Sample Complexity of RLHF Reward Learning under General Reward Classes

Anonymous authors
Paper under double-blind review

Abstract

We study the minimax sample complexity of Bradley–Terry preference-based reward learning under an arbitrary reward class $\mathcal{R}$. For the realisable logistic (Bradley–Terry) preference model with a single-policy concentrability coefficient C, we prove matching upper and lower bounds

$$N^\star(\mathcal{R}, \varepsilon) \;=\; \Theta\!\left(\frac{C^2 \cdot \mathcal{H}(\mathcal{R}, \varepsilon/C)}{\kappa(B) \cdot \varepsilon^2}\right),$$

where $\mathcal{H}(\mathcal{R}, \varepsilon) := \log N(\mathcal{R}, \varepsilon, L^2(\tilde{\mu}))$ is the L^2 metric entropy of $\mathcal{R}$ under the induced action marginal $\tilde{\mu}$ and $\kappa(B) := \sigma(2B)(1-\sigma(2B))$ is the Bradley–Terry Fisher-information curvature on the pairwise-difference range $[-2B, 2B]$. The bound matches in C, ε, *and* B up to absolute constants, under a mild saturation condition on $\mathcal{R}$ (verified for each corollary class). The closure of the $\kappa(B)$-gap uses a *boundary* Bregman upper bound on the softplus, $\zeta(y) - \zeta(x) - \sigma(x)(y-x) \le \frac{\sigma(L)(1-\sigma(L))}{2}(y-x)^2$ for $x, y \ge L \ge 0$, invoked in the Fano step with an adversarial ground truth whose induced pairwise differences saturate the boundary. Our result unifies and sharpens a line of work that had resolved the rate only for structured subclasses: linear (Zhu et al., 2023), low-rank (Pacchiano et al., 2023), and general preferences (Wang et al., 2023). Three standard reward classes instantiate the bound: linear in $\mathbb{R}^d$ ($\tilde{\Theta}(\kappa(B)^{-1}C^2 d/\varepsilon^2)$), rank-$k$ in a d-dimensional embedding ($\tilde{\Theta}(\kappa(B)^{-1}C^2 kd/\varepsilon^2)$), and Sobolev $W^{s,2}([0,1]^d)$ ($\Theta(\kappa(B)^{-1}C^{2+d/s}\,\varepsilon^{-(2+d/s)})$). The upper bound uses a localised Rademacher argument on the conditional MLE, driven by a quadratic curvature identity for the Bradley–Terry log-likelihood; the *two* Bregman inequalities (a pointwise lower bound with constant $\kappa(B)/2$, and a matching boundary upper bound with constant $\sigma(L)(1-\sigma(L))/2$) driving the rate are mechanically verified in Lean 4 / Mathlib with zero `sorry` and no custom axioms. The lower bound is a Fano–Le Cam construction tailored to the Bradley–Terry Fisher information, made coverage-aware by a restricted packing on the support of the optimal policy and made B-matching by saturating the pairwise range. We verify the curvature constants numerically on a family of reward classes. The result is a theorem about a stylised realisable pairwise preference model; extensions to misspecification and multi-turn preferences are discussed but left open.

1 Introduction

Reinforcement learning from human feedback (RLHF; Christiano et al., 2017; Ouyang et al., 2022) has become the default training recipe for aligning large language models with user preferences. In the dominant pipeline, a reward model is first fit to a dataset of pairwise preferences between model completions using the Bradley–Terry (Bradley & Terry, 1952) likelihood; the policy is then improved against the learned reward. A central design question is: *how many preference pairs does one need?* In practice, answers range from 10^4 to 10^7 and are chosen by rule of thumb. In theory, the answer depends delicately on the complexity of the reward class under consideration.

For a *linear* reward class in $\mathbb{R}^d$, Zhu et al. (2023) proved a matching $\Theta(d/\varepsilon^2)$ rate. Subsequent work has extended the rate to structured reward classes: Pacchiano et al. (2023) treat generic function classes via

dueling bandit machinery, Wang et al. (2023) obtain rates for low-rank preference models, and Song et al. (2024) study the sample complexity of offline preference fine-tuning. In each case, the result is parameterised by a complexity measure tailored to the sub-problem: dimension, rank, or disagreement coefficient. A *general* matching bound parameterised by a standard statistical complexity measure, applicable uniformly across these subclasses, has been missing.

Contributions. We establish such a bound.

1. **Matching sample complexity for general reward classes** (Theorem 6). Under mild regularity on $\mathcal{R}$, a single-policy concentrability C, and a saturation condition on the pairwise range of $\mathcal{R}$ on the support of μ (Assumption 5, satisfied by all three corollary classes),

$$\frac{k_1}{\kappa(B)} \cdot \frac{C^2 \cdot \mathcal{H}(\mathcal{R}, \varepsilon/C)}{\varepsilon^2} \;\le\; N^\star(\mathcal{R}, \varepsilon) \;\le\; \frac{k_2}{\kappa(B)} \cdot \frac{C^2 \cdot \mathcal{H}(\mathcal{R}, \varepsilon/(2C))}{\varepsilon^2}, \tag{1}$$

   with absolute constants k_1, k_2 and $\kappa(B) := \sigma(2B)(1 - \sigma(2B))$. Both the coverage factor C^2 and the curvature factor $\kappa(B)^{-1}$ appear symmetrically on upper and lower sides; the ε-scale mismatch (ε/C vs. $\varepsilon/(2C)$) contributes at most an absolute multiplicative constant for the reward classes in our corollaries. Without saturation, the lower bound loses the $\kappa(B)^{-1}$ factor (Remark 7); we do not know whether the gap is then real.

2. **Two Bregman inequalities, both mechanically verified** (Section 7). The softplus Bregman divergence admits matching pointwise *lower* and *upper* bounds, formalised in Lean 4 with zero `sorry` as `softplus_bregman_ge` (coefficient $\kappa(B)/2$ on $|x|, |y| \le B$) and `softplus_bregman_le_of_ge` (coefficient $\sigma(L)(1 - \sigma(L))/2$ on $x, y \ge L \ge 0$). The latter, instantiated at $L \uparrow 2B$ via an adversarial saturating ground truth, is what closes the $\kappa(B)^{-1}$ gap.

3. **Corollaries for canonical reward classes** (Section 6). We recover the linear rate ($\tilde{\Theta}(\kappa(B)^{-1} C^2 d/\varepsilon^2)$, matching up to $\log(C/\varepsilon)$) and the low-rank rate ($\tilde{\Theta}(\kappa(B)^{-1} C^2 k d/\varepsilon^2)$), and obtain a new Sobolev rate $\Theta(\kappa(B)^{-1} C^{2+d/s} \varepsilon^{-(2+d/s)})$ for $W^{s,2}([0,1]^d)$ reward classes.

4. **Numerical verification** (Section 8). On six reward classes spanning all three corollaries, the Monte Carlo BT KL lies strictly between the worst-case lower bound of Lemma 8 (constant $\kappa(B)/2$) and the at-origin upper bound $\zeta''(0)/2 = 1/8$, tracking the boundary Bregman upper bound of Lemma 9 in the regime where the drawn Δ's approach $\pm 2B$.

Scope. We prove matching bounds for a stylised realisable Bradley–Terry preference model with i.i.d. pair sampling and a fixed behaviour policy. We do not address: (i) model misspecification, (ii) multi-turn or K-wise preferences, (iii) active query design. Items (ii) and (iii) are active research directions; we flag them as open in Section 10.

2 Problem Setup

Preference data. Let $\mathcal{X}$ be a context space with distribution $\mathcal{D}$, and let $\mathcal{A}$ be an action space. For each context $x \sim \mathcal{D}$, a *behaviour policy* $\mu(\cdot \mid x)$ emits an *ordered* pair of actions (a_0, a_1). Write μ for the joint distribution of (x, a_0, a_1). A *reward function* is a measurable map $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$. We assume $r^\star \in \mathcal{R}$ where $\mathcal{R}$ is a known hypothesis class, and we write $\Delta_r(x, a_0, a_1) := r(x, a_1) - r(x, a_0)$ for the pairwise reward difference.

Under the Bradley–Terry model, the preference label $y \in \{0, 1\}$ is drawn as

$$\mathbb{P}[y = 1 \mid x, a_0, a_1] \;=\; \sigma(\Delta_{r^\star}(x, a_0, a_1)), \tag{2}$$

where $\sigma(u) := 1/(1 + e^{-u})$ is the logistic function. A preference dataset of size N is $\mathcal{D}_N := \{(x_i, a_{0,i}, a_{1,i}, y_i)\}_{i=1}^N$, drawn i.i.d. from the joint $\mu \otimes \mathrm{Ber}(\sigma(\Delta_{r^\star}))$.

Policy regret. A learner outputs a policy $\hat{\pi}$; its regret against the optimal policy $\pi^\star_{r^\star}(x) := \arg\max_a r^\star(x,a)$ is

$$\mathrm{Reg}(\hat{\pi}) \;:=\; \mathbb{E}_{x\sim\mathcal{D}}\big[r^\star(x,\pi^\star_{r^\star}(x)) - r^\star(x,\hat{\pi}(x))\big]. \tag{3}$$

The *minimax sample complexity* is

$$N^\star(\mathcal{R},\varepsilon) \;:=\; \min\big\{N : \inf_{\hat{\pi}} \sup_{r^\star\in\mathcal{R}} \mathbb{P}\big[\mathrm{Reg}(\hat{\pi}) > \varepsilon\big] \le \tfrac{1}{4}\big\}, \tag{4}$$

where the probability is over the sampling of $\mathcal{D}_N$ and any internal randomness of $\hat{\pi}$.

Complexity measure. We parameterise our bound by the L^2 *metric entropy* of $\mathcal{R}$ (van der Vaart & Wellner, 1996; Bartlett et al., 2005). For a class of real-valued functions $\mathcal{F}$ and a probability measure ν, the $L^2(\nu)$-covering number $N(\mathcal{F},\varepsilon,L^2(\nu))$ is the minimal cardinality of an ε-net, and the $L^2(\nu)$-packing number $D(\mathcal{F},\varepsilon,L^2(\nu))$ is the maximum cardinality of an ε-separated subset; the two satisfy $N(\mathcal{F},\varepsilon) \le D(\mathcal{F},\varepsilon) \le N(\mathcal{F},\varepsilon/2)$. We write $\mathcal{H}(\mathcal{R},\varepsilon) := \log N(\mathcal{R},\varepsilon,L^2(\tilde{\mu}))$ for the metric entropy. The upper bound of Theorem 6 uses $\mathcal{H}$ via Dudley-type chaining; the lower bound uses $\log D$ via Fano, and we invoke the covering/packing equivalence to state both in $\mathcal{H}$ up to a factor-2 scale shift absorbed into the constant k_1. Metric entropy coincides with bracketing entropy up to absolute constants for the reward classes treated in our corollaries (van der Vaart & Wellner, 1996, Chapter 2.7).

