. It measures true reward loss relative to the safe baseline when all methods optimize the same incorrect proxy reward. Since the true optimal rare action y_1 is unidentifiable under the bad event, the relevant question is whether a method avoids over-concentrating on the falsely rewarded rare action y_2, not whether it discovers y_1.

Figures 5 and 6 evaluate the same misspecified offline preference instance from two complementary perspectives. The curves are computed exactly from the induced policy formulas. Figure 5 shows the action-level mechanism: because the fitted reward incorrectly ranks the rare bad action $y_{\mathrm{bad}} = y_2$ highest, the exponential KL update used by DPO assigns excessive probability to this action and removes too much mass from the safe baseline y_0. TDPO avoids this failure mode more effectively. Its finite-order Rényi regularization induces a thresholded power-law update, so a spurious reward

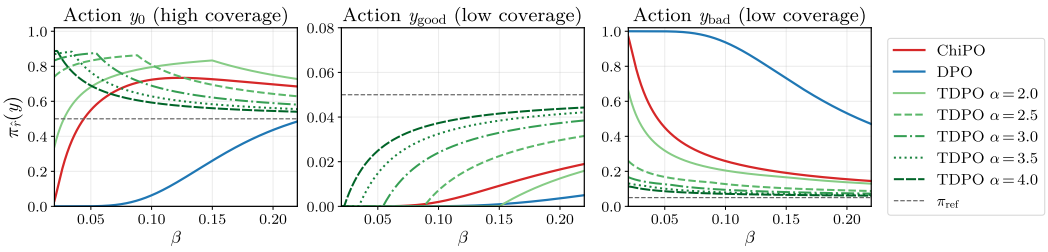

Figure 5: Action probabilities on the low-coverage example under reward misspecification ($r^\star = r_1$ but $\hat{r} = r_2$, $n = 10$). DPO strongly amplifies the spurious high reward on the rare bad action y_{bad}, whereas TDPO keeps this probability under tighter control through its finite-order Rényi, thresholded power-law update. Consequently, TDPO preserves substantially more mass on the safe baseline action y_0, demonstrating improved robustness to proxy-reward errors on poorly covered actions.

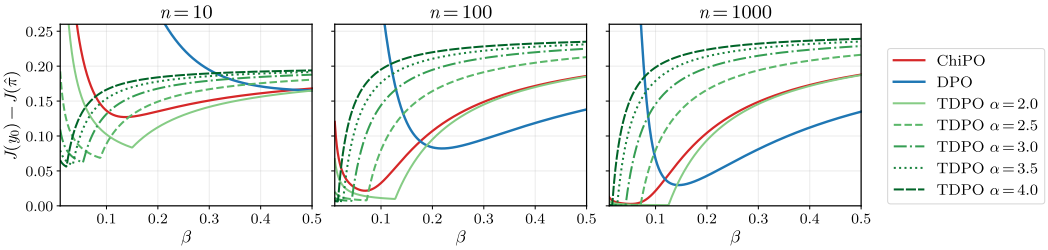

Figure 6: Baseline regret $J(y_0) - J(\hat{\pi})$ on the same example for $n \in \{10, 100, 1000\}$. Across coverage regimes, TDPO substantially reduces regret relative to KL-based DPO and yields a broader stable region over the regularization parameter β. The improvement comes from replacing exponential reweighting with a tunable power-law response, which suppresses overoptimization of low-coverage reward spikes while still allowing adaptation away from the reference policy.

spike on a low-coverage action is amplified much less aggressively than under exponential tilting. As a result, TDPO remains closer to the reference allocation on y_0 and places markedly less mass on y_{bad}. The limited probability assigned to the true good action $y_{\text{good}} = y_1$ is expected, since this action is indistinguishable under the bad event; the relevant goal here is therefore robustness against the false optimum rather than discovery of y_1.

