POINTWISE OR PAIRWISE: WHEN DO PAIRWISE LOSSES HELP REWARD LEARNING, PROVABLY?

Anonymous authors
Paper under double-blind review

ABSTRACT

Pairwise losses are increasingly used for reward learning even when pointwise rewards are observed, with mixed empirical results. *When and why do pairwise losses outperform pointwise losses?* We study this question in a grouped offline contextual-bandit setting allowing multiple actions per context, capturing many reward learning scenarios. We compare VALUE REGRESSION (VR), which regresses observed rewards pointwise, with VALUE DIFFERENCE REGRESSION (VDR), which regresses reward differences between a pair of actions sampled under the same context. We consider a semiparametric model where the mean reward is the sum of a learnable action-dependent component and an arbitrary context-dependent yet *action-independent* nuisance, capturing context-specific disturbances. Using a unified localized analysis, we prove finite-sample regression guarantees for finite and linear function classes and translate them into offline-regret bounds. For finite classes, VDR eliminates the misspecification term in the VR bound and improves a reward-scale-dependent error term by averaging over actions within each context, a benefit absent from the corresponding VR term. For linear classes, neither method uniformly dominates: within-context differencing removes nuisance-induced bias but may increase estimation variance relative to using absolute rewards when the misspecification is sufficiently low. This yields a feature geometry-dependent *bias–variance tradeoff*, which we corroborate with numerical experiments.

1 INTRODUCTION

Modern machine learning increasingly relies on learned reward functions to turn limited supervision into reusable objectives that can be optimized at scale. Examples include reinforcement learning from human feedback (RLHF) (Christiano et al., 2017), Best-of-N test-time scaling (Huang et al., 2025), reward- or value-guided reasoning and search (Guan et al., 2025; Li & Li, 2025), and, outside language modeling, automated essay scoring and speech-quality assessment (Cao et al., 2026; Xie et al., 2022). In many of these settings, supervision is *pointwise and numerical*, such as human ratings (Wang et al., 2024b), verifier scores (Lambert et al., 2025), or search-derived value estimates (Guan et al., 2025). In such cases, it is natural to regress directly on the observed rewards.

Several methods, on the other hand, *deliberately* construct pairwise targets, such as preferences or reward differences between actions under the same context, even when pointwise values are available. In language-model reasoning, Guan et al. (2025) convert search-derived Q-values into high-Q/low-Q pairs and use a Bradley–Terry loss (Bradley & Terry, 1952) to train a process reward model. Cui et al. (2024) construct preference pairs from numerical GPT-4 ratings and use a margin-adjusted Bradley–Terry loss to train a reward model. Beyond language modeling, pairwise squared losses on numerical label differences have been used for magnitude-preserving ranking (Cortes et al., 2007) and as a component of regression objectives (Zhu et al., 2025). Related pairwise objectives also appear even when numerical labels, such as essay scores and speech-quality scores, are available (Cao et al., 2026; Xie et al., 2022). In these examples, pairing is an *algorithmic choice* rather than an inherent property of the raw feedback.

Despite the growing use of pairwise losses, the empirical benefits of pairing are not consistent. Some studies report ablation results supporting the advantages of pairing over pointwise objectives. For example, pairwise preference training improves mathematical reasoning accuracy over pointwise Q-value regression (Guan et al., 2025) and speech quality ordering accuracy over pointwise score

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.

regression (Cao et al., 2026). However, some other studies report that pairing does not always improve over pointwise alternatives. In reward-learning ablations on Atari games, neither preference-based training nor pointwise regression to returns consistently yields better downstream returns (Christiano et al., 2017). In ranking experiments, either objective can outperform the other depending on the choice of kernel (Pahikkala et al., 2013). These mixed results motivate a more rigorous understanding of which properties of the learning problem determine whether pairing helps downstream decisions.

We therefore ask:

Given pointwise reward data, when does learning a reward function with pairwise losses lead to provably better downstream decisions than learning with pointwise losses?

We make significant progress towards answering this question through the lens of contextual bandits along with representative pairwise and pointwise losses.

1.1 Theoretical Setup: Grouped Offline Reward Learning

Let $\mathcal{X}$ be a context (prompt) space, $\mathcal{A}$ be an action (response) space with $|\mathcal{A}| < \infty$, and $\mathcal{F} \subseteq \{f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}\}$ be a value function space. $r^\star : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is an *unknown* reward function.

Grouped dataset. With a prompt distribution $\rho \in \Delta(\mathcal{X})$ and behavior policy $\pi_{\mathrm{ref}}(\cdot \mid x) \in \Delta(\mathcal{A})$, the learner is given a (passively) collected, *grouped* offline dataset

$$\mathcal{D} \coloneqq \{(x_i, a_{i,j}, r_{i,j})\}_{i\in[N_x],\, j\in[N_a]}, \qquad N \coloneqq N_x N_a,$$

where the contexts $x_i \sim \rho$ are i.i.d.; conditional on x_i, the actions $a_{i,j} \sim \pi_{\mathrm{ref}}(\cdot \mid x_i)$ are i.i.d.; and $r_{i,j} = r^\star(x_i, a_{i,j}) + \eta_{i,j}$ satisfying $|r_{i,j}| \le R_{\max}$ *a.s.* for a known $R_{\max} > 0$. Conditional on $\mathcal{Z} \coloneqq \{(x_i, a_{i,j})\}_{i,j}$, the $\eta_{i,j}$'s are independent, their conditional laws depend on $\mathcal{Z}$ only through $(x_i, a_{i,j})$, and they satisfy $\mathbb{E}[\eta_{i,j} \mid \mathcal{Z}] = 0$, $\mathbb{E}[\eta_{i,j}^2 \mid \mathcal{Z}] \le \sigma^2$, and $|\eta_{i,j}| \le 2R_{\max}$.[1] It may be that $N_a \ge 2$, as is the case when multiple candidate responses are sampled for reward modeling or group-based policy optimization (Cui et al., 2024; Guan et al., 2025; Shao et al., 2024). Henceforth, let $\mathbb{E}_{x,a} \coloneqq \mathbb{E}_{x\sim\rho, a\sim\pi_{\mathrm{ref}}(\cdot|x)}$ whenever there is no ambiguity.

Offline regret. The performance of a policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ is measured with the **offline regret**:

$$\mathrm{Reg}(\pi) \coloneqq V(r^\star) - V(\pi), \quad V(\pi) \coloneqq \mathbb{E}_{x\sim\rho, a\sim\pi(\cdot|x)}\left[r^\star(x,a)\right], \tag{1}$$

where $V(\pi)$ is the **value** of the policy π, and with a slight notation abuse, for $g \in \mathcal{F}$, $V(g)$ is the value of the Greedy policy $x \mapsto \arg\max_{a\in\mathcal{A}} g(x,a)$, following the standard convention in value-based offline RL (Chen & Jiang, 2019). Define the ***centering operator*** $\mathcal{C} : f \mapsto \mathcal{C}f$ as

$$\mathcal{C}f(x,a) \coloneqq f(x,a) - \mathbb{E}_{a\sim\pi_{\mathrm{ref}}(\cdot|x)}[f(x,a)],$$

which is action-gap-preserving in the sense that $\mathcal{C}f(x,a) - \mathcal{C}f(x,a') = f(x,a) - f(x,a')$. Then, based on standard coverage-based arguments (Rashidinejad et al., 2022), we show that a bound on the *centered prediction error* of a given value estimator $\hat{f}$, $\mathbb{E}[(\mathcal{C}\hat{f} - \mathcal{C}r^\star)^2]$, implies an offline regret bound via Greedy or Pessimism; see Proposition B.1 in Appendix B.

Weak realizability. We impose the following semiparametric-type weak realizability assumption on $r^\star$ (this subsumes realizability as it may be that $b^\star = 0$):

Assumption 1 (Weak Realizability)**.** *There exist $f^\star \in \mathcal{F}$ and $b^\star : \mathcal{X} \to \mathbb{R}$ such that $r^\star(x,a) = f^\star(x,a) + b^\star(x)$. In particular, we do not assume realizability on either $b^\star$ or $r^\star$; only on $f^\star$.*

As it may be that $r^\star \notin \mathcal{F}$, we define the **misspecification error**:

$$\varepsilon_{\mathrm{aprx}} \coloneqq \inf_{f\in\mathcal{F}} \mathbb{E}_{x,a}\left[(f(x,a) - r^\star(x,a))^2\right] \ge 0.$$

This arises in many scenarios. Under Bradley–Terry preferences (Bradley & Terry, 1952), adding a prompt-only term changes neither the preference model nor the optimal KL-regularized policy (Rafailov et al., 2023). Action-independent nuisances likewise arise in semiparametric contextual bandits and regression (Carranza et al., 2023; Foster & Syrgkanis, 2023; Greenewald et al., 2017; Kim & Paik, 2019; Kim et al., 2025; Krishnamurthy et al., 2018; Nie & Wager, 2021).

[1] Indeed, by triangle inequality, $|\eta_{i,j}| \le |r_{i,j}| + |r^\star(x_i, a_{i,j})| \le R_{\max} + |\mathbb{E}[r_{i,j} \mid x_i, a_{i,j}]| \le 2R_{\max}$.

Table 1: Centered squared prediction error bounds for VALUE REGRESSION (VR) and VALUE DIFFERENCE REGRESSION (VDR) over finite and linear classes. We omit universal constants; for $\mathcal{F}_{\mathrm{lin}}$, we also suppress $\log \frac{1}{\delta}$ factors. For $\mathcal{F}_{\mathrm{lin}}$, ε_{z} and $\varepsilon_{\mathrm{blk}}$ measure the geometric impact of misspecification on VR ($\varepsilon_{\mathrm{blk}} := \varepsilon_{\mathrm{x}} + \varepsilon_{\mathrm{a}}/N_a$ in Eqn. (5)). $\alpha_0 \geq 1$ (resp. $\alpha_{\mathrm{C}} \geq 1$) quantifies how well-conditioned the (resp. centered) features are; see Assumption 2 (resp. Assumption 3). $1 \leq d_{\mathrm{C}} \leq d$ is the dimension of the identifiable subspace for centered features.

	(VR)	(VDR)
$\lvert\mathcal{F}\rvert < \infty^{\dagger}$	$2\varepsilon_{\mathrm{aprx}} + \left(\frac{F^2 + FR_{\max}}{N_x} + \frac{\sigma^2}{N}\right)\log\frac{\lvert\mathcal{F}\rvert}{\delta}$ **(Eqn. (2) of Theorem 3.1)**	$\left(\frac{F^2}{N_x} + \frac{\sigma^2 + FR_{\max}}{N}\right)\log\frac{\lvert\mathcal{F}\rvert}{\delta}$ **(Eqn. (3) of Theorem 3.1)**
$\mathcal{F}_{\mathrm{lin}}{}^{\ddagger}$	$\varepsilon_{\mathrm{aprx}} + \frac{\varepsilon_{\mathrm{blk}} d}{N_x} + \frac{\varepsilon_{\mathrm{z}} d}{N_x^2} + \frac{\sigma^2 d}{N} + \frac{R_{\max}^2 \alpha_0^2 d}{N^2}$ **(Theorem 4.1)**	$\frac{\sigma^2 d_{\mathrm{C}}}{N} + \frac{R_{\max}^2 \alpha_{\mathrm{C}}^2 d_{\mathrm{C}}}{N^2}$ **(Theorem 4.2)**

† $F := \sup_{f\in\mathcal{F}} \|f\|_{\infty}$, and F may be much smaller than $R_{\max}$.

‡ N_x must be sufficiently large; see the theorems. Also, $\varepsilon_{\mathrm{z}} = \varepsilon_{\mathrm{blk}} = 0$ when $\varepsilon_{\mathrm{aprx}} = 0$.

Algorithms: VR and VDR. We now introduce the two value estimators studied in this paper, which formalize the pointwise and pairwise approaches described at the beginning. For the pointwise loss, we consider VALUE REGRESSION (VR), which directly regresses the observed rewards:

$$\hat{f}_N^{\mathrm{VR}} \in \arg\min_{f\in\mathcal{F}} \left\{ \widehat{\mathcal{L}}^{\mathrm{VR}}(f) := \frac{1}{N_x}\sum_{i=1}^{N_x} \frac{1}{N_a}\sum_{j=1}^{N_a} \big(f(x_i, a_{i,j}) - r_{i,j}\big)^2 \right\}. \tag{VR}$$

For the pairwise loss, there are many variations such as the Bradley–Terry loss (Bradley & Terry, 1952), margin loss (Chen et al., 2024), and squared difference loss (Xie et al., 2022). For ease of exposition, we choose a simple variant, VALUE DIFFERENCE REGRESSION (VDR), that regresses the reward difference: denoting $\Delta_{i,j,k}(f) := f(x_i, a_{i,j}) - f(x_i, a_{i,k})$ and $\widehat{\Delta}_{i,j,k}(r) := r_{i,j} - r_{i,k}$,

$$\hat{f}_N^{\mathrm{VDR}} \in \arg\min_{f\in\mathcal{F}} \left\{ \widehat{\mathcal{L}}^{\mathrm{VDR}}(f) := \frac{1}{N_x}\sum_{i=1}^{N_x} \frac{1}{\binom{N_a}{2}} \sum_{1\leq j<k\leq N_a} \Big(\Delta_{i,j,k}(f) - \widehat{\Delta}_{i,j,k}(r)\Big)^2 \right\}. \tag{VDR}$$

Using the centering identity of Zhu et al. (2025, Corollary 2), $\widehat{\mathcal{L}}^{\mathrm{VDR}}$ can be evaluated exactly in linear time by centering residuals within each context (the essence is the identity $\mathbb{E}[(X-Y)^2] = 2\,\mathbb{E}[(X - \mathbb{E}[X])^2]$ where X and Y are i.i.d). Pair-selection strategies are also used in preference learning (Guan et al., 2025; Pukdee et al., 2026), but are outside our scope.

1.2 OUR CONTRIBUTIONS

Our main departure from the prior literature is the *grouped* offline dataset $\mathcal{D}$: standard formulations typically require $N_a = 1$, whereas we allow for $N_a \geq 2$ in which case the N triplets become *non-i.i.d.* Thus, statistical accuracy may depend on not only N but also N_x; the term with N_x governs generalization *across contexts*, while the term with N_a governs concentration of *within-context* variations. We focus on finite-sample guarantees with sharp dependence on these two sample sizes N_x and N_a, as well as on the complexity of $\mathcal{F}$ ($\log|\mathcal{F}|$, d), misspecification-related terms ($\varepsilon_{\mathrm{aprx}}, \varepsilon_{\mathrm{z}}, \varepsilon_{\mathrm{x}}, \varepsilon_{\mathrm{a}}$), the reward scale $R_{\max}$, and feature geometry-related terms ($\alpha_0, \alpha_{\mathrm{C}}, d_{\mathrm{C}}$).

Our contributions are as follows, along with the rest of the paper's organization:

Grouped finite-sample analysis (Section 2). Building on localized squared-loss analyses (Bartlett et al., 2005; Liang et al., 2015), we develop a common blockwise analysis recipe for VR and VDR that distinguishes terms governed by $1/N_x$ from those governed by $1/N$.

Finite classes (Section 3, first row of Table 1). Under weak realizability, VDR eliminates the $\varepsilon_{\mathrm{aprx}}$ appearing in the VR guarantee and improves the $FR_{\max}$ term from the context scale $1/N_x$ to the full-sample scale $1/N$. We complement this by reporting an instance where VR fails yet VDR succeeds in achieving $\widetilde{\mathcal{O}}(1/\sqrt{N})$ offline regret.

Linear classes (Section 4, second row of Table 1). Next, we consider the linear class $\mathcal{F}_{\mathrm{lin}} := \{(x,a) \mapsto \langle \boldsymbol{\phi}(x,a), \boldsymbol{\theta}\rangle : \boldsymbol{\theta} \in \mathbb{R}^d\}$ with a known feature map $\boldsymbol{\phi} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$. We show that *neither method uniformly dominates*: their relative performance depends on the *feature geometry*. While VDR eliminates the misspecification terms $\varepsilon_{\mathrm{aprx}}, \varepsilon_{\mathrm{z}}, \varepsilon_{\mathrm{x}}, \varepsilon_{\mathrm{a}}$ present in the VR bound, their N^{-2} terms scale with different geometric quantities: α_0^2 for VR, determined by the raw features, and α_{C}^2 for VDR, determined by feature differences. This yields a geometric bias–variance tradeoff: VDR removes nuisance-induced misspecification error (*bias*), but its *variance* depends on the geometry of feature differences, which can be less favorable than the raw feature geometry governing VR. Numerical experiments on toy and real-world LLM tasks illustrate this tradeoff.

2 A Localized Analysis for Grouped Regression

Let $\boldsymbol{r} = (r_{i,j})_{i\in[N_x], j\in[N_a]}$ denote the vector of realized (noisy) rewards. We consider the following generic empirical and population squared losses:

$$\widehat{\mathcal{L}}(f) := \frac{1}{N_x}\sum_{i=1}^{N_x} \left\|\mathcal{T}_i(f - \boldsymbol{r})\right\|_2^2, \qquad \mathcal{L}(f) := \mathbb{E}\left[\left\|\mathcal{T}_i(f - r^\star)\right\|_2^2\right],$$

where $\mathcal{T} : \mathbb{R}^{N_a} \to \mathbb{R}^m$ is an *algorithm-dependent linear transformation*, $\mathcal{T}_i f := \mathcal{T}\left((f(x_i, a_{i,1}), \ldots, f(x_i, a_{i,N_a}))\right)$, and $\mathcal{T}_i \boldsymbol{r} := \mathcal{T}\left((r_{i,1}, \ldots, r_{i,N_a})\right)$. VR and VDR correspond to $\mathcal{T}^{\mathrm{VR}} : \boldsymbol{u} \mapsto \frac{1}{\sqrt{N_a}}\boldsymbol{u}$ and $\mathcal{T}^{\mathrm{VDR}} : \boldsymbol{u} \mapsto \binom{N_a}{2}^{-1/2}(u_j - u_k)_{1\le j<k\le N_a}$, respectively, for $\boldsymbol{u} \in \mathbb{R}^{N_a}$.

We now define the following notion of empirical complexity measure:

Definition 2.1 (Localized block-offset modulus)**.** *For a comparator $f^\circ \in \mathcal{F}$ and $\lambda, r \ge 0$, define the **localized block-offset modulus** $\Psi(r;\lambda,f^\circ)$ as follows:*

$$\Psi(r;\lambda,f^\circ) := \sup_{\substack{h\in\mathcal{F}-f^\circ \\ \mathcal{L}(f^\circ+h)-\mathcal{L}(f^\circ)\le r}} \left[\left(\mathcal{L}(f^\circ+h) - \mathcal{L}(f^\circ)\right) - \left(\widehat{\mathcal{L}}(f^\circ+h) - \widehat{\mathcal{L}}(f^\circ)\right) - \lambda\widehat{Q}(h)\right],$$

where $\mathcal{F} - f^\circ := \{f - f^\circ : f \in \mathcal{F}\}$ and $\widehat{Q}(h) := \frac{1}{N_x}\sum_{i=1}^{N_x}\|\mathcal{T}_i h\|_2^2$.

We now present a fixed-point inequality for the excess loss, whose proof is deferred to Appendix C.2:

Proposition 2.1 (Localized fixed-point inequality)**.** *Fix any comparator $f^\circ \in \mathcal{F}$ and any estimator $\hat{f}_N \in \mathcal{F}$. Suppose that the following two conditions hold: denoting $h := \hat{f}_N - f^\circ$,*

(i) There is a constant $\lambda \ge 0$ such that our estimator $\hat{f}_N$ satisfies

$$\widehat{\mathcal{L}}(\hat{f}_N) - \widehat{\mathcal{L}}(f^\circ) \le -\lambda\widehat{Q}(\hat{f}_N - f^\circ);$$

(ii) There is a deterministic $\psi : \mathbb{R} \to \mathbb{R}$ such that $\Psi(r;\lambda,f^\circ) \le \psi(r)$ for all $r \ge 0$.

Then, we have the following fixed-point inequality for the population excess loss:

$$\mathcal{L}(\hat{f}_N) - \mathcal{L}(f^\circ) \le \sup\{r \ge 0 : r \le \psi(r)\}.$$

Although seemingly abstract, Proposition 2.1 provides a common recipe for deriving tight statistical rates for VR and VDR over generic $\mathcal{F}$: verify conditions (i) and (ii). In our applications, condition (i) follows deterministically from the estimator's loss geometry, whereas condition (ii) is established (with high probability) using problem-specific concentrations.

Relation to prior work. The fixed-point perspective in Proposition 2.1 follows the classical localization of empirical and Rademacher processes (Bartlett et al., 2005; Bousquet, 2002; Bousquet et al., 2002; Koltchinskii, 2011; Koltchinskii & Panchenko, 2000), all of which are based on Talagrand's celebrated concentration inequalities (Talagrand, 1995; 1996a;b). For squared loss, related formulations include offset Rademacher process (Liang et al., 2015; Rakhlin & Sridharan, 2014), quadratic–multiplier and small-ball methods (Mendelson, 2014; 2018), and empirical-entropy characterizations of sharp rates (Rakhlin et al., 2017). In our modulus, $-\lambda\widehat{Q}(h)$ corresponds to the quadratic term of the offset Rademacher process (Liang et al., 2015; Rakhlin & Sridharan, 2014).

Our main technical distinction is how we verify condition (ii) under grouped sampling. Working directly with the *unsymmetrized* block-offset modulus makes the two-level concentration structure explicit. For finite classes, we use scalar blockwise concentration and union bounds; for linear classes, we combine a lower-isometry event with direct vector and matrix concentration. This preserves the two sampling scales, separating contributions by N_x^{-1} and N^{-1} with relevant dependencies.

3 Finite Class: Nuisance Cancellation and Gains from $N_a \geq 2$

For finite classes, we consider $F := \sup_{f\in\mathcal{F}} \|f\|_\infty$, which may satisfy $F \ll R_{\max}$. Since $f^\star \in \mathcal{F}$, F controls the within-context action gaps, whereas $R_{\max}$ and $\varepsilon_{\text{aprx}}$ may also reflect the nuisance $b^\star$. With this, we present the error bounds for finite classes: assuming $N_a \geq 2$ for VDR,

Theorem 3.1 (VR & VDR Bound for Finite $\mathcal{F}$)**.** *Let $1 \leq |\mathcal{F}| < \infty$, $\delta \in (0,1)$, and suppose that Assumption 1 holds. Then, each of the following holds with probability at least $1-\delta$:*

$$\mathbb{E}_{x,a}\left[\left(\mathcal{C}\hat{f}_N^{\text{VR}} - \mathcal{C}r^\star\right)^2\right] \leq 2\varepsilon_{\text{aprx}} + \left(\frac{22F^2 + 4FR_{\max}}{N_x} + \frac{18\sigma^2}{N}\right)\log\frac{|\mathcal{F}|}{\delta}, \tag{2}$$

$$\mathbb{E}_{x,a}\left[\left(\mathcal{C}\hat{f}_N^{\text{VDR}} - \mathcal{C}r^\star\right)^2\right] \leq \frac{8}{3}\left(\frac{7F^2}{N_x} + \frac{12\sigma^2 + 8FR_{\max}}{N}\right)\log\frac{2|\mathcal{F}|}{\delta}. \tag{3}$$

Proof. Here, we showcase the localized analysis recipe (Section 2). As the analysis for VDR proceeds similarly, we only show the proof of VR here; VDR's proof is presented in Appendix D.2.

We set $f^\circ_{\text{VR}} \in \arg\min_{f\in\mathcal{F}} \mathcal{L}^{\text{VR}}(f)$, which satisfies $\mathcal{L}^{\text{VR}}(f^\circ_{\text{VR}}) = \varepsilon_{\text{aprx}}$, i.e.,

$$\mathbb{E}_{x,a}\left[(\mathcal{C}\hat{f}_N^{\text{VR}} - \mathcal{C}r^\star)^2\right] \leq \mathbb{E}_{x,a}\left[(\hat{f}_N^{\text{VR}} - r^\star)^2\right] = \mathcal{L}^{\text{VR}}(\hat{f}_N^{\text{VR}}) = \varepsilon_{\text{aprx}} + \mathcal{L}^{\text{VR}}(\hat{f}_N^{\text{VR}}) - \mathcal{L}^{\text{VR}}(f^\circ_{\text{VR}}), \tag{4}$$

where the first inequality follows because the variance is bounded by the second moment: $\mathbb{E}[(\mathcal{C}g(x,a))^2] = \mathbb{E}_x[\text{Var}[g(x,a) \mid x]] \leq \mathbb{E}[g(x,a)^2]$. As $\hat{f}_N^{\text{VR}}$ is an empirical risk minimizer, condition (i) of Proposition 2.1 holds with $\lambda = 0$. As for (ii), the following lemma, proved in Appendix D.1.1, provides such $\psi(\cdot)$ that upper bounds the localized modulus with high probability:

Lemma 3.1. *Denote $t_\delta := \log\frac{|\mathcal{F}|}{\delta}$ and $v_\delta := \left(\frac{F^2}{N_x} + \frac{\sigma^2}{N}\right)t_\delta$. Then, with probability at least $1-\delta$:*

$$\Psi^{\text{VR}}(r; 0, f^\circ_{\text{VR}}) \leq \psi(r) := 4\sqrt{v_\delta\left(r + 2\varepsilon_{\text{aprx}}\right)} + \frac{8F(F+R_{\max})t_\delta}{3N_x}, \quad \forall r \geq 0.$$

We remark that the above lemma is obtained via *one-sided* Bernstein's inequality (Bernstein, 1927) over the blocks $i \in [N_x]$, attaining per-block variance of $F^2 + \sigma^2 N_a^{-1}$ instead of the naïve $F^2 + \sigma^2$.

The above lemma and Proposition 2.1 imply $\mathcal{L}^{\text{VR}}(\hat{f}_N^{\text{VR}}) - \mathcal{L}^{\text{VR}}(f^\circ_{\text{VR}}) \leq \bar{r}$, where

$$\bar{r} := \sup\left\{r \geq 0 : r \leq 4\sqrt{v_\delta(r + 2\varepsilon_{\text{aprx}})} + \frac{8F(F+R_{\max})t_\delta}{3N_x}\right\}.$$

Using Young's inequality to separate $4\sqrt{v_\delta(r+2\varepsilon_{\text{aprx}})} \leq \frac{1}{3}(r + 2\varepsilon_{\text{aprx}}) + 12v_\delta$ and rearranging, we obtain $\bar{r} \leq \varepsilon_{\text{aprx}} + 18v_\delta + \frac{4F(F+R_{\max})t_\delta}{N_x}$. We conclude by combining this with Eqn. (4). □

Comparison and discussion. The terms scaling with N_x^{-1} reflect variation across contexts; additional action samples reduce within-context fluctuations, but they do not provide information about unseen contexts without further structure on $\mathcal{F}$. VR retains the approximation error $\varepsilon_{\text{aprx}}$, even when this is caused entirely by the action-independent $b^\star$. In contrast, VDR removes such dependency and moves the $FR_{\max}$ factor from N_x^{-1} to N^{-1}, which can be substantial when $R_{\max} \gg F$.

Also, note that Eqn. (2) is a *non-exact* oracle inequality (Lecué & Mendelson, 2012; Massart & Nédélec, 2006), as the coefficient of $\varepsilon_{\text{aprx}}$ is two, not one. Indeed, it is known that ERM over an arbitrary finite $\mathcal{F}$ cannot admit an exact oracle inequality with the optimal rate of

$\mathcal{O}(\log(|\mathcal{F}|/\delta)/N)$ (Lecué & Rigollet, 2014), while alternate aggregation-type estimators (Audibert, 2007; 2009; Lecué & Rigollet, 2014; Liang et al., 2015) can. In Appendix E, we prove an oracle inequality for the STAR ESTIMATOR (Audibert, 2007; 2009), adapted to our grouped dataset setting.