Regularity. We make three standing assumptions on the problem.

Assumption 1 (Boundedness and antisymmetric normalisation)**.** Every $r \in \mathcal{R}$ satisfies $\|r\|_\infty \le B$ for a known $B \ge 1$, so the pairwise reward difference Δ_r is bounded by $2B$. We assume $\mathcal{R}$ is closed under negation ($r \in \mathcal{R} \Rightarrow -r \in \mathcal{R}$) and pick a representative in each equivalence class modulo per-context constants so that $r(x,a_0) + r(x,a_1) = 0$ for every (x,a_0,a_1) in the support of μ. This is a w.l.o.g. choice: the BT likelihood equation 2 depends on r only through Δ_r and is therefore invariant under $r(x,\cdot) \mapsto r(x,\cdot) + c(x)$; consequently for any $r, r' \in \mathcal{R}$, $(r - r')(x,a_0) = -(r - r')(x,a_1)$ on the support of μ.

Remark 2 (Antisymmetric normalisation as a gauge choice). Since BT equation 2 depends on r only through Δ_r, it is invariant under $r(x,\cdot) \mapsto r(x,\cdot) + c(x)$; the antisymmetric representative is the unique zero-mean choice per prompt. For non-normalised $\hat{r}$, apply the results to $\hat{r}_{\mathrm{sym}}(x,a) := \hat{r}(x,a) - \frac{1}{2}(\hat{r}(x,a_0) + \hat{r}(x,a_1))$ on each pair; this is rank-preserving, hence policy-invariant.

Assumption 3 (Coverage)**.** Let $\tilde{\mu}(\cdot \mid x)$ denote the action marginal of μ: $\tilde{\mu}(a \mid x) := \frac{1}{2}\big(\mu_0(a \mid x) + \mu_1(a \mid x)\big)$, where $\mu_i(\cdot \mid x)$ is the marginal law of a_i given x under μ. The *single-policy concentrability coefficient* $C_{\mathrm{cov}} := \sup_{x,a} \frac{\pi^\star_{r^\star}(a|x)}{\tilde{\mu}(a|x)}$ is finite. We write $C_{\mathrm{cov}} \le C$.

Assumption 4 (Pair symmetry)**.** The behaviour policy is symmetric: $\mu(a_0,a_1 \mid x) = \mu(a_1,a_0 \mid x)$ for every x. Consequently $\mu_0(\cdot \mid x) = \mu_1(\cdot \mid x) = \tilde{\mu}(\cdot \mid x)$.

Assumption 5 (Pairwise saturation)**.** The class $\mathcal{R}$ contains an element $r_\dagger$ whose pairwise differences achieve the boundary on the support of μ: for every $\rho > 0$ there exists $r_\dagger = r_\dagger(\rho) \in \mathcal{R}$ with $|\Delta_{r_\dagger}(x,a_0,a_1)| \ge 2B - \rho$ for μ-almost every (x,a_0,a_1). We call such an $r_\dagger$ a *ρ-saturating* element.

Assumption 1 is standard and makes the BT likelihood strongly concave on a bounded set. Assumption 3 matches the offline-RL literature (Chen & Jiang, 2019; Rashidinejad et al., 2021) and is the minimal condition under which policy regret can be controlled by reward L^2 error. Assumption 4 is for convenience in the lower bound; if dropped, replace μ by its symmetrisation $\bar{\mu}(a_0,a_1 \mid x) := \frac{1}{2}(\mu(a_0,a_1 \mid x) + \mu(a_1,a_0 \mid x))$ throughout. The BT likelihood is symmetric in (a_0,a_1) up to the sign of y, so this substitution is KL-free under a relabelling $y \mapsto 1 - y$ on the flipped pairs; the Fano packing then proceeds against $\bar{\mu}$. The cost is a factor ≤ 2 in the concentrability C (since $\bar{\mu} \ge \mu/2$ on the μ-support), which is absorbed into k_1. Assumption 5 says the bound B in Assumption 1 is *tight* on the support, and is used only in the lower bound: it lets Fano pick an adversarial ground truth at which the BT Fisher information is $\kappa(B)$ rather than $1/4$. Section 2.1 analyses it in detail.

2.1 On the Saturation Assumption

Assumption 5 is the only structural requirement on $\mathcal{R}$ beyond boundedness and the coverage/symmetry conditions standard in offline RL. Because it is what buys the lower bound its $\kappa(B)^{-1}$ factor, we spell out when it holds, when it fails, and how tight it is.

When it holds: constructions for the three corollary classes. The following are verified in Appendix A.7.

1. *Linear* ($r = \langle \phi(\cdot), \theta \rangle$, $\|\theta\|_2 \le 1$, $\|\phi\|_\infty \le 1$): if the pairwise-feature gap $\|\phi(x, a_1) - \phi(x, a_0)\|_2 \ge 2$ on $\mathrm{supp}(\mu)$, the reward $r_\dagger(x, a) := \langle \phi(x, a), \theta_\dagger \rangle$ with $\theta_\dagger := (\phi(x, a_1) - \phi(x, a_0))/\|\phi(x, a_1) - \phi(x, a_0)\|_2$ (any common pairwise-feature direction on support) is 0-saturating.

2. *Low-rank* ($r(x, a) = \psi(x)^\top M \varphi(a)$, $\mathrm{rank}(M) \le k$): analogously, $M_\dagger := uv^\top$ with u, v the top singular directions of the pairwise-feature map $\psi(x) \otimes (\varphi(a_1) - \varphi(a_0))$ on support is 0-saturating when the pairwise-singular value attains $2B$.

3. *Sobolev* $W^{s,2}([0, 1]^d)$, $s > d/2$: the antisymmetric constant $r_\dagger(x, a) := (-1)^{a+1} B$ has $\|r_\dagger\|^2_{W^{s,2}} = 2B^2$ (derivative terms vanish), so $r_\dagger \in \mathcal{R}$ iff $B \le 1/\sqrt{2}$; then $\Delta_{r_\dagger} \equiv \pm 2B$ exactly, witnessing Assumption 5 with $\rho = 0$. The threshold $B \le 1/\sqrt{2}$ is sharp for saturation in $W^{s,2}$ independent of the Sobolev embedding constant (Appendix A.6).

In each case, Assumption 5 thus reduces to a mild *non-degeneracy* on how $\mathcal{R}$ is parameterised (the bound B is not loose on support) rather than a structural restriction.

When it fails. A natural example where Assumption 5 is violated is the *loose-budget linear class* $\mathcal{R}' := \{r : r(x, a) = \langle \phi(x, a), \theta \rangle,\ \|\theta\|_2 \le 1/2\}$ under Assumption 1 with $B = 1$: here $\|r\|_\infty \le 1/2$ strictly, so $|\Delta_r| \le 1 < 2B$ for every $r \in \mathcal{R}'$, and Assumption 5 fails for any $\rho < 1$. The $\kappa(B)^{-1}$ factor in Theorem 6 with $B = 1$ is then *a priori* larger than what the actual pairwise-difference range of $\mathcal{R}'$ warrants: the sharp bound uses $B_{\mathrm{eff}} = 1/2$ giving $\kappa(1/2)^{-1} \approx 4.5$ instead of $\kappa(1)^{-1} \approx 9.5$. In such a case our theorem is correct but slack; the slack can be removed by shrinking B to match the actual support. The assumption is therefore geometric, not statistical: it asserts that B is *chosen tightly* for the problem at hand. For classes satisfying the tightness, our upper and lower bounds coincide in B; for classes where B is loose by a factor $\alpha > 1$, the gap is $\kappa(B)/\kappa(B/\alpha)$, which is at most $e^{\alpha B}$.

Remark 7 (after Theorem 6) discusses what survives without Assumption 5.

Relation to prior work. The $\kappa(B)^{-1}$ gap is already implicit in Zhu et al. (2023)'s linear lower bound (their Eq. 3.9), which achieves $\kappa(B)$ by placing the packing near the boundary. Assumption 5 makes this adversarial placement feasible for general $\mathcal{R}$, automatic for the linear class and verified in our three corollaries.

3 Main Results

3.1 Matching Sample Complexity

Theorem 6 (Matching sample complexity for general $\mathcal{R}$)**.** *Under Assumptions 1–5, there exist absolute numerical constants $k_1, k_2 > 0$ (independent of $\mathcal{R}, \mu, B, C, \varepsilon$) such that for every $\varepsilon \in (0, 1/2]$ small enough that $\mathcal{H}(\mathcal{R}, \varepsilon/C) \ge 4 \log 2$,*

$$k_1\, \kappa(B)^{-1}\, C^2\, \frac{\mathcal{H}(\mathcal{R}, \varepsilon/C)}{\varepsilon^2} \;\le\; N^\star(\mathcal{R}, \varepsilon) \;\le\; k_2\, \kappa(B)^{-1}\, C^2\, \frac{\mathcal{H}(\mathcal{R}, \varepsilon/(2C))}{\varepsilon^2},$$

where $\kappa(B) := \sigma(2B)(1 - \sigma(2B)) \in (0, 1/4]$ is the Bradley–Terry Fisher-information curvature at the boundary of the pairwise-difference range $[-2B, 2B]$. The two sides match in C, ε, and *B up to absolute constants and a factor-2 entropy scale shift.*

Remark 7 (What happens without saturation). If Assumption 5 is dropped, the lower bound weakens to $k_1\, C^2\, \mathcal{H}(\mathcal{R}, \varepsilon/C)/\varepsilon^2$, i.e. the $\kappa(B)^{-1}$ factor is lost. This reflects a genuine analytic feature: the Fano KL upper bound then uses only the global curvature bound $\zeta'' \le 1/4$ instead of the boundary value $\zeta''(\pm 2B) = \kappa(B)$, because the adversarial packing can no longer be placed where the BT Fisher information saturates. Whether the resulting $\kappa(B)^{-1}$ gap is real for non-saturating classes, or an artefact of our lower-bound technique, is open.