Figure 6 confirms that this action-level stability translates into lower baseline regret. Across all three coverage levels and a broad range of β, TDPO improves substantially over DPO, whose regret remains high when the KL update overoptimizes the erroneous proxy signal. The TDPO family also reveals the practical role of the Rényi order α: larger orders produce more conservative tail responses, giving a tunable bias–robustness tradeoff under reward misspecification. Although ChiPO can be a strong robust baseline in this synthetic construction, TDPO achieves a comparable anti-overoptimization effect within a pure Rényi framework and retains the direct-preference optimization interface. These results support the main claim that TDPO narrows the gap between standard DPO and more pessimistic policy-shaping methods by replacing exponential KL geometry with tail-aware power-law control.

I Local natural-gradient interpretation of the Rényi order

This subsection explains the sentence in the main text: for a fixed infinitesimal trust-region radius, increasing the Rényi order α shrinks the admissible natural-gradient step by the factor $\alpha^{-1/2}$. The argument is a direct consequence of the local Rényi–KL expansion.

Let $\{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ be a regular dominated parametric model, and let $I(\theta)$ be the Fisher information matrix. For a small local parameter perturbation h, Theorem 5.4 gives

$$D_\alpha(P_{\theta+h} \| P_\theta) = \frac{\alpha}{2} h^\top I(\theta) h + o(\|h\|^2),$$

whereas $D_{\mathrm{KL}}(P_{\theta+h}\|P_\theta) = \frac{1}{2}h^\top I(\theta)h + o(\|h\|^2)$. Thus, at infinitesimal scale, Rényi divergence has the same Fisher quadratic form $\|h\|_{I(\theta)} := (h^\top I(\theta)h)^{1/2}$ as KL divergence, but with curvature multiplied by α.

Fix a small trust-region radius $\rho > 0$. Under a local KL constraint

$$D_{\mathrm{KL}}(P_{\theta+h}\|P_\theta) \le \rho,$$

the preceding expansion gives, to second order, $\frac{1}{2}h^\top I(\theta)h \le \rho$, or equivalently

$$\|h\|_{I(\theta)} \le \sqrt{2\rho}.$$

Under the order-α Rényi constraint

$$D_\alpha(P_{\theta+h}\|P_\theta) \le \rho,$$

the local constraint becomes $\frac{\alpha}{2}h^\top I(\theta)h \le \rho$, hence

$$\|h\|_{I(\theta)} \le \sqrt{\frac{2\rho}{\alpha}} = \alpha^{-1/2}\sqrt{2\rho}.$$

Therefore, for the same infinitesimal divergence radius ρ, the Rényi trust region has Fisher-radius smaller than the KL trust region by the factor

$$\alpha^{-1/2}.$$

One can also see this by the local natural-gradient optimization problem. The first-order expansion of the reward objective:

$$J(\theta + h) = J(\theta) + g^\top h + o(\|h\|), \qquad g := \nabla_\theta J(\theta),$$

Under the local KL trust region, the infinitesimal update solves

$$\max_h \; g^\top h \quad \text{subject to} \quad \frac{1}{2}h^\top I(\theta)h \le \rho.$$

The Lagrangian is

$$\mathcal{L}_{\mathrm{KL}}(h, \lambda) = g^\top h - \lambda\left(\frac{1}{2}h^\top I(\theta)h - \rho\right),$$

and the first-order condition gives $g - \lambda I(\theta)h = 0$. Hence, the optimizer $h = \lambda^{-1}I(\theta)^{-1}g$ is proportional to the natural-gradient direction $I(\theta)^{-1}g$. Imposing the active constraint $\frac{1}{2}h^\top I(\theta)h = \rho$ yields $\lambda = \sqrt{\frac{g^\top I(\theta)^{-1}g}{2\rho}}$ and

$$h^\star_{\mathrm{KL}} = \sqrt{2\rho}\,\frac{I(\theta)^{-1}g}{\left(g^\top I(\theta)^{-1}g\right)^{1/2}}.$$