Failure of VR. We complement this with a separation result, whose proof is in Appendix G.1:

Theorem 3.2 (Failure of VR). *There exists a finite-class instance satisfying Assumption 1 such that, for every fixed $N_a \geq 1$, $\delta \in (0,1)$ and sufficiently large N_x, $\mathrm{Reg}(\hat{\pi}_N^{\text{VR-GREEDY}}) = \Omega(R_{\max})$, with probability at least $1-\delta$. Moreover, on the same instance with $N_a \geq 2$, $\mathrm{Reg}(\hat{\pi}_N^{\text{VDR-GREEDY}}) = 0$, with probability tending to one exponentially fast in N_x.*

4 LINEAR CLASSES: A GEOMETRIC BIAS–VARIANCE TRADEOFF

Unlike the finite-class setting, we show a geometric bias–variance tradeoff between VR and VDR. For clarity of exposition, we omit universal constants and $\log\frac{1}{\delta}$ dependencies. Throughout, we use the random variables $X \sim \rho$ and $A, A', A_1, \ldots, A_{N_a} \overset{i.i.d.}{\sim} \pi_{\mathrm{ref}}(\cdot \mid X)$. We denote $\mathrm{Cov}(\boldsymbol{G}) := \mathbb{E}[(\boldsymbol{G} - \mathbb{E}\boldsymbol{G})^2]$ for a random self-adjoint matrix $\boldsymbol{G}$. All proofs are presented in Appendix F.

4.1 VR FOR LINEAR CLASS

Following the bounded statistical-leverage and approximation-error conditions of Hsu et al. (2014, Conditions 1 and 3), we similarly impose the following assumptions:

Assumption 2. $\boldsymbol{\Sigma} := \mathbb{E}[\boldsymbol{\phi}(X,A)\boldsymbol{\phi}(X,A)^\top] \succ \mathbf{0}$. *Also, there exists $\alpha_0 \geq 1$ such that, for the **whitened feature** $\boldsymbol{z}(x,a) := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\phi}(x,a)$,*

$$\operatorname{ess\,sup} \left\|\boldsymbol{z}(X,A)\right\|_2 \leq \alpha_0\sqrt{d}, \qquad 0 \leq \varepsilon_{\mathrm{z}} := d^{-1} \operatorname{ess\,sup} \left\|e^\circ(X,A)\boldsymbol{z}(X,A)\right\|_2^2 < \infty,$$

where $e^\circ(x,a) := f^\circ(x,a) - r^\star(x,a)$ with $f^\circ \in \arg\min_{f\in\mathcal{F}_{\mathrm{lin}}} \mathbb{E}_{x,a}[(f(x,a) - r^\star(x,a))^2]$.

Remark 1 (Relation to Design Objective). *The condition $\operatorname{ess\,sup} \left\|\boldsymbol{z}(X,A)\right\|_2 \leq \alpha_0\sqrt{d}$ is closely related to the G-optimal design (Pukelsheim, 2006): $\left\|\boldsymbol{z}(x,a)\right\|_2^2 = \left\|\boldsymbol{\phi}(x,a)\right\|_{\boldsymbol{\Sigma}^{-1}}^2 \leq \alpha_0^2 d$. Although ρ and π_{ref} are fixed in our setting, in the active learning scenario, by jointly optimizing for $(\rho, \pi_{\mathrm{ref}})$, one could achieve $\alpha_0 = 1$ by the Kiefer–Wolfowitz theorem (Kiefer & Wolfowitz, 1960).*

We define the random covariance matrix $\boldsymbol{G}_{N_a} := \frac{1}{N_a}\sum_{j=1}^{N_a} \boldsymbol{z}(X,A_j)\boldsymbol{z}(X,A_j)^\top$, and non-random, across- and within-context residual quantities

$$\varepsilon_{\mathrm{x}} := \frac{1}{d}\mathbb{E}_X\left[\left\|\mathbb{E}_A[e^\circ \boldsymbol{z} \mid X]\right\|_2^2\right], \quad \varepsilon_{\mathrm{a}} := \frac{1}{d}\mathbb{E}_X\left[\mathbb{E}_A\left[\left\|e^\circ\boldsymbol{z} - \mathbb{E}_A[e^\circ\boldsymbol{z} \mid X]\right\|_2^2 \mid X\right]\right], \tag{5}$$

The upright subscripts x and a label the between-context and within-context components, respectively; both quantities are population scalars.

With this, we finally present the VR bound for the linear class.

Theorem 4.1 (VR Bound for $\mathcal{F}_{\mathrm{lin}}$). *Let $\delta \in (0,1)$, and suppose that Assumptions 1 and 2 hold and $N_x \geq \max\left\{18\left\|\mathrm{Cov}(\boldsymbol{G}_{N_a})\right\|_{\mathrm{op}}, 2\right\}\log\frac{3d}{\delta}$. Then, with probability at least $1-\delta$,*[a]

$$\mathbb{E}_{x,a}\left[(\mathcal{C}\hat{f}_N^{\mathrm{VR}} - \mathcal{C}r^\star)^2\right] \leq \varepsilon_{\mathrm{aprx}} + \widetilde{\mathcal{O}}\left(\frac{\varepsilon_{\mathrm{x}} d}{N_x} + \frac{\varepsilon_{\mathrm{z}} d}{N_x^2} + \frac{(\varepsilon_{\mathrm{a}} + \sigma^2)d}{N} + \frac{R_{\max}^2\alpha_0^2 d}{N^2}\right).$$

[a] The $\frac{\sigma^2 d}{N}$ term can be sharpened to $\frac{\sigma^2 d_{\mathrm{eff}}}{N}$ for a tighter geometric quantity $d_{\mathrm{eff}} \leq d$ by retaining the centered-feature geometry; see Appendix F.1.3.

Proof Sketch. Because of the non-i.i.d. nature of our problem setup, we *cannot* directly invoke the existing result for ordinary least squares (OLS) from Hsu et al. (2014, Theorem 1). Instead, we follow the same recipe as laid out in Section 2, carefully combining it with *vector/matrix* Bernstein

inequalities (Hsu et al., 2012a; Tropp, 2015) over the blocks $i \in [N_x]$ and a variance computation that separates the feature-diversity and misspecification terms. □

Requirement on N_x. Let $\boldsymbol{G}(X) := \mathbb{E}_A[\boldsymbol{z}\boldsymbol{z}^\top \mid X]$. By the law of total covariance, the block Gram variance decomposes as

$$\mathrm{Cov}(\boldsymbol{G}_{N_a}) = \mathrm{Cov}_X(\boldsymbol{G}(X)) + \frac{1}{N_a}\mathbb{E}_X\left[\mathrm{Cov}_A(\boldsymbol{z}\boldsymbol{z}^\top \mid X)\right].$$

The first term captures variation in the conditional feature covariance across contexts and is unaffected by repeated actions, whereas the second captures within-context variation and decays with N_a. Thus, when $\boldsymbol{G}(X)$ is nearly constant, repeated actions improve concentration of the empirical Gram matrix much as additional independent contexts would; otherwise, sufficiently many distinct contexts remain necessary for lower isometry. Because these quantities are defined after whitening, they are invariant to invertible linear reparameterizations of the features and can be large when feature directions are represented unevenly across contexts. In the special case $\boldsymbol{G}(X) = \boldsymbol{I}_d$ almost surely, $\left\|\mathrm{Cov}(\boldsymbol{G}_{N_a})\right\|_{\mathrm{op}} \le (\alpha_0^2 d - 1)/N_a$, so the sample-size requirement in Theorem 4.1 is satisfied if $N_x \gtrsim \log\frac{d}{\delta}$ and $N \gtrsim \alpha_0^2 d \log\frac{d}{\delta}$. Importantly, the block Gram variance enters only the sample-size condition ensuring lower isometry, not the error bound itself.

Interpreting the error bound. The conditional mean $\mathbb{E}_A[e^\circ \boldsymbol{z} \mid X]$ is a context-specific misspecification score and need not vanish: the population normal equation guarantees only that $\mathbb{E}_X[\mathbb{E}_A[e^\circ \boldsymbol{z} \mid X]] = \boldsymbol{0}$. Consequently, the across-context term $\frac{\varepsilon_{\mathrm{x}} d}{N_x}$ can persist as $N_a \to \infty$, whereas the within-context term $\frac{\varepsilon_{\mathrm{a}} d}{N}$ is averaged over all observations and vanishes in the same limit. We also note that $\varepsilon_{\mathrm{x}} = 0$ does *not* imply $\varepsilon_{\mathrm{a}} = 0$ and vice-versa.

The leverage bound yields $\varepsilon_{\mathrm{x}} + \varepsilon_{\mathrm{a}}/N_a \le \alpha_0^2 \varepsilon_{\mathrm{aprx}}$, and hence the following *crude* upper bound:

$$\mathbb{E}_{x,a}\left[(\mathcal{C}\hat{f}_N^{\mathrm{VR}} - \mathcal{C}r^\star)^2\right] \le \left(1 + \widetilde{\mathcal{O}}\left(\frac{\alpha_0^2 d}{N_x}\right)\right)\varepsilon_{\mathrm{aprx}} + \widetilde{\mathcal{O}}\left(\frac{\varepsilon_{\mathrm{z}} d}{N_x^2}\right) + \widetilde{\mathcal{O}}\left(\frac{\sigma^2 d}{N} + \frac{R_{\max}^2 \alpha_0^2 d}{N^2}\right).$$

Thus, for fixed N_x, the oracle inequality loses exactness and retains a misspecification-related bottleneck of order N_x^{-2}. The almost-sure quantity ε_{z} cannot be controlled by the L_2 error $\varepsilon_{\mathrm{aprx}}$ alone, but satisfies $\varepsilon_{\mathrm{z}} \le \alpha_0^2 \left\|e^\circ\right\|_\infty^2$.

Finally, in the *well-specified* setting, $\varepsilon_{\mathrm{aprx}} = \varepsilon_{\mathrm{z}} = \varepsilon_{\mathrm{x}} = \varepsilon_{\mathrm{a}} = 0$ in Theorem 4.1, and the error bound becomes $\widetilde{\mathcal{O}}\left(\frac{\sigma^2 d}{N} + \frac{R_{\max}^2 \alpha_0^2 d}{N^2}\right)$. Thus, in the well-specified case, no leading estimation term scales only as N_x^{-1}, and $R_{\max}$ and α_0 appear only in the lower-order Bernstein correction.

Relation to random-design least squares. When $N_a = 1$, the residual and noise terms recover the first-order dependence of classical random-design OLS, including the tight well-specified rate $\sigma^2 d/N$ under homoscedastic noise (Hsu et al., 2014, Theorem 1 and Remark 10). Our contribution is the grouped case $N_a \ge 2$, where the decompositions above identify which fluctuations scale with N_x and which benefit from all $N = N_x N_a$ observations. We discuss connections to misspecified linear regression (Amortila et al., 2024; Maran & Szepesvári, 2025) more broadly in Appendix A(**3**).

4.2 VDR for Linear Class

Define the *centered feature* $\overline{\boldsymbol{\phi}}(x,a) := \boldsymbol{\phi}(x,a) - \mathbb{E}_{a\sim\pi_{\mathrm{ref}}(\cdot|x)}[\boldsymbol{\phi}(x,a)]$, and its covariance as $\boldsymbol{\Sigma}_{\mathrm{C}} := \mathbb{E}_{x,a}\left[\overline{\boldsymbol{\phi}}(x,a)\overline{\boldsymbol{\phi}}(x,a)^\top\right]$. Let $d_{\mathrm{C}} := \mathrm{rank}(\boldsymbol{\Sigma}_{\mathrm{C}})$. Whenever $d_{\mathrm{C}} \ge 1$, let $\boldsymbol{U}_{\mathrm{C}} \in \mathbb{R}^{d\times d_{\mathrm{C}}}$ satisfy $\boldsymbol{U}_{\mathrm{C}}^\top \boldsymbol{U}_{\mathrm{C}} = \boldsymbol{I}_{d_{\mathrm{C}}}$ and $\mathrm{col}(\boldsymbol{U}_{\mathrm{C}}) = \mathrm{range}(\boldsymbol{\Sigma}_{\mathrm{C}})$, and define

$$\overline{\boldsymbol{\Sigma}}_{\mathrm{C}} := \boldsymbol{U}_{\mathrm{C}}^\top \boldsymbol{\Sigma}_{\mathrm{C}} \boldsymbol{U}_{\mathrm{C}} \succ \boldsymbol{0}, \qquad \boldsymbol{z}_{\mathrm{C}}(x,a) := \overline{\boldsymbol{\Sigma}}_{\mathrm{C}}^{-1/2} \boldsymbol{U}_{\mathrm{C}}^\top \overline{\boldsymbol{\phi}}(x,a) \in \mathbb{R}^{d_{\mathrm{C}}}.$$

Then, $\mathbb{E}[\boldsymbol{z}_{\mathrm{C}}\boldsymbol{z}_{\mathrm{C}}^\top] = \boldsymbol{I}_{d_{\mathrm{C}}}$ and $\mathbb{E}_A[\boldsymbol{z}_{\mathrm{C}}(X,A) \mid X] = \boldsymbol{0}$.

Unlike VR, VDR does not identify directions of the parameter $\boldsymbol{\theta}$ that add the same value to every action at a given context, since its loss depends only on within-context differences. Thus, we state the design assumption on its identifiable subspace as follows:

Assumption 3. *$d_{\mathrm{C}} \ge 1$, and there is a $\alpha_{\mathrm{C}} \ge 1$ s.t.* $\operatorname{ess\,sup} \left\|\boldsymbol{z}_{\mathrm{C}}(X,A) - \boldsymbol{z}_{\mathrm{C}}(X,A')\right\|_2 \le \alpha_{\mathrm{C}}\sqrt{2d_{\mathrm{C}}}$.

For $N_a \geq 2$, denoting $\overline{\boldsymbol{z}}_{\mathrm{C},N_a}(X) \coloneqq \frac{1}{N_a}\sum_{j=1}^{N_a} \boldsymbol{z}_{\mathrm{C}}(X, A_j)$, define the random, *sample* covariance matrix $\boldsymbol{G}^{\mathrm{C}}_{N_a} \coloneqq \frac{1}{N_a-1}\sum_{j=1}^{N_a} \left(\boldsymbol{z}_{\mathrm{C}}(X, A_j) - \overline{\boldsymbol{z}}_{\mathrm{C},N_a}(X)\right)\left(\boldsymbol{z}_{\mathrm{C}}(X, A_j) - \overline{\boldsymbol{z}}_{\mathrm{C},N_a}(X)\right)^\top$.

Theorem 4.2 (VDR Bound for $\mathcal{F}_{\mathrm{lin}}$)**.** *Let $N_a \geq 2$, $\delta \in (0,1)$, and suppose that Assumptions 1 and 3 hold, and $N_x \geq \max\left\{18\left\|\mathrm{Cov}(\boldsymbol{G}^{\mathrm{C}}_{N_a})\right\|_{\mathrm{op}}, 2\right\}\log\frac{2d_{\mathrm{C}}}{\delta}$. Then, w.p. at least $1-\delta$,*

$$\mathbb{E}_{x,a}\left[\left(\mathcal{C}\hat{f}^{\mathrm{VDR}}_N - \mathcal{C}r^\star\right)^2\right] \leq \widetilde{\mathcal{O}}\left(\frac{\sigma^2 d_{\mathrm{C}}}{N} + \frac{R^2_{\max}\alpha^2_{\mathrm{C}} d_{\mathrm{C}}}{N^2}\right).$$

Relation to orthogonalized regression. VDR shares the nuisance-cancellation principle of orthogonalized estimation in semiparametric contextual bandits and regression (Chernozhukov et al., 2018; Choi et al., 2023; Foster & Syrgkanis, 2023; Kim & Paik, 2019; Kim et al., 2025; Krishnamurthy et al., 2018; Robinson, 1988). These works typically observe a single action–reward outcome per observed context and construct an orthogonal score via centering or residualization. Our setting instead observes $N_a \geq 2$ action–reward samples per observed context and forms within-context pairs; our analysis separates fluctuations across contexts from those that average over the repeated actions.

Interpreting the centered-feature geometry. d_{C} is the *dimension of the feature subspace generated by within-context action variation*; intercepts and other action-invariant directions lie in $\ker(\boldsymbol{\Sigma}_{\mathrm{C}})$. For every $\boldsymbol{v} \in \ker(\boldsymbol{\Sigma}_{\mathrm{C}})$, $\mathbb{E}[(\boldsymbol{v}^\top\mathcal{C}\boldsymbol{\phi})^2] = 0$, so $\boldsymbol{v}^\top\mathcal{C}\boldsymbol{\phi} = 0$ almost surely. Thus, VDR identifies $\boldsymbol{\theta}^\star$ only modulo this kernel, which suffices to identify $\mathcal{C}f^\star$ and all action gaps on the support of π_{ref}. In general, $d_{\mathrm{C}} = d - \dim\ker(\boldsymbol{\Sigma}_{\mathrm{C}})$: it equals d without action-invariant directions, $d-1$ with only an intercept, and less with additional such directions. We assume $d_{\mathrm{C}} \geq 1$ (otherwise all modeled action gaps vanish on the behavior-policy support). Finally, α_{C} controls the maximal leverage of *differences* of whitened centered features on the identifiable subspace. When both design conditions hold, α_{C} can be smaller or larger than α_0; neither condition dominates, as we show next.

4.3 Comparing VR and VDR: Feature Geometry Perspective

Comparing the bounds. Under Assumption 1, the action-independent nuisance $b^\star(x)$ cancels from every within-context reward difference. Consequently, the VDR bound (Theorem 4.2) contains no misspecification-related terms, whereas the VR bound (Theorem 4.1) retains $\varepsilon_{\mathrm{aprx}}$ and the residual terms ε_{x}, ε_{a}, and ε_{z}, i.e., VDR's bound outperforms that of VR under large misspecification.

In the well-specified case ($\varepsilon_{\mathrm{aprx}} = \varepsilon_{\mathrm{x}} = \varepsilon_{\mathrm{a}} = \varepsilon_{\mathrm{z}} = 0$), the comparison is instead governed by the feature geometry. The displayed bounds reduce to $\widetilde{\mathcal{O}}(\sigma^2 d/N + R^2_{\max}\alpha_0^2 d/N^2)$ for VR and $\widetilde{\mathcal{O}}(\sigma^2 d_{\mathrm{C}}/N + R^2_{\max}\alpha^2_{\mathrm{C}} d_{\mathrm{C}}/N^2)$ for VDR. Since $d_{\mathrm{C}} \leq d$, the leading variance term in the displayed VDR bound is no larger; with the sharper VR bound of Appendix F.1.3, however, it becomes $\sigma^2 d_{\mathrm{eff}}/N$ with $d_{\mathrm{eff}} \leq d_{\mathrm{C}}$. The lower-order terms have no analogous uniform ordering across feature geometries, and hence upper bound-wise, neither estimator dominates one another.

To concretely show this, we provide two simple examples where α_{C} can be arbitrarily larger than α_0, and vice versa. For $p \in (0,1)$, let $X = 1$, $A \sim \mathrm{Bernoulli}(p)$, and set $\boldsymbol{\phi}(A) = 1 + A \in \mathbb{R}$. Then, $\boldsymbol{\Sigma} = 1+3p$, and $\boldsymbol{\Sigma}_{\mathrm{C}} = p(1-p)$, which gives $\alpha_0 = 2/\sqrt{1+3p} = \Theta(1)$ and $\alpha_{\mathrm{C}} = 1/\sqrt{2p(1-p)} = \Theta(p^{-1/2})$. Thus $\alpha_{\mathrm{C}}/\alpha_0 \to \infty$ as $p \to 0^+$. Second, let $X \sim \mathrm{Bernoulli}(p)$, $A \sim \mathrm{Unif}(\{-1,+1\})$, and set $\boldsymbol{\phi}(X,A) = (X,A) \in \mathbb{R}^2$. Here $\boldsymbol{\Sigma} = \mathrm{diag}(p,1)$, so $\alpha_0 = \sqrt{(1+p)/(2p)} = \Theta(p^{-1/2})$. Centering gives $\overline{\boldsymbol{\phi}} = (0,A)$, so $\boldsymbol{z}_{\mathrm{C}} = A$ and $\alpha_{\mathrm{C}} = \sqrt{2}$. Therefore $\alpha_0/\alpha_{\mathrm{C}} \to \infty$ as $p \to 0^+$.

Synthetic experiment. We consider linear contextual bandits with five-dimensional contexts and three-dimensional cube actions, indexed by a sampling probability $\delta \in [0, 1/2]$ (distinct from the confidence level above) and an action-independent nuisance magnitude $\beta \in [0,2]$. The true value depends only on the first action coordinate. At each context, the behavior policy favors one sign of this coordinate and samples the opposite with probability δ. When $\beta = 0$ and δ is small, VR can exploit absolute rewards while VDR has few comparisons between actions with different true values. Increasing β misspecifies VR through an action-independent shift that cancels from VDR.

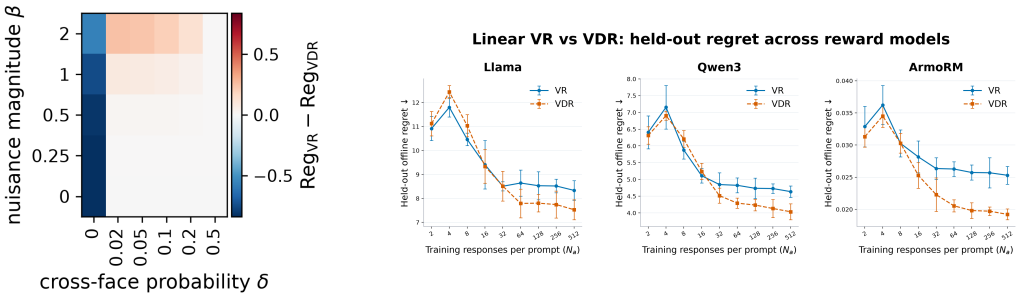

Figure 1: **Left:** synthetic regret difference $\mathrm{Reg}_{\mathrm{VR}} - \mathrm{Reg}_{\mathrm{VDR}}$ as the cross-face probability δ and nuisance magnitude β vary; negative values favor VR and positive values favor VDR. **Right:** real-data regret versus the number of labeled responses per training prompt; lower is better. Error bars are pointwise 95% Student-t intervals over five seeds.

We fit both methods with $(N_x, N_a) = (200, 8)$ and evaluate greedy-policy regret over 8 independent instances; details and ablations are deferred to Appendix H.1. The left panel of Figure 1 shows that when $\beta = 0$, VR has lower regret in every instance for every tested $\delta > 0$. At fixed $\delta = 0.02$, $\mathrm{Reg}_{\mathrm{VR}} - \mathrm{Reg}_{\mathrm{VDR}}$ changes from negative to positive as β grows. These results illustrate the bias–variance tradeoff: VR benefits from absolute rewards under correct specification when informative within-context comparisons are rare, while VDR becomes preferable as nuisance-induced bias grows.

LLM experiment. We compare ℓ_2-regularized linear VR and VDR scorers on frozen LLM representations in a response-selection task. We collect 1,000 prompts from WildChat, UltraFeedback, GSM8K, MATH, MBPP, HelpSteer2, and TL;DR, and cache 1,000 Llama-3.2-3B-Instruct responses per prompt. Each response is represented by its 3,072-dimensional mean-pooled final-layer hidden state. Keeping the prompts, responses, and representations fixed, we repeat the experiment with rewards from Skywork-Reward-V2-Llama-3.1-8B, Skywork-Reward-V2-Qwen3-8B, and ArmoRM-Llama3-8B-v0.1. Across five seeds, both methods use the same 800 training prompts and sampled responses, are evaluated on 200 held-out prompts, and vary the number of labeled responses per training prompt from $N_a = 2$ to 512; full details are deferred to Appendix H.2.

The right panel of Figure 1 shows that for $N_a \geq 64$, VDR has significantly lower mean regret for all three reward models (see paired intervals in Figure 4, Appendix H.2). At $N_a = 512$, the reductions are 9.7%, 13.0%, and 24.0% for Llama, Qwen3, and ArmoRM, respectively. Thus the preferred objective changes with the amount of within-prompt supervision. Although outside our theoretical scope, we also report a controlled MCTS math reasoning experiment in Appendix H.3, where a binary-preference variant of VDR achieves higher observed final-answer accuracy than VR with both linear and MLP heads.

5 Conclusion and Future Work

We establish finite-sample prediction guarantees for Value Regression (VR) and Value Difference Regression (VDR) with grouped offline data (which imply offline-regret bounds under coverage). Our analysis distinguishes two sampling scales: N_x governs generalization across contexts, while N_a controls the averaging of within-context variability. For finite classes, VDR avoids the approximation-error term in the VR bound and improves a reward-scale term through repeated action observations. For linear classes, the comparison depends on feature geometry: differencing removes nuisance-induced bias, while fitting absolute rewards can reduce estimation variance under correct specification. Our experiments illustrate the resulting bias–variance tradeoff, with neither method uniformly preferred.

We conclude with some potential future directions. One is to assess whether the weak realizability assumption holds in LLM applications and other domains, and to identify weaker assumptions under which meaningful guarantees remain possible. Other directions include determining whether analogous tradeoffs arise under the Bradley–Terry model with logistic losses, understanding how pair selection affects VDR, and extending the analysis beyond pairwise comparisons (e.g., ranking).

AI Disclosure Statement

In this work, we used generative AI tools to generate synthetic datasets; help develop theoretical models or conceptual frameworks; formulate mathematical claims and provide critical ingredients for their proofs; assist in writing proofs; propose or refine hypotheses; design or provide feedback on research methodology or experiments; implement methods; assist with translation; support qualitative and thematic data analysis; and interpret results. Dataset cleaning and reformatting were not applicable to this work. We also used generative AI tools to create or modify scientific figures or images; suggest experimental parameters; create or edit software code; draft portions of the paper; summarize or analyze existing literature; discover research topics or identify gaps; brainstorm; source or search for information; improve the paper's readability; identify relevant literature; suggest the paper's structure; and propose titles or keywords. We reviewed all AI-assisted work: all LLM-generated proofs were verified and polished for readability by at least two authors, all cited literature was manually checked, and LLM-generated code was verified and tested for correctness by two authors. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

References

Philip Amortila, Nan Jiang, and Csaba Szepesvári. The Optimal Approximation Factors in Misspecified Off-Policy Value Function Estimation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 768–790. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/amortila23a.html.

Philip Amortila, Tongyi Cao, and Akshay Krishnamurthy. Mitigating Covariate Shift in Misspecified Regression with Applications to Reinforcement Learning. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 130–160. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/amortila24a.html.

Patrice Assouad. Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique*, 1983.

Jean-Yves Audibert. Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems*, volume 20, pp. 41–48. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/hash/ef575e8837d065a1683c022d2077d342-Abstract.html.

Jean-Yves Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591 – 1646, 2009. doi:10.1214/08-AOS623.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program Synthesis with Large Language Models. *arXiv preprint arXiv:2108.07732*, 2021. URL https://arxiv.org/abs/2108.07732.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497 – 1537, 2005. doi:10.1214/009053605000000282.

Sergei N. Bernstein. Sur l'extension du théoréme limite du calcul des probabilités aux sommes de quantités dépendantes. *Mathematische Annalen*, 97(1):1–59, Dec 1927. doi:10.1007/BF01447859.

Olivier Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*. PhD thesis, École Polytechnique, 2002.

Olivier Bousquet, Vladimir Koltchinskii, and Dmitriy Panchenko. Some Local Measures of Complexity of Convex Hulls and Generalization Bounds. In *Computational Learning Theory*, pp. 59–73, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. doi:10.1007/3-540-45435-7_5.

Ralph Allan Bradley and Milton E. Terry. Rank Analysis of Incomplete Block Designs: The Method of Paired Comparisons. *Biometrika*, 39(3-4):324–345, 1952. doi:10.1093/biomet/39.3-4.324.

Yifei Cao, Changhao Jiang, Jiabao Zhuang, Jiajun Sun, Ming Zhang, Zhiheng Xi, Hui Li, Shihan Dou, Yuran Wang, Yunke Zhang, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. From Scores to Preferences: Redefining Evaluation Paradigm for Speech Quality Reward Modeling. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 32731–32749, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi:10.18653/v1/2026.findings-acl.1638.

Aldo Gael Carranza, Sanath Kumar Krishnamurthy, and Susan Athey. Flexible and Efficient Contextual Bandits with Heterogeneous Treatment Effect Oracles. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7190–7212. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/carranza23a.html.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. Step-level Value Preference Optimization for Mathematical Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7889–7903, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.findings-emnlp.463.

Jinglin Chen and Nan Jiang. Information-Theoretic Considerations in Batch Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1042–1051. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/chen19e.html.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 02 2018. doi:10.1111/ectj.12097.