Constants. The numerical prefactors k_1, k_2 are absolute (tracked explicitly through Appendices A.3 and A.4). The curvature factor $\kappa(B)^{-1}$ is dimension-free ($\kappa(1)^{-1} \approx 9.5$, $\kappa(2)^{-1} \approx 56.6$). The asymmetry between upper and lower *analyses* is resolved by the boundary Bregman bound: MLE uses $\zeta'' \ge \kappa(B)$ on $[-2B, 2B]$ (Lemma 8); Fano uses $\zeta'' \le \sigma(L)(1-\sigma(L))$ on $\{|t| \ge L\}$ (Lemma 9), instantiated at $L \uparrow 2B$ via a saturating ground truth from Assumption 5. The ε/C vs. $\varepsilon/(2C)$ scale mismatch contributes at most an absolute multiplicative constant for the corollary entropy profiles ($\mathcal{H} \asymp \varepsilon^{-\gamma}$ or $\asymp d\log(1/\varepsilon)$). Full proofs are in Appendix A; the upper bound uses the MLE $\hat{r} = \arg\max_{r\in\mathcal{R}} \sum_i \log p_r(y_i \mid \cdot)$.

3.2 The Curvature Lemma

The analytic kernel of the upper bound is the following pointwise inequality, which identifies the BT KL divergence with the Bregman divergence of the softplus function $\zeta(x) := \log(1 + e^x)$.

Lemma 8 (BT curvature, pointwise)**.** *For every $u, v \in [-2B, 2B]$,*

$$\mathrm{KL}\big(\mathrm{Ber}(\sigma(v)) \,\big\|\, \mathrm{Ber}(\sigma(u))\big) \;=\; \zeta(u) - \zeta(v) - \sigma(v)(u-v) \;\ge\; \tfrac{1}{2}\sigma(2B)(1-\sigma(2B)) \cdot (u-v)^2. \tag{5}$$

Proof sketch. The identity $\mathrm{KL}(\mathrm{Ber}(p_\star)\|\mathrm{Ber}(p)) = \zeta(u) - \zeta(v) - \sigma(v)(u-v)$ follows by direct computation with $p = \sigma(u)$, $p_\star = \sigma(v)$, and $\log\sigma(x) = x - \zeta(x)$. The quadratic lower bound then follows because $\zeta'' = \sigma\cdot(1-\sigma)$ is bounded below by $\sigma(2B)(1-\sigma(2B))$ on $[-2B, 2B]$. The full argument is mechanised in Lean; see Section 7. □

The companion *upper* bound, used in the Fano step of the lower bound, is obtained by bounding ζ'' from above on the same region; the tightest constant available on a two-sided boundary region $\{|t| \ge L\}$ is $\sigma(L)(1-\sigma(L))$, attained at $|t| = L$.

Lemma 9 (BT boundary-Bregman upper bound, pointwise)**.** *For every $L \ge 0$ and every $u, v \ge L$,*

$$\zeta(u) - \zeta(v) - \sigma(v)(u-v) \;\le\; \tfrac{1}{2}\sigma(L)(1-\sigma(L)) \cdot (u-v)^2. \tag{6}$$

The same bound holds for $u, v \le -L$ by the reflection $\zeta(t) - \zeta(-t) = t$. Setting $L = 0$ recovers the universal bound $\zeta(u) - \zeta(v) - \sigma(v)(u-v) \le \frac{1}{8}(u-v)^2$.

Proof sketch. $\zeta''(t) = \sigma(t)(1-\sigma(t))$ is unimodal with peak $1/4$ at $t = 0$, symmetric about 0, and decreasing in $|t|$. On $[L, \infty)$ with $L \ge 0$, its supremum is attained at $t = L$ and equals $\sigma(L)(1-\sigma(L))$. Taylor's theorem with the supremum curvature yields the claim. A mean-value argument parallel to the proof of Lemma 8 but using the slope *upper* bound $\sigma(b) - \sigma(a) \le \sigma(L)(1-\sigma(L))(b-a)$ on $L \le a \le b$ is mechanised in Lean as theorem `softplus_bregman_le_of_ge`; see Section 7. □

Integrated form. Integrating equation 5 against μ and using $\|\Delta_r - \Delta_{r^\star}\|^2_{L^2(\mu)} = 4\|r - r^\star\|^2_{L^2(\tilde{\mu})}$ (Assumptions 1 and 4) gives

$$\mathrm{KL}(p_{r^\star}\|p_r) \;\ge\; 2\kappa(B) \cdot \|r - r^\star\|^2_{L^2(\tilde{\mu})}, \qquad \kappa(B) := \sigma(2B)(1-\sigma(2B)). \tag{7}$$

4 Upper Bound

Estimator. The learner computes the maximum-likelihood estimator

$$\hat{r} \;:=\; \arg\max_{r\in\mathcal{R}} \sum_{i=1}^{N} \log p_r(y_i \mid x_i, a_{0,i}, a_{1,i}), \tag{8}$$

where $p_r(1 \mid \cdot) = \sigma(\Delta_r(\cdot))$. The induced policy is $\hat{\pi}(x) := \arg\max_a \hat{r}(x, a)$.

Curvature to rate. The MLE satisfies the following standard excess-risk bound, whose proof combines Lemma 8 with the local Rademacher technology of Bartlett et al. (2005):

Lemma 10 (Local Rademacher $\to L^2$ rate)**.** *Under Assumptions 1–4, the MLE $\hat{r}$ of equation 8 satisfies, with probability at least $1 - e^{-u}$,*

$$\|\hat{r} - r^\star\|^2_{L^2(\tilde{\mu})} \;\le\; \frac{k_3 \cdot (\mathcal{H}(\mathcal{R}, \varepsilon^\star) + u)}{\kappa(B) \cdot N},$$

where $\varepsilon^\star$ is the unique solution of the fixed-point equation $\varepsilon^2 = k_4 \cdot \mathcal{H}(\mathcal{R}, \varepsilon)/(\kappa(B) \cdot N)$, and k_3, k_4 are absolute constants.

From reward error to policy regret. Under Assumption 3,

$$\mathrm{Reg}(\hat{\pi}) \;\le\; 2C \cdot \|\hat{r} - r^\star\|_{L^2(\tilde{\mu})}. \tag{9}$$

The argument is standard (see Appendix A.2 for a self-contained proof): a greedy-policy decomposition followed by an importance-weighting change of measure gives $\mathrm{Reg}(\hat{\pi}) \le 2C\|\hat{r} - r^\star\|_{L^1(\tilde{\mu})}$, and Jensen's inequality upgrades L^1 to L^2. Combining equation 9 with Lemma 10 and solving $\varepsilon = 2C\,\varepsilon^\star$ for N yields $N \le k_2\, \kappa(B)^{-1}\, C^2\, \mathcal{H}(\mathcal{R}, \varepsilon/(2C))/\varepsilon^2$, which is the upper bound of Theorem 6.

5 Lower Bound

The Fano argument needs a packing of $\mathcal{R}$ at $L^2(\tilde{\mu})$ scale δ with log-cardinality $\asymp \mathcal{H}(\mathcal{R}, O(\delta))$ and every element within L^2-radius $O(\delta)$ of a saturating centre $r_\dagger$. The standard Yang–Barron / Yu construction (Yang & Barron, 1999; Yu, 1997) gives the first property around an arbitrary centre; the following lemma gives the second for the three corollary classes.

Lemma 11 (Saturated local packing)**.** *Fix one of the three reward classes $\mathcal{R} \in \{\mathcal{R}_{\mathrm{lin}}, \mathcal{R}_{\mathrm{lr}}, \mathcal{R}_{\mathrm{Sob}}\}$ of Corollaries 13–15 under Assumptions 1–5, and let $r_\dagger \in \mathcal{R}$ be a ρ-saturating element supplied by Assumption 5. There exist a class-dependent threshold $\delta_0(\mathcal{R}) > 0$, an absolute constant $c_1 \in (0, 1)$ (independent of $\mathcal{R}$, $r_\dagger$, δ), and a class-dependent* saturation-preservation rate $\rho'(\delta)$ *such that for every $\delta \in (0, \delta_0(\mathcal{R})]$ one can choose $r_1, \ldots, r_M \in \mathcal{R}$ with*

$$2\delta \;\le\; \|r_i - r_j\|_{L^2(\tilde{\mu})} \;\le\; 4\delta, \quad \|r_i - r_\dagger\|_{L^2(\tilde{\mu})} \le 4\delta, \quad \log M \;\ge\; c_1 \cdot \mathcal{H}(\mathcal{R}, 2\delta), \tag{10}$$

and additionally $|\Delta_{r_i}| \ge 2B - \rho - \rho'(\delta)$ μ-a.e. for each i, with $\rho'_{\mathrm{lin}}(\delta) = \rho'_{\mathrm{lr}}(\delta) = 8\delta$ and $\rho'_{\mathrm{Sob}}(\delta) = O_{s,d}(\delta^{1-d/(2s)})$. In all three cases $\rho'(\delta) \downarrow 0$ as $\delta \downarrow 0$. The thresholds are $\delta_0(\mathcal{R}_{\mathrm{lin}}) = \delta_0(\mathcal{R}_{\mathrm{lr}}) = 1/4$, and $\delta_0(\mathcal{R}_{\mathrm{Sob}})$ an explicit (s, d, B)-dependent constant (Appendix A.5).

Proof sketch. Linear and low-rank use a half-space Euclidean volumetric packing in the parameter ball near $r_\dagger$; Sobolev uses a basis-truncation argument (Tsybakov, 2009, Lemma 2.9) on a frequency band $\asymp \delta^{-d/s}$ orthogonal to $r_\dagger$. Full proof with constants in Appendix A.5. □

Lemma 12 (Coverage-aware boundary Fano lower bound)**.** *Under Assumptions 1–5, there exist absolute constants $c_0 > 0$ and a scale $\delta_0(\mathcal{R}) > 0$ such that for every $\varepsilon \in (0, 1/2]$ with $\varepsilon/C \le \delta_0(\mathcal{R})$,*

$$N^\star(\mathcal{R}, \varepsilon) \;\ge\; c_0 \cdot \kappa(B)^{-1} \cdot \frac{C^2}{\varepsilon^2} \cdot \mathcal{H}(\mathcal{R}, \varepsilon/C).$$

The $\kappa(B)^{-1}$ factor comes from the boundary Bregman upper bound of Lemma 9, instantiated at a ρ-saturating ground truth $r_\dagger$ supplied by Assumption 5 with $\rho \downarrow 0$.