Under the local order-α Rényi trust region, the corresponding problem is

$$\max_h \; g^\top h \qquad \text{subject to} \qquad \frac{\alpha}{2}h^\top I(\theta)h \le \rho.$$

Equivalently, $h^\top I(\theta)h \le \frac{2\rho}{\alpha}$. The same Lagrange multiplier calculation gives

$$h^\star_\alpha = \sqrt{\frac{2\rho}{\alpha}}\,\frac{I(\theta)^{-1}g}{\left(g^\top I(\theta)^{-1}g\right)^{1/2}}.$$

Therefore

$$h^\star_\alpha = \alpha^{-1/2}h^\star_{\mathrm{KL}}.$$

This is the precise meaning of the statement that increasing α shrinks the admissible natural-gradient step by the factor $\alpha^{-1/2}$ when the trust-region radius is fixed.

This scaling also explains the robustness interpretation. In practice, the reward gradient is not observed exactly. Instead one uses an estimated gradient

$$\widehat{g} = g + \xi,$$

where ξ denotes the reward-model or preference-estimation error. The KL local update based on $\widehat{g}$ is

$$\widehat{h}_{\mathrm{KL}} = \sqrt{2\rho}\,\frac{I(\theta)^{-1}\widehat{g}}{\left(\widehat{g}^{\top} I(\theta)^{-1}\widehat{g}\right)^{1/2}},$$

whereas the order-α Rényi update is

$$\widehat{h}_{\alpha} = \sqrt{\frac{2\rho}{\alpha}}\,\frac{I(\theta)^{-1}\widehat{g}}{\left(\widehat{g}^{\top} I(\theta)^{-1}\widehat{g}\right)^{1/2}} = \alpha^{-1/2}\widehat{h}_{\mathrm{KL}}.$$

Thus the component of the update driven by the estimation error ξ is also reduced by the same local factor $\alpha^{-1/2}$. Informally, at a fixed infinitesimal trust-region radius, a larger Rényi order makes the policy update less able to move aggressively in a spurious natural-gradient direction created by small reward-estimation errors.

J Limits of quadratic mixed regularization

J.1 Proof of Theorem 5.3

Proof. Fix $\epsilon > 0$ and $\alpha > 2$. We construct a family of strictly positive pairs (P_q, Q_q), verify its divergence budget, and choose an explicit member whose moment exceeds an arbitrary $M > 0$. Every expectation below is a sum over two points.

Step 1: Choose the parameters and verify the probabilities. Choose any fixed c such that

$$0 < c < \sqrt{2\epsilon}, \qquad \Delta := \epsilon - \frac{c^2}{2} > 0. \tag{29}$$

The positive number Δ is the slack in the divergence budget. Define

$$q_{\mathrm{valid}} := \min\left\{\frac{1}{2},\ c^2,\ \frac{1}{4c^2}\right\} > 0. \tag{30}$$

For any $q \in (0, q_{\mathrm{valid}})$, set $p_q = c\sqrt{q}$ and define

$$Q_q\{1\} = q, \quad Q_q\{0\} = 1 - q, \qquad P_q\{1\} = p_q, \quad P_q\{0\} = 1 - p_q.$$

These are probability measures with strictly positive masses. Indeed, $0 < q < 1/2$ gives $q > 0$ and $1 - q > 0$, while

$$0 < p_q = c\sqrt{q} < c\sqrt{\frac{1}{4c^2}} = \frac{1}{2} < 1.$$

Thus $1 - p_q > 0$ as well. Moreover, $q < c^2$ implies $\sqrt{q} < c$, so

$$p_q = c\sqrt{q} > \sqrt{q}\sqrt{q} = q. \tag{31}$$

We use the fixed event $A = \{1\}$. Its reference mass $Q_q(A) = q$ tends to zero along the family; the space and the event do not change.