Young-Geun Choi, Gi-Soo Kim, Seunghoon Paik, and Myunghee Cho Paik. Semi-parametric contextual bandits with graph-Laplacian regularization. *Information Sciences*, 645:119367, 2023. doi:10.1016/j.ins.2023.119367.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4302–4310. Curran Associates, Inc., 2017. URL https://arxiv.org/abs/1706.03741.

Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and Empirical Minimization of U-statistics. *The Annals of Statistics*, 36(2):844 – 874, 2008. doi:10.1214/009052607000000910.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Corinna Cortes, Mehryar Mohri, and Ashish Rastogi. Magnitude-Preserving Ranking Algorithms. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pp. 169–176, New York, NY, USA, 2007. Association for Computing Machinery. doi:10.1145/1273496.1273518.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. ULTRAFEEDBACK: Boosting Language Models with Scaled AI Feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9722–9744. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/cui24f.html.

Simon S. Du, Sham M. Kakade, Ruosong Wang, and Lin F. Yang. Is a Good Representation Sufficient for Sample Efficient Reinforcement Learning? In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1genAVKPB.

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error Propagation for Approximate Policy and Value Iteration. In *Advances in Neural Information Processing Systems*, volume 23, pp. 568 – 576. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/hash/65cc2c8205a05d7379fa3a6386f710e1-Abstract.html.

Dylan J. Foster and Vasilis Syrgkanis. Orthogonal Statistical Learning. *The Annals of Statistics*, 51 (3):879 – 908, 2023. doi:10.1214/23-AOS2258.

Ragnar Frisch and Frederick V. Waugh. Partial Time Regressions as Compared with Individual Trends. *Econometrica*, 1(4):387–401, 1933. doi:10.2307/1907330.

Zhaolin Gao, Jonathan D Chang, Wenhao Zhan, Owen Oertell, Gokul Swamy, Kianté Brantley, Thorsten Joachims, J Andrew Bagnell, Jason D Lee, and Wen Sun. REBEL: Reinforcement Learning via Regressing Relative Rewards. *Advances in Neural Information Processing Systems*, 37:52354–52400, 2024. URL https://openreview.net/forum?id=yxjWAJzUyV.

Kristjan Greenewald, Ambuj Tewari, Susan Murphy, and Predag Klasnja. Action Centered Contextual Bandits. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5979–5987. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/4fa177df22864518b2d7818d4db5db2d-Paper.pdf.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20640–20661. PMLR, 13–19 Jul 2025. URL https://openreview.net/forum?id=5zwF1GizFa.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(none):1 – 6, 2012a. doi:10.1214/ECP.v17-2079.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random Design Analysis of Ridge Regression. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pp. 9.1–9.24. PMLR, 2012b. URL https://proceedings.mlr.press/v23/hsu12.html.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random Design Analysis of Ridge Regression. *Foundations of Computational Mathematics*, 14(3):569–600, Jun 2014. doi:10.1007/s10208-014-9192-1.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J Foster. Is Best-of-N the Best of Them? Coverage, Scaling, and Optimality in Inference-Time Alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25075–25126. PMLR, 13–19 Jul 2025. URL https://openreview.net/forum?id=QnjfkhrbYK.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, Yunlong Feng, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. Qwen2.5-Coder Technical Report. *arXiv preprint arXiv:2409.12186*, 2024. URL https://arxiv.org/abs/2409.12186.

J. Kiefer and J. Wolfowitz. The Equivalence of Two Extremum Problems. *Canadian Journal of Mathematics*, 12:363–366, 1960. doi:10.4153/CJM-1960-030-4.

Gi-Soo Kim and Myunghee Cho Paik. Contextual Multi-armed Bandit Algorithm for Semiparametric Reward Model. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3389–3397. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kim19d.html.

Seok-Jin Kim, Gi-Soo Kim, and Min-hwan Oh. Experimental Design for Semiparametric Bandits. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 3215–3252. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/kim25a.html.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École d'Été de Probabilités de Saint-Flour XXXVIII-2008*. Lecture Notes in Mathematics. Springer Berlin, Heidelberg, 2011.

Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher Processes and Bounding the Risk of Function Learning. In *High Dimensional Probability II*, pp. 443–457, Boston, MA, 2000. Birkhäuser Boston. doi:10.1007/978-1-4612-1358-1_29.

Akshay Krishnamurthy, Zhiwei Steven Wu, and Vasilis Syrgkanis. Semiparametric Contextual Bandits. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2776–2785. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/krishnamurthy18a.html.

Sanath Kumar Krishnamurthy, Vitor Hadad, and Susan Athey. Adapting to misspecification in contextual bandits with offline regression oracles. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5805–5814. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/krishnamurthy21a.html.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing Frontiers in Open Language Model Post-Training. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=i1uGbfHHpH.

Tor Lattimore, Csaba Szepesvári, and Gellert Weisz. Learning with Good Feature Representations in Bandits and in RL with a Generative Model. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5662–5670. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/lattimore20a.html.

Guillaume Lecué and Shahar Mendelson. General nonexact oracle inequalities for classes with a subexponential envelope. *The Annals of Statistics*, 40(2):832 – 860, 2012. doi:10.1214/11-AOS965.

Guillaume Lecué and Philippe Rigollet. Optimal learning with Q-aggregation. *The Annals of Statistics*, 42(1):211 – 224, 2014. doi:10.1214/13-AOS1190.

Yunwen Lei, Antoine Ledent, and Marius Kloft. Sharper Generalization Bounds for Pairwise Learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21236–21246. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f3173935ed8ac4bf073c1bcd63171f8a-Abstract.html.

Wendi Li and Yixuan Li. Process Reward Model with Q-value Rankings. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wQEdh2cgEk.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with Square Loss: Localization through Offset Rademacher Complexity. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1260–1285, Paris, France, 03–06 Jul 2015. PMLR. URL https://proceedings.mlr.press/v40/Liang15.html.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, and Yang Liu. Skywork-Reward-V2: Scaling Preference Data Curation via Human-AI Synergy. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ofgxkMLqic.

Dong C. Liu and Jorge Nocedal. On the Limited Memory BFGS Method for Large Scale Optimization. *Mathematical Programming*, 45(1):503–528, 1989. doi:10.1007/BF01589116.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Davide Maran and Csaba Szepesvári. Beyond Least Squares: Uniform Approximation and the Hidden Cost of Misspecification. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 168329–168377. Curran Associates, Inc., 2025. URL https://openreview.net/forum?id=2T2zMiqcY6.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34 (5):2326 – 2366, 2006. doi:10.1214/009053606000000786.

Mathematical Association of America. American Mathematics Competitions: AMC 12 and AIME, 2024. URL https://maa.org/student-programs/amc/.

Shahar Mendelson. Learning without Concentration. In *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pp. 25–39, Barcelona, Spain, 13–15 Jun 2014. PMLR. URL https://proceedings.mlr.press/v35/mendelson14.html.

Shahar Mendelson. Learning without concentration for general loss functions. *Probability Theory and Related Fields*, 171(1):459–502, Jun 2018. doi:10.1007/s00440-017-0784-y.

Meta. Llama-3.2-3B-Instruct Model Card. Hugging Face model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct.

Yair Mundlak. On the Pooling of Time Series and Cross Section Data. *Econometrica*, 46(1):69–85, 1978. doi:10.2307/1913646.

Rémi Munos. Error Bounds for Approximate Policy Iteration. In *Proceedings of the Twentieth International Conference on Machine Learning*, ICML'03, pp. 560–567. AAAI Press, 2003. URL https://cdn.aaai.org/ICML/2003/ICML03-074.pdf.

Rémi Munos. Performance Bounds in L_p-norm for Approximate Value Iteration. *SIAM Journal on Control and Optimization*, 46(2):541–561, 2007. doi:10.1137/040614384.

X Nie and S Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2): 299–319, 06 2021. doi:10.1093/biomet/asaa076.

Tapio Pahikkala, Antti Airola, Michiel Stock, Bernard De Baets, and Willem Waegeman. Efficient regularized least-squares algorithms for conditional ranking on relational data. *Machine Learning*, 93(2):321–356, 2013. doi:10.1007/s10994-013-5354-7.

Rattana Pukdee, Maria-Florina Balcan, and Pradeep Ravikumar. Reward Learning from Best-of-N Preference Data: Targets, Tradeoffs, and Design Principles. *arXiv preprint arXiv:2605.30619*, 2026. URL https://arxiv.org/abs/2605.30619.

Friedrich Pukelsheim. *Optimal Design of Experiments*, volume 50 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), 2006.

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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Alexander Rakhlin and Karthik Sridharan. Online Non-Parametric Regression. In *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pp. 1232–1264, Barcelona, Spain, 13–15 Jun 2014. PMLR. URL https://proceedings.mlr.press/v35/rakhlin14.html.

Alexander Rakhlin, Karthik Sridharan, and Alexandre B. Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789 – 824, 2017. doi:10.3150/14-BEJ679.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessimism. *IEEE Transactions on Information Theory*, 68(12):8156–8196, 2022. doi:10.1109/TIT.2022.3185139.

P. M. Robinson. Root-N-Consistent Semiparametric Regression. *Econometrica*, 56(4):931–954, 1988. doi:10.2307/1912705.

J. Jon Ryu, Jeongyeol Kwon, Benjamin Koppe, and Kwang-Sung Jun. Improved Offline Contextual Bandits with Second-Order Bounds: Betting and Freezing. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5015–5053. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/ryu25a.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Paul Speckman. Kernel Smoothing in Partial Linear Models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(3):413–436, 1988. doi:10.1111/j.2517-6161.1988.tb01738.x.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html.

Adith Swaminathan and Thorsten Joachims. Counterfactual Risk Minimization: Learning from Logged Bandit Feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 814–823, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/swaminathan15.html.

Michel Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Études Scientifiques*, 81(1):73–205, Dec 1995. doi:10.1007/BF02699376.

Michel Talagrand. A new look at independence. *The Annals of Probability*, 24(1):1 – 34, 1996a. doi:10.1214/aop/1042644705.

Michel Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126 (3):505–563, Nov 1996b. doi:10.1007/s002220050108.

Joel A. Tropp. An Introduction to Matrix Concentration Inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015. doi:10.1561/2200000048.

Benjamin Van Roy and Shi Dong. Comments on the Du-Kakade-Wang-Yang Lower Bounds. *arXiv preprint arXiv:1911.07910*, 2019. URL https://arxiv.org/abs/1911.07910.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10582–10592, Miami, Florida, USA, 2024a. Association for Computational Linguistics. doi:10.18653/v1/2024.findings-emnlp.620.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. HelpSteer 2: Open-source dataset for training top-performing reward models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 1474–1501. Curran Associates, Inc., 2024b. URL https://openreview.net/forum?id=PvVKUFhaNy.

Sebastian Johann Wetzel, Kevin Ryczko, Roger Gordon Melko, and Isaac Tamblyn. Twin neural network regression. *Applied AI Letters*, 3(4):e78, 2022. doi:10.1002/ail2.78.

Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qu. Automated Essay Scoring via Pairwise Contrastive Regression. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 2724–2733, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.240/.

Bin Yu. Assouad, Fano, and Le Cam. In David Pollard, Erik Torgersen, and Grace L. Yang (eds.), *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pp. 423–435. Springer New York, New York, NY, 1997. doi:10.1007/978-1-4612-1880-7_29.

Zhiwei Zhang, Hui Liu, Xiaomin Li, Zhenwei Dai, Jingying Zeng, Fali Wang, Minhua Lin, Ramraj Chandradevan, Linlin Wu, Zhen Li, Chen Luo, Zongyu Wu, Xianfeng Tang, Qi He, and Suhang Wang. Bradley-Terry and Multi-Objective Reward Modeling Are Complementary. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3QHKJcwnpb.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled Reinforcement Learning with Human Feedback from Pairwise or K-wise Comparisons. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 43037–43067. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zhu23f.html.

Dixian Zhu, Tianbao Yang, and Livnat Jerby. Gradient Aligned Regression via Pairwise Losses. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80259–80281. PMLR, 13–19 Jul 2025. URL https://openreview.net/forum?id=1bnH3Q3zL6.

CONTENTS

1 Introduction **1**

1.1 Theoretical Setup: Grouped Offline Reward Learning 2

1.2 Our Contributions 3

2 A Localized Analysis for Grouped Regression **4**

3 Finite Class: Nuisance Cancellation and Gains from $N_a \geq 2$ **5**

4 Linear Classes: A Geometric Bias–Variance Tradeoff **6**

4.1 VR for Linear Class 6

4.2 VDR for Linear Class 7

4.3 Comparing VR and VDR: Feature Geometry Perspective 8

5 Conclusion and Future Work **9**

A Related Work **19**

B Estimation Error to Offline Regret **21**

B.1 Coverage Coefficients and Regret Guarantees 21

B.2 Proof of Proposition B.1: Offline Regret of GREEDY and PESSIMISM 21

C Auxiliary Results for Grouped Regression **24**

C.1 VDR Population Loss and Centered Prediction Error 24

C.2 Proof of Proposition 2.1: Localized Fixed-Point Inequality 24

D Deferred Proofs from Section 3: Finite Class **25**

D.1 VALUE REGRESSION (VR) 25

D.1.1 Proof of Lemma 3.1: Localized Modulus for VR in Finite Class 25

D.2 VALUE DIFFERENCE REGRESSION (VDR) 26

D.2.1 Proof of the VDR Bound in Theorem 3.1 26

D.2.2 Proof of Lemma D.1: Localized Modulus for VDR in Finite Class 27

E Exact Oracle Inequality for Finite Class via the STAR ESTIMATOR **29**

E.1 STAR ESTIMATOR and Its Guarantee 29

E.2 Proof of Theorem E.1: STAR ESTIMATOR Bound for Finite Class 29

E.3 Proof of Lemma E.1: Localized Modulus for the STAR ESTIMATOR in Finite Class 30

F Deferred Proofs from Section 4: Linear Class **33**

F.1 VALUE REGRESSION (VR) 33

F.1.1 Proof of Theorem 4.1: VR Bound for Linear $\mathcal{F}_{\text{lin}}$ 33

F.1.2 Proof of Lemma F.1: Localized Modulus for Linear VR 34

F.1.3 Alternate Error Bound for Linear VR 36
F.2 VALUE DIFFERENCE REGRESSION (VDR) 38
F.2.1 Proof of Theorem 4.2: VDR Bound for Linear $\mathcal{F}_{\text{lin}}$ 38
F.2.2 Proof of Lemma F.2 39

G Lower Bounds **41**
G.1 Proof of Theorem 3.2: Failure Mode of VR 41
G.2 Agnostic Minimax Lower Bound 43
G.3 Proof of Theorem G.1: Agnostic Minimax Lower Bound 45

H Deferred Experimental Details **50**
H.1 Controlled Synthetic Linear-Bandit Experiment 50
H.2 LLM Response-Selection Experiment 54
H.3 MCTS-Based Mathematical Reasoning Experiment 55

A Related Work

We review related work on reward modeling objectives, regression under misspecification, statistical guarantees for pointwise/pairwise regression, and decision-making in contextual bandits.

(1) Reward modeling objectives. Reward estimators can be trained to predict absolute numerical values, numerical differences between alternatives (Wetzel et al., 2022; Xie et al., 2022), or preferences. These objectives impose different requirements: pointwise Value Regression (VR) fits individual reward labels, Value Difference Regression (VDR) fits their numerical gaps, and preference-ranking objectives fit comparison outcomes without directly matching observed gap magnitudes. Numerical-gap objectives have been studied in magnitude-preserving ranking and conditional ranking (Cortes et al., 2007; Pahikkala et al., 2013). Recent works have also demonstrated that combining pointwise regression with pairwise losses can improve predictive performance (Zhu et al., 2025) and value-guided reasoning (Chen et al., 2024). Preference-based reward modeling instead commonly fits comparison probabilities using the Bradley–Terry model (Bradley & Terry, 1952; Christiano et al., 2017; Zhu et al., 2023), with related comparison objectives used in LLM reward models (Guan et al., 2025; Zhang et al., 2026). These findings show that there is an intricate distinction between absolute prediction, gap prediction, and downstream decision quality. Our work studies VR and VDR constructed from the same numerical observations, separating their statistical behavior under a shared, *grouped* sampling model.

(2) Semiparametric estimation. Our weak realizability assumption has the semiparametric form $r^\star(x,a) = f^\star(x,a)+b^\star(x)$, where $f^\star \in \mathcal{F}$ captures action-dependent variation and $b^\star$ is an unknown action-independent nuisance function; neither $b^\star$ nor the complete reward function must belong to $\mathcal{F}$. For linear $\mathcal{F}$, our assumption has the structure of a classical partially linear model (Speckman, 1988), with a finite-dimensional action-dependent component and a nonparametric context-dependent nuisance. Paralleling the partialling-out construction of Robinson (1988), our VDR eliminate the nuisance $b^\star$ and target the within-context differences of $f^\star$. These differences identify its centered component on the behavior-policy support, while action-invariant parameter directions remain unidentified. The related principle of removing shared variation also appears in partial regression and panel-data within transformations (Frisch & Waugh, 1933; Mundlak, 1978). Modern approaches use orthogonal scores and nuisance estimation to reduce sensitivity to nuisance-estimation errors (Chernozhukov et al., 2018; Foster & Syrgkanis, 2023); heterogeneous-treatment-effect estimation similarly isolates treatment contrasts through residualized objectives (Nie & Wager, 2021). In sequential decision-making, action-centered and semiparametric contextual bandits allow complex action-independent baselines while modeling action effects (Choi et al., 2023; Greenewald et al., 2017; Kim & Paik, 2019; Krishnamurthy et al., 2018). Carranza et al. (2023) develop contextual bandit algorithms based on heterogeneous-treatment-effect oracles, while Kim et al. (2025) use experimental design and orthogonalized regression for semiparametric bandits. We share the motivation of removing decision-irrelevant nuisance variation, but our setting instead compares pointwise regression with regression on within-context reward differences, examining how the regression objective and repeated action observation affect estimation and decision guarantees.

(3) Misspecified linear regression. Random-design linear regression under misspecification has been studied under different prediction and distribution-shift criteria. For out-of-sample L_2 prediction, Hsu et al. (2012b; 2014) analyze ordinary least squares and ridge regression, quantifying the effects of approximation residuals and empirical-covariance errors. Our VR analysis (Theorem 4.1) belongs to this line: when $N_a = 1$, its bound has the same first-order structure, while for $N_a \geq 2$ it further separates residual fluctuations across contexts from those averaged over all observations. Under adversarial covariate shift, Amortila et al. (2024) show that standard ERM can amplify L_∞ misspecification through distribution mismatch and develop a disagreement-based regression procedure that avoids this amplification. Maran & Szepesvári (2025) instead study uniform prediction for misspecified random-design linear models, identifying the Lebesgue constant of the population L_2 projection as the sharp amplification factor for uniform approximation error in L_∞. Related work characterizes optimal approximation factors for misspecified linear off-policy value-function estimation (Amortila et al., 2023). For downstream decisions, Lattimore et al. (2020) quantify the effect of uniform linear approximation error on bandit and RL guarantees, while Krishnamurthy et al. (2021) develop contextual bandit algorithms whose regret adapts to average reward-model

misspecification using offline regression oracles. Our VR analysis extends the prediction-error perspective to repeated observations within contexts. The VDR comparison concerns a narrower form of misspecification: action-independent residuals can be removed before regression by differencing, whereas action-dependent residuals generally remain. Related coverage-dependent approximation errors and decision lower bounds are discussed in Appendix G.2.

(4) Statistical guarantees with grouped observations. Classical pairwise-learning theory addresses the dependence caused by reusing observations across pairs, using U-statistic and stability analyses (Clémençon et al., 2008; Lei et al., 2020). Our observations have an additional grouped structure: N_a action–reward observations share each of N_x contexts, and within-context differencing cancels the common nuisance. Treating each context block as an independent observation permits block-level analysis, but exploiting repeated action samples requires tracking within-block variation, which we do throughout our proofs.

Our analysis also connects to localization methods for squared-loss learning. Local Rademacher complexities characterize estimation rates through critical radii (Bartlett et al., 2005; Bousquet et al., 2002; Koltchinskii, 2011), while offset Rademacher complexity uses a negative quadratic term to control stochastic fluctuations (Liang et al., 2015). The separation of quadratic and multiplier processes is also central to small-ball analyses of least squares (Mendelson, 2014). Our localized block-offset modulus follows these principles but is defined directly through the unsymmetrized excess-loss process. Evaluating it with the grouped sampling structure preserves the distinction between fluctuations across contexts and within contexts.

(5) Offline contextual bandits and RL. Offline contextual bandits study policy selection and learning from previously collected action–reward observations. Prior work addresses logging-policy mismatch through propensity-weighted objectives and variance control (Swaminathan & Joachims, 2015), and establishes policy guarantees through pessimism under limited coverage (Rashidinejad et al., 2022). In batch RL, Chen & Jiang (2019) study the roles of distribution coverage and function approximation in sample-efficient learning. Related analyses of approximate policy and value iteration connect value-function approximation errors to policy performance (Farahmand et al., 2010; Munos, 2003). More recently, Ryu et al. (2025) develop betting-based confidence bounds for off-policy selection and a freezing approach to off-policy learning, obtaining guarantees that adapt to second-order information. Related policy-optimization methods also use comparison objectives: DPO applies a preference loss to policy log-ratios (Rafailov et al., 2023), whereas REBEL fits differences of policy log-ratios onto numerical reward differences (Gao et al., 2024). Our work studies a complementary question: how pointwise versus pairwise value regression affects regret under structured misspecification and repeated action observations.

B Estimation Error to Offline Regret

B.1 Coverage Coefficients and Regret Guarantees

In this appendix, we present two offline RL algorithms along with their offline regret guarantees. To do so, we first define two notions of **coverage coefficients**, as commonly done in offline RL literature (Chen & Jiang, 2019; Rashidinejad et al., 2022). First, the *policy-specific coverage*: for each deterministic $\pi : \mathcal{X} \to \mathcal{A}$,

$$C(\pi) := \mathbb{E}_{x\sim\rho}\left[\frac{1}{\pi_{\mathrm{ref}}(\pi(x) \mid x)}\right], \quad C^\star := C(\pi^\star),$$

where $\pi^\star$ is a fixed deterministic optimal policy, $\pi^\star(x) \in \arg\max_{a\in\mathcal{A}} r^\star(x,a)$, and we take the convention that $1/0 = \infty$. Next, the *($\mathcal{F}$-restricted) all-policy coverage*:

$$C^\infty_{\mathcal{F}} := \mathbb{E}_{x\sim\rho}\left[\max_{a\in\mathcal{A}_{\mathcal{F}}(x)} \frac{1}{\pi_{\mathrm{ref}}(a \mid x)}\right], \quad \mathcal{A}_{\mathcal{F}}(x) := \left\{a \in \mathcal{A} : \exists f \in \mathcal{F} \text{ s.t. } f(x,a) \geq \max_{b\in\mathcal{A}} f(x,b)\right\}.$$

Note that such a restricted variant of the all-policy coefficient in an $\mathcal{F}$-dependent manner has been considered before, e.g., Chen & Jiang (2019, Appendix G). For every policy satisfying $\pi(x) \in \mathcal{A}_{\mathcal{F}}(x)$, one has $C(\pi) \leq C^\infty_{\mathcal{F}}$. There is no general ordering between $C(\pi)$ and $C^\star = C(\pi^\star)$ for arbitrary π; we can nevertheless construct instances in which $C^\star \ll C^\infty_{\mathcal{F}}$ (e.g., Remark 3).

We now present offline regret upper bounds for the Greedy and Pessimism policies:

Proposition B.1. *Let $\hat{f}_N \in \mathcal{F}$ be our estimator, and let $\mathcal{C} : g \mapsto \mathcal{C}g$ be the centering operator $\mathcal{C}g(x,a) := g(x,a) - \mathbb{E}_{a'\sim\pi_{\mathrm{ref}}(\cdot|x)}[g(x,a')]$, which preserves the action gap. Suppose that Assumption 1 holds and that $\mathbb{E}_{x,a}[(\mathcal{C}\hat{f}_N - \mathcal{C}r^\star)^2] \leq \gamma_N^2$. Then,*

- *The* Greedy *policy satisfies*

$$\mathrm{Reg}(\hat{\pi}_N^{\textsc{Greedy}}) \leq 2\sqrt{C^\infty_{\mathcal{F}} - 1}\,\gamma_N, \quad \hat{\pi}_N^{\textsc{Greedy}}(x) := \arg\max_{a\in\mathcal{A}} \hat{f}_N(x,a);$$

- *The* Pessimism *policy satisfies*

$$\mathrm{Reg}(\hat{\pi}_N^{\textsc{Pessimism}}) \leq 2\sqrt{C^\star - 1}\,\gamma_N,$$
$$\hat{\pi}_N^{\textsc{Pessimism}} \in \arg\max_{\pi:\mathcal{X}\to\mathcal{A}} \left\{\mathbb{E}_{x\sim\rho}\left[\hat{f}_N(x,\pi(x))\right] - \gamma_N\sqrt{C(\pi) - 1}\right\}.$$

Proof Sketch. We utilize the conditional mean-zero property of the centering operator and the Cauchy–Schwarz inequality. Although this is standard in offline RL (Rashidinejad et al., 2022), we provide the full proof in Appendix B.2. □

Implementation details. Despite the benefit of improving the coverage from $C^\infty_{\mathcal{F}}$ to $C^\star$, Pessimism introduces some implementation difficulties. Pessimism requires computing $\mathbb{E}_{x\sim\rho}$ and knowing γ_N and $C(\pi)$ for each π. The former can be approximated to an arbitrary precision via sampling from ρ. The latter requires knowledge of π_{ref} and γ_N, which may be problematic.

For instance, consider the error bound of VR for finite class (Theorem 3.1), where $\gamma_N^2 = 2\varepsilon_{\mathrm{aprx}} + \left(\frac{22F^2+4FR_{\max}}{N_x} + \frac{18\sigma^2}{N}\right)\log\frac{|\mathcal{F}|}{\delta}$. In this case, the learner needs a useful *a priori* upper bound on $\varepsilon_{\mathrm{aprx}}$ in addition to knowing π_{ref}. A worst-case bound of $\varepsilon_{\mathrm{aprx}} \leq 2F^2 + 2R_{\max}^2$ may be utilized, but it can be too crude especially under small misspecification.

B.2 Proof of Proposition B.1: Offline Regret of Greedy and Pessimism

Recall that $r^\star(x,a) = f^\star(x,a) + b^\star(x)$ by Assumption 1, so $f^\star(x,a) - f^\star(x,a') = r^\star(x,a) - r^\star(x,a')$ for any context $x \in \mathcal{X}$. Thus, $r^\star$ and $f^\star$ share the optimal policy $\pi^\star(x) \in \arg\max_{a\in\mathcal{A}} r^\star(x,a) = \arg\max_{a\in\mathcal{A}} f^\star(x,a)$, and $\pi^\star(x) \in \mathcal{A}_{\mathcal{F}}(x)$. For notational simplicity, we denote $\Delta(x,a) := \mathcal{C}\hat{f}_N(x,a) - \mathcal{C}r^\star(x,a)$. By the definition of $\mathcal{C}$, for every $x \in \mathcal{X}$,

$\mathbb{E}_{a\sim\pi_{\mathrm{ref}}(\cdot|x)}[\Delta(x,a) \mid x] = 0$ and

$$\mathbb{E}_{x,a}[\Delta(x,a)^2] = \mathbb{E}_x\left[\mathrm{Var}_{a\sim\pi_{\mathrm{ref}}(\cdot|x)}\big(\hat{f}_N(x,a) - r^\star(x,a) \mid x\big)\right] \le \gamma_N^2. \tag{6}$$

If the relevant coverage coefficient is infinite, the corresponding bound is vacuous; hence, in what follows, we only use the inverse-propensity identities in the finite-coverage case.