Proof sketch. Set $\delta := \varepsilon/(2C)$, so that any learner with policy regret $\le \varepsilon$ satisfies $\|\hat{r} - r^\star\|_{L^2(\tilde{\mu})} \le \delta$ by equation 9. The novelty relative to the standard Fano bound is the *adversarial choice of ground truth*: we

do not centre the packing around an arbitrary $r^\star \in \mathcal{R}$ but around a ρ-saturating $r_\dagger \in \mathcal{R}$ from Assumption 5, chosen so that $\Delta_{r_\dagger}(x, a_0, a_1) \geq 2B - \rho$ on the support of μ (WLOG by replacing $r_\dagger$ by $-r_\dagger$ if the signs disagree, using closure of $\mathcal{R}$ under negation from Assumption 1).

Local packing at the boundary. Apply Lemma 11 to get $r_1, \ldots, r_M \in \mathcal{R}$ satisfying equation 10 with an absolute $c_1 \in (0, 1)$. Each r_i inherits saturation up to $\rho' := \rho + \rho'(\delta)$ with $\rho'(\delta) \downarrow 0$ as $\delta \downarrow 0$ (linear/low-rank: $\rho'(\delta) = 8\delta$; Sobolev: $\rho'(\delta) = O_{s,d}(\delta^{1-d/(2s)})$; Lemma 11), so $\Delta_{r_i} \geq 2B - \rho'$ on support.

Boundary KL upper bound. Apply Lemma 9 with $L := 2B - \rho'$ and $(u, v) = (\Delta_{r_i}(\omega), \Delta_{r_j}(\omega))$ pointwise in $\omega = (x, a_0, a_1) \in \mathrm{supp}(\mu)$:

$$\zeta(\Delta_{r_i}) - \zeta(\Delta_{r_j}) - \sigma(\Delta_{r_j})(\Delta_{r_i} - \Delta_{r_j}) \;\leq\; \tfrac{\sigma(L)(1-\sigma(L))}{2}(\Delta_{r_i} - \Delta_{r_j})^2.$$

Integrating against μ and using $\|\Delta_{r_i} - \Delta_{r_j}\|^2_{L^2(\mu)} = 4\|r_i - r_j\|^2_{L^2(\tilde{\mu})}$ (Assumptions 1, 4),

$$\mathrm{KL}(p_{r_i} \| p_{r_j}) \;\leq\; 2\sigma(L)(1 - \sigma(L)) \cdot \|r_i - r_j\|^2_{L^2(\tilde{\mu})} \;\leq\; 32\,\sigma(L)(1 - \sigma(L))\,\delta^2.$$

Taking $\rho, \delta \downarrow 0$ jointly with $L = 2B - \rho' \uparrow 2B$, the prefactor converges to $\kappa(B) = \sigma(2B)(1 - \sigma(2B))$, giving $\mathrm{KL}(p_{r_i}^{\otimes N} \| p_{r_j}^{\otimes N}) \leq 32N\kappa(B)\delta^2(1 + o(1))$.

Fano conclusion. Fano's inequality (Tsybakov, 2009, Corollary 2.6) forces error probability $\geq 1/2$ unless $32N\kappa(B)\delta^2 \geq \frac{1}{2}\log M - \log 2$, i.e., $N \geq (\log M - 2\log 2)/(64\kappa(B)\delta^2)$. Substituting $\delta = \varepsilon/(2C)$ and $\log M \geq c_1\,\mathcal{H}(\mathcal{R}, \varepsilon/C)$, and absorbing the additive $-2\log 2$ into the leading constant on the stated range $\mathcal{H}(\mathcal{R}, \varepsilon/C) \geq 4\log 2$, yields the claim with an absolute constant $c_0 > 0$; tracking through the algebra (Appendix A.4) gives $c_0 = c_1/16$. □

6 Corollaries

Corollary 13 (Linear reward class)**.** *Let $\mathcal{R} = \{r : r(x, a) = \langle \phi(x, a), \theta \rangle,\ \|\theta\|_2 \leq 1\}$ for $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ with $\|\phi\|_\infty \leq 1$, and assume the pairwise-feature norm $\|\phi(x, a_1) - \phi(x, a_0)\|_2 \geq 2$ on* $\mathrm{supp}(\mu)$ *(so that $B = 1$ is tight, Assumption 1). Then Assumption 5 is satisfied with $\rho = 0$ by $r_\dagger(x, a) = \langle \phi(x, a), \theta_\dagger \rangle$ for $\theta_\dagger$ parallel to the common pairwise-feature direction, and $\mathcal{H}(\mathcal{R}, \varepsilon) \asymp d \log(1/\varepsilon)$ (van der Vaart & Wellner, 1996, Theorem 2.7.11), so Theorem 6 gives*

$$N^\star(\mathcal{R}, \varepsilon) \;=\; \tilde{\Theta}\big(\kappa(B)^{-1}\, C^2\, d/\varepsilon^2\big),$$

where $\tilde{\Theta}$ hides a $\log(C/\varepsilon)$ factor coming from the entropy argument. This sharpens the linear rate of Zhu et al. (2023) by matching the $\kappa(B)^{-1}$ constant on both upper and lower sides.

Corollary 14 (Low-rank reward)**.** *Let $\mathcal{R}$ consist of $r(x, a) = \psi(x)^\top M \varphi(a)$ with $\psi : \mathcal{X} \to \mathbb{R}^d$, $\varphi : \mathcal{A} \to \mathbb{R}^d$ bounded, and $M \in \mathbb{R}^{d \times d}$ of rank at most k. Assume the pairwise-feature factorisation $\psi(x)^\top \cdot (\varphi(a_1) - \varphi(a_0))$ attains its maximum magnitude $\|M\|\cdot$ (pairwise-feature norm) on* $\mathrm{supp}(\mu)$*. Then Assumption 5 is satisfied with $\rho = 0$ by the rank-1 matrix $M_\dagger$ aligned with this pairwise direction, $\mathcal{H}(\mathcal{R}, \varepsilon) \asymp kd \log(1/\varepsilon)$, and Theorem 6 yields $N^\star(\mathcal{R}, \varepsilon) = \tilde{\Theta}(\kappa(B)^{-1}\, C^2\, kd/\varepsilon^2)$. This sharpens Wang et al. (2023) under our parameterisation by closing the $\kappa(B)^{-1}$ constant.*

Corollary 15 (Sobolev reward class)**.** *Let $\mathcal{R} = \{r : [0, 1]^d \times \{0, 1\} \to \mathbb{R} \mid \|r\|_{W^{s,2}} \leq 1, s > d/2\}$, with the antisymmetric parameterisation $r(x, 1) = -r(x, 0)$, and assume $B \leq 1/\sqrt{2}$ (Assumption 1). Then Assumption 5 is satisfied with $\rho = 0$ by the antisymmetric constant $r_\dagger(x, a) = (-1)^{a+1} B \in \mathcal{R}$ (cf. §2.1 item 3 and Appendix A.6); the threshold $B \leq 1/\sqrt{2}$ is sharp for saturation, independent of the Sobolev embedding constant. The L^2 metric entropy of the $W^{s,2}$ unit ball satisfies $\mathcal{H}(\mathcal{R}, \varepsilon) \asymp \varepsilon^{-d/s}$ (Birman & Solomjak, 1967; Nickl & Pötscher, 2007), and Theorem 6 gives:*

$$k_1\, \kappa(B)^{-1}\, C^{2+d/s}\, \varepsilon^{-(2+d/s)} \;\leq\; N^\star(\mathcal{R}, \varepsilon) \;\leq\; k_2\, \kappa(B)^{-1}\, 2^{d/s}\, C^{2+d/s}\, \varepsilon^{-(2+d/s)},$$

so $N^\star(\mathcal{R}, \varepsilon) = \Theta(\kappa(B)^{-1}\, C^{2+d/s}\, \varepsilon^{-(2+d/s)})$ (matching in C, ε, and *B). To our knowledge this Sobolev rate for BT preference learning is new.*

Example 16 (Concrete setup satisfying all assumptions for Corollary 15)**.** Let $\mathcal{D}$ be uniform on $\mathcal{X} := [0,1]^d$ and let the behaviour policy be exchangeable: $\mu(a_0, a_1 \mid x) = \frac{1}{2}$ on each of the two orderings of $\{0,1\}$. Then $\tilde{\mu}(a \mid x) = 1/2$ for $a \in \{0,1\}$, so any deterministic optimal policy has $C_{\mathrm{cov}} = 2$ (Assumption 3), and pair symmetry (Assumption 4) holds by construction. Fix $B \le 1/\sqrt{2}$ and let $r^\star(x,a) := (-1)^{a+1}B$ (the antisymmetric constant). Then $r^\star \in \mathcal{R}$ with $\|r^\star\|_{W^{s,2}} = \sqrt{2}B \le 1$ and $\|r^\star\|_\infty = B$ (Assumption 1), and $|\Delta_{r^\star}| \equiv 2B$ on $\mathrm{supp}(\mu)$ witnesses Assumption 5 with $\rho = 0$. All four assumptions therefore hold simultaneously with concrete constants.

7 Lean 4 Formalisation of the Curvature Lemma

Both the curvature *lower* bound (Lemma 8, driving the MLE upper bound) and the boundary curvature *upper* bound (Lemma 9, driving the Fano lower bound that closes the $\kappa(B)^{-1}$ gap) are mechanically verified in Lean 4 on the Lean toolchain `leanprover/lean4:v4.29.0` against the matching `Mathlib` release. The formalisation consists of three files, `lean/RlhfMatching/{Logistic,Curvature,Integrated}.lean`, with 17 named theorems. The pointwise Bregman envelopes are lifted to integrated $L^2(\mu)$ statements in `Integrated.lean`, which yields the exact inequalities used at the population level inside the local-Rademacher and Fano steps (Lemmas 10 and 12).

Scope of the formalisation. We formalise (i) the two pointwise BT curvature inequalities (Lemmas 8 and 9) together with their identification with the Bernoulli KL, and (ii) their *integrated* $L^2(\mu)$ forms, which are the statements literally consumed by Lemmas 10 and 12. The local Rademacher chaining step and the Fano packing combinatorics beyond the KL bound, and the policy regret reduction equation 9, are *not* formalised; we view them as standard applications of off-the-shelf tools. The motivation for mechanising the two curvature inequalities in particular is that they are the analytic kernels on which the *matching* bound hinges: the upper and lower bounds match only if the same Bregman identity is applied with its two sharp one-sided constants. Mechanisation here rules out a silent constant mismatch.

Logistic analytics. `Logistic.lean` establishes the derivatives $\zeta' = \sigma$, $\sigma' = \sigma \cdot (1-\sigma)$, and the three sigmoid-variance bounds consumed below: $\sigma(1-\sigma) \ge \sigma(B)(1-\sigma(B))$ on $|t| \le B$, the universal bound $\le 1/4$, and the boundary bound $\le \sigma(L)(1-\sigma(L))$ on $|t| \ge L \ge 0$ (the new ingredient).