Step 2: Compute and normalize the density ratio. Since Q_q gives positive mass to both points, $P_q \ll Q_q$. The density $U_q = dP_q/dQ_q$ takes the values

$$U_q(1) = \frac{p_q}{q} = \frac{c}{\sqrt{q}}, \qquad U_q(0) = \frac{1 - p_q}{1 - q} = \frac{1 - c\sqrt{q}}{1 - q}. \tag{32}$$

Both are positive and finite for each admissible q, and

$$\mathbb{E}_{Q_q} U_q = q\frac{p_q}{q} + (1 - q)\frac{1 - p_q}{1 - q} = p_q + (1 - p_q) = 1.$$

Step 3: Calculate the χ divergence on both points. Subtracting one inside each square in the definition gives

$$\begin{aligned}\chi^2(P_q\|Q_q) &= q\left(\frac{p_q}{q}-1\right)^2+(1-q)\left(\frac{1-p_q}{1-q}-1\right)^2\\ &= q\left(\frac{p_q-q}{q}\right)^2+(1-q)\left(\frac{q-p_q}{1-q}\right)^2\\ &= \frac{(p_q-q)^2}{q}+\frac{(p_q-q)^2}{1-q}\\ &= (p_q-q)^2\left(\frac{1}{q}+\frac{1}{1-q}\right)=\frac{(p_q-q)^2}{q(1-q)}.\end{aligned} \tag{33}$$

Substitute $p_q = c\sqrt{q}$ and cancel a factor q:

$$\chi^2(P_q\|Q_q)=\frac{(c\sqrt{q}-q)^2}{q(1-q)}=\frac{(c-\sqrt{q})^2}{1-q}\longrightarrow c^2 \qquad (q\downarrow 0). \tag{34}$$

This limit follows because $\sqrt{q}\to 0$ and $1-q\to 1$. We also need an upper bound valid before taking the limit:

$$\begin{aligned}\chi^2(P_q\|Q_q)-c^2 &= \frac{c^2-2c\sqrt{q}+q-c^2(1-q)}{1-q}=\frac{-2c\sqrt{q}+(1+c^2)q}{1-q}\\ &\le \frac{(1+c^2)q}{1-q}\le 2(1+c^2)q.\end{aligned} \tag{35}$$

Dropping $-2c\sqrt{q}\le 0$ gives the first inequality. The second uses $q<1/2$ and therefore $1/(1-q)\le 2$.

Step 4: Expand and bound both terms of the KL. By its definition on a two-point space,

$$D_{\mathrm{KL}}(P_q\|Q_q)=\mathbb{E}_{Q_q}[U_q\log U_q]=p_q\log\frac{p_q}{q}+(1-p_q)\log\frac{1-p_q}{1-q}=:T_1(q)+T_0(q). \tag{36}$$

The first summand is nonnegative because $p_q>q$. Since $p_q<1$, we have $p_q/q\le 1/q$, so

$$0\le T_1(q)=c\sqrt{q}\log\frac{c}{\sqrt{q}}\le c\sqrt{q}\log\frac{1}{q}. \tag{37}$$

To verify that this upper bound tends to zero, observe that for $t\ge 1$,

$$\log t=\int_1^t\frac{dx}{x}\le\int_1^t x^{-3/4}\,dx=4(t^{1/4}-1)\le 4t^{1/4}.$$

The integral comparison uses $x^{-1}\le x^{-3/4}$ for $x\ge 1$. Taking $t=1/q$ and multiplying by $c\sqrt{q}$ yields

$$0\le T_1(q)\le 4c\,q^{1/4}\longrightarrow 0. \tag{38}$$

For the complement summand, $p_q>q$ gives $0<(1-p_q)/(1-q)<1$, so $T_0(q)\le 0$. Define

$$s_q:=\frac{p_q-q}{1-q}.$$

Its numerator is positive and smaller than $1-q$ because $p_q<1$. Thus $0<s_q<1$, and $1-s_q=(1-p_q)/(1-q)$. For $s\in(0,1)$,