Part 1: offline regret of the GREEDY policy. Since $\mathcal{C}$ preserves the action gap, for any $f \in \mathcal{F}$, f and $\mathcal{C}f$ have the same GREEDY policy $\hat{\pi}$ since $f(x,a) - f(x,a') = \mathcal{C}f(x,a) - \mathcal{C}f(x,a')$, and $\hat{\pi}(x) \in \mathcal{A}_{\mathcal{F}}(x)$. Let $\hat{\pi}(x) := \hat{\pi}_N^{\mathrm{GREEDY}}(x) = \arg\max_{a\in\mathcal{A}} \hat{f}_N(x,a)$, then $\hat{\pi}(x) \in \mathcal{A}_{\mathcal{F}}(x)$. For any context $x \in \mathcal{X}$, the suboptimality gap can be bounded by the maximum estimation error across all actions in $\mathcal{A}_{\mathcal{F}}(x)$ as:

$$\begin{aligned}
&r^\star(x,\pi^\star(x)) - r^\star(x,\hat{\pi}(x)) \\
&= \mathcal{C}r^\star(x,\pi^\star(x)) - \mathcal{C}r^\star(x,\hat{\pi}(x)) \\
&= \Big(\mathcal{C}r^\star(x,\pi^\star(x)) - \mathcal{C}\hat{f}_N(x,\pi^\star(x))\Big) + \Big(\mathcal{C}\hat{f}_N(x,\pi^\star(x)) - \mathcal{C}\hat{f}_N(x,\hat{\pi}(x))\Big) \\
&\quad + \Big(\mathcal{C}\hat{f}_N(x,\hat{\pi}(x)) - \mathcal{C}r^\star(x,\hat{\pi}(x))\Big) \\
&\le \Big(\mathcal{C}r^\star(x,\pi^\star(x)) - \mathcal{C}\hat{f}_N(x,\pi^\star(x))\Big) + \Big(\mathcal{C}\hat{f}_N(x,\hat{\pi}(x)) - \mathcal{C}r^\star(x,\hat{\pi}(x))\Big) \\
&\le 2\max_{a\in\mathcal{A}_{\mathcal{F}}(x)} |\Delta(x,a)|,
\end{aligned}$$

where the first inequality holds since $\hat{\pi}$ is the GREEDY policy of $\mathcal{C}\hat{f}_N$ as well as of $\hat{f}_N$, and the last inequality holds since $\pi^\star(x), \hat{\pi}(x) \in \mathcal{A}_{\mathcal{F}}(x)$. By Eqn. (6), for every $a \in \mathcal{A}_{\mathcal{F}}(x)$,

$$\begin{aligned}
|\Delta(x,a)| &= \left|\mathbb{E}_{a'\sim\pi_{\mathrm{ref}}(\cdot|x)}\left[\left(\frac{\mathbb{1}\{a'=a\}}{\pi_{\mathrm{ref}}(a \mid x)} - 1\right)\Delta(x,a') \,\middle|\, x\right]\right| \\
&\le \sqrt{\frac{1}{\pi_{\mathrm{ref}}(a \mid x)} - 1}\sqrt{\mathbb{E}_{a'\sim\pi_{\mathrm{ref}}(\cdot|x)}[\Delta(x,a')^2 \mid x]}.
\end{aligned}$$

Taking the expectation over $x \sim \rho$ and applying the Cauchy–Schwarz inequality ($\mathbb{E}[YZ] \le \sqrt{\mathbb{E}[Y^2]\mathbb{E}[Z^2]}$), we obtain:

$$\begin{aligned}
\mathrm{Reg}(\hat{\pi}_N^{\mathrm{GREEDY}}) &= \mathbb{E}_{x\sim\rho}\left[r^\star\big(x,\pi^\star(x)\big) - r^\star\big(x,\hat{\pi}(x)\big)\right] \\
&\le 2\mathbb{E}_{x\sim\rho}\left[\sqrt{\max_{a\in\mathcal{A}_{\mathcal{F}}(x)}\left\{\frac{1}{\pi_{\mathrm{ref}}(a \mid x)} - 1\right\}}\sqrt{\mathbb{E}_{a\sim\pi_{\mathrm{ref}}}[\Delta(x,a)^2 \mid x]}\right] \\
&\le 2\sqrt{\mathbb{E}_{x\sim\rho}\left[\max_{a\in\mathcal{A}_{\mathcal{F}}(x)}\left\{\frac{1}{\pi_{\mathrm{ref}}(a \mid x)} - 1\right\}\right]}\sqrt{\mathbb{E}_{x\sim\rho,a\sim\pi_{\mathrm{ref}}}[\Delta(x,a)^2]} \\
&\le 2\sqrt{C_{\mathcal{F}}^{\infty} - 1}\gamma_N.
\end{aligned}$$

Part 2: offline regret of the PESSIMISM policy. First, for any deterministic policy $\pi : \mathcal{X} \to \mathcal{A}$, we bound the policy evaluation error. Let $p_\pi(x) := \pi_{\mathrm{ref}}(\pi(x) \mid x)$. By Eqn. (6) and the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\Big|\mathbb{E}_{x\sim\rho}\big[\Delta(x,\pi(x))\big]\Big| &= \left|\mathbb{E}_{x,a}\left[\left(\frac{\mathbb{1}\{a=\pi(x)\}}{p_\pi(x)} - 1\right)\Delta(x,a)\right]\right| \\
&\le \sqrt{\mathbb{E}_{x,a}\left[\left(\frac{\mathbb{1}\{a=\pi(x)\}}{p_\pi(x)} - 1\right)^2\right]}\sqrt{\mathbb{E}_{x,a}[\Delta(x,a)^2]} \\
&\le \gamma_N\sqrt{C(\pi) - 1}.
\end{aligned}$$

Thus, for any policy π, the evaluation error is bounded by:

$$\left|\mathbb{E}_{x\sim\rho}\left[\mathcal{C}\hat{f}_N(x,\pi(x))\right]-\mathbb{E}_{x\sim\rho}\left[\mathcal{C}r^\star\big(x,\pi(x)\big)\right]\right|=\left|\mathbb{E}_{x\sim\rho}\left[\Delta\big(x,\pi(x)\big)\right]\right|\leq\gamma_N\sqrt{C(\pi)-1}. \quad (7)$$

Recall that the pessimistic objective is $\mathbb{E}_{x\sim\rho}[\hat{f}_N(x,\pi(x))]-\gamma_N\sqrt{C(\pi)-1}$. Since $\mathcal{C}$ preserves the action-gap for the fixed context x, we have

$$\begin{aligned}\hat{\pi}_N^{\textsc{Pessimism}}&\in\arg\max_{\pi:\mathcal{X}\to\mathcal{A}}\left\{\mathbb{E}_{x\sim\rho}\left[\hat{f}_N(x,\pi(x))\right]-\gamma_N\sqrt{C(\pi)-1}\right\}\\&=\arg\max_{\pi:\mathcal{X}\to\mathcal{A}}\left\{\mathbb{E}_{x\sim\rho}\left[\mathcal{C}\hat{f}_N(x,\pi(x))\right]-\gamma_N\sqrt{C(\pi)-1}\right\}.\end{aligned}$$

Thus, by the optimality of $\hat{\pi}_N^{\textsc{Pessimism}}$ w.r.t. this objective, it holds

$$\mathbb{E}_{x\sim\rho}\left[\mathcal{C}\hat{f}_N(x,\hat{\pi}_N^{\textsc{Pessimism}}(x))\right]-\gamma_N\sqrt{C(\hat{\pi}_N^{\textsc{Pessimism}})-1}\geq\mathbb{E}_{x\sim\rho}\left[\mathcal{C}\hat{f}_N(x,\pi^\star(x))\right]-\gamma_N\sqrt{C(\pi^\star)-1}. \quad (8)$$

Together with the bound implied by Eqn. (7),

$$\begin{aligned}&\mathrm{Reg}(\hat{\pi}_N^{\textsc{Pessimism}})\\&=V(\pi^\star)-V(\hat{\pi}_N^{\textsc{Pessimism}})\\&=\mathbb{E}_{x\sim\rho}\left[r^\star\big(x,\pi^\star(x)\big)-r^\star\big(x,\hat{\pi}_N^{\textsc{Pessimism}}(x)\big)\right]\\&=\mathbb{E}_{x\sim\rho}\left[\mathcal{C}r^\star\big(x,\pi^\star(x)\big)-\mathcal{C}r^\star\big(x,\hat{\pi}_N^{\textsc{Pessimism}}(x)\big)\right]\\&\leq\mathbb{E}_{x\sim\rho}\left[\mathcal{C}r^\star\big(x,\pi^\star(x)\big)\right]-\left(\mathbb{E}_{x\sim\rho}\left[\mathcal{C}\hat{f}_N\big(x,\hat{\pi}_N^{\textsc{Pessimism}}(x)\big)\right]-\gamma_N\sqrt{C(\hat{\pi}_N^{\textsc{Pessimism}})-1}\right)\\&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(Eqn. (7))}\\&\leq\mathbb{E}_{x\sim\rho}\left[\mathcal{C}r^\star\big(x,\pi^\star(x)\big)\right]-\left(\mathbb{E}_{x\sim\rho}\left[\mathcal{C}\hat{f}_N\big(x,\pi^\star(x)\big)\right]-\gamma_N\sqrt{C(\pi^\star)-1}\right)\qquad\text{(Eqn. (8))}\\&\leq\left|\mathbb{E}_{x\sim\rho}\left[\mathcal{C}\hat{f}_N\big(x,\pi^\star(x)\big)\right]-\mathbb{E}_{x\sim\rho}\left[\mathcal{C}r^\star\big(x,\pi^\star(x)\big)\right]\right|+\gamma_N\sqrt{C(\pi^\star)-1}\\&\leq\gamma_N\sqrt{C(\pi^\star)-1}+\gamma_N\sqrt{C(\pi^\star)-1}=2\sqrt{C^\star-1}\gamma_N.\qquad\qquad\qquad\qquad\qquad\qquad\text{(Eqn. (7))}\end{aligned}$$

which completes the proof for both policies. □

C Auxiliary Results for Grouped Regression

For a comparator $f^\circ \in \mathcal{F}$ and a perturbation $h \in \mathcal{F} - f^\circ$, define the population and empirical excess losses by

$$\Delta(h; f^\circ) := \mathcal{L}(f^\circ + h) - \mathcal{L}(f^\circ), \qquad \widehat{\Delta}(h; f^\circ) := \widehat{\mathcal{L}}(f^\circ + h) - \widehat{\mathcal{L}}(f^\circ).$$

C.1 VDR Population Loss and Centered Prediction Error

Recall that $\mathcal{C}f(x,a) = f(x,a) - \mathbb{E}_{a' \sim \pi_{\mathrm{ref}}(\cdot|x)}[f(x,a')]$. The following identity will be used in both the finite- and linear-class VDR proofs.

Lemma C.1 (VDR population-loss identity)**.** *For every $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$,*

$$\mathcal{L}^{\mathrm{VDR}}(f) = 2\mathbb{E}_{x,a}\left[\left(\mathcal{C}f(x,a) - \mathcal{C}r^\star(x,a)\right)^2\right]. \tag{9}$$

Under Assumption 1, $\mathcal{C}f^\star = \mathcal{C}r^\star$ and $\mathcal{L}^{\mathrm{VDR}}(f^\star) = 0$. In particular,

$$\mathbb{E}_{x,a}\left[\left(\mathcal{C}\hat{f}_N^{\mathrm{VDR}}(x,a) - \mathcal{C}r^\star(x,a)\right)^2\right] = \frac{1}{2}\left(\mathcal{L}^{\mathrm{VDR}}(\hat{f}_N^{\mathrm{VDR}}) - \mathcal{L}^{\mathrm{VDR}}(f^\star)\right) = \frac{1}{2}\mathcal{L}^{\mathrm{VDR}}(\hat{f}_N^{\mathrm{VDR}}). \tag{10}$$

Proof. Let $A, A' \sim \pi_{\mathrm{ref}}(\cdot \mid X)$ independently and set $g = f - r^\star$. Conditional on X,

$$\begin{aligned}\mathbb{E}\left[\left(g(X,A) - g(X,A')\right)^2 \,\middle|\, X\right] &= 2\mathbb{E}\left[\left(g(X,A) - \mathbb{E}[g(X,A) \mid X]\right)^2 \,\middle|\, X\right] \\ &= 2\mathbb{E}\left[\left(\mathcal{C}g(X,A)\right)^2 \mid X\right].\end{aligned}$$

Taking expectation over X proves Eqn. (9). Under Assumption 1, $f^\star(x,a) - r^\star(x,a) = -b^\star(x)$ is action-independent, which gives $\mathcal{C}f^\star = \mathcal{C}r^\star$ and the remaining claims. □

C.2 Proof of Proposition 2.1: Localized Fixed-Point Inequality

Fix the comparator $f^\circ \in \mathcal{F}$ and denote

$$h := \hat{f}_N - f^\circ, \qquad \hat{r} := \Delta(h; f^\circ) = \mathcal{L}(\hat{f}_N) - \mathcal{L}(f^\circ).$$

By condition (i) and the definition of the empirical excess loss,

$$\widehat{\Delta}(h; f^\circ) = \widehat{\mathcal{L}}(\hat{f}_N) - \widehat{\mathcal{L}}(f^\circ) \le -\lambda\widehat{Q}(h).$$

Therefore,

$$\begin{aligned}\hat{r} &= \left[\Delta(h; f^\circ) - \widehat{\Delta}(h; f^\circ) - \lambda\widehat{Q}(h)\right] + \left[\widehat{\Delta}(h; f^\circ) + \lambda\widehat{Q}(h)\right] \\ &\le \Delta(h; f^\circ) - \widehat{\Delta}(h; f^\circ) - \lambda\widehat{Q}(h).\end{aligned} \tag{11}$$

Suppose first that $\hat{r} > 0$. Since $h \in \mathcal{F} - f^\circ$ and $\Delta(h; f^\circ) = \hat{r}$, the function h belongs to the set over which $\Psi(\hat{r}; \lambda, f^\circ)$ takes its supremum. Hence, by (11),

$$\hat{r} \le \Psi(\hat{r}; \lambda, f^\circ) \le \psi(\hat{r}),$$

where the second inequality follows from condition (ii). Therefore,

$$\hat{r} \in \{r \ge 0 : r \le \psi(r)\} \Longrightarrow \mathcal{L}(\hat{f}_N) - \mathcal{L}(f^\circ) = \hat{r} \le \sup\{r \ge 0 : r \le \psi(r)\}.$$

It remains to consider $\hat{r} \le 0$. Since $0 \in \mathcal{F} - f^\circ$ and

$$\Delta(0; f^\circ) = \widehat{\Delta}(0; f^\circ) = \widehat{Q}(0) = 0,$$

we have $\Psi(0; \lambda, f^\circ) \ge 0$. Condition (ii) therefore implies $\psi(0) \ge 0$, so that

$$0 \in \{r \ge 0 : r \le \psi(r)\}.$$

Consequently,

$$\mathcal{L}(\hat{f}_N) - \mathcal{L}(f^\circ) = \hat{r} \le 0 \le \sup\{r \ge 0 : r \le \psi(r)\}.$$

This completes the proof. □

D Deferred Proofs from Section 3: Finite Class

D.1 Value Regression (VR)

D.1.1 Proof of Lemma 3.1: Localized Modulus for VR in Finite Class

Recall that

$$f^\circ := f^\circ_{\mathrm{VR}} \in \arg\min_{g\in\mathcal{F}} \mathbb{E}_{x,a}\left[(g-r^\star)^2\right], \qquad \mathbb{E}_{x,a}\left[(f^\circ-r^\star)^2\right] = \varepsilon_{\mathrm{aprx}}.$$

Let

$$t_\delta := \log\frac{|\mathcal{F}|}{\delta}, \qquad v_\delta := \left(\frac{F^2}{N_x}+\frac{\sigma^2}{N}\right)t_\delta.$$

For any $f\in\mathcal{F}$, let

$$h := f-f^\circ, \qquad \Delta_f := \Delta(h;f^\circ) = \mathbb{E}_{x,a}\left[(f-r^\star)^2\right] - \varepsilon_{\mathrm{aprx}}.$$

Since f° is a population minimizer, $\Delta_f \geq 0$.

Since $\lambda = 0$, the localized block-offset modulus (Definition 2.1) specialized to VR is simply

$$\Psi(r;0,f^\circ) = \sup_{\substack{f\in\mathcal{F}\\ \Delta_f\leq r}}\left[\Delta_f - \left(\widehat{\mathcal{L}}(f)-\widehat{\mathcal{L}}(f^\circ)\right)\right]. \tag{12}$$

It therefore suffices to uniformly control the deviation between the population and empirical excess losses over $f\in\mathcal{F}$.

More precisely, we will prove that, with probability at least $1-\delta$, simultaneously for all $f\in\mathcal{F}$,

$$\Delta_f - \left(\widehat{\mathcal{L}}(f)-\widehat{\mathcal{L}}(f^\circ)\right) \leq 4\sqrt{v_\delta\left(\Delta_f+2\varepsilon_{\mathrm{aprx}}\right)} + \frac{8F(F+R_{\max})t_\delta}{3N_x}. \tag{13}$$

Once this uniform deviation is established, the desired modulus bound follows by restricting to $\Delta_f\leq r$.

Reduction to blockwise concentration. For each $f\in\mathcal{F}$, define

$$U_{i,j}(f) := \big(f(x_i,a_{i,j})-r_{i,j}\big)^2 - \big(f^\circ(x_i,a_{i,j})-r_{i,j}\big)^2, \qquad U_i(f) := \frac{1}{N_a}\sum_{j=1}^{N_a} U_{i,j}(f).$$

Then $\{U_i(f)\}_{i=1}^{N_x}$ are independent across $i\in[N_x]$, and

$$\mathbb{E}[U_i(f)] = \Delta_f, \qquad \frac{1}{N_x}\sum_{i=1}^{N_x} U_i(f) = \widehat{\mathcal{L}}(f)-\widehat{\mathcal{L}}(f^\circ). \tag{14}$$

Range and two-level block variance. We next bound the range and variance of $U_i(f)$, the two inputs needed for blockwise Bernstein concentration. Let

$$q_f(x,a) := (f(x,a)-r^\star(x,a))^2 - (f^\circ(x,a)-r^\star(x,a))^2.$$

Writing $h=f-f^\circ$ and using $r_{i,j}=r^\star(x_i,a_{i,j})+\eta_{i,j}$ gives

$$U_{i,j}(f) = q_f(x_i,a_{i,j}) - 2h(x_i,a_{i,j})\eta_{i,j}. \tag{15}$$

Since $\|f\|_\infty, \|f^\circ\|_\infty \leq F$ and $|r_{i,j}|\leq R_{\max}$, one can show that

$$|U_{i,j}(f)| \leq 4F(F+R_{\max}). \tag{16}$$

Moreover,

$$\begin{aligned} q_f(x,a)^2 &= h(x,a)^2\left((f(x,a)-r^\star(x,a)) + (f^\circ(x,a)-r^\star(x,a))\right)^2 \\ &\leq 8F^2\left((f(x,a)-r^\star(x,a))^2 + (f^\circ(x,a)-r^\star(x,a))^2\right), \end{aligned} \tag{17}$$

while

$$h(x,a)^2 \leq 2\left((f(x,a)-r^\star(x,a))^2 + (f^\circ(x,a)-r^\star(x,a))^2\right). \tag{18}$$

Since $\mathbb{E}[\eta_{i,j} \mid x_i, a_{i,j}] = 0$ and $\mathbb{E}[\eta_{i,j}^2 \mid x_i, a_{i,j}] \le \sigma^2$, Eqns. (15)–(18) imply

$$\begin{aligned}\mathbb{E}[U_{i,j}(f)^2] &\le 8(F^2+\sigma^2)\mathbb{E}_{x,a}\left[(f-r^\star)^2 + (f^\circ - r^\star)^2\right] \\ &= 8(F^2+\sigma^2)\left(\Delta_f + 2\varepsilon_{\mathrm{aprx}}\right).\end{aligned}$$

To retain the benefit of the N_a repeated action samples, define

$$m_f(x) \coloneqq \mathbb{E}_{a\sim\pi_{\mathrm{ref}}(\cdot|x)}[q_f(x,a)].$$

By Jensen's inequality and Eqn. (17),

$$\mathrm{Var}(m_f(X)) \le 8F^2\left(\Delta_f + 2\varepsilon_{\mathrm{aprx}}\right).$$

Conditional on x_i, the variables $\{U_{i,j}(f)\}_{j=1}^{N_a}$ are independent with common conditional mean $m_f(x_i)$. Hence, by the law of total variance,

$$\begin{aligned}\mathrm{Var}(U_i(f)) &= \left(1-\frac{1}{N_a}\right)\mathrm{Var}(m_f(X)) + \frac{1}{N_a}\mathrm{Var}(U_{i,1}(f)) \\ &\le 8\left(F^2 + \frac{\sigma^2}{N_a}\right)\left(\Delta_f + 2\varepsilon_{\mathrm{aprx}}\right).\end{aligned}$$

Uniform concentration and localization. Finally, Eqn. (16) implies

$$\mathbb{E}[U_i(f)] - U_i(f) \le 8F(F+R_{\max}) \qquad \text{almost surely.}$$

Applying the one-sided Bernstein inequality to the independent mean-zero variables $\{\mathbb{E}[U_i(f)] - U_i(f)\}_{i=1}^{N_x}$ gives, for every fixed $f \in \mathcal{F}$, with probability at least $1-e^{-t}$,

$$\Delta_f - \frac{1}{N_x}\sum_{i=1}^{N_x} U_i(f) \le 4\sqrt{\left(\frac{F^2}{N_x} + \frac{\sigma^2}{N_x N_a}\right)\left(\Delta_f + 2\varepsilon_{\mathrm{aprx}}\right)t} + \frac{8F(F+R_{\max})t}{3N_x}. \tag{19}$$

Taking $t = t_\delta$ and applying a union bound over $f \in \mathcal{F}$, we obtain an event of probability at least $1-\delta$ on which Eqn. (19) holds simultaneously for every $f \in \mathcal{F}$.

Since $N = N_x N_a$, Eqns. (14) and (19) give precisely the uniform deviation in Eqn. (13). Restricting Eqn. (13) to $\Delta_f \le r$ and taking the supremum in Eqn. (12) yields

$$\Psi(r; 0, f^\circ_{\mathrm{VR}}) \le 4\sqrt{v_\delta\left(r + 2\varepsilon_{\mathrm{aprx}}\right)} + \frac{8F(F+R_{\max})t_\delta}{3N_x}, \qquad \forall r \ge 0,$$

which proves the claim. □

D.2 Value Difference Regression (VDR)

D.2.1 Proof of the VDR Bound in Theorem 3.1

We take $f^\circ_{\mathrm{VDR}} = f^\star$ and $\lambda = 0$. The proof has three steps: empirical risk minimization verifies condition (i) of Proposition 2.1, the modulus lemma below verifies condition (ii) and determines the fixed point, and Eqn. (10) converts the resulting excess-loss bound into centered prediction error.

Since $\hat{f}_N^{\mathrm{VDR}}$ is an empirical risk minimizer, condition (i) holds. For condition (ii), let us first define

$$t_\delta \coloneqq \log\frac{2|\mathcal{F}|}{\delta}, \quad \beta_\delta \coloneqq \left[\frac{F^2}{N_x}\left(\frac{8}{3} + \frac{32\lfloor N_a^2/4\rfloor}{N_a(N_a-1)}\right) + \frac{16\sigma^2}{N_x(N_a-1)} + \frac{64FR_{\max}}{3N}\right]t_\delta.$$

The following lemma, proved in Appendix D.2.2, controls the localized modulus.

Lemma D.1. *With probability at least* $1-\delta$,

$$\Psi(r; 0, f^\circ_{\mathrm{VDR}}) \le \frac{3}{4}r + \frac{1}{2}\beta_\delta, \qquad \forall r \ge 0. \tag{20}$$

On the event in Lemma D.1, condition (ii) holds. Therefore, Proposition 2.1 gives

$$\mathcal{L}^{\mathrm{VDR}}(\hat{f}_N^{\mathrm{VDR}}) - \mathcal{L}^{\mathrm{VDR}}(f^\star) \le \sup\left\{r \ge 0 : r \le \frac{3}{4}r + \frac{1}{2}\beta_\delta\right\} = 2\beta_\delta.$$

By Lemma C.1,

$$\mathbb{E}_{x,a}\left[\left(\mathcal{C}\hat{f}_N^{\mathrm{VDR}} - \mathcal{C}r^\star\right)^2\right] = \frac{1}{2}\mathcal{L}^{\mathrm{VDR}}(\hat{f}_N^{\mathrm{VDR}}) \le \beta_\delta.$$

It remains only to compare β_δ with the displayed rate in Theorem 3.1. Since $N_a \ge 2$ and $N = N_x N_a$,

$$\frac{\lfloor N_a^2/4 \rfloor}{N_a(N_a-1)} \le \frac{1}{2}, \qquad \frac{1}{N_x(N_a-1)} = \frac{N_a}{N(N_a-1)} \le \frac{2}{N}.$$

Consequently,

$$\begin{aligned}\beta_\delta &\le \left(\frac{56F^2}{3N_x} + \frac{32\sigma^2}{N} + \frac{64FR_{\max}}{3N}\right)\log\frac{2|\mathcal{F}|}{\delta} \\ &= \frac{8}{3}\left(\frac{7F^2}{N_x} + \frac{12\sigma^2 + 8FR_{\max}}{N}\right)\log\frac{2|\mathcal{F}|}{\delta},\end{aligned}$$

which proves Eqn. (3) of Theorem 3.1. □

D.2.2 Proof of Lemma D.1: Localized Modulus for VDR in Finite Class

Here, let $f^\circ := f^\circ_{\mathrm{VDR}} = f^\star$, and for any $f \in \mathcal{F}$, write $g := f - f^\star$.

Define the population transformed squared error

$$q_f := \Delta(g; f^\star) = \mathcal{L}^{\mathrm{VDR}}(f) = 2\mathbb{E}_{x,a}\left[(\mathcal{C}g(x,a))^2\right] = 2\mathbb{E}_x\left[\mathrm{Var}_{a\sim\pi_{\mathrm{ref}}(\cdot|x)}\big(g(x,a)\big)\right].$$

By the definition of the localized modulus, it is enough to prove that, with probability at least $1-\delta$, simultaneously for all $f \in \mathcal{F}$,

$$q_f - \widehat{\Delta}(g; f^\star) \le \frac{3}{4}q_f + \frac{1}{2}\beta_\delta. \tag{21}$$

Indeed, restricting this inequality to $q_f \le r$ and taking the supremum immediately gives Eqn. (20).

For the empirical quantities, let us denote

$$g_{i,j} := g(x_i, a_{i,j}), \qquad \bar{g}_i := \frac{1}{N_a}\sum_{j=1}^{N_a} g_{i,j}, \qquad \widetilde{g}_{i,j} := g_{i,j} - \bar{g}_i.$$

We will repeatedly use the pairwise-centering identity

$$\sum_{j<k}(u_j - u_k)^2 = N_a\sum_{j=1}^{N_a}(u_j - \bar{u})^2. \tag{22}$$

It gives the empirical transformed squared error

$$\widehat{q}_f := \frac{1}{N_x\binom{N_a}{2}}\sum_{i=1}^{N_x}\sum_{j<k}(g_{i,j} - g_{i,k})^2 = \frac{2}{N_x(N_a-1)}\sum_{i,j}\widetilde{g}_{i,j}^2.$$

Since $r^\star = f^\star + b^\star$ and the action-independent $b^\star(x_i)$ cancels from every within-context difference, another application of Eqn. (22) gives

$$\widehat{\Delta}(g; f^\star) = \widehat{q}_f - Z_f, \qquad Z_f := \frac{4}{N_x(N_a-1)}\sum_{i,j}\widetilde{g}_{i,j}\eta_{i,j}. \tag{23}$$

Thus, the pointwise deviation has the master decomposition

$$q_f - \widehat{\Delta}(g; f^\star) = (q_f - \widehat{q}_f) + Z_f.$$

We control its two terms separately.