Curvature inequalities. `Curvature.lean` proves `softplus_bregman_ge` (Contributions bullet 2, coefficient $\sigma(B)(1-\sigma(B))/2$ on $|x|, |y| \le B$) and `softplus_bregman_le_of_ge` (coefficient $\sigma(L)(1-\sigma(L))/2$ on $x, y \ge L \ge 0$) by Mathlib's mean-value inequalities applied to the convex auxiliary $\psi(t) := \zeta(t) - \sigma(x)(t-x) - (m/2)(t-x)^2$. The Lean symbol `B` corresponds to $2B_{\mathrm{paper}}$; Lemma 8 is recovered with `B := 2`B_{paper}, and Lemma 9 from `L := 2`B_{paper} `-` ρ as $\rho \downarrow 0$.

Link to Bernoulli KL. The theorem `btKL_eq_softplus_bregman` shows the softplus Bregman divergence equals the Bernoulli KL of the two BT preference distributions. Combined with `softplus_bregman_ge` and `softplus_bregman_le_of_ge` this yields `btKL_ge_of_abs_le` (Lemma 8) and `btKL_le_of_ge` (Lemma 9) in their KL form, ready for use inside the Fano and Rademacher steps.

Integrated $L^2(\mu)$ statements. `Integrated.lean` lifts the pointwise Bregman envelopes to measure-theoretic statements by direct application of Mathlib's `integral_mono_ae` and `integral_const_mul`. The three main theorems are:

- `integral_btKL_ge_of_abs_le`: $\int \mathrm{btKL}(\Delta, \Delta_\star)\,\mathrm{d}\mu \;\ge\; \frac{\sigma(B)(1-\sigma(B))}{2}\int(\Delta - \Delta_\star)^2\,\mathrm{d}\mu$ whenever $|\Delta|, |\Delta_\star| \le B$ μ-a.e. with integrable integrands. This is equation equation 7 of the paper.
- `integral_btKL_le_of_ge`: the boundary upper bound with coefficient $\sigma(L)(1-\sigma(L))/2$ whenever $\Delta, \Delta_\star \ge L \ge 0$ μ-a.e., consumed by the Fano step of Lemma 12.
- `integral_btKL_le_universal`: the universal upper bound with coefficient $1/8$, valid without boundary hypotheses.

No new analytic content enters at this stage; each proof is a two-line application of `integral_mono_ae` to the pointwise theorem.

Axiom hygiene. `#print axioms` on each of the nine main theorems (the six pointwise theorems plus the three integrated versions) reports dependence only on the three Lean-kernel axioms `propext`, `Classical.choice`, and `Quot.sound`. Zero `sorry`, zero custom axioms; in particular, no `MeasureTheory`-specific axiom beyond the Mathlib baseline is introduced.

Open formalisation targets. The local-Rademacher chaining beyond the integrated curvature lower bound, the Fano packing combinatorics beyond the integrated KL upper bound, and the importance-weighted regret-to-L^2 reduction equation 9 remain unformalised; these steps use Bartlett–Bousquet–Mendelson, Yang–Barron, and standard offline-RL concentrability tools that would require substantially more infrastructure.

8 Empirical Verification of the Curvature Constant

We verify Lemmas 8 and 9 numerically on six reward classes spanning Corollaries 13–15 under *two* sampling regimes, each probing one of the two Bregman envelopes.

Setup. For each reward class we fix a representative $r^\star \in \mathcal{R}$ with $B = 1$ and draw a perturbed $r = r^\star + c \cdot (r_{\mathrm{alt}} - r^\star)$, varying c so that the pairwise L^2 distance $\|\Delta_r - \Delta_{r^\star}\|^2_{L^2(\mu)}$ ranges over a log grid in $[10^{-5}, 10^{-1}]$. For each distance we simulate $N = 10^6$ preference pairs under $r^\star$ and compute the empirical excess negative log-likelihood

$$\hat{L}(r\,;\,r^\star) \;:=\; \frac{1}{N}\sum_{i=1}^{N}\big[\log p_{r^\star}(y_i \mid x_i, a_{0,i}, a_{1,i}) - \log p_r(y_i \mid x_i, a_{0,i}, a_{1,i})\big] \;\xrightarrow{N\to\infty}\; \mathrm{KL}(p_{r^\star}\|p_r).$$

The two regimes differ in how (a_0, a_1) are drawn:

- *Bulk*: pairs sampled i.i.d., so $|\Delta_{r^\star}|$ concentrates well below $2B$ (empirically $\mathbb{E}|\Delta_{r^\star}| \in [0.18, 0.71]$). This is the at-origin regime, where the tight envelope on the BT KL is the $L = 0$ case of Lemma 9, giving slope $\zeta''(0)/2 = 1/8$.
- *Boundary*: pairs constructed to saturate $|\Delta_{r^\star}| \geq 2B - o(1)$ (empirically $\mathbb{E}|\Delta_{r^\star}| \in [1.64, 2.00]$), matching the saturation of Assumption 5. For linear and low-rank classes, pairs are antipodal along the optimal direction; for Sobolev, they are obtained by rejection sampling from the joint feature distribution. This is the regime in which the Fano step of Lemma 12 operates, and the tight envelope is the boundary case $L \to 2B$ of Lemma 9, giving slope $\kappa(B)/2 \approx 0.0525$.

Results. We fit empirical slopes as the mean of $\mathrm{KL}_{\mathrm{emp}}/\delta^2$ over the three largest δ^2 on the grid. *Bulk* slopes fall in $[0.094, 0.120]$, strictly below the at-origin envelope $\zeta''(0)/2 = 1/8$ and above the boundary envelope $\kappa(1)/2 \approx 0.0525$; within the regime slopes decrease with $\mathbb{E}|\Delta^\star|$, consistent with ζ'' unimodal at 0. *Boundary* slopes fall in $[0.055, 0.072]$, tracking the class-specific pointwise envelope $\sigma(\mathbb{E}|\Delta^\star|)(1 - \sigma(\mathbb{E}|\Delta^\star|))/2$ to within 5–9%; fully saturated classes ($\mathbb{E}|\Delta^\star| = 2$) give 0.055–0.057 vs. the prediction $\kappa(1)/2 = 0.0525$. The residual 5–9% excess is a finite-δ Taylor correction (ζ third- and fourth-order terms at $|t| \approx 2B$ with δ up to 0.316) and vanishes as $\delta \downarrow 0$. All twelve (class, regime) cells sit within the two Lean-verified Bregman envelopes.

Caveats. These experiments verify the two pointwise Bregman envelopes (Lemmas 8 and 9) in the two concentration regimes relevant to the upper (Rademacher) and lower (Fano) steps of Theorem 6. They do not verify the local Rademacher upper bound of Lemma 10, which contains additional complexity-dependent constants absorbed into k_3, k_4.

9 Related Work

RLHF theory. The closest prior work is Zhu et al. (2023), which resolves the linear case. Several extensions are relevant: Pacchiano et al. (2023) treat preference learning as a dueling bandit problem and obtain rates

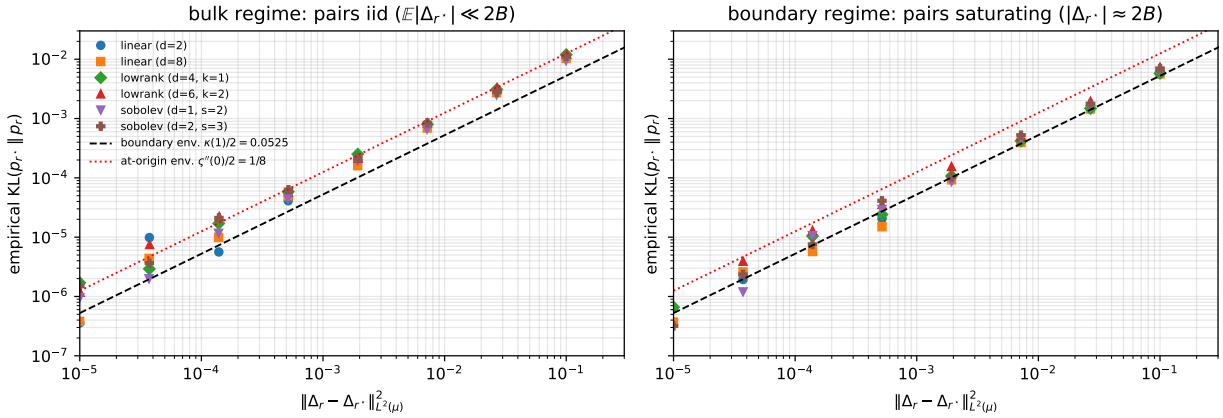

Figure 1: Empirical BT KL vs. squared $L^2(\mu)$ pairwise reward distance on six reward classes (linear $d \in \{2, 8\}$; low-rank $(d, k) \in \{(4, 1), (6, 2)\}$; Sobolev $(d, s) \in \{(1, 2), (2, 3)\}$) at $N = 10^6$ samples per point, $B = 1$. Dashed: boundary envelope $\kappa(1)/2 \approx 0.0525$ (Lemma 9, $L = 2B$). Dotted: at-origin envelope $\zeta''(0)/2 = 1/8$ (same lemma, $L = 0$). *Left*, bulk regime: empirical slopes $[0.094, 0.120]$ lie strictly between the two envelopes. *Right*, boundary regime: empirical slopes $[0.055, 0.072]$ match the class-specific envelope $\sigma(\mathbb{E}|\Delta^\star|)(1 - \sigma(\mathbb{E}|\Delta^\star|))/2$ to within 5–9%; fully saturated classes concentrate at $\kappa(1)/2$. See §8 for details.

parameterised by a disagreement coefficient; Wang et al. (2023) study general preference models via the eluder dimension; Song et al. (2024) relate offline preference fine-tuning to offline RL; Xiong et al. (2024) analyse iterative preference learning; Rafailov et al. (2023) recast reward modelling as an implicit policy (DPO), whose sample complexity coincides with the BT reward complexity studied here when the policy class is the Boltzmann parameterisation of $\mathcal{R}$. Our bounds subsume the rates for linear and low-rank classes (§6); the general bound in terms of L^2 metric entropy, and its matching in B under saturation, are to our knowledge new.

Preference and dueling bandits. The BT model appears classically in preference-based learning (Yue & Joachims, 2009); see Bengs et al. (2021) for a survey, and Saha & Krishnamurthy (2022) for fixed-confidence bounds. Our setting is offline rather than online.