$$0\le -\log(1-s)=\int_0^s\frac{dt}{1-t}\le\frac{s}{1-s},$$

since the integrand is at most $(1-s)^{-1}$ throughout $[0,s]$. Consequently,

$$\begin{aligned}|T_0(q)| &= (1-p_q)[-\log(1-s_q)]\\ &\le (1-p_q)\frac{s_q}{1-s_q}\end{aligned}$$

$$= (1-p_q)\frac{(p_q-q)/(1-q)}{(1-p_q)/(1-q)} = p_q - q \le p_q = c\sqrt{q} \longrightarrow 0. \tag{39}$$

Both terms in (36) therefore tend to zero.

For the subsequent budget estimate, nonnegativity can also be checked directly. The inequality $\log t \le t-1$ for $t>0$ follows because $t-1-\log t$ has derivative $1-1/t$ and minimum zero at $t=1$. Apply this inequality to q/p_q and $(1-q)/(1-p_q)$:

$$\begin{aligned} -D_{\mathrm{KL}}(P_q\|Q_q) &= p_q \log\frac{q}{p_q} + (1-p_q)\log\frac{1-q}{1-p_q} \\ &\le p_q\left(\frac{q}{p_q}-1\right) + (1-p_q)\left(\frac{1-q}{1-p_q}-1\right) \\ &= (q-p_q) + (p_q-q) = 0. \end{aligned}$$

Since $T_0(q) \le 0$, we have proved the nonasymptotic bound

$$0 \le D_{\mathrm{KL}}(P_q\|Q_q) \le T_1(q) \le 4c\,q^{1/4}. \tag{40}$$

Step 5: Verify the mixed budget with an explicit cutoff. Equations (34) and (40) give

$$D_{\mathrm{mix}}(P_q\|Q_q) \longrightarrow \frac{c^2}{2} < \epsilon.$$

The phrase *for sufficiently small* q can be made fully explicit. Define

$$q_* := \min\left\{q_{\mathrm{valid}},\ \frac{\Delta}{2(1+c^2)},\ \left(\frac{\Delta}{8c}\right)^4\right\} > 0. \tag{41}$$

Every entry is positive because $c>0$ and $\Delta>0$. For any $0<q\le q_*/2$, the bounds above apply and give

$$\begin{aligned} D_{\mathrm{mix}}(P_q\|Q_q) &\le \frac{c^2}{2} + (1+c^2)q + 4c\,q^{1/4} \\ &\le \frac{c^2}{2} + \frac{\Delta}{2} + \frac{\Delta}{2} = \frac{c^2}{2} + \Delta = \epsilon. \end{aligned} \tag{42}$$

The second line uses $q \le \Delta/[2(1+c^2)]$ for the linear term and $q \le (\Delta/(8c))^4$ for the fourth-root term. Also $q \le q_*/2 < q_{\mathrm{valid}}$, so the pair remains strictly positive.

Step 6: Calculate the higher moment and select a pair for each M. Using both density values in (32) gives

$$\begin{aligned} \mathbb{E}_{Q_q} U_q^\alpha &= q\left(\frac{p_q}{q}\right)^\alpha + (1-q)\left(\frac{1-p_q}{1-q}\right)^\alpha \\ &= \frac{p_q^\alpha}{q^{\alpha-1}} + \frac{(1-p_q)^\alpha}{(1-q)^{\alpha-1}} \\ &= c^\alpha q^{1-\alpha/2} + \frac{(1-c\sqrt{q})^\alpha}{(1-q)^{\alpha-1}} \ge c^\alpha q^{1-\alpha/2}. \end{aligned} \tag{43}$$

The last step discards a nonnegative summand. Set

$$\rho := \frac{\alpha-2}{2} > 0.$$

Then $1-\alpha/2 = -\rho$, so this lower bound is $c^\alpha q^{-\rho}$ and tends to infinity as $q \downarrow 0$.