Step 1: control of the empirical quadratic term. For each block, define

$$A_i(f) \coloneqq \frac{1}{\binom{N_a}{2}} \sum_{j<k} (g_{i,j} - g_{i,k})^2.$$

Then $\{A_i(f)\}_{i=1}^{N_x}$ are independent, $\mathbb{E}[A_i(f)] = q_f$, and

$$0 \le A_i(f) \le K_{N_a}, \qquad K_{N_a} \coloneqq \frac{32F^2 \lfloor N_a^2/4 \rfloor}{N_a(N_a-1)}. \tag{24}$$

Indeed, $g = f - f^\star$ takes values in an interval of length at most $4F$, and the maximum empirical variance of N_a numbers in an interval of length $4F$ gives Eqn. (24). Moreover,

$$q_f \le 8F^2, \qquad \mathrm{Var}(A_i(f)) \le K_{N_a} q_f.$$

Hence, a one-sided Bernstein inequality followed by a union bound over $f \in \mathcal{F}$ gives, with probability at least $1 - \delta/2$, simultaneously for all $f \in \mathcal{F}$,

$$q_f - \widehat{q}_f \le \sqrt{\frac{2K_{N_a} q_f t_\delta}{N_x}} + \frac{8F^2 t_\delta}{3N_x}.$$

Using $\sqrt{ab} \le (a+b)/2$ gives[2]

$$\frac{1}{2} q_f \le \widehat{q}_f + d_\delta, \qquad d_\delta \coloneqq \left(K_{N_a} + \frac{8}{3}F^2 \right) \frac{t_\delta}{N_x}. \tag{25}$$

Step 2: control of the reward-noise term. Let

$$c_\delta \coloneqq \frac{\sigma^2 t_\delta}{N_x(N_a-1)}, \qquad b_\delta \coloneqq \frac{32FR_{\max} t_\delta}{3N}.$$

Conditional on all sampled contexts and actions, the variables $\{\widetilde{g}_{i,j}\eta_{i,j}\}_{i,j}$ are independent and mean zero. Moreover,

$$\sum_{i,j} \mathbb{E}\left[\widetilde{g}_{i,j}^2 \eta_{i,j}^2 \,\middle|\, \{x_k, a_{k,\ell}\}_{k,\ell} \right] \le \frac{\sigma^2 N_x (N_a - 1)}{2} \widehat{q}_f,$$

and, since g takes values in an interval of length at most $4F$,

$$|\widetilde{g}_{i,j}\eta_{i,j}| \le 8FR_{\max} \frac{N_a - 1}{N_a}.$$

Thus, conditional Bernstein and a union bound over $f \in \mathcal{F}$ imply, with conditional probability at least $1 - \delta/2$, simultaneously for all $f \in \mathcal{F}$,

$$Z_f \le 4\sqrt{c_\delta \widehat{q}_f} + b_\delta. \tag{26}$$

Step 3: combining the two events. Work on the intersection of Eqns. (25) and (26), which has probability at least $1 - \delta$. By Eqns. (23) and (26),

$$\begin{aligned}
q_f - \widehat{\Delta}(g; f^\star) &\le q_f - \widehat{q}_f + 4\sqrt{c_\delta \widehat{q}_f} + b_\delta \\
&\le q_f - \frac{1}{2}\widehat{q}_f + 8c_\delta + b_\delta && \text{(Young's inequality: } 4\sqrt{c_\delta \widehat{q}_f} \le \tfrac{1}{2}\widehat{q}_f + 8c_\delta.\text{)} \\
&\le \frac{3}{4} q_f + \frac{1}{2} d_\delta + 8c_\delta + b_\delta. && \text{(Eqn. (25))} \\
&= \frac{3}{4} q_f + \frac{1}{2}\beta_\delta && \text{(Definitions of } K_{N_a}, d_\delta, c_\delta, \text{ and } b_\delta\text{)}
\end{aligned}$$

Thus, the preceding display proves Eqn. (21). Restricting it to $q_f \le r$ and taking the supremum in the definition of $\Psi(r; 0, f^\circ_{\mathrm{VDR}})$ yields

$$\Psi(r; 0, f^\circ_{\mathrm{VDR}}) \le \frac{3}{4} r + \frac{1}{2}\beta_\delta, \qquad \forall r \ge 0.$$

This proves the claim. □

[2] Note that this is the lower isometry condition of Liang et al. (2015, Definition 5).

E Exact Oracle Inequality for Finite Class via the Star Estimator

E.1 Star Estimator and Its Guarantee

Here, we show that a different squared-loss estimator, namely the Star Estimator (Audibert, 2007; 2009; Liang et al., 2015), admits an exact oracle inequality in our problem setting, unlike the usual ERM, which achieves a non-exact oracle inequality (see the VR bound of Theorem 3.1).

Recall that $1 \le |\mathcal{F}| < \infty$ with $\sup_{f\in\mathcal{F}} \|f\|_\infty \le F$, and that the population and empirical VR losses are defined as

$$\mathcal{L}^{\mathrm{VR}}(f) := \mathbb{E}_{x,a}\left[(f - r^\star)^2\right], \quad \widehat{\mathcal{L}}^{\mathrm{VR}}(f) := \frac{1}{N}\sum_{i=1}^{N_x}\sum_{j=1}^{N_a}(f(x_i,a_{i,j}) - r_{i,j})^2.$$

We now recall the Star Estimator (Star), first introduced by Audibert (2007) and later revisited by Liang et al. (2015):

Definition E.1 (Star Estimator). *For a given class $\mathcal{F}$ and function g, define the **star hull***

$$\mathrm{star}(\mathcal{F}, g) := \{\lambda g + (1-\lambda) f : f \in \mathcal{F}, \lambda \in [0,1]\}.$$

The Star Estimator $\hat{f}_N^{\mathrm{Star}}$ *is the following two-step squared-loss estimator:*

$$\hat{f}_N^{\mathrm{Star}} \in \operatorname*{arg\,min}_{f\in\mathrm{star}(\mathcal{F},\hat{g}_N)} \widehat{\mathcal{L}}^{\mathrm{VR}}(f), \quad \textit{where} \quad \hat{g}_N \in \operatorname*{arg\,min}_{f\in\mathcal{F}} \widehat{\mathcal{L}}^{\mathrm{VR}}(f).$$

Although $\mathrm{star}(\mathcal{F}, \hat{g}_N)$ is data-dependent, the output $\hat{f}_N^{\mathrm{Star}}$ belongs to the deterministic function class

$$\mathcal{H}_{\mathrm{star}} := \bigcup_{g\in\mathcal{F}} \mathrm{star}(\mathcal{F}, g) = \{\lambda g + (1-\lambda) f : g, f \in \mathcal{F}, \lambda \in [0,1]\}. \tag{27}$$

Since $\mathcal{H}_{\mathrm{star}} \subseteq \mathrm{conv}(\mathcal{F})$, we immediately have $\sup_{g\in\mathcal{H}_{\mathrm{star}}} \|g\|_\infty \le F$.

We now state the error bound for Star in the finite class, whose proof is presented in Appendix E.2:

Theorem E.1 (Star Estimator Bound for Finite $\mathcal{F}$). *Let $\delta \in (0,1)$, $t_x := \log \frac{2|\mathcal{F}|^2(N_x+1)}{\delta}$, and*

$$\rho_\delta := \left(\frac{1331}{27}F^2 + \frac{260}{3}F R_{\max} + 38 R_{\max}^2\right)\frac{t_x}{N_x} + 38\sigma^2\frac{t_x}{N}.$$

Then, with probability at least $1-\delta$,

$$\mathbb{E}_{x,a}\left[\left(\mathcal{C}\hat{f}_N^{\mathrm{Star}} - \mathcal{C}r^\star\right)^2\right] \le \varepsilon_{\mathrm{aprx}} + \rho_\delta = \varepsilon_{\mathrm{aprx}} + \widetilde{\mathcal{O}}\left(\frac{F^2 + R_{\max}^2}{N_x} + \frac{\sigma^2}{N}\right).$$

Recall the error bound guarantee of VR (Theorem 3.1): $2\varepsilon_{\mathrm{aprx}} + \widetilde{\mathcal{O}}\left(\frac{F^2 + F R_{\max}}{N_x} + \frac{\sigma^2}{N}\right)$. Comparing the two, Star reduces the coefficient of $\varepsilon_{\mathrm{aprx}}$ from two to one, at the price of $R_{\max}^2$ instead of $F R_{\max}$ in the N_x^{-1} term. Order-wise, the two bounds are meaningfully different when $R_{\max} \gg F$ and $R_{\max}^2/N_x \gg \varepsilon_{\mathrm{aprx}}, \sigma^2/N$, up to logarithmic factors, in which case the VR bound is smaller than the Star bound.

E.2 Proof of Theorem E.1: Star Estimator Bound for Finite Class

As done for VR, we set our comparator as $f^\circ \in \arg\min_{f\in\mathcal{F}} \left\{\mathcal{L}^{\mathrm{VR}}(f) := \mathbb{E}\left[(f - r^\star)^2\right]\right\}$, which satisfies $\mathcal{L}^{\mathrm{VR}}(f^\circ) = \varepsilon_{\mathrm{aprx}}$, i.e.,

$$\mathbb{E}_{x,a}\left[(\hat{f}_N^{\mathrm{Star}} - r^\star)^2\right] = \varepsilon_{\mathrm{aprx}} + \underbrace{\mathcal{L}^{\mathrm{VR}}(\hat{f}_N^{\mathrm{Star}}) - \mathcal{L}^{\mathrm{VR}}(f^\circ)}_{(*)}. \tag{28}$$

We again use Proposition 2.1 to tightly upper-bound $(*)$.

For $h \in \mathcal{H}_{\text{star}} - f^\circ$, define as before

$$\Delta(h; f^\circ) := \mathcal{L}^{\text{VR}}(f^\circ + h) - \mathcal{L}^{\text{VR}}(f^\circ),$$

$$\widehat{\Delta}(h; f^\circ) := \widehat{\mathcal{L}}^{\text{VR}}(f^\circ + h) - \widehat{\mathcal{L}}^{\text{VR}}(f^\circ), \qquad \widehat{Q}(h) := \frac{1}{N}\sum_{i,j} h(x_i, a_{i,j})^2.$$

First, unlike VR, Liang et al. (2015, Lemma 1) showed that the STAR ESTIMATOR satisfies condition (i) of Proposition 2.1 with *nonzero* $\lambda = \frac{1}{18}$, *deterministically*:[3]

$$\widehat{\Delta}(\hat{f}_N^{\text{STAR}} - f^\circ; f^\circ) \le -\frac{1}{18}\widehat{Q}(\hat{f}_N^{\text{STAR}} - f^\circ).$$

For the STAR ESTIMATOR, we use the localized block-offset modulus with respect to $\mathcal{H}_{\text{star}}$, not $\mathcal{F}$, whose definition we recall here:

$$\Psi^{\text{STAR}}(r) := \sup_{\substack{h \in \mathcal{H}_{\text{star}} - f^\circ \\ \Delta(h; f^\circ) \le r}} \left[\Delta(h; f^\circ) - \widehat{\Delta}(h; f^\circ) - \frac{1}{18}\widehat{Q}(h)\right]. \tag{29}$$

The following lemma, whose proof is provided in the next section, verifies condition (ii):

Lemma E.1. *Let ρ_δ be as defined in Theorem E.1. With probability at least $1 - \delta$,*

$$\Psi^{\text{STAR}}(r) \le \rho_\delta, \qquad \forall r \ge 0. \tag{30}$$

On the event (30), the fixed-point argument of Proposition 2.1 gives

$$\mathcal{L}^{\text{VR}}(\hat{f}_N^{\text{STAR}}) - \mathcal{L}^{\text{VR}}(f^\circ) \le \sup\{r \ge 0 : r \le \rho_\delta\} = \rho_\delta.$$

Combining Eqn. (28) with the fact that the variance is bounded by the second moment[4] gives,

$$\mathbb{E}_{x,a}\left[\left(\mathcal{C}\hat{f}_N^{\text{STAR}} - \mathcal{C}r^\star\right)^2\right] \le \mathbb{E}_{x,a}\left[\left(\hat{f}_N^{\text{STAR}} - r^\star\right)^2\right] \le \varepsilon_{\text{aprx}} + \rho_\delta.$$

This concludes the proof. □

E.3 PROOF OF LEMMA E.1: LOCALIZED MODULUS FOR THE STAR ESTIMATOR IN FINITE CLASS

To derive a uniform high-probability bound over $\mathcal{H}_{\text{star}}$, we first construct a finite net of $\mathcal{H}_{\text{star}}$. For each element of the net, we apply Bernstein's inequality to control the loss difference and the squared distance from f°. A union bound makes these bounds hold simultaneously over the net. We then extend them to every $s \in \mathcal{H}_{\text{star}}$ using the approximation property of the net. Finally, combining these bounds yields a uniform bound on the localized modulus.

Step 1: construct a finite net over $\mathcal{H}_{\text{star}}$. We construct a finite net of the deterministic class $\mathcal{H}_{\text{star}}$ in Eqn. (27). Let

$$\mathcal{H}_{\text{star},N_x} := \left\{(1-\alpha)f + \alpha g : f, g \in \mathcal{F},\ \alpha \in \left\{0, \frac{1}{N_x}, \dots, 1\right\}\right\}.$$

Then $|\mathcal{H}_{\text{star},N_x}| \le |\mathcal{F}|^2(N_x+1)$, and for any given $s \in \mathcal{H}_{\text{star}}$, there exist $\alpha_s \in [0,1]$ and $f_1, f_2 \in \mathcal{F}$ such that $s = (1-\alpha_s)f_1 + \alpha_s f_2$.

Let $\bar{\alpha} := \lfloor N_x \alpha_s \rfloor / N_x$ and $h = (1-\bar{\alpha})f_1 + \bar{\alpha} f_2$. Then since $|\alpha_s - \bar{\alpha}| \le 1/N_x$,

$$\begin{aligned}\|s - h\|_\infty &= \|(\bar{\alpha} - \alpha_s)f_1 + (\alpha_s - \bar{\alpha})f_2\|_\infty \\ &\le \|(\bar{\alpha} - \alpha_s)f_1\|_\infty + \|(\alpha_s - \bar{\alpha})f_2\|_\infty \le \frac{2F}{N_x}.\end{aligned}$$

[3] Precisely, $\mathcal{F}$ in Proposition 2.1 should change to $\mathcal{H}_{\text{star}}$ for this analysis.

[4] $\mathbb{E}[(\mathcal{C}g(x,a))^2] = \mathbb{E}_x[\text{Var}[g(x,a) \mid x]] \le \mathbb{E}[g(x,a)^2]$

Therefore, for a given $s \in \mathcal{H}_{\text{star}}$, there exists $h \in \mathcal{H}_{\text{star},N_x}$ such that

$$\|s - h\|_\infty \le \frac{2F}{N_x}. \tag{31}$$

Step 2: concentration bound via Bernstein's inequality. Fix $h \in \mathcal{H}_{\text{star},N_x}$ and denote

$$d_h := h - f^\circ, \qquad D_h := \mathbb{E}_{x,a}[d_h(x,a)^2], \qquad \widehat{D}_h := \frac{1}{N}\sum_{i,j} d_h(x_i,a_{i,j})^2,$$

as well as

$$\ell_h(x,a) := (h(x,a) - r^\star(x,a))^2 - (f^\circ(x,a) - r^\star(x,a))^2,$$
$$\Delta_h := \mathcal{L}^{\text{VR}}(h) - \mathcal{L}^{\text{VR}}(f^\circ), \qquad \widehat{\Delta}_h := \widehat{\mathcal{L}}^{\text{VR}}(h) - \widehat{\mathcal{L}}^{\text{VR}}(f^\circ).$$

We first obtain uniform one-sided concentration bounds for $\Delta_h - \widehat{\Delta}_h$ and $D_h - \widehat{D}_h$. For each within-block observation, define the loss difference and the corresponding block average by

$$Z_{i,j}(h) := \big(h(x_i,a_{i,j}) - r_{i,j}\big)^2 - \big(f^\circ(x_i,a_{i,j}) - r_{i,j}\big)^2, \quad Z_i(h) := \frac{1}{N_a}\sum_{j=1}^{N_a} Z_{i,j}(h).$$

Then the variables $\{Z_i(h)\}_{i=1}^{N_x}$ are independent, with $\mathbb{E}[Z_i(h)] = \Delta_h$ and $\frac{1}{N_x}\sum_i Z_i(h) = \widehat{\Delta}_h$. Since every function in $\mathcal{H}_{\text{star}}$ and f° is bounded by F and $|r_{i,j}| \le R_{\max}$,

$$|Z_{i,j}(h)| \le F(F + 4R_{\max}). \tag{32}$$

Moreover, $\ell_h(x,a) = d_h(x,a)\big(h(x,a) + f^\circ(x,a) - 2r^\star(x,a)\big)$, so

$$\ell_h(x,a)^2 \le 4(F + R_{\max})^2 d_h(x,a)^2. \tag{33}$$

Since $Z_{i,j}(h) = \ell_h(x_i,a_{i,j}) - 2d_h(x_i,a_{i,j})\eta_{i,j}$, the conditional mean-zero property of the reward noise gives

$$\mathbb{E}[Z_{i,j}(h)^2] \le 4\big((F + R_{\max})^2 + \sigma^2\big)D_h.$$

Let $m_h(x) := \mathbb{E}_{a\sim\pi_{\text{ref}}(\cdot|x)}[\ell_h(x,a)]$. By Jensen's inequality and Eqn. (33),

$$\begin{aligned}\text{Var}(m_h(X)) &\le \mathbb{E}_X[m_h(X)^2] = \mathbb{E}_X[(\mathbb{E}_a[\ell_h(X,a)])^2] \\ &\le \mathbb{E}_X[\mathbb{E}_a[\ell_h(X,a)^2]] \le 4(F + R_{\max})^2\mathbb{E}_{X,a}[d_h(X,a)^2] \\ &\le 4(F + R_{\max})^2 D_h\end{aligned}$$

Conditional on x_i, the variables $\{Z_{i,j}(h)\}_{j=1}^{N_a}$ are independent with common conditional mean $m_h(x_i)$. Therefore, the law of total variance gives

$$\begin{aligned}\text{Var}(Z_{1,1}(h)) &= \text{Var}(\mathbb{E}[Z_{1,1}(h)|X]) + \mathbb{E}[\text{Var}(Z_{1,1}(h)|X)] \\ &= \text{Var}(m_h(X)) + \mathbb{E}[\text{Var}(Z_{1,1}(h)|X)]\end{aligned} \tag{34}$$

Applying the law of total variance again and using Eqn. (34),

$$\begin{aligned}\text{Var}(Z_i(h)) &= \text{Var}(\mathbb{E}[Z_i(h)|X]) + \mathbb{E}[\text{Var}(Z_i(h)|X)] \\ &= \text{Var}(m_h(X)) + \frac{1}{N_a}\mathbb{E}\left[\text{Var}(Z_{1,1}(h)\big|X)\right] \\ &= \left(1 - \frac{1}{N_a}\right)\text{Var}(m_h(X)) + \frac{1}{N_a}\text{Var}(Z_{1,1}(h)) \\ &\le 4\left((F + R_{\max})^2 + \frac{\sigma^2}{N_a}\right)D_h.\end{aligned}$$

Define

$$a := t_x\left(\frac{(F + R_{\max})^2}{N_x} + \frac{\sigma^2}{N}\right), \qquad b := \frac{t_x}{N_x}.$$

By one-sided Bernstein's inequality, Eqn. (32), and a union bound over $h \in \mathcal{H}_{\text{star},N_x}$, with probability at least $1 - \delta/2$, simultaneously for all $h \in \mathcal{H}_{\text{star},N_x}$,

$$\Delta_h - \widehat{\Delta}_h \le \sqrt{8aD_h} + \frac{2F(F + 4R_{\max})b}{3}. \tag{35}$$

We also require a one-sided comparison between the population and empirical squared distances. For each block, let

$$V_i(h) := \frac{1}{N_a}\sum_{j=1}^{N_a} d_h(x_i, a_{i,j})^2.$$

Since $|d_h| \le 2F$, $0 \le V_i(h) \le 4F^2$, $\mathbb{E}[V_i(h)] = D_h$, and $\mathrm{Var}(V_i(h)) \le 4F^2 D_h$.

A second one-sided Bernstein inequality and union bound therefore give, with probability at least $1 - \delta/2$, simultaneously for all $h \in \mathcal{H}_{\mathrm{star},N_x}$,

$$D_h - \widehat{D}_h \le \sqrt{8F^2 b D_h} + \frac{4F^2 b}{3}. \tag{36}$$

Step 3: upper bound on the localized modulus. We then work on the intersection of Eqn. (35) and Eqn. (36), which has probability at least $1 - \delta$.

Let $s \in \mathcal{H}_{\mathrm{star}}$ be arbitrary and choose $h \in \mathcal{H}_{\mathrm{star},N_x}$, whose existence is proved in Eqn. (31). Since s, h, f° are all bounded by F, while $|r^\star|, |r_{i,j}| \le R_{\max}$, we have:

$$\begin{aligned}
|\Delta(s - f^\circ; f^\circ) - \Delta_h| &\le \frac{4F(F + R_{\max})}{N_x}, \\
\left|\widehat{\Delta}(s - f^\circ; f^\circ) - \widehat{\Delta}_h\right| &\le \frac{4F(F + R_{\max})}{N_x}, \\
\left|\widehat{Q}(s - f^\circ) - \widehat{D}_h\right| &\le \frac{8F^2}{N_x}.
\end{aligned} \tag{37}$$

Then, with $c_0 := \frac{1}{18}$, combining Eqn. (35)– (37) gives the following: for every $s \in \mathcal{H}_{\mathrm{star}}$,

$$\begin{aligned}
&\Delta(s - f^\circ; f^\circ) - \widehat{\Delta}(s - f^\circ; f^\circ) - c_0\widehat{Q}(s - f^\circ) \\
&\le -c_0 D_h + \left(\sqrt{8a} + c_0\sqrt{8F^2 b}\right)\sqrt{D_h} + \frac{4c_0 F^2 b}{3} + \frac{2F(F + 4R_{\max})b}{3} \\
&\quad + \frac{8F(F + R_{\max}) + 8c_0 F^2}{N_x}.
\end{aligned} \tag{38}$$

The right-hand side is now independent of the population excess loss of s. Maximizing its first two terms over $\sqrt{D_h} \ge 0$ and using $c_0 = 1/18$ gives

$$\sup_{D_h \ge 0}\left\{-c_0 D_h + \left(\sqrt{8a} + c_0\sqrt{8F^2 b}\right)\sqrt{D_h}\right\} = 36a + \frac{1}{9}F^2 b + 4F\sqrt{ab}.$$

Using $4F\sqrt{ab} \le 2a + 2F^2 b$ and $t_x > 1$, we obtain from Eqn. (38)

$$\begin{aligned}
&\Delta(s - f^\circ; f^\circ) - \widehat{\Delta}(s - f^\circ; f^\circ) - c_0\widehat{Q}(s - f^\circ) \\
&\le 38 t_x\left(\frac{(F + R_{\max})^2}{N_x} + \frac{\sigma^2}{N}\right) + \left(\frac{59}{27}F^2 + \frac{2F(F + 4R_{\max})}{3}\right)\frac{t_x}{N_x} \\
&\quad + \left(\frac{4}{9}F^2 + 8F(F + R_{\max})\right)\frac{t_x}{N_x}.
\end{aligned}$$

Expanding $(F + R_{\max})^2$ and collecting terms yields exactly

$$\begin{aligned}
&\Delta(s - f^\circ; f^\circ) - \widehat{\Delta}(s - f^\circ; f^\circ) - \frac{1}{18}\widehat{Q}(s - f^\circ) \\
&\le \underbrace{\left(\frac{1331}{27}F^2 + \frac{260}{3}F R_{\max} + 38{R_{\max}}^2\right)\frac{t_x}{N_x} + 38\sigma^2\frac{t_x}{N}}_{=\rho_\delta}.
\end{aligned} \tag{39}$$

Since Eqn. (39) holds for every $s \in \mathcal{H}_{\mathrm{star}}$, restricting to $\Delta(s - f^\circ; f^\circ) \le r$ and taking the supremum in Eqn. (29) gives

$$\Psi^{\mathrm{STAR}}(r) \le \rho_\delta, \qquad \forall r \ge 0.$$

This proves the lemma. □

F Deferred Proofs from Section 4: Linear Class

F.1 Value Regression (VR)

F.1.1 Proof of Theorem 4.1: VR Bound for Linear $\mathcal{F}_{\text{lin}}$

Write $f^\circ_{\text{VR}}(x,a) = f^\circ(x,a) = \langle \boldsymbol{\phi}(x,a), \boldsymbol{\theta}^\circ \rangle$ for the population least-squares projection defined in Assumption 2, where

$$\boldsymbol{\theta}^\circ \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathbb{E}_{x,a}\left[\left(\langle \boldsymbol{\phi}(x,a), \boldsymbol{\theta}\rangle - r^\star(x,a)\right)^2\right].$$

Throughout this subsection, $\mathcal{L}^{\text{VR}}$ and $\widehat{\mathcal{L}}^{\text{VR}}$ denote the generic losses in Section 2 instantiated with $\mathcal{T} = \mathcal{T}^{\text{VR}}$. For $h = f - f^\circ_{\text{VR}}$, the corresponding population and empirical excess losses are

$$\begin{aligned}\Delta(h; f^\circ_{\text{VR}}) &:= \mathcal{L}^{\text{VR}}(f^\circ_{\text{VR}} + h) - \mathcal{L}^{\text{VR}}(f^\circ_{\text{VR}}) \\ &= \mathbb{E}_{x,a}\left[(f^\circ_{\text{VR}} + h - r^\star)^2 - (f^\circ_{\text{VR}} - r^\star)^2\right],\end{aligned}$$

and

$$\begin{aligned}\widehat{\Delta}(h; f^\circ_{\text{VR}}) &:= \widehat{\mathcal{L}}^{\text{VR}}(f^\circ_{\text{VR}} + h) - \widehat{\mathcal{L}}^{\text{VR}}(f^\circ_{\text{VR}}) \\ &= \frac{1}{N_x N_a}\sum_{i=1}^{N_x}\sum_{j=1}^{N_a}\left[\left(f^\circ_{\text{VR}}(x_i,a_{i,j}) + h(x_i,a_{i,j}) - r_{i,j}\right)^2 - \left(f^\circ_{\text{VR}}(x_i,a_{i,j}) - r_{i,j}\right)^2\right].\end{aligned}$$

The empirical quadratic term in the localized modulus is

$$\widehat{Q}(h) = \widehat{Q}^{\text{VR}}(h) := \frac{1}{N_x N_a}\sum_{i=1}^{N_x}\sum_{j=1}^{N_a} h(x_i,a_{i,j})^2.$$

Thus, for this proof,

$$\Psi(r; 1, f^\circ_{\text{VR}}) = \sup_{\substack{h\in\mathcal{F}_{\text{lin}} - f^\circ_{\text{VR}} \\ \Delta(h; f^\circ_{\text{VR}})\le r}}\left[\Delta(h; f^\circ_{\text{VR}}) - \widehat{\Delta}(h; f^\circ_{\text{VR}}) - \widehat{Q}^{\text{VR}}(h)\right].$$

For the proof, let us denote $\varepsilon_{\text{blk}} := \varepsilon_{\text{x}} + \frac{\varepsilon_{\text{a}}}{N_a}$, $t_\delta := \log\frac{3}{\delta}$, and

$$\Gamma_\delta := 3\sqrt{\frac{d\,\varepsilon_{\text{blk}}}{N_x}}\left(1+\sqrt{8t_\delta}\right) + \frac{4\sqrt{\varepsilon_{\text{z}} d}\,t_\delta}{N_x} + \sqrt{\frac{3\sigma^2 d}{N}}\left(1+\sqrt{8t_\delta}\right) + \frac{8R_{\max}\alpha_0\sqrt{d}\,t_\delta}{N}. \tag{40}$$

We now state the following lemma, whose proof is deferred to Appendix F.1.2, that bounds the localized modulus of VR in $\mathcal{F}_{\text{lin}}$:

Lemma F.1 (Localized modulus for linear VR). *Suppose the assumptions of Theorem 4.1 hold. Then, with probability at least $1-\delta$,*

$$\Psi(r; 1, f^\circ_{\text{VR}}) \le \frac{2}{3}r + \frac{1}{3}\Gamma_\delta^2, \qquad \forall r \ge 0.$$

The proof of Theorem 4.1 then proceeds using the same recipe.