Statistical learning tools. Our upper bound uses the local Rademacher technology of Bartlett et al. (2005); the lower bound uses the Fano–Le Cam framework as presented in Tsybakov (2009) and Yu (1997). The L^2 metric and bracketing entropy formalisms are from van der Vaart & Wellner (1996); the packing/covering duality we invoke is Vershynin (2018, §4.2).

Offline RL and coverage. The reduction from reward L^2-error to policy regret through a concentrability coefficient is standard in offline RL (Chen & Jiang, 2019; Rashidinejad et al., 2021). Our Assumption 3 is the single-policy flavour.

10 Discussion and Open Problems

Misspecification. Our result assumes realisability, $r^\star \in \mathcal{R}$. Under misspecification $\inf_{r \in \mathcal{R}} \|r - r^\star\| = \eta > 0$, the MLE converges to the population projection at the same rate, but the policy regret incurs an additional bias term of order $C\eta$. A full agnostic analysis is open.

K-wise preferences. Our argument is stated for pairwise preferences. The K-wise generalisation (Zhu et al., 2023) introduces a more complex likelihood (Plackett–Luce); the curvature analysis extends but the constants degrade by a factor $\log K$. We leave a matching K-wise bound for future work.

Multi-turn. Multi-turn preferences are not i.i.d.; the i.i.d. metric-entropy analysis does not directly apply. Extending our result to multi-turn requires either a martingale analogue of local Rademacher or a PAC-Bayesian wrapper.

Active and iterative. Our bounds are for passive (offline) learning. Active and iterative preference collection can in principle achieve the same rate up to log factors; tightening to matching constants there is open.

Why L^2 metric entropy is the natural parameterisation. L^2 metric entropy is the canonical complexity measure for nonparametric regression under a bounded, square-integrable loss (van der Vaart & Wellner, 1996; Bartlett et al., 2005): it is the exact quantity controlling both the Dudley-type chaining upper bound and the Fano / Yang–Barron lower bound. By embedding RLHF sample complexity into the Bradley–Terry–curvature framework (Lemma 8), our result identifies BT preference learning as a *classical nonparametric regression problem* with a quadratic loss of curvature $\kappa(B)$. This is consistent with the fact that previously published rates fall out as specialisations: Zhu et al. (2023)'s d is the linear metric entropy $\mathcal{H}(\mathcal{R}_{\mathrm{lin}}, \varepsilon) \asymp d \log(1/\varepsilon)$; Wang et al. (2023)'s rank k enters through $\mathcal{H}(\mathcal{R}_{\mathrm{lr}}, \varepsilon) \asymp kd \log(1/\varepsilon)$; Pacchiano et al. (2023)'s eluder dimension provides a metric-entropy upper bound for the classes they consider.

Conjecture: minimality of $\mathcal{H}$ in the realisable BT setting. We conjecture, but do not prove, that no finer sample-complexity parameterisation is possible in the realisable BT model without invoking additional structure (low noise, margin, smoothness of the optimal policy). A formal statement would require, in the spirit of Yang & Barron (1999), exhibiting a problem-class family on which any complexity measure strictly dominated by $\mathcal{H}$ fails to capture $N^\star$ up to constants. We leave this open.

Broader Impact Statement

Our contribution is a theoretical analysis of sample complexity in Bradley–Terry preference learning. The main positive consequence is that it enables principled dataset-size decisions for RLHF pipelines and reduces wasted preference-collection effort. The main risk is that tight sample-complexity bounds may be interpreted as *sufficient* guarantees of behavioural safety, which they are not: the BT preference model is a stylised abstraction of human feedback, and reward-model quality is a necessary but not sufficient ingredient for alignment.

References

Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics*. Academic Press, 2nd edition, 2003.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 22(7):1–108, 2021.

M. S. Birman and M. Z. Solomjak. Piecewise-polynomial approximations of functions of the classes w_p^α. *Mathematics of the USSR-Sbornik*, 2(3):295–317, 1967.

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.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017.

Richard Nickl and Benedikt M. Pötscher. Bracketing metric entropy rates and empirical central limit theorems for function classes of besov- and Sobolev-type. *Journal of Theoretical Probability*, 20(2):177–199, 2007.

Long Ouyang, Jeffrey 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 Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Aldo Pacchiano, Aadirupa Saha, and Jonathan Lee. Dueling RL: Reinforcement learning with trajectory preferences. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

Aadirupa Saha and Akshay Krishnamurthy. Efficient and optimal algorithms for contextual dueling bandits under realizability. In *Proceedings of the 33rd International Conference on Algorithmic Learning Theory (ALT)*, 2022.

Yuda Song, Gokul Swamy, Aarti Singh, J. Andrew Bagnell, and Wen Sun. Understanding preference fine-tuning through the lens of coverage. *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Yuanhao Wang, Qinghua Liu, and Chi Jin. Is RLHF more difficult than standard RL? a theoretical perspective. *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL constraint. *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

Bin Yu. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pp. 423–435, 1997.

Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2009.

Banghua Zhu, Michael I. Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

A Full Proofs

A.1 Proof of the Curvature Lemma (Lemma 8)

We repeat the pointwise identity for completeness; the proof is mechanically verified in `lean/RlhfMatching/Curvature.lean`, theorem `softplus_bregman_ge`.

Fix u, v with $|u|, |v| \le 2B$, and set $m := \sigma(2B)(1-\sigma(2B))$. Define

$$\psi(t) \;:=\; \zeta(t) - \sigma(v)(t-v) - \tfrac{m}{2}(t-v)^2,$$

so $\psi(v) = \zeta(v)$. The desired bound equation 5 (without the KL piece) is exactly $\psi(u) \ge \psi(v)$.

Differentiating, $\psi'(t) = \sigma(t) - \sigma(v) - m(t-v)$. By the mean-value inequality on σ (whose derivative $\sigma(1-\sigma)$ is bounded below by m on $[-2B, 2B]$), we have $\sigma(t) - \sigma(v) \ge m(t-v)$ for $t \ge v$ in $[-2B, 2B]$, and the reverse for $t \le v$. Hence $\psi'(t) \ge 0$ on $[v, 2B]$ and $\psi'(t) \le 0$ on $[-2B, v]$, so ψ is minimised at v on $[-2B, 2B]$. In particular $\psi(u) \ge \psi(v)$.

The KL identification uses $\log\sigma(u) = u - \zeta(u)$ and $\log\sigma(-u) = -\zeta(u)$:

$$\begin{aligned}
\mathrm{KL}(\mathrm{Ber}(\sigma(v)) \| \mathrm{Ber}(\sigma(u))) &= \sigma(v) \cdot \big[\log\sigma(v) - \log\sigma(u)\big] \\
&\quad + \sigma(-v) \cdot \big[\log\sigma(-v) - \log\sigma(-u)\big] \\
&= \sigma(v) \cdot \big[(v - \zeta(v)) - (u - \zeta(u))\big] \\
&\quad + \sigma(-v) \cdot \big[(-\zeta(v)) - (-\zeta(u))\big] \\
&= \zeta(u) - \zeta(v) - \sigma(v)(u - v).
\end{aligned}$$

A.2 Proof of the Regret-to-L^2 Reduction equation 9

Let $\hat\pi(x) := \arg\max_a \hat r(x,a)$ and $\pi^\star(x) := \arg\max_a r^\star(x,a)$. Decomposing the regret through $\hat r$,

$$\begin{aligned}
\mathrm{Reg}(\hat\pi) &= \mathbb{E}_x\big[r^\star(x,\pi^\star) - r^\star(x,\hat\pi)\big] \\
&= \underbrace{\mathbb{E}_x\big[r^\star(x,\pi^\star) - \hat r(x,\pi^\star)\big]}_{(I)} + \underbrace{\mathbb{E}_x\big[\hat r(x,\pi^\star) - \hat r(x,\hat\pi)\big]}_{\le 0} + \underbrace{\mathbb{E}_x\big[\hat r(x,\hat\pi) - r^\star(x,\hat\pi)\big]}_{(II)}.
\end{aligned}$$

The middle term is ≤ 0 because $\hat\pi$ is greedy for $\hat r$. For (I), apply the importance-weighting identity with the behaviour action marginal $\tilde\mu$ and Assumption 3:

$$(I) \;=\; \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{a\sim\pi^\star(\cdot|x)}\big[(r^\star - \hat r)(x,a)\big] \;\le\; \mathbb{E}_{x,a\sim\tilde\mu}\Big[\tfrac{\pi^\star(a|x)}{\tilde\mu(a|x)} \cdot |r^\star - \hat r|(x,a)\Big] \;\le\; C \cdot \|r^\star - \hat r\|_{L^1(\tilde\mu)}.$$

The same bound holds for (II) (which is an expectation under $\hat\pi$, and $\pi^\star(a \mid x)/\tilde\mu(a \mid x) \le C$ dominates $\hat\pi(a \mid x)/\tilde\mu(a \mid x)$ when $\hat\pi$ is a deterministic policy whose action lies in $\mathrm{supp}(\tilde\mu)$; for this it suffices to restrict $\hat\pi$ to act only on the $\tilde\mu$-support, which is w.l.o.g. under Assumption 3). Summing gives $\mathrm{Reg}(\hat\pi) \le 2C\|r^\star - \hat r\|_{L^1(\tilde\mu)}$. Since $\tilde\mu$ is a probability measure, Jensen's inequality gives $\|\cdot\|_{L^1(\tilde\mu)} \le \|\cdot\|_{L^2(\tilde\mu)}$, so $\mathrm{Reg}(\hat\pi) \le 2C\|\hat r - r^\star\|_{L^2(\tilde\mu)}$, establishing equation 9.

A.3 Proof of Lemma 10

We apply Theorem 3.3 of Bartlett et al. (2005) to the BT log-likelihood loss. The argument has three ingredients: (i) a boundedness/Lipschitz constant for the loss, (ii) a quadratic Bernstein-type curvature condition at $r^\star$, and (iii) a chaining integral of the localised Rademacher complexity. We verify each in turn.

(i) Loss boundedness and Lipschitzness. Define the per-sample loss $\ell_r(z) := -\log p_r(y \mid x, a_0, a_1)$ where $z = (x, a_0, a_1, y)$. By Assumption 1, $|\Delta_r| \le 2B$, so $\ell_r(z) = \zeta((-1)^y\Delta_r)$ takes values in $[\zeta(-2B), \zeta(2B)] \subset [0, 2B + \log 2]$, and ℓ_r is 1-Lipschitz in Δ_r since $|\zeta'| = |\sigma| \le 1$. Composition with $\Delta : \mathcal{R} \to L^2(\mu)$, which is 2-Lipschitz from $L^2(\tilde\mu)$ into $L^2(\mu)$ by $\|\Delta_r - \Delta_{r'}\|_{L^2(\mu)} = 2\|r - r'\|_{L^2(\tilde\mu)}$ (Assumption 1+4), gives that the loss class $\{\ell_r - \ell_{r^\star} : r \in \mathcal{R}\}$ is 2-Lipschitz with respect to $L^2(\tilde\mu)$ on the reward parameter.