More explicitly, for an arbitrary target $M>0$, set

$$q_M := \frac{1}{2}\min\left\{q_*,\ \left(\frac{c^\alpha}{M}\right)^{1/\rho}\right\}, \qquad p_M := c\sqrt{q_M}. \tag{44}$$

All quantities in the minimum are positive, so $q_M>0$. Because $q_M \le q_*/2$, Step 5 gives a valid pair and $D_{\mathrm{mix}}(P_{q_M}\|Q_{q_M}) \le \epsilon$. Moreover,

$$q_M \le \frac{1}{2}\left(\frac{c^\alpha}{M}\right)^{1/\rho} \quad\Longrightarrow\quad \mathbb{E}_{Q_{q_M}} U_{q_M}^\alpha \ge c^\alpha q_M^{-\rho} \ge 2^\rho M > M.$$

Here taking the power $-\rho<0$ reverses the inequality for q_M. Since M was arbitrary, the feasible moment values are unbounded above. Their supremum is $+\infty$. □

J.2 Proof of Equation (13)

Proof. Let $U = \frac{d\pi}{d\pi_0}$, $p = \pi(A)$, $q = \pi_0(A)$. Since $D_{\mathrm{KL}}(\pi\|\pi_0) \geq 0$, the mixed-divergence budget implies $\frac{1}{2}\chi^2(\pi\|\pi_0) \leq D_{\mathrm{mix}}(\pi\|\pi_0) \leq \epsilon$. Therefore,

$$\mathbb{E}_{\pi_0}(U-1)^2 = \chi^2(\pi\|\pi_0) \leq 2\epsilon. \tag{45}$$

Because $U = d\pi/d\pi_0$ is a probability density, $\mathbb{E}_{\pi_0}U = 1$, and hence $\mathbb{E}_{\pi_0}(U-1) = 0$. It follows that

$$p - q = \mathbb{E}_{\pi_0}[U\mathbf{1}_A] - \mathbb{E}_{\pi_0}[\mathbf{1}_A] = \mathbb{E}_{\pi_0}[(U-1)\mathbf{1}_A] = \mathbb{E}_{\pi_0}[(U-1)(\mathbf{1}_A - q)].$$

Applying the Cauchy–Schwarz inequality gives

$$|p-q| \leq \left(\mathbb{E}_{\pi_0}(U-1)^2\right)^{1/2}\left(\mathbb{E}_{\pi_0}(\mathbf{1}_A - q)^2\right)^{1/2}.$$

Since $\mathbf{1}_A$ is a Bernoulli random variable with mean q,

$$\mathbb{E}_{\pi_0}(\mathbf{1}_A - q)^2 = q(1-q)^2 + (1-q)q^2 = q(1-q).$$

Combining this identity with (45), we obtain

$$|\pi(A) - \pi_0(A)| \leq \sqrt{2\epsilon\,\pi_0(A)(1-\pi_0(A))}.$$

Dropping the lower side of this two-sided bound yields $\pi(A) \leq \pi_0(A) + \sqrt{2\epsilon\pi_0(A)(1-\pi_0(A))}$. □

K The Rényi Order as a Design Choice

The preference environment consists of the prompt distribution, the reference policy, the underlying reward, and the Bradley–Terry sampling rule. None of these is defined by the Rényi order α. Holding the environment fixed while changing α leaves the distribution of preference data unchanged; it changes the regularized objective used to construct a policy. Thus α is a design choice, rather than an unknown parameter of the environment to be recovered from preference observations.

The order controls how strongly TDPO penalizes large policy-to-reference density ratios. Its best choice depends on the reward structure, data coverage, and where reward estimation errors may induce overoptimization. These features do not yield a generally optimal order from the preference law alone. We therefore analyze each fixed choice of α; comparing orders requires a common policy-performance criterion, rather than comparing the values of their different regularized objectives.