Let

$$\hat{\boldsymbol{\theta}}^{\text{VR}}_N \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{N_x}\sum_{i=1}^{N_x}\frac{1}{N_a}\sum_{j=1}^{N_a}\left(\langle \boldsymbol{\theta}, \boldsymbol{\phi}(x_i,a_{i,j})\rangle - r_{i,j}\right)^2.$$

Let $\widehat{h}_{\text{VR}} := \hat{f}^{\text{VR}}_N - f^\circ_{\text{VR}}$. The OLS normal equation gives

$$\widehat{\Delta}(\widehat{h}_{\text{VR}}; f^\circ_{\text{VR}}) = \widehat{\mathcal{L}}^{\text{VR}}(\hat{f}^{\text{VR}}_N) - \widehat{\mathcal{L}}^{\text{VR}}(f^\circ_{\text{VR}}) = -\widehat{Q}^{\text{VR}}(\widehat{h}_{\text{VR}}).$$

Thus condition (i) of Proposition 2.1 holds with $\lambda = 1$. By Lemma F.1, condition (ii) holds on an event of probability at least $1-\delta$. On this event,

$$\mathcal{L}^{\text{VR}}(\hat{f}^{\text{VR}}_N) - \mathcal{L}^{\text{VR}}(f^\circ_{\text{VR}}) \le \sup\left\{r \ge 0 : r \le \frac{2}{3}r + \frac{1}{3}\Gamma_\delta^2\right\} = \Gamma_\delta^2.$$

Since $\mathcal{L}^{\mathrm{VR}}(f^\circ_{\mathrm{VR}}) = \varepsilon_{\mathrm{aprx}}$ and we have $\mathbb{E}[(\mathcal{C}g(x,a))^2] = \mathbb{E}_x[\mathrm{Var}[g(x,a) \mid x]] \le \mathbb{E}[g(x,a)^2]$, the original argument gives

$$\mathbb{E}_{x,a}\left[\left(\mathcal{C}\hat{f}^{\mathrm{VR}}_N - \mathcal{C}r^\star\right)^2\right] \le \mathbb{E}_{x,a}\left[\left(\hat{f}^{\mathrm{VR}}_N - r^\star\right)^2\right] \le \varepsilon_{\mathrm{aprx}} + \Gamma^2_\delta. \tag{41}$$

Finally, since $\varepsilon_{\mathrm{blk}} = \varepsilon_{\mathrm{x}} + \varepsilon_{\mathrm{a}}/N_a$ and $N = N_x N_a$, we have that

$$\Gamma^2_\delta = \widetilde{\mathcal{O}}\left(\frac{\varepsilon_{\mathrm{x}} d}{N_x} + \frac{\varepsilon_{\mathrm{z}} d}{N_x^2} + \frac{(\varepsilon_{\mathrm{a}} + \sigma^2)d}{N} + \frac{R^2_{\max}\alpha_0^2 d}{N^2}\right). \tag{42}$$

Combining Eqns. (41)–(42) proves Theorem 4.1. □

F.1.2 Proof of Lemma F.1: Localized Modulus for Linear VR

Write

$$\boldsymbol{q}(x,a) \coloneqq e^\circ(x,a)\boldsymbol{z}(x,a), \qquad \boldsymbol{m}(x) \coloneqq \mathbb{E}_{A\sim\pi_{\mathrm{ref}}(\cdot|x)}[\boldsymbol{q}(x,A)].$$

Since $\boldsymbol{\theta}^\circ$ is the population least-squares projection over $\mathbb{R}^d$, its normal equation gives

$$\mathbb{E}_{x,a}[\boldsymbol{\phi}(x,a)e^\circ(x,a)] = \mathbf{0}, \qquad \text{and hence,} \qquad \mathbb{E}_X[\boldsymbol{m}(X)] = \mathbf{0}. \tag{43}$$

For the observed dataset, we define the following random quantities: for each $i \in [N_x]$,

$$\boldsymbol{G}_i \coloneqq \frac{1}{N_a}\sum_{j=1}^{N_a} \boldsymbol{z}(x_i,a_{i,j})\boldsymbol{z}(x_i,a_{i,j})^\top, \qquad \boldsymbol{q}_i \coloneqq \frac{1}{N_a}\sum_{j=1}^{N_a}\boldsymbol{q}(x_i,a_{i,j}),$$

and "score vectors" averaged over $i \in [N_x]$:

$$\boldsymbol{H} \coloneqq \frac{1}{N_x}\sum_{i=1}^{N_x}\boldsymbol{G}_i, \qquad \boldsymbol{t}_{\mathrm{mis}} \coloneqq \frac{1}{N_x}\sum_{i=1}^{N_x}\boldsymbol{q}_i, \qquad \boldsymbol{t}_\eta \coloneqq \frac{1}{N}\sum_{i=1}^{N_x}\sum_{j=1}^{N_a}\boldsymbol{z}(x_i,a_{i,j})\eta_{i,j}. \tag{44}$$

We now express Δ, $\widehat{\Delta}$, and $\widehat{Q}^{\mathrm{VR}}$ in the whitened coordinates. Fix $h = f - f^\circ_{\mathrm{VR}} \in \mathcal{F}_{\mathrm{lin}} - f^\circ_{\mathrm{VR}}$, and write

$$\boldsymbol{w} \coloneqq \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ), \qquad h(x,a) = \boldsymbol{z}(x,a)^\top \boldsymbol{w}.$$

The population normal equation gives

$$\Delta(h; f^\circ_{\mathrm{VR}}) = \mathbb{E}_{x,a}[h(x,a)^2] = \|\boldsymbol{w}\|_2^2.$$

Moreover, $\widehat{Q}^{\mathrm{VR}}(h) = \boldsymbol{w}^\top \boldsymbol{H}\boldsymbol{w}$, and direct expansion of the empirical excess loss gives

$$\widehat{\Delta}(h; f^\circ_{\mathrm{VR}}) = \boldsymbol{w}^\top \boldsymbol{H}\boldsymbol{w} + 2\boldsymbol{w}^\top \boldsymbol{t}_{\mathrm{mis}} - 2\boldsymbol{w}^\top \boldsymbol{t}_\eta.$$

Consequently, the process to be controlled is

$$\Delta(h; f^\circ_{\mathrm{VR}}) - \widehat{\Delta}(h; f^\circ_{\mathrm{VR}}) - \widehat{Q}^{\mathrm{VR}}(h) = \|\boldsymbol{w}\|_2^2 - 2\boldsymbol{w}^\top \boldsymbol{H}\boldsymbol{w} + 2\boldsymbol{w}^\top(\boldsymbol{t}_\eta - \boldsymbol{t}_{\mathrm{mis}}). \tag{45}$$

It is therefore enough to show, uniformly over h, that the right-hand side of Eqn. (45) is at most $2\Delta(h; f^\circ_{\mathrm{VR}})/3 + \Gamma^2_\delta/3$; restricting this bound to $\Delta(h; f^\circ_{\mathrm{VR}}) \le r$ concludes the proof.

We now prove the uniform bound through a lower-isometry event for $\boldsymbol{H}$ and concentration of the two score vectors, $\boldsymbol{t}_{\mathrm{mis}}$ and $\boldsymbol{t}_\eta$. We use the vector Bernstein inequality of Hsu et al. (2012a, Proposition 1.2): if independent[5] mean-zero random vectors $\boldsymbol{y}_k$ satisfy $\sum_k \mathbb{E}\|\boldsymbol{y}_k\|_2^2 \le v$ and $\|\boldsymbol{y}_k\|_2 \le L$ almost surely, then, with probability at least $1 - e^{-t}$,

$$\left\|\sum_k \boldsymbol{y}_k\right\|_2 \le \sqrt{v}\left(1 + \sqrt{8t}\right) + \frac{4}{3}Lt. \tag{46}$$

This lets us avoid a separate covering argument.

Recall from Section 4.1 that $\boldsymbol{G}_{N_a} = \frac{1}{N_a}\sum_{j=1}^{N_a}\boldsymbol{z}(X,A_j)\boldsymbol{z}(X,A_j)^\top$ where $\boldsymbol{z}(x,a) = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\phi}(x,a)$ and $\boldsymbol{\Sigma} = \mathbb{E}_{x,a}[\boldsymbol{\phi}(x,a)\boldsymbol{\phi}(x,a)^\top]$.

[5]The original statement holds for martingale difference vector sequences; for our purpose, it suffices to apply it to independent vector sequences, as done in Hsu et al. (2014).

Step 1: lower isometry of the empirical Gram matrix. The matrices $\boldsymbol{G}_i$ are i.i.d. copies of $\boldsymbol{G}_{N_a}$ and satisfy $\mathbb{E}[\boldsymbol{G}_i] = \boldsymbol{I}_d$. Moreover, $\boldsymbol{G}_i \succeq \boldsymbol{0}$, and hence $\lambda_{\max}(\boldsymbol{I}_d - \boldsymbol{G}_i) \le 1$.

The one-sided matrix Bernstein inequality (Tropp, 2015, Theorem 6.6.1) therefore gives, with probability at least $1 - \delta/3$,

$$\lambda_{\max}(\boldsymbol{I}_d - \boldsymbol{H}) \le \sqrt{\frac{2\left\|\mathrm{Cov}(\boldsymbol{G}_{N_a})\right\|_{\mathrm{op}} \log(3d/\delta)}{N_x}} + \frac{2\log(3d/\delta)}{3N_x} \le \frac{2}{3},$$

where the last step uses the sample-size condition from Theorem 4.1. Thus, we have[6]

$$\boldsymbol{H} \succeq \frac{1}{3}\boldsymbol{I}_d. \tag{47}$$

Step 2: control of the misspecification score. The blocks are independent, and Eqn. (43) implies $\mathbb{E}[\boldsymbol{q}_i] = \boldsymbol{0}$. Moreover, by conditioning on the context,

$$\mathbb{E}\left[\|\boldsymbol{q}_i\|_2^2\right] = d\left(\varepsilon_{\mathrm{x}} + \frac{\varepsilon_{\mathrm{a}}}{N_a}\right) = d\varepsilon_{\mathrm{blk}}. \tag{48}$$

Assumption 2 gives $\|\boldsymbol{q}_i\|_2 \le \sqrt{\varepsilon_{\mathrm{z}} d}$. Applying Eqn. (46) to the independent block averages and using Eqn. (48), we obtain, with probability at least $1 - \delta/3$,

$$\|\boldsymbol{t}_{\mathrm{mis}}\|_2 \le \underbrace{\sqrt{\frac{d\varepsilon_{\mathrm{blk}}}{N_x}}\left(1 + \sqrt{8t_\delta}\right) + \frac{4\sqrt{\varepsilon_{\mathrm{z}} d}\, t_\delta}{3N_x}}_{=:B_{\mathrm{mis}}}. \tag{49}$$

Step 3: control of the reward-noise score. On the event in Eqn. (47), condition on the complete context–action design $\{(x_i, a_{i,j})\}_{i,j}$. The vectors

$$\boldsymbol{y}_{i,j} := \frac{1}{N}\boldsymbol{H}^{-1/2}\boldsymbol{z}(x_i, a_{i,j})\eta_{i,j}$$

are conditionally independent and mean zero. Since $\sum_{i,j} \boldsymbol{z}_{i,j}\boldsymbol{z}_{i,j}^\top = N\boldsymbol{H}$,

$$\sum_{i,j} \mathbb{E}\left[\left\|\boldsymbol{y}_{i,j}\right\|_2^2 \,\middle|\, \{(x_i, a_{i,j})\}_{i,j}\right] \le \frac{\sigma^2}{N^2}\,\mathrm{tr}\left(\boldsymbol{H}^{-1}\sum_{i,j}\boldsymbol{z}_{i,j}\boldsymbol{z}_{i,j}^\top\right) = \frac{\sigma^2 d}{N},$$

where we denote $\boldsymbol{z}_{i,j} := \boldsymbol{z}(x_i, a_{i,j})$. Also, using $|\eta_{i,j}| \le 2R_{\max}$, Assumption 2, and Eqn. (47),

$$\left\|\boldsymbol{y}_{i,j}\right\|_2 \le \frac{2R_{\max}\alpha_0\sqrt{3d}}{N}.$$

A conditional application of Eqn. (46) therefore yields, with conditional probability at least $1 - \delta/3$,

$$\left\|\boldsymbol{t}_\eta\right\|_{\boldsymbol{H}^{-1}} \le \underbrace{\sigma\sqrt{\frac{d}{N}}\left(1 + \sqrt{8t_\delta}\right) + \frac{8R_{\max}\alpha_0\sqrt{3d}\, t_\delta}{3N}}_{=:B_\eta}.$$

Here $\|\boldsymbol{v}\|_{\boldsymbol{H}^{-1}} := \sqrt{\boldsymbol{v}^\top\boldsymbol{H}^{-1}\boldsymbol{v}}$. The conditional failure probability remains at most $\delta/3$ after averaging over the design. Hence, by a union bound, the three events above hold simultaneously with probability at least $1 - \delta$.

Step 4: combining the three events. Let $D := \left\|\boldsymbol{t}_\eta - \boldsymbol{t}_{\mathrm{mis}}\right\|_{\boldsymbol{H}^{-1}}$. Cauchy–Schwarz and AM-GM inequality gives

$$2\boldsymbol{w}^\top(\boldsymbol{t}_\eta - \boldsymbol{t}_{\mathrm{mis}}) \le 2\|\boldsymbol{w}\|_{\boldsymbol{H}}\left\|\boldsymbol{t}_\eta - \boldsymbol{t}_{\mathrm{mis}}\right\|_{\boldsymbol{H}^{-1}} \le \|\boldsymbol{w}\|_{\boldsymbol{H}}^2 + D^2.$$

Together with Eqn. (45) and (47), this yields

$$\begin{aligned}\Delta(h; f^\circ_{\mathrm{VR}}) - \widehat{\Delta}(h; f^\circ_{\mathrm{VR}}) - \widehat{Q}^{\mathrm{VR}}(h) &\le \|\boldsymbol{w}\|_2^2 - \|\boldsymbol{w}\|_{\boldsymbol{H}}^2 + D^2 \\ &\le \frac{2}{3}\Delta(h; f^\circ_{\mathrm{VR}}) + D^2.\end{aligned} \tag{50}$$

[6] This is precisely the lower isometry condition of Liang et al. (2015, Definition 5).

On the same event,

$$D \le \sqrt{3}\,\|\boldsymbol{t}_{\mathrm{mis}}\|_2 + \|\boldsymbol{t}_\eta\|_{\boldsymbol{H}^{-1}} \le \sqrt{3}B_{\mathrm{mis}} + B_\eta = \frac{\Gamma_\delta}{\sqrt{3}},$$

where the last equality follows from Eqn. (40). Thus $D^2 \le \Gamma_\delta^2/3$. Restricting Eqn. (50) to $\Delta(h; f^\circ_{\mathrm{VR}}) \le r$ and taking the supremum in the definition of the localized block-offset modulus gives

$$\Psi(r; 1, f^\circ_{\mathrm{VR}}) \le \frac{2}{3}r + \frac{1}{3}\Gamma_\delta^2, \qquad \forall r \ge 0.$$

This proves the claim. □

F.1.3 Alternate Error Bound for Linear VR

Here, we present an alternate error bound for VR in $\mathcal{F}_{\mathrm{lin}}$. To do so, we introduce additional notation regarding centered feature geometry.

With $\overline{\boldsymbol{\phi}}(x,a) := \boldsymbol{\phi}(x,a) - \mathbb{E}_{a\sim\pi_{\mathrm{ref}}(\cdot|x)}[\boldsymbol{\phi}(x,a)]$, define the following quantities (some of which appear in Section 4.2):

$$\boldsymbol{\Sigma}_{\mathrm{C}} := \mathbb{E}_{x,a}\left[\overline{\boldsymbol{\phi}}(x,a)\overline{\boldsymbol{\phi}}(x,a)^\top\right], \qquad \boldsymbol{M}_{\mathrm{C}} := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_{\mathrm{C}}\boldsymbol{\Sigma}^{-1/2},$$
$$\kappa_{\mathrm{C}} := \|\boldsymbol{M}_{\mathrm{C}}\|_{\mathrm{op}}, \qquad d_{\mathrm{eff}} := \mathrm{tr}(\boldsymbol{M}_{\mathrm{C}}).$$

Then we have that $\mathbf{0} \preceq \boldsymbol{M}_{\mathrm{C}} \preceq \boldsymbol{I}_d$, and hence, $0 \le \kappa_{\mathrm{C}} \le 1$ and $0 \le d_{\mathrm{eff}} \le d$. Also define the centered bias

$$\varepsilon_{\mathrm{VR},C} := \mathbb{E}_{x,a}\left[\left(\mathcal{C}f^\circ_{\mathrm{VR}} - \mathcal{C}r^\star\right)^2\right] = \mathbb{E}_X\left[\mathrm{Var}_A\left(e^\circ(X,A) \mid X\right)\right] \le \varepsilon_{\mathrm{aprx}}. \tag{51}$$

For notational convenience, let

$$\Gamma_{\delta,\mathrm{C}} := 3\sqrt{\frac{\kappa_{\mathrm{C}}\,\varepsilon_{\mathrm{blk}}\,d}{N_x}}\left(1+\sqrt{8t_\delta}\right) + \frac{4\sqrt{\kappa_{\mathrm{C}}\varepsilon_{\mathrm{z}}d}\,t_\delta}{N_x} + \sqrt{\frac{3\sigma^2 d_{\mathrm{eff}}}{N}}\left(1+\sqrt{8t_\delta}\right) + \frac{8R_{\max}\alpha_0\sqrt{\kappa_{\mathrm{C}}d}\,t_\delta}{N} \tag{52}$$

and

$$\mathfrak{R}^{\mathrm{VR}}_{N,\mathrm{C}} := \frac{\kappa_{\mathrm{C}}\varepsilon_{\mathrm{x}}d}{N_x} + \frac{\kappa_{\mathrm{C}}\varepsilon_{\mathrm{z}}d}{N_x^2} + \frac{\kappa_{\mathrm{C}}\varepsilon_{\mathrm{a}}d + \sigma^2 d_{\mathrm{eff}}}{N} + \frac{R^2_{\max}\alpha_0^2\kappa_{\mathrm{C}}d}{N^2}. \tag{53}$$

We now state the alternate error bound:

Theorem F.1 (Alternate Error Bound for Linear VR). *Suppose that the assumptions of Theorem 4.1 hold. Then, with probability at least* $1-\delta$,

$$\mathbb{E}_{x,a}\left[\left(\mathcal{C}\hat{f}^{\mathrm{VR}}_N - \mathcal{C}r^\star\right)^2\right] \le \left(\sqrt{\varepsilon_{\mathrm{VR},C}} + \Gamma_{\delta,\mathrm{C}}\right)^2$$
$$\le \varepsilon_{\mathrm{VR},C} + \widetilde{\mathcal{O}}\left(\sqrt{\varepsilon_{\mathrm{VR},C}\,\mathfrak{R}^{\mathrm{VR}}_{N,\mathrm{C}}} + \mathfrak{R}^{\mathrm{VR}}_{N,\mathrm{C}}\right).$$

Discussion. We compare the above with the original error bound of VR (Theorem 4.1) and the error bound of VDR (Theorem 4.2). First, in the well-specified case ($\varepsilon_{\mathrm{aprx}} = 0$), the alternate error bound of VR is $\widetilde{\mathcal{O}}\left(\frac{\sigma^2 d_{\mathrm{eff}}}{N} + \frac{R^2_{\max}\alpha_0^2\kappa_{\mathrm{C}}d}{N^2}\right)$, while the original error bound is $\widetilde{\mathcal{O}}\left(\frac{\sigma^2 d}{N} + \frac{R^2_{\max}\alpha_0^2 d}{N^2}\right)$ and the error bound of VDR is $\widetilde{\mathcal{O}}\left(\frac{\sigma^2 d_{\mathrm{C}}}{N} + \frac{R^2_{\max}\alpha_{\mathrm{C}}^2 d_{\mathrm{C}}}{N^2}\right)$. As $d_{\mathrm{eff}} \le d$ and $\kappa_{\mathrm{C}} \le 1$, the alternate error bound is tighter than the original error bound, and the gap can be arbitrarily large. For example, let $X, A \in \{-1,+1\}^d$ be independent Rademacher vectors and define $\boldsymbol{\phi}(X,A) = \sqrt{1-\tau}\,X + \sqrt{\tau}\,A$ for some $\tau \in (0,1]$. Then $\boldsymbol{\Sigma} = \boldsymbol{I}_d$ and $\boldsymbol{\Sigma}_{\mathrm{C}} = \tau\boldsymbol{I}_d$, so that $d_{\mathrm{eff}} = \tau d$ and $\kappa_{\mathrm{C}} = \tau$, while $\alpha_0 = \mathcal{O}(1)$. Thus, the alternate error bound improves both terms of the original error bound by a factor of τ.

Even compared with VDR, the leading N^{-1} term of the alternate error bound is no larger than the leading N^{-1} term of the VDR bound, since $d_{\mathrm{eff}} \le d_{\mathrm{C}}$.[7] This comparison is strict whenever

[7] Indeed, $0 \preceq \boldsymbol{M}_{\mathrm{C}} \preceq \boldsymbol{I}_d$ and $\boldsymbol{\Sigma} \succ \mathbf{0}$, and hence $d_{\mathrm{eff}} = \mathrm{tr}(\boldsymbol{M}_{\mathrm{C}}) \le \mathrm{rank}(\boldsymbol{M}_{\mathrm{C}}) = \mathrm{rank}(\boldsymbol{\Sigma}_{\mathrm{C}}) = d_{\mathrm{C}}$. Moreover, $\sigma^2 d_{\mathrm{C}}/(N_x(N_a-1)) = \frac{N_a}{N_a-1}\sigma^2 d_{\mathrm{C}}/N$, where $N_a/(N_a-1) \le 2$ for $N_a \ge 2$.

$d_{\mathrm{eff}} < d_{\mathrm{C}}$. On the other hand, the lower-order N^{-2} terms of the alternate VR and VDR bounds are not uniformly ordered, since α_0 and α_{C} are not comparable in general.

There may also be cases in which $0 = \varepsilon_{\mathrm{VR},C} < \varepsilon_{\mathrm{aprx}}$, in which case the gap is even larger in favor of the alternate error bound. For example, let X and A be independent Rademacher random variables, let $\boldsymbol{\phi}(X, A) = A$, and let $r^\star(X, A) = X$. This satisfies weak realizability with $f^\star = 0$ and $b^\star(X) = X$. Moreover, the population VR projection is $f^\circ_{\mathrm{VR}} = 0$, and hence $e^\circ(X, A) = -X$. It follows that $\varepsilon_{\mathrm{aprx}} = 1$ but $\varepsilon_{\mathrm{VR},C} = 0$, since e° is action-independent. In addition, $\varepsilon_{\mathrm{x}} = 0$ and $\varepsilon_{\mathrm{a}} = \varepsilon_{\mathrm{z}} = 1$. Thus, in the noiseless case, the alternate bound is $\widetilde{\mathcal{O}}(N^{-1} + N_x^{-2})$, whereas the original bound contains the constant approximation-error term $\varepsilon_{\mathrm{aprx}} = 1$.

Finally, when $\varepsilon_{\mathrm{VR},C} > 0$ is constant, the square-root interaction in the alternate bound may yield a slower estimation remainder. For example, if $\kappa_{\mathrm{C}}\varepsilon_{\mathrm{x}} d/N_x$ is the dominant term in $\mathfrak{R}^{\mathrm{VR}}_{N,\mathrm{C}}$, then the alternate bound scales, ignoring lower-order terms, as $\varepsilon_{\mathrm{VR},C} + \widetilde{\mathcal{O}}\left(\sqrt{\frac{\kappa_{\mathrm{C}}\varepsilon_{\mathrm{VR},C}\varepsilon_{\mathrm{x}} d}{N_x}}\right)$, whereas the original bound scales as $\varepsilon_{\mathrm{aprx}} + \widetilde{\mathcal{O}}\left(\frac{\varepsilon_{\mathrm{x}} d}{N_x}\right)$. Consequently, when $\varepsilon_{\mathrm{VR},C} = \varepsilon_{\mathrm{aprx}} > 0$, the estimation remainder in the alternate bound is worse order-wise. When $\varepsilon_{\mathrm{VR},C} < \varepsilon_{\mathrm{aprx}}$, however, neither bound uniformly dominates: the smaller centered approximation error of the alternate bound must be compared with its slower square-root remainder.

We now present the proof of the alternate error bound:

Proof of Theorem F.1. We retain the notation from the proof of Lemma F.1. By the arguments in **Steps 1–2** of that proof, Eqn. (47) and Eqn. (49) hold simultaneously with probability at least $1-2\delta/3$. On the lower-isometry event, condition on the complete context–action design and define

$$\boldsymbol{y}^{\mathrm{C}}_{i,j} \coloneqq \frac{1}{N}\boldsymbol{M}_{\mathrm{C}}^{1/2}\boldsymbol{H}^{-1}\boldsymbol{z}(x_i, a_{i,j})\eta_{i,j}.$$

These vectors are conditionally independent and mean zero with conditional second moments bounded as

$$\begin{aligned}\sum_{i,j}\mathbb{E}\left[\left\|\boldsymbol{y}^{\mathrm{C}}_{i,j}\right\|_2^2 \,\middle|\, \{(x_i, a_{i,j})\}_{i,j}\right] &\le \frac{\sigma^2}{N^2}\operatorname{tr}\left(\boldsymbol{M}_{\mathrm{C}}^{1/2}\boldsymbol{H}^{-1}\left(\sum_{i,j}\boldsymbol{z}_{i,j}\boldsymbol{z}_{i,j}^\top\right)\boldsymbol{H}^{-1}\boldsymbol{M}_{\mathrm{C}}^{1/2}\right)\\ &= \frac{\sigma^2}{N}\operatorname{tr}(\boldsymbol{H}^{-1}\boldsymbol{M}_{\mathrm{C}}) \le \frac{3\sigma^2 d_{\mathrm{eff}}}{N}.\end{aligned}$$

Moreover,

$$\left\|\boldsymbol{y}^{\mathrm{C}}_{i,j}\right\|_2 \le \frac{6R_{\max}\alpha_0\sqrt{\kappa_{\mathrm{C}} d}}{N},$$

where we used $\left\|\boldsymbol{H}^{-1}\right\|_{\mathrm{op}} \le 3$, $\|\boldsymbol{M}_{\mathrm{C}}\|_{\mathrm{op}} = \kappa_{\mathrm{C}}$, and $\|\boldsymbol{z}\|_2 \le \alpha_0\sqrt{d}$. Thus, a conditional application of vector Bernstein's inequality (Eqn. (46)) gives, with conditional probability at least $1 - \delta/3$,

$$\left\|\boldsymbol{M}_{\mathrm{C}}^{1/2}\boldsymbol{H}^{-1}\boldsymbol{t}_\eta\right\|_2 \le \underbrace{\sqrt{\frac{3\sigma^2 d_{\mathrm{eff}}}{N}}\left(1 + \sqrt{8t_\delta}\right) + \frac{8R_{\max}\alpha_0\sqrt{\kappa_{\mathrm{C}} d}\, t_\delta}{N}}_{=:B_{\eta,\mathrm{C}}}. \tag{54}$$

The conditional failure probability remains at most $\delta/3$ after averaging over the design. Hence, the three events above hold simultaneously with probability at least $1 - \delta$.