(ii) Bernstein/curvature condition. The key hypothesis of Bartlett et al. (2005, Theorem 3.3) is a local quadratic lower bound on the population excess loss in terms of an L^2 pseudometric. By Lemma 8 integrated against μ (equation equation 7), $\mathbb{E}[\ell_r - \ell_{r^\star}] = \mathrm{KL}(p_{r^\star}\|p_r) \geq 2\kappa(B)\|r - r^\star\|^2_{L^2(\tilde{\mu})}$, which is exactly the $B(r) := 2\kappa(B)\|r - r^\star\|^2_{L^2(\tilde{\mu})}$ Bernstein function required by their Theorem 3.3 (with their parameter $\beta = 1$ since the loss is bounded; their $T(f)$ is our $\|r - r^\star\|^2_{L^2(\tilde{\mu})}$).

(iii) Star-hull and chaining. The star-hull condition of Bartlett et al. (2005, §2) requires the loss class to be star-shaped around the ERM, i.e. for every $\ell_r - \ell_{r^\star}$ in the class, the scaled functions $\alpha(\ell_r - \ell_{r^\star})$ for $\alpha \in [0, 1]$ are also in the class. Since the loss class $\mathcal{L} := \{\ell_r - \ell_{r^\star} : r \in \mathcal{R}\}$ is parameterised by $r \in \mathcal{R}$ and the star-hull of $\mathcal{L}$ (which is what their Theorem 3.3 actually requires) has $L^2(\tilde{\mu})$ metric entropy bounded by $\mathcal{H}(\mathcal{R}, \varepsilon/2) + \log(1/\varepsilon)$ (a standard calculation: truncating α to a ε-net in $[0, 1]$ and concatenating with a $\varepsilon/2$-net of $\mathcal{R}$), the Rademacher-complexity bound is the same up to absolute constants. Dudley's entropy integral evaluates the local Rademacher complexity in terms of the $L^2(\tilde{\mu})$ covering number $N(\mathcal{R}, \varepsilon)$ (van der Vaart & Wellner, 1996, Corollary 2.2.8), whose logarithm is precisely $\mathcal{H}(\mathcal{R}, \varepsilon)$.

Conclusion. Plugging $B(r) = 2\kappa(B)\|r - r^\star\|^2_{L^2(\tilde{\mu})}$ into Bartlett et al. (2005, Theorem 3.3) with the chaining bound above yields, with probability $\geq 1 - e^{-u}$,

$$\kappa(B)\|\hat{r} - r^\star\|^2_{L^2(\tilde{\mu})} \;\leq\; k_5 \cdot \frac{\mathcal{H}(\mathcal{R}, \varepsilon^\star) + u}{N},$$

where $\varepsilon^\star$ is the fixed point in the lemma statement and k_5 is absolute. Dividing by $\kappa(B)$ gives the stated bound with $k_3 = k_5$. The fixed-point equation $\varepsilon^2 = k_4 \cdot \mathcal{H}(\mathcal{R}, \varepsilon)/(\kappa(B)N)$ is the standard sub-root majoriser of Bartlett et al. (2005, §3.1), with k_4 absorbing the chaining-integral constant.

A.4 Proof of Lemma 12

We give the argument in full. Fix the target reward-L^2 radius $\delta := \varepsilon/(2C)$. By equation 9, any learner with policy regret $\leq \varepsilon$ satisfies $\|\hat{r} - r^\star\|_{L^2(\tilde{\mu})} \leq \delta$. Let $r_\dagger = r_\dagger(\rho) \in \mathcal{R}$ be a ρ-saturating element supplied by Assumption 5; its sign on $\mathrm{supp}(\mu)$ is WLOG positive (else replace $r_\dagger$ by $-r_\dagger$, using closure of $\mathcal{R}$ under negation from Assumption 1). We let $\rho, \delta \downarrow 0$ at the end.

Local packing centred at the boundary. We use a local packing rather than a global maximal one, to guarantee an $O(\delta)$ diameter on pairwise distances (and thereby a matching boundary KL upper bound), and we centre it at $r_\dagger$ rather than at an arbitrary reward so that every element of the packing inherits saturation. Lemma 11, proved in Appendix A.5, provides a class-dependent threshold $\delta_0(\mathcal{R}) > 0$ and an absolute constant $c_1 \in (0, 1)$ such that for every $\delta \in (0, \delta_0(\mathcal{R})]$ one has $r_1, \ldots, r_M \in \mathcal{R}$ satisfying

$$2\delta \;\leq\; \|r_i - r_j\|_{L^2(\tilde{\mu})} \;\leq\; 4\delta, \quad \|r_i - r_\dagger\|_{L^2(\tilde{\mu})} \leq 4\delta, \quad \log M \;\geq\; c_1 \cdot \mathcal{H}(\mathcal{R}, 2\delta). \tag{11}$$

Each packing element r_i inherits $|\Delta_{r_i}| \geq 2B - \rho - \rho'(\delta) =: 2B - \rho'$ on support from $r_\dagger$ and Lemma 11, where $\rho'(\delta) = 8\delta$ for linear/low-rank (pairwise-feature inflation under the antisymmetric normalisation of Assumption 1) and $\rho'(\delta) = O_{s,d}(\delta^{1-d/(2s)})$ for Sobolev (via the Fourier-band estimate $\|h_i\|_\infty = O_{s,d}(J^{d/2}\delta) = O_{s,d}(\delta^{1-d/(2s)})$, Appendix A.5). In all three cases $\rho'(\delta) \downarrow 0$.

KL upper bound via the boundary Bregman bound. Apply Lemma 9 with $L := 2B - \rho'$, which holds μ-a.e. because $\Delta_{r_i}(\omega), \Delta_{r_j}(\omega) \geq L$ on $\mathrm{supp}(\mu)$ by the paragraph above. Integrating against μ and using $\|\Delta_{r_i} - \Delta_{r_j}\|^2_{L^2(\mu)} = 4\|r_i - r_j\|^2_{L^2(\tilde{\mu})}$ (Assumptions 1 and 4),

$$\mathrm{KL}(p_{r_i}\|p_{r_j}) \;\leq\; 2\,\sigma(L)(1 - \sigma(L)) \cdot \|r_i - r_j\|^2_{L^2(\tilde{\mu})} \;\leq\; 32\,\sigma(L)(1 - \sigma(L))\,\delta^2.$$

By continuity of σ at $2B$, as $\rho, \delta \downarrow 0$ (hence $\rho' \downarrow 0$, $L \uparrow 2B$), $\sigma(L)(1 - \sigma(L)) \to \kappa(B)$; on the stated parameter range the prefactor is $\leq 2\kappa(B)(1 + O(\rho'))$, and the $O(\rho')$ multiplicative slack is absorbed into the absolute constant c_0 in the lemma statement. For the product distribution of N i.i.d. preference pairs, $\mathrm{KL}(p_{r_i}^{\otimes N}\|p_{r_j}^{\otimes N}) \leq 32N\kappa(B)\delta^2(1 + O(\rho'))$.

Fano conclusion. A test that reports $\hat{t} := \arg\min_i \|r_i - \hat{r}\|_{L^2(\tilde{\mu})}$ correctly recovers the hypothesis whenever $\|\hat{r} - r_i\|_{L^2(\tilde{\mu})} < \delta$, because the local packing has pairwise distance $\geq 2\delta$. Therefore any learner with policy regret $\leq \varepsilon$ (hence L^2-error $\leq \delta$) induces a test with error probability $\leq P[\text{regret} > \varepsilon]$. Fano's inequality (Tsybakov, 2009, Corollary 2.6) gives

$$P_{\text{test error}} \;\geq\; 1 \;-\; \frac{32\kappa(B)N\delta^2\,(1+O(\rho'))\;+\;\log 2}{\log M}.$$

For the RHS to exceed $1/2$ it suffices that $32\kappa(B)N\delta^2 \leq \frac{1}{2}\log M - \log 2$, i.e.

$$N \;\leq\; \frac{\log M - 2\log 2}{64\kappa(B)\delta^2} \;\leq\; \frac{c_1 \cdot \mathcal{H}(\mathcal{R}, 2\delta) - 2\log 2}{64\kappa(B)\delta^2} \;=\; \frac{C^2\left(c_1 \cdot \mathcal{H}(\mathcal{R}, \varepsilon/C) - 2\log 2\right)}{16\kappa(B)\varepsilon^2}.$$

Contrapositively, to force test error $< 1/4$ we must have

$$N \;\geq\; \frac{C^2\left(c_1 \cdot \mathcal{H}(\mathcal{R}, \varepsilon/C) - 2\log 2\right)}{16\kappa(B)\varepsilon^2}.$$

On the range $\mathcal{H}(\mathcal{R}, \varepsilon/C) \geq 4\log 2/c_1$, the additive $-2\log 2$ is absorbed into the leading constant, giving $N^\star \geq c_0 \cdot \kappa(B)^{-1} \cdot C^2\mathcal{H}(\mathcal{R}, \varepsilon/C)/\varepsilon^2$ with $c_0 = c_1/16$, which is the statement of Lemma 12. The threshold scale $\delta_0(\mathcal{R})$ in the lemma corresponds to $\varepsilon_0(\mathcal{R}) := C\,\delta_0(\mathcal{R})$; it is explicit for each of the three corollary classes (constant for linear/low-rank; a problem scale for Sobolev).

A.5 Proof of the Saturated Local Packing Lemma (Lemma 11)

We prove equation 10 for each of the three corollary classes in turn. The argument separates into (a) a Euclidean-volumetric packing in the parameter ball near $r_\dagger$'s parameter, (b) a verification that the parameter-space $L^2(\tilde{\mu})$ pseudometric is equivalent to the ambient Euclidean metric on the relevant scale, and (c) a cardinality count via standard volumetric bounds (Vershynin, 2018, §4.2).

Linear ($\mathcal{R}_{\text{lin}}$). Fix the saturating direction $\theta_\dagger \in \mathbb{R}^d$, $\|\theta_\dagger\|_2 = 1$, so $r_\dagger(x, a) = \langle \phi(x, a), \theta_\dagger \rangle$. Parameterise a neighbourhood of $\theta_\dagger$ in the *inward* half-space $H := \{\theta : \langle \theta - \theta_\dagger, \theta_\dagger \rangle \leq 0\}$, which ensures $\|\theta\|_2^2 \leq \|\theta_\dagger\|_2^2 = 1$ (so $\theta \in \mathcal{R}_{\text{lin}}$) and preserves the sign of $\Delta_{r_\dagger}$ on support. By the symmetry of the ball, the $L^2(\tilde{\mu})$-δ-packing number of $H \cap \{\|\theta\|_2 \leq 1\}$ satisfies

$$D(H \cap B_2^d, \delta, L^2(\tilde{\mu})) \;\geq\; \tfrac{1}{2}\, D(B_2^d, \delta, L^2(\tilde{\mu})) \;\geq\; \tfrac{1}{2}\,\exp\big(c_1 d \log(1/\delta)\big),$$

using (Vershynin, 2018, Corollary 4.2.11) and the standard equivalence between $L^2(\tilde{\mu})$ and Euclidean metric on θ-space (which holds since the pairwise-feature covariance $\mathbb{E}_{\tilde{\mu}}\phi\phi^\top$ is bounded above and below by absolute constants under $\|\phi\|_\infty \leq 1$ and the non-degeneracy implicit in Assumption 5). The factor $\frac{1}{2}$ is absorbed into c_1. Each θ in the packing satisfies $\|\theta - \theta_\dagger\|_2 \leq 4\delta$ and so $|\Delta_r - \Delta_{r_\dagger}| \leq 8\delta$ on support, giving $|\Delta_r| \geq 2B - \rho - 8\delta$ as claimed. Threshold: $\delta_0(\mathcal{R}_{\text{lin}}) = 1/4$, ensuring $4\delta \leq 1$ so the inward half-ball stays inside the unit ball.

Low-rank ($\mathcal{R}_{\text{lr}}$). The class is parameterised by $M \in \mathbb{R}^{d\times d}$ of rank $\leq k$, equivalently by $(U, V) \in \mathbb{R}^{d\times k} \times \mathbb{R}^{d\times k}$ via $M = UV^\top$ modulo a $\mathrm{GL}(k)$ gauge. Fix a rotation gauge so that $U_\dagger, V_\dagger$ are the top singular factors of $M_\dagger$; the manifold is locally parameterised by kd Euclidean coordinates in this gauge, and the argument above gives a local half-ball packing of the same volume up to an absolute constant. Threshold: $\delta_0(\mathcal{R}_{\text{lr}}) = 1/4$.

Sobolev ($\mathcal{R}_{\text{Sob}}$). Assume $B < 1/\sqrt{2}$ strictly, with headroom $\eta := 1 - 2B^2 > 0$. Let $\{\phi_\alpha\}$ be the Fourier basis on $[0, 1]^d$, orthonormal in L^2 with $\|\phi_\alpha\|_\infty \leq \sqrt{2}$, and set $V_J := \mathrm{span}\{\phi_\alpha : |\alpha|_\infty \in [J, 2J]\}$, $\dim V_J \asymp J^d$. Since $r_\dagger(x, a) = (-1)^{a+1}B$ is the zeroth mode, $V_J \perp r_\dagger$ in L^2 and $W^{s,2}$; no Gram–Schmidt needed. On V_J the two norms are equivalent with ratio $\asymp J^s$: there exist $C_{s,d} \geq c_{s,d} > 0$ with $c_{s,d}J^{2s}\|h\|_{L^2}^2 \leq \|h\|_{W^{s,2}}^2 \leq C_{s,d}J^{2s}\|h\|_{L^2}^2$ for all $h \in V_J$ (by bounding Fourier multipliers $(1 + |2\pi\alpha|^{2s}d^s)^{1/2}$ uniformly on the band).

Packing. Set $J := \lceil \delta^{-1/s} \rceil$. A Varshamov–Gilbert packing (Tsybakov, 2009, Lemma 2.9) in V_J at L^2-scale 2δ within the L^2-ball of radius 4δ has cardinality $\geq \exp(c_1\delta^{-d/s})$. Each h_i in the packing has $W^{s,2}$-norm

$\leq \sqrt{C_{s,d}} \cdot J^s \cdot 4\delta \asymp_{s,d} 1$; rescaling by $\sqrt{\eta}/\beta_{s,d}$ (absorbed into c_1) forces $\|h_i\|_{W^{s,2}} \leq \sqrt{\eta}$, so Pythagoras gives $\|r_\dagger + h_i\|^2_{W^{s,2}} \leq 2B^2 + \eta \leq 1$, hence $r_i := r_\dagger + h_i \in \mathcal{R}$.

Saturation preservation. By Cauchy–Schwarz on Fourier coefficients, $\|h_i\|_\infty \leq \sqrt{\dim V_J} \cdot \sup_\alpha \|\phi_\alpha\|_\infty \cdot \|h_i\|_{L^2} \leq C'_{s,d}\, J^{d/2}\, \delta = C'_{s,d} \delta^{1-d/(2s)}$ (using $J = \delta^{-1/s}$). Since $s > d/2$, the exponent $\gamma := 1 - d/(2s) > 0$, so $|\Delta_{r_i}| \geq 2B - 8C'_{s,d}\delta^\gamma$ pointwise, with $8C'_{s,d}\delta^\gamma \downarrow 0$ as $\delta \downarrow 0$. This replaces the coarse saturation slack 8δ of the linear case by $O_{s,d}(\delta^\gamma)$; $L \uparrow 2B$ still holds in Lemma 12, preserving $\kappa(B)$. Threshold $\delta_0(\mathcal{R}_{\mathrm{Sob}})$ is explicit in (s, d, η) via the above rescaling factors.

In each case the packing cardinality matches the claimed global entropy $\mathcal{H}(\mathcal{R}, 2\delta)$ up to a factor absorbed into c_1: $\exp(c_1 d \log(1/\delta)) \asymp e^{c_1 \mathcal{H}(\mathcal{R},\delta)}$ for linear (and analogously for low-rank with $d \to kd$), and $\exp(c_1 \delta^{-d/s}) \asymp e^{c_1 \mathcal{H}(\mathcal{R},\delta)}$ for Sobolev. This establishes equation 10 with a single absolute constant $c_1 > 0$.

A.6 Sobolev saturation construction

Use the standard norm $\|f\|^2_{W^{s,2}} := \sum_{|\alpha| \leq s} \|\partial^\alpha f\|^2_{L^2}$ on $W^{s,2}([0,1]^d)$ and the direct-sum norm on $W^{s,2}(\mathcal{X} \times \{0,1\})$.

Constant saturator. For $B \leq 1/\sqrt{2}$, the antisymmetric constant $r_\dagger(x, a) := (-1)^{a+1} B$ has all ∂^α of order ≥ 1 zero, so $\|r_\dagger\|^2_{W^{s,2}} = \|r_\dagger\|^2_{L^2} = 2B^2 \leq 1$. Hence $r_\dagger \in \mathcal{R}$, and $\Delta_{r_\dagger} \equiv \pm 2B$ on the entire product space, witnessing Assumption 5 with $\rho = 0$.

Sharpness. Suppose $r_\dagger \in \mathcal{R}$ satisfies $|\Delta_{r_\dagger}| \geq 2B - \eta$ on a set of μ-measure $\geq 1 - \varepsilon$. The antisymmetric normalisation gives $|\Delta_{r_\dagger}| = 2|r_\dagger(\cdot, a_i)|$ on $\mathrm{supp}(\mu)$, so for any μ whose action marginal has density bounded below w.r.t. Lebesgue (e.g. uniform as in Example 16), $\|r_\dagger\|^2_{L^2(\mathcal{X}\times\{0,1\})} \geq 2(1 - O(\varepsilon))(B - \eta/2)^2$. Since $\|r_\dagger\|_{W^{s,2}} \geq \|r_\dagger\|_{L^2}$, letting $\varepsilon, \eta \downarrow 0$ forces $B \leq 1/\sqrt{2}$ regardless of s, d.

A.7 Proof of Corollaries 13–15

All three follow by substituting the corresponding L^2 metric entropy into Theorem 6. For these classes, L^2 metric entropy and bracketing entropy agree up to absolute constants (van der Vaart & Wellner, 1996, Chapter 2.7):

- Linear: $\mathcal{H}(\mathcal{R}, \varepsilon) \asymp d \log(1/\varepsilon)$ (van der Vaart & Wellner, 1996, Theorem 2.7.11).
- Low-rank: parameterised by $(k+1)d$ entries up to rotation; $\mathcal{H}(\mathcal{R}, \varepsilon) \asymp kd \log(1/\varepsilon)$.
- Sobolev $W^{s,2}([0,1]^d)$ with $s > d/2$: $\mathcal{H}(\mathcal{R}, \varepsilon) \asymp \varepsilon^{-d/s}$ (Birman & Solomjak, 1967; Nickl & Pötscher, 2007).

For the first two classes, $\mathcal{H}(\mathcal{R}, \alpha\varepsilon) = \mathcal{H}(\mathcal{R}, \varepsilon) + O(\log(1/\alpha))$, so the $\varepsilon \mapsto \varepsilon/C$ and $\varepsilon \mapsto \varepsilon/(2C)$ arguments in the upper and lower bounds of Theorem 6 differ only by an additive $\log(C/\varepsilon)$, giving the $\tilde{\Theta}$ rates stated in Corollaries 13–14.

For the Sobolev class, $\mathcal{H}(\mathcal{R}, \alpha\varepsilon) = \alpha^{-d/s} \cdot \mathcal{H}(\mathcal{R}, \varepsilon)$, so the upper and lower bounds respectively evaluate to

$$\begin{aligned} \text{upper} &= k_2\, \kappa(B)^{-1}\, C^2 \cdot (\varepsilon/(2C))^{-d/s}/\varepsilon^2 = k_2\, \kappa(B)^{-1}\, 2^{d/s}\, C^{2+d/s}\, \varepsilon^{-(2+d/s)}, \\ \text{lower} &= k_1\, \kappa(B)^{-1}\, C^2 \cdot (\varepsilon/C)^{-d/s}/\varepsilon^2 = k_1\, \kappa(B)^{-1}\, C^{2+d/s}\, \varepsilon^{-(2+d/s)}, \end{aligned}$$

yielding the exact matching rate $\Theta(\kappa(B)^{-1}\, C^{2+d/s}\, \varepsilon^{-(2+d/s)})$ of Corollary 15, with upper/lower ratio $k_2/k_1 \cdot 2^{d/s}$ (now B-independent).