Similarly, the empirical normal equation and Eqn. (44) give

$$\widehat{\boldsymbol{w}} \coloneqq \boldsymbol{\Sigma}^{1/2}\left(\hat{\boldsymbol{\theta}}^{\mathrm{VR}}_N - \boldsymbol{\theta}^\circ\right) \Longrightarrow \boldsymbol{H}\widehat{\boldsymbol{w}} = \boldsymbol{t}_\eta - \boldsymbol{t}_{\mathrm{mis}}.$$

Hence, by Eqn. (47), (49), (54),

$$\begin{aligned}\mathbb{E}_{x,a}\left[\left(\mathcal{C}\hat{f}^{\mathrm{VR}}_N - \mathcal{C}f^\circ_{\mathrm{VR}}\right)^2\right]^{1/2} &= \left\|\boldsymbol{M}_{\mathrm{C}}^{1/2}\widehat{\boldsymbol{w}}\right\|_2\\ &\le \left\|\boldsymbol{M}_{\mathrm{C}}^{1/2}\boldsymbol{H}^{-1}\boldsymbol{t}_{\mathrm{mis}}\right\|_2 + \left\|\boldsymbol{M}_{\mathrm{C}}^{1/2}\boldsymbol{H}^{-1}\boldsymbol{t}_\eta\right\|_2\\ &\le 3\sqrt{\kappa_{\mathrm{C}}}\, B_{\mathrm{mis}} + B_{\eta,\mathrm{C}} = \Gamma_{\delta,\mathrm{C}}.\end{aligned} \tag{55}$$

The triangle inequality, Eqn. (51), and Eqn. (55) imply

$$\mathbb{E}_{x,a}\left[\left(\mathcal{C}\hat{f}_N^{\mathrm{VR}} - \mathcal{C}r^\star\right)^2\right]^{1/2} \le \mathbb{E}_{x,a}\left[\left(\mathcal{C}\hat{f}_N^{\mathrm{VR}} - \mathcal{C}f_{\mathrm{VR}}^\circ\right)^2\right]^{1/2} + \mathbb{E}_{x,a}\left[\left(\mathcal{C}f_{\mathrm{VR}}^\circ - \mathcal{C}r^\star\right)^2\right]^{1/2}$$
$$\le \sqrt{\varepsilon_{\mathrm{VR},C}} + \Gamma_{\delta,\mathrm{C}}.$$

Finally, Eqns. (52)–(53) give $\Gamma^2_{\delta,\mathrm{C}} = \widetilde{\mathcal{O}}(\mathfrak{R}^{\mathrm{VR}}_{N,\mathrm{C}})$. This proves Theorem F.1. □

F.2 Value Difference Regression (VDR)

F.2.1 Proof of Theorem 4.2: VDR Bound for Linear $\mathcal{F}_{\mathrm{lin}}$

Throughout this subsection, $\mathcal{L}^{\mathrm{VDR}}$ and $\widehat{\mathcal{L}}^{\mathrm{VDR}}$ denote the generic losses in Section 2 instantiated with $\mathcal{T} = \mathcal{T}^{\mathrm{VDR}}$, and the comparator is $f^\circ_{\mathrm{VDR}} = f^\star$. For $h = f - f^\star$, let $h_{i,j} := h(x_i, a_{i,j})$. Under Assumption 1, the action-independent nuisance cancels from every within-context difference, so

$$\Delta(h; f^\star) := \mathcal{L}^{\mathrm{VDR}}(f^\star + h) - \mathcal{L}^{\mathrm{VDR}}(f^\star)$$
$$= \mathbb{E}_{x,a,a'}\left[\left(h(x,a) - h(x,a')\right)^2\right], \tag{56}$$

and

$$\widehat{\Delta}(h; f^\star) := \widehat{\mathcal{L}}^{\mathrm{VDR}}(f^\star + h) - \widehat{\mathcal{L}}^{\mathrm{VDR}}(f^\star)$$
$$= \frac{1}{N_x\binom{N_a}{2}} \sum_{i=1}^{N_x} \sum_{1\le j<k\le N_a} \left[\left(h_{i,j} - h_{i,k} - (\eta_{i,j} - \eta_{i,k})\right)^2 - (\eta_{i,j} - \eta_{i,k})^2\right].$$

The empirical quadratic term is

$$\widehat{Q}(h) = \widehat{Q}^{\mathrm{VDR}}(h) := \frac{1}{N_x\binom{N_a}{2}} \sum_{i=1}^{N_x} \sum_{1\le j<k\le N_a} (h_{i,j} - h_{i,k})^2. \tag{57}$$

Thus, in this subsection,

$$\Psi(r; 1, f^\star) = \sup_{\substack{h\in\mathcal{F}_{\mathrm{lin}} - f^\star \\ \Delta(h;f^\star)\le r}} \left[\Delta(h; f^\star) - \widehat{\Delta}(h; f^\star) - \widehat{Q}^{\mathrm{VDR}}(h)\right].$$

For the proof, let us denote $t_\delta := \log\frac{2}{\delta}$ and Define

$$\Gamma^{\mathrm{VDR}}_\delta := \sqrt{\frac{3\sigma^2 d_{\mathrm{C}}}{N_x(N_a-1)}}\left(1 + \sqrt{8t_\delta}\right) + \frac{8\sqrt{2}\, R_{\max}\alpha_{\mathrm{C}}\sqrt{d_{\mathrm{C}}}\, t_\delta}{N}.$$

Lemma F.2 (Localized modulus for linear VDR)**.** *Suppose the assumptions of Theorem 4.2 hold. Then, with probability at least* $1-\delta$*,*

$$\Psi(r; 1, f^\star) \le \frac{2}{3}r + \frac{2}{3}(\Gamma^{\mathrm{VDR}}_\delta)^2, \qquad \forall r \ge 0.$$

Again, we follow the same recipe.

Since $\mathcal{F}_{\mathrm{lin}}$ is parametrized by $\boldsymbol{\theta} \in \mathbb{R}^d$, let $\widehat{h}_{\mathrm{VDR}} := \hat{f}_N^{\mathrm{VDR}} - f^\star$. The normal equation for linear VDR gives

$$\widehat{\Delta}(\widehat{h}_{\mathrm{VDR}}; f^\star) = \widehat{\mathcal{L}}^{\mathrm{VDR}}(\hat{f}_N^{\mathrm{VDR}}) - \widehat{\mathcal{L}}^{\mathrm{VDR}}(f^\star) = -\widehat{Q}^{\mathrm{VDR}}(\widehat{h}_{\mathrm{VDR}}).$$

Thus condition (i) of Proposition 2.1 holds with $\lambda = 1$. By Lemma F.2, condition (ii) holds on an event of probability at least $1-\delta$. On this event,

$$\mathcal{L}^{\mathrm{VDR}}(\hat{f}_N^{\mathrm{VDR}}) - \mathcal{L}^{\mathrm{VDR}}(f^\star) \le \sup\left\{r \ge 0 : r \le \frac{2}{3}r + \frac{2}{3}(\Gamma^{\mathrm{VDR}}_\delta)^2\right\}$$
$$= 2(\Gamma^{\mathrm{VDR}}_\delta)^2.$$

Lemma C.1 therefore yields

$$\mathbb{E}_{x,a}\left[\left(\mathcal{C}\hat{f}_N^{\mathrm{VDR}} - \mathcal{C}r^\star\right)^2\right] \le (\Gamma_\delta^{\mathrm{VDR}})^2 = \widetilde{\mathcal{O}}\left(\frac{\sigma^2 d_{\mathrm{C}}}{N_x(N_a-1)} + \frac{R_{\max}^2 \alpha_{\mathrm{C}}^2 d_{\mathrm{C}}}{N^2}\right).$$

This proves Theorem 4.2. □

F.2.2 Proof of Lemma F.2

Under Assumption 1, let us write $f^\star(x,a) = \langle \boldsymbol{\phi}(x,a), \boldsymbol{\theta}^\star\rangle$ for some $\boldsymbol{\theta}^\star \in \mathbb{R}^d$. we first recall the centered features as defined in Section 4.2.

The centered feature is $\overline{\boldsymbol{\phi}}(x,a) := \boldsymbol{\phi}(x,a) - \mathbb{E}_{a\sim\pi_{\mathrm{ref}}(\cdot|x)}[\boldsymbol{\phi}(x,a)]$, and its covariance is $\boldsymbol{\Sigma}_{\mathrm{C}} := \mathbb{E}_{x,a}\left[\overline{\boldsymbol{\phi}}(x,a)\overline{\boldsymbol{\phi}}(x,a)^\top\right]$. Let $d_{\mathrm{C}} := \mathrm{rank}(\boldsymbol{\Sigma}_{\mathrm{C}})$. Whenever $d_{\mathrm{C}} \ge 1$, let $\boldsymbol{U}_{\mathrm{C}} \in \mathbb{R}^{d\times d_{\mathrm{C}}}$ satisfy $\boldsymbol{U}_{\mathrm{C}}^\top \boldsymbol{U}_{\mathrm{C}} = \boldsymbol{I}_{d_{\mathrm{C}}}$ and $\mathrm{col}(\boldsymbol{U}_{\mathrm{C}}) = \mathrm{range}(\boldsymbol{\Sigma}_{\mathrm{C}})$, and define

$$\overline{\boldsymbol{\Sigma}}_{\mathrm{C}} := \boldsymbol{U}_{\mathrm{C}}^\top \boldsymbol{\Sigma}_{\mathrm{C}} \boldsymbol{U}_{\mathrm{C}} \succ \boldsymbol{0}, \qquad \boldsymbol{z}_{\mathrm{C}}(x,a) := \overline{\boldsymbol{\Sigma}}_{\mathrm{C}}^{-1/2} \boldsymbol{U}_{\mathrm{C}}^\top \overline{\boldsymbol{\phi}}(x,a) \in \mathbb{R}^{d_{\mathrm{C}}}.$$

Then, $\mathbb{E}[\boldsymbol{z}_{\mathrm{C}}\boldsymbol{z}_{\mathrm{C}}^\top] = \boldsymbol{I}_{d_{\mathrm{C}}}$ and $\mathbb{E}_A[\boldsymbol{z}_{\mathrm{C}}(X,A) \mid X] = \boldsymbol{0}$.

For each observed block $i \in [N_x]$, define

$$\overline{\boldsymbol{z}}_{i,\mathrm{C}} := \frac{1}{N_a}\sum_{j=1}^{N_a} \boldsymbol{z}_{\mathrm{C}}(x_i, a_{i,j}), \quad \widetilde{\boldsymbol{z}}_{i,j} := \frac{1}{\sqrt{2}}\left(\boldsymbol{z}_{\mathrm{C}}(x_i,a_{i,j}) - \overline{\boldsymbol{z}}_{i,\mathrm{C}}\right), \quad \boldsymbol{G}_i^{\mathrm{C}} := \frac{2}{N_a - 1}\sum_{j=1}^{N_a} \widetilde{\boldsymbol{z}}_{i,j}\widetilde{\boldsymbol{z}}_{i,j}^\top, \tag{58}$$

and

$$\boldsymbol{H} := \frac{1}{N_x}\sum_{i=1}^{N_x} \boldsymbol{G}_i^{\mathrm{C}}, \quad \boldsymbol{s} := \frac{2}{N_x(N_a-1)}\sum_{i=1}^{N_x}\sum_{j=1}^{N_a} \widetilde{\boldsymbol{z}}_{i,j}\eta_{i,j}. \tag{59}$$

Using the empirical Gram matrix and score in Eqn. (59), we now express the quantities recalled in Eqns. (56)–(57) in the identifiable centered-feature coordinates. Fix any $h = f - f^\star$, where

$$f(x,a) = \langle \boldsymbol{\phi}(x,a), \boldsymbol{\theta}\rangle, \qquad \boldsymbol{w} := \sqrt{2}\,\overline{\boldsymbol{\Sigma}}_{\mathrm{C}}^{1/2}\boldsymbol{U}_{\mathrm{C}}^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^\star) \in \mathbb{R}^{d_{\mathrm{C}}}.$$

For every $\boldsymbol{v} \in \ker(\boldsymbol{\Sigma}_{\mathrm{C}})$, $\mathbb{E}[(\boldsymbol{v}^\top \mathcal{C}\boldsymbol{\phi})^2] = 0$; hence $\overline{\boldsymbol{\phi}} \in \mathrm{range}(\boldsymbol{\Sigma}_{\mathrm{C}})$ almost surely. It follows that

$$\left(\boldsymbol{\phi}(x,a) - \boldsymbol{\phi}(x,a')\right)^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^\star) = \frac{1}{\sqrt{2}}\left(\boldsymbol{z}_{\mathrm{C}}(x,a) - \boldsymbol{z}_{\mathrm{C}}(x,a')\right)^\top \boldsymbol{w} \qquad \text{almost surely.}$$

By Lemma C.1,

$$\Delta(h; f^\star) = \mathbb{E}_{x,a,a'}\left[\left(h(x,a) - h(x,a')\right)^2\right] = \|\boldsymbol{w}\|_2^2. \tag{60}$$

Moreover,

$$\widehat{Q}^{\mathrm{VDR}}(h) = \boldsymbol{w}^\top \boldsymbol{H}\boldsymbol{w}.$$

Since the action-independent nuisance $b^\star(x_i)$ cancels from every within-context difference, direct expansion of the empirical VDR excess loss gives

$$\widehat{\Delta}(h; f^\star) = \boldsymbol{w}^\top \boldsymbol{H}\boldsymbol{w} - 2\boldsymbol{w}^\top \boldsymbol{s}.$$

Therefore, the process to be controlled is

$$\Delta(h; f^\star) - \widehat{\Delta}(h; f^\star) - \widehat{Q}^{\mathrm{VDR}}(h) = \|\boldsymbol{w}\|_2^2 - 2\boldsymbol{w}^\top \boldsymbol{H}\boldsymbol{w} + 2\boldsymbol{w}^\top \boldsymbol{s}. \tag{61}$$

It is therefore enough to show, uniformly over h, that the right-hand side of Eqn. (61) is at most $2\Delta(h; f^\star)/3 + 2(\Gamma_\delta^{\mathrm{VDR}})^2/3$; restricting this bound to $\Delta(h; f^\star) \le r$ concludes the proof.

Similarly, we prove the uniform bound through a lower-isometry event for $\boldsymbol{H}$ and concentration of $\boldsymbol{s}$ in the corresponding empirical norm.

Step 1: lower isometry of the empirical Gram matrix. The matrices $\boldsymbol{G}_i^{\mathrm{C}}$ are i.i.d. copies of $\boldsymbol{G}_{N_a}^{\mathrm{C}}$ and satisfy $\mathbb{E}[\boldsymbol{G}_i^{\mathrm{C}}] = \boldsymbol{I}_{d_{\mathrm{C}}}$. Moreover, $\boldsymbol{G}_i^{\mathrm{C}} \succeq \boldsymbol{0}$, and hence $\lambda_{\max}(\boldsymbol{I}_{d_{\mathrm{C}}} - \boldsymbol{G}_i^{\mathrm{C}}) \le 1$.

The one-sided matrix Bernstein inequality (Tropp, 2015, Theorem 6.6.1) therefore gives, with probability at least $1-\delta/2$,

$$\lambda_{\max}(\boldsymbol{I}_{d_{\mathrm{C}}} - \boldsymbol{H}) \le \sqrt{\frac{2\left\|\mathrm{Cov}(\boldsymbol{G}^{\mathrm{C}}_{N_a})\right\|_{\mathrm{op}} \log(2d_{\mathrm{C}}/\delta)}{N_x}} + \frac{2\log(2d_{\mathrm{C}}/\delta)}{3N_x} \le \frac{2}{3},$$

where the last step uses the sample-size condition in Theorem 4.2. Thus,

$$\boldsymbol{H} \succeq \frac{1}{3}\boldsymbol{I}_{d_{\mathrm{C}}}. \tag{62}$$

Step 2: control of the reward-noise score. On the event in Eqn. (62), condition on the complete context–action design $\{(x_i, a_{i,j})\}_{i,j}$. Then, the vectors

$$\boldsymbol{y}_{i,j} := \frac{2}{N_x(N_a-1)}\boldsymbol{H}^{-1/2}\widetilde{\boldsymbol{z}}_{i,j}\eta_{i,j}$$

are conditionally independent and mean zero. By Eqn. (58),

$$\sum_{i,j}\mathbb{E}\left[\left\|\boldsymbol{y}_{i,j}\right\|_2^2 \,\middle|\, \{(x_i,a_{i,j})\}_{i,j}\right] \le \frac{4\sigma^2}{N_x^2(N_a-1)^2}\,\mathrm{tr}\left(\boldsymbol{H}^{-1}\sum_{i,j}\widetilde{\boldsymbol{z}}_{i,j}\widetilde{\boldsymbol{z}}_{i,j}^{\top}\right) = \frac{2\sigma^2 d_{\mathrm{C}}}{N_x(N_a-1)}.$$

On the other hand, Eqn. (58) gives

$$\widetilde{\boldsymbol{z}}_{i,j} = \frac{1}{N_a}\sum_{k\neq j}\frac{\boldsymbol{z}_{\mathrm{C}}(x_i,a_{i,j}) - \boldsymbol{z}_{\mathrm{C}}(x_i,a_{i,k})}{\sqrt{2}}.$$

Thus, the triangle inequality and Assumption 3 imply

$$\left\|\widetilde{\boldsymbol{z}}_{i,j}\right\|_2 \le \frac{N_a-1}{N_a}\alpha_{\mathrm{C}}\sqrt{d_{\mathrm{C}}}.$$

Using $|\eta_{i,j}| \le 2R_{\max}$ and $\left\|\boldsymbol{H}^{-1/2}\right\|_{\mathrm{op}} \le \sqrt{3}$ on Eqn. (62), we obtain

$$\left\|\boldsymbol{y}_{i,j}\right\|_2 \le \frac{4\sqrt{3}\,R_{\max}\alpha_{\mathrm{C}}\sqrt{d_{\mathrm{C}}}}{N_xN_a} = \frac{4\sqrt{3}\,R_{\max}\alpha_{\mathrm{C}}\sqrt{d_{\mathrm{C}}}}{N}.$$

Applying the vector Bernstein inequality (Eqn. (46)) conditionally on the design gives, with conditional probability at least $1-\delta/2$,

$$\|\boldsymbol{s}\|_{\boldsymbol{H}^{-1}} \le \sqrt{\frac{2\sigma^2 d_{\mathrm{C}}}{N_x(N_a-1)}}\left(1+\sqrt{8t_\delta}\right) + \frac{16\sqrt{3}\,R_{\max}\alpha_{\mathrm{C}}\sqrt{d_{\mathrm{C}}}\,t_\delta}{3N} = \sqrt{\frac{2}{3}}\,\Gamma^{\mathrm{VDR}}_\delta. \tag{63}$$

The conditional failure probability remains at most $\delta/2$ after averaging over the design. Hence, Eqns. (62)–(63) hold simultaneously with probability at least $1-\delta$.

Step 3: combining the two events. Let $D := \|\boldsymbol{s}\|_{\boldsymbol{H}^{-1}}$. By Cauchy–Schwarz and AM-GM inequalities,

$$2\boldsymbol{w}^{\top}\boldsymbol{s} \le 2\,\|\boldsymbol{w}\|_{\boldsymbol{H}}\,\|\boldsymbol{s}\|_{\boldsymbol{H}^{-1}} \le \|\boldsymbol{w}\|_{\boldsymbol{H}}^2 + D^2.$$

Hence, using $\boldsymbol{H} \succeq \boldsymbol{I}_{d_{\mathrm{C}}}/3$ and Eqn. (60),

$$\begin{aligned}\Delta(h;f^\star) - \widehat{\Delta}(h;f^\star) - \widehat{Q}^{\mathrm{VDR}}(h) &\le \|\boldsymbol{w}\|_2^2 - \boldsymbol{w}^{\top}\boldsymbol{H}\boldsymbol{w} + D^2 \\ &\le \frac{2}{3}\Delta(h;f^\star) + D^2 \\ &\le \frac{2}{3}\Delta(h;f^\star) + \frac{2}{3}(\Gamma^{\mathrm{VDR}}_\delta)^2.\end{aligned}$$

Restricting to $\Delta(h;f^\star) \le r$ and taking the supremum in the definition of $\Psi(r;1,f^\star)$ yields

$$\Psi(r;1,f^\star) \le \frac{2}{3}r + \frac{2}{3}(\Gamma^{\mathrm{VDR}}_\delta)^2, \qquad \forall r \ge 0.$$

This proves the claim. □

G LOWER BOUNDS

We first give an estimator-specific separation: under Assumption 1, VR can incur constant regret while VDR followed by greedy selection succeeds on the same instance. We then remove Assumption 1 and establish an agnostic minimax lower bound for arbitrary learners. Appendix G.1 proves the separation; Appendix G.2 states the minimax result along with a comparison to prior lower bounds. Lastly, Appendix G.3 proves it.

Throughout, $N = N_x N_a$ is the total number of observations in the grouped dataset. Unless otherwise specified, $\mathbb{E}_{x,a}$ denotes $\mathbb{E}_{x\sim\rho,a\sim\pi_{\mathrm{ref}}(\cdot|x)}$. All policy values and regrets are evaluated using the true reward $r^\star$.

G.1 PROOF OF THEOREM 3.2: FAILURE MODE OF VR

The proof first identifies a predictor g that fits absolute rewards better than $f^\star$ but reverses the true action ordering. We then show that empirical VR selects g with high probability by concentrating over the N_x independent context blocks. On that event, greedy selection chooses the suboptimal action. Finally, we show that VDR followed by greedy selection succeeds on the same instance when $N_a \geq 2$.

Step 1: construction and population failure. First, recall that the approximation error is defined as:

$$\varepsilon_{\mathrm{aprx}} := \inf_{f\in\mathcal{F}} \mathbb{E}_{x\sim\rho,a\sim\pi_{\mathrm{ref}}(\cdot|x)} \left[(f(x,a) - r^\star(x,a))^2\right].$$

While $\varepsilon_{\mathrm{aprx}}$ measures how well the model class approximates the *absolute reward values*, greedy decision making depends only on the *relative ordering of actions* within each context. Consider the weak realizability model

$$r^\star(x,a) = f^\star(x,a) + b^\star(x)$$

and choose noiseless observations, so $r_{i,j} = r^\star(x_i, a_{i,j})$ and $\eta_{i,j} = 0$. The action-independent baseline $b^\star(x)$ does not itself change the true action ordering. Nevertheless, ordinary VALUE REGRESSION can still fail when three ingredients are present: (1) the baseline varies across contexts, (2) the behavior policy correlates actions with contexts, and (3) the value class cannot represent the context-dependent baseline. In this case, different actions are observed under different mixtures of contexts, so the baseline can be absorbed into the fitted action values and reverse their ordering.

To make this precise, consider two contexts $\mathcal{X} = \{x_1, x_2\}$, with $\rho(x_1) = \rho(x_2) = \frac{1}{2}$, and two actions $\mathcal{A} = \{a_1, a_2\}$. Fix a baseline magnitude $B > 0$ and an action gap $\Delta > 0$, to be specified below, and define

$$f^\star(x, a_1) = \Delta, \quad f^\star(x, a_2) = 0, \quad \forall x \in \mathcal{X},$$

and let

$$b^\star(x_1) = B, \quad b^\star(x_2) = 0.$$

Now one can readily see that $r^\star(x,a) = f^\star(x,a) + b^\star(x)$, and a_1 is the optimal action at both contexts, therefore

$$r^\star(x, a_1) - r^\star(x, a_2) = \Delta.$$

Now choose a behavior policy that correlates actions with contexts:

$$\pi_{\mathrm{ref}}(a_i \mid x_j) := p\mathbb{1}\{i = j\} + (1-p)\mathbb{1}\{i \neq j\}, \quad p \in (0, 1/2).$$

Here one can observe that the policy makes the two actions appear in systematically different contexts. In particular,

$$\begin{aligned} \pi_{\mathrm{ref}}(a_1 \mid x_1) &= p, & \pi_{\mathrm{ref}}(a_1 \mid x_2) &= 1-p \\ \pi_{\mathrm{ref}}(a_2 \mid x_1) &= 1-p, & \pi_{\mathrm{ref}}(a_2 \mid x_2) &= p. \end{aligned}$$

Since $\rho(x_1) = \rho(x_2) = \frac{1}{2}$, $\Pr(A = a_1) = \Pr(A = a_2) = \frac{1}{2}$, and therefore by Bayes' rule,

$$\begin{aligned} \Pr(X = x_1 \mid A = a_1) &= p, & \Pr(X = x_2 \mid A = a_1) &= 1-p \\ \Pr(X = x_1 \mid A = a_2) &= 1-p, & \Pr(X = x_2 \mid A = a_2) &= p. \end{aligned}$$

Hence the action-wise conditional mean rewards under the behavior distribution are

$$\mathbb{E}[r^\star(x,a) \mid A = a_1] = p(B+\Delta) + (1-p)\Delta = \Delta + pB,$$
$$\mathbb{E}[r^\star(x,a) \mid A = a_2] = (1-p)B + p \cdot 0 = (1-p)B.$$

Motivated by this, consider the predictor g such that

$$g(x,a_1) \coloneqq \Delta + pB, \quad g(x,a_2) \coloneqq (1-p)B, \quad \forall x \in \mathcal{X},$$

and take the finite value function class $\mathcal{F} \coloneqq \{f^\star, g\}$.

We now compare the population squared risks of the two candidates. To this end, define the population squared-loss objective

$$\mathcal{L}^{\mathrm{VR}}(f) \coloneqq \mathbb{E}_{x\sim\rho, a\sim\pi_{\mathrm{ref}}(\cdot|x)}\left[\big(f(x,a) - r^\star(x,a)\big)^2\right],$$

and let $f^\circ_{\mathrm{VR}} \in \arg\min_{f\in\mathcal{F}} \mathcal{L}^{\mathrm{VR}}(f)$ denote the population VALUE REGRESSION solution. For g, a direct calculation yields:

$$\mathcal{L}^{\mathrm{VR}}(g) = \frac{p}{2}(1-p)^2B^2 + \frac{1-p}{2}p^2B^2 + \frac{1-p}{2}p^2B^2 + \frac{p}{2}(1-p)^2B^2 = p(1-p)B^2.$$

On the other hand, since $f^\star(x,a) - r^\star(x,a) = -b^\star(x)$,

$$\mathcal{L}^{\mathrm{VR}}(f^\star) = \mathbb{E}_x[b^\star(x)^2] = \frac{B^2}{2}.$$

Since $p(1-p) \le \frac{1}{4} < \frac{1}{2}$, we have $\mathcal{L}^{\mathrm{VR}}(g) < \mathcal{L}^{\mathrm{VR}}(f^\star)$ and therefore

$$f^\circ_{\mathrm{VR}} = g, \quad \varepsilon_{\mathrm{aprx}} = p(1-p)B^2.$$

Furthermore, we note that

$$\begin{aligned} g(x,a_2) - g(x,a_1) &= (1-p)B - (\Delta + pB) \\ &= (1-2p)B - \Delta. \end{aligned}$$

Hence, whenever $(1-2p)B > \Delta$, the VR solution g prefers a_2 at every context, while a_1 is optimal at every context.

For a concrete instance at the natural reward scale, take

$$p = \frac{1}{4}, \quad B = \frac{3}{4}R_{\max}, \quad \Delta = \frac{1}{8}R_{\max}.$$

Then a straightforward yet tedious calculation yields:

$$B + \Delta = \frac{7}{8}R_{\max} \le R_{\max},$$

so all rewards are in $[0, R_{\max}]$, while

$$(1-2p)B = \frac{3}{8}R_{\max} > \frac{1}{8}R_{\max} = \Delta,$$

so the population VR chooses the wrong action. Moreover, $\varepsilon_{\mathrm{aprx}} = p(1-p)B^2 = \frac{27}{256}R_{\max}^2$, implying that it is of order $R_{\max}^2$, while the induced GREEDY policy incurs

$$\mathrm{Reg}(f^\circ_{\mathrm{VR}}) = \Delta = \frac{1}{8}R_{\max}.$$

Step 2: empirical VR selects the same wrong predictor. It remains to transfer the strict population comparison to empirical VR. We keep the instance and the finite class $\mathcal{F} = \{f^\star, g\}$ unchanged. The grouped dataset contains $N = N_x N_a$ observations, but the N_a observations within a block share the random context x_i. We therefore apply concentration to the N_x independent blocks, retaining the within-block average over all N_a actions. For block i, define

$$D_i = \frac{1}{N_a}\sum_{j=1}^{N_a}\left[\big(g(x_i,a_{i,j}) - r_{i,j}\big)^2 - \big(f^\star(x_i,a_{i,j}) - r_{i,j}\big)^2\right].$$

Then $D_1, \ldots, D_{N_x}$ are independent and

$$\widehat{\mathcal{L}}^{\mathrm{VR}}(g) - \widehat{\mathcal{L}}^{\mathrm{VR}}(f^\star) = \frac{1}{N_x}\sum_{i=1}^{N_x} D_i,$$

$$\mathbb{E}[D_i] = \mathcal{L}^{\mathrm{VR}}(g) - \mathcal{L}^{\mathrm{VR}}(f^\star) = -\frac{45}{256}R_{\max}^2.$$

The single-observation loss differences at (x_1, a_1), (x_1, a_2), (x_2, a_1), and (x_2, a_2) are, respectively, $(-63, -135, 9, 81)R_{\max}^2/256$. Every block average therefore satisfies

$$-\frac{135}{256}R_{\max}^2 \le D_i \le \frac{81}{256}R_{\max}^2.$$

Applying Hoeffding's inequality to the N_x independent blocks gives

$$\begin{aligned}\Pr\left(\widehat{\mathcal{L}}^{\mathrm{VR}}(g) \ge \widehat{\mathcal{L}}^{\mathrm{VR}}(f^\star)\right) &= \Pr\left(\frac{1}{N_x}\sum_{i=1}^{N_x} D_i \ge 0\right) \\ &\le \exp\left(-\frac{2N_x(45R_{\max}^2/256)^2}{(216R_{\max}^2/256)^2}\right) \\ &= \exp\left(-\frac{25N_x}{288}\right).\end{aligned}$$

Consequently, on the event $\mathcal{E} = \left\{\widehat{\mathcal{L}}^{\mathrm{VR}}(g) < \widehat{\mathcal{L}}^{\mathrm{VR}}(f^\star)\right\}$, the unique empirical VR minimizer is g, regardless of the ERM tie-breaking rule. The event has probability at least $1 - \exp(-25N_x/288)$, uniformly over $N_a \ge 1$. On this event, greedy selection from the empirical VR predictor chooses a_2 at both contexts. Consequently,

$$\mathrm{Reg}\left(\widehat{\pi}_N^{\text{VR-GREEDY}}\right) = \Delta = \frac{R_{\max}}{8}.$$

In conclusion, the VR claim of Theorem 3.2 holds with probability at least $1 - \delta$ whenever $N_x \ge \frac{288}{25}\log\frac{1}{\delta}$.

Step 3: VDR **succeeds on the same instance.** For $N_a \ge 2$, one block containing both actions is enough to distinguish $f^\star$ from g using reward differences. In particular, the baseline cancels from every within-context difference, so $\widehat{\mathcal{L}}^{\mathrm{VDR}}(f^\star) = 0$. In contrast,

$$\begin{aligned}g(x, a_1) - g(x, a_2) &= -\frac{1}{4}R_{\max}, \\ r^\star(x, a_1) - r^\star(x, a_2) &= \frac{1}{8}R_{\max}.\end{aligned}$$

Therefore, if any context block contains both actions, g has strictly positive VDR loss and $f^\star$ is the unique empirical VDR minimizer. For either context, the probability that all N_a sampled actions agree is $p^{N_a} + (1-p)^{N_a}$. Thus, for any ERM tie-breaking rule, independence across blocks gives

$$\Pr\left(\widehat{f}_N^{\mathrm{VDR}} \ne f^\star\right) \le \left[p^{N_a} + (1-p)^{N_a}\right]^{N_x} \le \left(\frac{5}{8}\right)^{N_x}.$$

Therefore, VDR followed by GREEDY action selection has zero regret with probability at least $1 - (5/8)^{N_x}$. This completes the proof.

G.2 AGNOSTIC MINIMAX LOWER BOUND

We now ask whether the approximation and statistical terms in the regret bound can be improved without Assumption 1. The answer is negative in a worst-case sense: for every learner, some admissible true reward forces regret of order $\sqrt{C\varepsilon} + R_{\max}\sqrt{C\log|\mathcal{F}|/N}$ with probability bounded away from zero. Unlike the preceding example, the learner may output any policy, not only one obtained from a predictor in $\mathcal{F}$. We write $\mathrm{Reg}_{r^\star}$ when the dependence of regret on the true reward needs to be explicit.

For comparison, combining the finite-class VR prediction bound with the PESSIMISM policy guarantee in Appendix B yields, under the corresponding implementation assumptions, a regret upper bound whose leading terms are

$$\sqrt{C^\star\varepsilon_{\mathrm{aprx}}} + R_{\max}\sqrt{\frac{C^\star\log|\mathcal{F}|}{N}}$$

when N_a is bounded. The lower bound below shows that these two dependencies are unavoidable in a worst-case sense over a coverage-budget class.

Theorem G.1 (Agnostic minimax lower bound). *Let $C \geq 2$, $R_{\max} > 0$, and $d \geq 1$ be an integer, and let $\varepsilon \in \left[0, \frac{R_{\max}^2}{32C}\right]$. For every pair of positive integers N_x, N_a with $N = N_x N_a \geq Cd$, there exists an instance $(\mathcal{X}, \mathcal{A}, \rho, \pi_{\mathrm{ref}}, \mathcal{F}, \mathcal{R}^\star)$ with $(|\mathcal{A}|, |\mathcal{F}|, |\mathcal{X}|) = \left(2, 2^d, d + \max\left\{1, \left\lceil \frac{43N\varepsilon}{R_{\max}^2} \right\rceil\right\}\right)$. All observed rewards and all functions in $\mathcal{F}$ take values in $[0, R_{\max}]$, and the following hold.*

(i) *The coverage coefficients satisfy $C_{\mathcal{F}}^{\infty} = C$ and $C^\star(r^\star) \leq C$ for every $r^\star \in \mathcal{R}^\star$. Every true reward has approximation error exactly ε:*

$$\inf_{f \in \mathcal{F}} \mathbb{E}_{x,a}\left[\big(f(X,A) - r^\star(X,A)\big)^2\right] = \varepsilon,$$

with the infimum attained. Conditional on the complete context–action design, the reward noise variables are independent, have mean zero, and are bounded in absolute value by $R_{\max}$.

(ii) *The expected minimax regret satisfies*

$$\inf_{\mathsf{L}} \max_{r^\star \in \mathcal{R}^\star} \mathbb{E}_{r^\star}^{\mathsf{L}}\left[\mathrm{Reg}_{r^\star}(\widehat{\pi}_N)\right] \geq \frac{1}{48}\left(\sqrt{C\varepsilon} + R_{\max}\sqrt{\frac{C \log_2 |\mathcal{F}|}{N}}\right).$$

(iii) *The constant-probability version also holds:*

$$\inf_{\mathsf{L}} \max_{r^\star \in \mathcal{R}^\star} \Pr_{r^\star}^{\mathsf{L}}\left(\mathrm{Reg}_{r^\star}(\widehat{\pi}_N) \geq \frac{1}{96}\left[\sqrt{C\varepsilon} + R_{\max}\sqrt{\frac{C \log_2 |\mathcal{F}|}{N}}\right]\right) \geq \frac{1}{8}.$$

Here the infimum is over all measurable learners L that map the grouped dataset to a possibly randomized policy $\widehat{\pi}_N : \mathcal{X} \to \Delta(\mathcal{A})$. The outer expectations and probabilities include both the data and the learner's internal randomness.

Related work. Coverage-dependent amplification of prediction error appears in the error-propagation analyses of approximate dynamic programming and in batch RL (Chen & Jiang, 2019; Farahmand et al., 2010; Munos, 2003; 2007). Amortila et al. (2023) characterize optimal approximation factors for misspecified linear off-policy value-function estimation, which is distinct from the offline policy-learning problem considered here. Our statistical term is related to the single-policy-concentrability rates of Rashidinejad et al. (2022), but their pointwise coverage condition differs from our expectation-form coefficient. Theorem G.1 establishes a lower bound over a coverage-budget class; it does not identify $C^\star$ with $C_{\mathcal{F}}^{\infty}$ at every member of the family.

For the misspecification term, two related lines of work are particularly relevant. First, lower bounds for misspecified linear models use small uniform approximation errors to make identifying a near-optimal action difficult (Du et al., 2020; Lattimore et al., 2020; Van Roy & Dong, 2019). Our construction shares the idea of distributing uncertainty over many coordinates, but uses a finite function class, offline observations, and an average squared-error budget. Second, other lower bounds apply to restricted classes of learners or policies. Krishnamurthy et al. (2021, Theorem 2) prove an average-misspecification lower bound for randomized policies induced by kernels in the convex hull of a model-induced kernel class. Amortila et al. (2024, Proposition 2.2) establish an asymptotic ERM lower bound under uniform misspecification; their subsequent discussion gives an obstruction for proper learners under L_2 misspecification. Here the lower bound applies to arbitrary measurable learners, including learners that do not output a function in $\mathcal{F}$. The number of misspecified contexts may grow with N, so this is a finite-sample minimax statement rather than an asymptotic impossibility result at a fixed finite instance.

Remark 2 (No conflict with L_∞-based possibility results). *Amortila et al. (2024) show that disagreement-based regression can avoid coverage amplification relative to the uniform misspecification budget*

$$\varepsilon_\infty = \inf_{f \in \mathcal{F}} \|f - r^\star\|_\infty.$$

This does not contradict Theorem G.1, which is parameterized by average squared approximation error. On our instances, the approximation error of the matching in-class function is supported on

the pairs (y_k, a_2), whose total probability under $\rho \otimes \pi_{\mathrm{ref}}$ is $1/(2C)$. Consequently,

$$\varepsilon_\infty = \sqrt{2C\varepsilon_{\mathrm{aprx}}}.$$

The misspecification contribution to our lower bound is therefore of order ε_∞ itself, not an additional coverage amplification of that uniform budget. Thus the lower bound in terms of $(C, \varepsilon_{\mathrm{aprx}})$ does not imply adaptation to the finer L_∞ approximation budget.

G.3 Proof of Theorem G.1: Agnostic Minimax Lower Bound

Fix the parameters in Theorem G.1 and write $\bar{\varepsilon} = \varepsilon / R_{\max}^2$. For any fixed learner, it suffices to find a true reward for which regret is large. We encode the action preference at each context by an unknown sign: positive favors a_2, and negative favors a_1. At zero gap, either action is optimal and that coordinate contributes no regret. Considering all sign assignments gives a hypercube, reducing policy learning to recovering the unknown action-preference signs. The reduction has two properties: First, expected regret equals a weighted sum of sign-error probabilities, whose weights sum to S. Second, data distributions for neighboring sign vectors have total variation distance at most $1/2$. Weighted Assouad testing then gives a true reward with expected regret at least $S/4$. Because regret is always at most S, the same true reward has regret at least $S/8$ with probability at least $1/8$. The construction below makes these thresholds at least the two rates in the theorem.

We first specify the instance and derive this regret reduction. We then verify the approximation and coverage requirements and bound the information in the grouped data. Constants are not optimized.

Step 1: construct the two context groups. The first block will carry d signs represented in $\mathcal{F}$; the second will carry m signs absent from the modeled action differences. We use normalized gaps Δ and Δ_0 to control their statistical and approximation contributions, respectively. Choose

$$\Delta = \frac{1}{3}\sqrt{\frac{Cd}{N}} \le \frac{1}{3},$$
$$\Delta_0 = \sqrt{8C\bar{\varepsilon}} \le \frac{1}{2},$$
$$m = \max\left\{1, \lceil 43N\bar{\varepsilon} \rceil\right\}.$$

The inequalities use $N \ge Cd$ and $C\bar{\varepsilon} \le 1/32$. The choice of m spreads the misspecification over enough contexts to keep each sign difficult to identify. When $\varepsilon = 0$, we have $\Delta_0 = 0$, and the misspecification block contributes no regret.

Set $\mathcal{X} = \{x_1, \dots, x_d\} \cup \{y_1, \dots, y_m\}$, $\mathcal{A} = \{a_1, a_2\}$, and choose the context distribution and behavior policy as:

$$\rho(x_j) = \frac{1}{2d}, \qquad \rho(y_k) = \frac{1}{2m}, \qquad \pi_{\mathrm{ref}}(a_2 \mid x) = \frac{1}{C} \quad \forall x \in \mathcal{X}.$$

The x-contexts yield the statistical lower bound, while the y-contexts yield the misspecification lower bound. Since $C \ge 2$, we have

$$\pi_{\mathrm{ref}}(a_1 \mid x) = 1 - \frac{1}{C} \ge \frac{1}{C} = \pi_{\mathrm{ref}}(a_2 \mid x).$$

For each $\sigma \in \{-1, +1\}^d$, define

$$f_\sigma(x_j, a_1) = \frac{R_{\max}}{2},$$
$$f_\sigma(x_j, a_2) = \frac{R_{\max}}{2}(1 + \sigma_j \Delta),$$
$$f_\sigma(y_k, a_1) = f_\sigma(y_k, a_2) = \frac{R_{\max}}{2}.$$

For $\tau \in \{-1, +1\}^m$, let $r^\star_{\sigma,\tau}$ agree with f_σ everywhere except at the pairs (y_k, a_2), where

$$r^\star_{\sigma,\tau}(y_k, a_2) = \frac{R_{\max}}{2}(1 + \tau_k \Delta_0).$$

Set

$$\mathcal{F} \coloneqq \{f_\sigma : \sigma \in \{-1,+1\}^d\},$$
$$\mathcal{R}^\star \coloneqq \{r^\star_{\sigma,\tau} : (\sigma,\tau) \in \{-1,+1\}^d \times \{-1,+1\}^m\}.$$

Then $|\mathcal{F}| = 2^d$, so $d = \log_2 |\mathcal{F}|$. All function values and true mean rewards lie in $[R_{\max}/4, 3R_{\max}/4]$. When $\varepsilon > 0$, the classes $\mathcal{F}$ and $\mathcal{R}^\star$ are disjoint; when $\varepsilon = 0$, $\mathcal{R}^\star = \mathcal{F}$ as sets.

Generate the contexts and actions according to the grouped sampling scheme. Conditional on the complete context–action design, let the rewards be independent with

$$r_{i,j} = R_{\max} B_{i,j}, \qquad B_{i,j} \sim \text{Bernoulli}\left(\frac{r^\star(x_i, a_{i,j})}{R_{\max}}\right).$$

Then the conditional reward mean is $r^\star(x_i, a_{i,j})$, and

$$\eta_{i,j} = r_{i,j} - r^\star(x_i, a_{i,j})$$

is conditionally mean zero and bounded in absolute value by $R_{\max}$.

Step 2: reduce regret to sign recovery. The construction makes the optimal action depend on one sign at each context. This turns the learner's regret into a weighted testing loss. Write $a^{+1} = a_2$ and $a^{-1} = a_1$. Under $r^\star_{\sigma,\tau}$, the optimal action at x_j is a^{σ_j}, with gap $R_{\max}\Delta/2$. At y_k, action a^{τ_k} is optimal, with gap $R_{\max}\Delta_0/2$; when $\Delta_0 = 0$, both actions are optimal and this gap is zero. For every possibly randomized policy $\widehat{\pi} : \mathcal{X} \to \Delta(\mathcal{A})$,

$$\text{Reg}_{\sigma,\tau}(\widehat{\pi}) = \sum_{j=1}^{d} \alpha_j \widehat{\pi}(a^{-\sigma_j} \mid x_j) + \sum_{k=1}^{m} \beta_k \widehat{\pi}(a^{-\tau_k} \mid y_k), \tag{64}$$

where

$$\alpha_j = \frac{R_{\max}\Delta}{4d}, \qquad \beta_k = \frac{R_{\max}\Delta_0}{4m}.$$

Each weight is the probability of its context multiplied by the action gap, so a wrong sign contributes exactly the corresponding regret. In particular, every policy satisfies $0 \le \text{Reg}_{\sigma,\tau}(\widehat{\pi}) \le S$, where

$$S = \sum_{j=1}^{d} \alpha_j + \sum_{k=1}^{m} \beta_k = \frac{R_{\max}(\Delta + \Delta_0)}{4}.$$

To lower-bound the weighted error, we use the following weighted version of Assouad's lemma (Assouad, 1983; Yu, 1997). In particular, we compare pairs of true rewards that differ at only one context. If their data distributions are close, a learner cannot reliably choose the correct action under both rewards.

Lemma G.1 (Weighted Assouad's lemma). *Let $\{P_\omega : \omega \in \{-1,+1\}^K\}$ be probability measures on a common measurable space, and let $w_1, \dots, w_K \ge 0$. Suppose that, for each coordinate ℓ, there is $\delta_\ell \in [0,1]$ such that*

$$D_{\text{TV}}(P_\omega, P_{\omega'}) \le \delta_\ell$$

whenever ω and ω' differ only at coordinate ℓ. Then every possibly randomized estimator $\widehat{\omega} \in \{-1,+1\}^K$ based on an observation from P_ω satisfies

$$\max_{\omega \in \{-1,+1\}^K} \sum_{\ell=1}^{K} w_\ell \Pr_\omega(\widehat{\omega}_\ell \ne \omega_\ell) \ge \frac{1}{2} \sum_{\ell=1}^{K} w_\ell (1 - \delta_\ell).$$

Proof. Pair each ω with its neighbor obtained by flipping coordinate ℓ. The sum of their testing error probabilities is at least $1 - D_{\text{TV}}(P_\omega, P_{\omega'})$. Averaging over all sign vectors gives coordinate-wise error at least $(1-\delta_\ell)/2$. Multiply by w_ℓ, sum over ℓ, and use that the maximum is at least the uniform average. Randomized tests obey the same inequality because an independent random seed does not change total variation. □

To apply the lemma to an arbitrary learner, condition on its output policy and draw $A_j \sim \widehat{\pi}_N(\cdot \mid x_j)$. Set $\widehat{\sigma}_j = +1$ if $A_j = a_2$, and set $\widehat{\sigma}_j = -1$ otherwise. Define $\widehat{\tau}_k$ analogously at each y_k. Equation (64)

shows that the weighted testing risk equals the learner's expected regret. This reduction imposes no restriction to predictors in $\mathcal{F}$ or to policies induced by such predictors. It remains to verify the theorem's budgets and the neighboring-distribution condition in Lemma G.1.

Step 3: verify coverage and exact misspecification. We now check that the instance belongs to the class in the theorem. At each x_j, both actions maximize some member of $\mathcal{F}$. At every y_k, both actions tie under every member of $\mathcal{F}$. Thus $\mathcal{A}_{\mathcal{F}}(x) = \mathcal{A}$ at every context, and

$$C_{\mathcal{F}}^{\infty} = \mathbb{E}_X \left[\max_{a \in \mathcal{A}} \frac{1}{\pi_{\mathrm{ref}}(a \mid X)} \right] = C.$$

For every deterministic policy π, including any optimal policy,

$$C(\pi) \le C.$$

This verifies the coverage-budget assertion in part (i). The precise dependence of $C^\star$ on the true reward is recorded in Remark 3 below.

Exact misspecification. For any $r^\star = r^\star_{\sigma,\tau}$ and any $f_{\sigma'} \in \mathcal{F}$,

$$\begin{aligned}\mathbb{E}_{x,a}\left[(f_{\sigma'} - r^\star)^2\right] &= \sum_{j=1}^{d} \frac{1}{2d}\frac{1}{C}\frac{R_{\max}^2\Delta^2}{4}(\sigma'_j - \sigma_j)^2 + \sum_{k=1}^{m} \frac{1}{2m}\frac{1}{C}\frac{R_{\max}^2\tau_k^2\Delta_0^2}{4} \\ &= \frac{R_{\max}^2\Delta^2}{2Cd} d_{\mathrm{H}}(\sigma', \sigma) + \frac{R_{\max}^2\Delta_0^2}{8C},\end{aligned}$$

where d_{H} denotes the unnormalized Hamming distance. The infimum is attained at f_σ, and hence

$$\varepsilon_{\mathrm{aprx}} = \frac{R_{\max}^2\Delta_0^2}{8C} = \varepsilon.$$

This completes part (i).

Uniform approximation error. The same construction gives

$$\inf_{f \in \mathcal{F}} \|f - r^\star\|_\infty = \frac{R_{\max}\Delta_0}{2} = \sqrt{2C\varepsilon}.$$

In particular, every candidate has this absolute error on every (y_k, a_2), and f_σ agrees with $r^\star_{\sigma,\tau}$ on all other pairs.

Step 4: bound the information in the grouped data. We next show that changing one sign changes the data distribution only slightly. Although the independent sampling units are context blocks, conditional independence of the rewards lets the KL divergence scale with the total number of observations $N = N_x N_a$. Let $P_{\sigma,\tau}$ denote the joint law of the grouped dataset under $r^\star_{\sigma,\tau}$. The N triplets are not generally independent, but the context–action design has the same law under every hypothesis. Conditional on this design, the reward law is a product by construction. Consequently, writing $u = (\sigma, \tau)$ and $v = (\sigma', \tau')$, the KL chain rule gives

$$\begin{aligned}D_{\mathrm{KL}}(P_u \| P_v) &= \mathbb{E}_{\mathrm{design}} \sum_{i=1}^{N_x}\sum_{j=1}^{N_a} D_{\mathrm{KL}}\left(\frac{r_u^\star(x_i, a_{i,j})}{R_{\max}}, \frac{r_v^\star(x_i, a_{i,j})}{R_{\max}}\right) \\ &= N\mathbb{E}_{x,a}\left[D_{\mathrm{KL}}\left(\frac{r_u^\star(X, A)}{R_{\max}}, \frac{r_v^\star(X, A)}{R_{\max}}\right)\right].\end{aligned}$$

The final equality uses linearity of expectation and the common marginal law of each observed context–action pair. It does not require the N triplets to be independent. For Bernoulli means $p, q \in (0, 1)$, we use

$$D_{\mathrm{KL}}(p, q) \le \frac{(p-q)^2}{q(1-q)},$$

which follows by applying $\log t \le t - 1$ to the two terms in Bernoulli KL. We also use Pinsker's inequality,

$$D_{\mathrm{TV}}(P, Q) \le \sqrt{\frac{D_{\mathrm{KL}}(P\|Q)}{2}}.$$

σ-neighbors. Write $\sigma^{\oplus j}$ for σ with its jth sign flipped. If two sign vectors differ only at coordinate $j \in [d]$, the reward laws differ only at (x_j, a_2), an event of probability $1/(2dC)$. At that pair, the Bernoulli parameters are $(1+\Delta)/2$ and $(1-\Delta)/2$. Since $\Delta^2 = Cd/(9N) \le 1/9$,

$$\begin{aligned} D_{\mathrm{KL}}(P_{\sigma,\tau} \| P_{\sigma^{\oplus j},\tau}) &\le \frac{N}{2dC} \frac{4\Delta^2}{1-\Delta^2} \\ &\le \frac{N}{2dC} \frac{4Cd}{9N} \frac{9}{8} \\ &= \frac{1}{4}. \end{aligned}$$

Therefore,

$$D_{\mathrm{TV}}(P_{\sigma,\tau}, P_{\sigma^{\oplus j},\tau}) \le \sqrt{\frac{1}{8}} < \frac{1}{2}.$$

τ-neighbors. Write $\tau^{\oplus k}$ for τ with its kth sign flipped. If the sign vectors differ only at coordinate $k \in [m]$, the reward laws differ only at (y_k, a_2), an event of probability $1/(2mC)$. Since $\Delta_0^2 = 8C\bar{\varepsilon} \le 1/4$ and $m \ge (128/3)N\bar{\varepsilon}$,

$$\begin{aligned} D_{\mathrm{KL}}(P_{\sigma,\tau} \| P_{\sigma,\tau^{\oplus k}}) &\le \frac{N}{2mC} \frac{4\Delta_0^2}{1-\Delta_0^2} \\ &\le \frac{N}{2mC} \frac{4 \cdot 8C\bar{\varepsilon}}{3/4} \\ &= \frac{64N\bar{\varepsilon}}{3m} \\ &\le \frac{1}{2}. \end{aligned}$$

Thus their total variation distance is at most $1/2$. The condition on m follows from $m \ge 43N\bar{\varepsilon}$ and $43 \ge 128/3$. When $\varepsilon = 0$, neighboring laws in the τ coordinates coincide, and the bound is immediate.

Step 5: conclude the expectation and probability bounds. Apply Lemma G.1 with $K = d + m$ and parameter $\omega = (\sigma, \tau)$. Use weights

$$(\alpha_1, \dots, \alpha_d, \beta_1, \dots, \beta_m)$$

and $\delta_\ell = 1/2$ for every coordinate. Use the sign estimator from Step 2, whose weighted testing risk equals the expected policy regret. Hence, for every learner,

$$\begin{aligned} \max_{\sigma,\tau} \mathbb{E}_{\sigma,\tau} \left[\mathrm{Reg}_{\sigma,\tau}(\widehat{\pi}_N) \right] &\ge \frac{1}{4} \left(\sum_{j=1}^{d} \alpha_j + \sum_{k=1}^{m} \beta_k \right) \\ &= \frac{S}{4} \\ &= \frac{R_{\max}(\Delta + \Delta_0)}{16} \\ &= \frac{R_{\max}}{48} \sqrt{\frac{Cd}{N}} + \frac{\sqrt{C\varepsilon}}{4\sqrt{2}} \\ &\ge \frac{1}{48} \left(\sqrt{C\varepsilon} + R_{\max} \sqrt{\frac{Cd}{N}} \right). \end{aligned}$$

Substituting $d = \log_2 |\mathcal{F}|$ and taking the infimum over learners proves part (ii). The argument also allows the learner to know $(\mathcal{X}, \mathcal{A}, \rho, \pi_{\mathrm{ref}}, \mathcal{F}, \mathcal{R}^\star, \varepsilon, N_x, N_a)$ in advance, but not the unknown true reward.

Constant-probability lower bound. Fix a learner, choose $(\sigma^\star, \tau^\star)$ attaining the preceding maximum, and write

$$Z = \mathrm{Reg}_{\sigma^\star,\tau^\star}(\widehat{\pi}_N).$$

The randomness in Z includes the dataset and any internal randomness used to produce the policy. Action randomization is already averaged in the definition of policy regret. By Eqn. (64), $0 \le Z \le S$ almost surely and $\mathbb{E}[Z] \ge S/4$. For every $t \in (0, S)$,

$$\begin{aligned}\mathbb{E}[Z] &\le t \Pr(Z < t) + S \Pr(Z \ge t) \\ &\le t + S \Pr(Z \ge t).\end{aligned}$$

Taking $t = S/8$ gives

$$\Pr(Z \ge S/8) \ge \frac{S/4 - S/8}{S} = \frac{1}{8}.$$

Moreover,

$$\begin{aligned}\frac{S}{8} &= \frac{R_{\max}(\Delta + \Delta_0)}{32} \\ &= \frac{R_{\max}}{96}\sqrt{\frac{Cd}{N}} + \frac{\sqrt{C\varepsilon}}{8\sqrt{2}} \\ &\ge \frac{1}{96}\left(\sqrt{C\varepsilon} + R_{\max}\sqrt{\frac{Cd}{N}}\right).\end{aligned}$$

Taking the infimum over learners proves part (iii). This completes the proof.

Remark 3 (The expectation-form coverage coefficient)**.** *The behavior policy is identical across contexts, so the constructed family satisfies $C_{\mathcal{F}}^{\infty} = C$. For an optimal deterministic policy, write*

$$t = \rho\{x : \pi^\star(x) = a_2\}.$$

Then

$$C^\star = tC + (1-t)\frac{C}{C-1} \le C.$$

For $\varepsilon > 0$, the optimal actions are unique and

$$t = \frac{1}{2d}\sum_{j=1}^{d} \mathbf{1}\{\sigma_j = +1\} + \frac{1}{2m}\sum_{k=1}^{m} \mathbf{1}\{\tau_k = +1\}.$$

Thus $C^\star$ varies with the true reward, whereas $C_{\mathcal{F}}^{\infty}$ does not. For example, when $\varepsilon > 0$ and all signs are negative, $C^\star = C/(C-1)$ while $C_{\mathcal{F}}^{\infty} = C$, so the two coefficients can differ arbitrarily as C grows. This observation does not assert that this particular member attains the minimax lower bound: a learner that always chooses a_1 is optimal on it. The theorem establishes a worst-case lower bound over the class $C^\star \le C$, not a hard-instance lower bound with prescribed $C^\star = C \ll C_{\mathcal{F}}^{\infty}$.

H Deferred Experimental Details

Here, we provide the missing details from Section 4.3. We first give the complete construction and diagnostic results for the controlled synthetic experiment, then describe the LLM response-selection experiment, and finally present the downstream MCTS mathematical-reasoning experiment as promised. The first two experiments use the squared-loss VR and VDR objectives studied in the main text. The MCTS experiment instead uses a binary-preference variant of VDR and provides a complementary evaluation of relative supervision over absolute supervision.

H.1 Controlled Synthetic Linear-Bandit Experiment

Section 4.3 identifies two effects that can decide between VR and VDR for linear classes. First, even when the model is well specified, VDR learns only from within-context differences and can be hurt when the behavior policy rarely compares the actions that matter. Second, an action-independent nuisance $b^\star$ misspecifies VR but cancels from every VDR target. This experiment constructs a family of linear contextual bandits in which the two effects are controlled by separate parameters: the cross-face probability δ for the first and the nuisance magnitude β for the second. A third parameter, the alignment ρ, controls the direction of the true value. We first specify the problem, then explain which quantities of the theory each parameter controls, and finally describe the estimators, protocol, and results. Throughout this subsection, ρ and δ denote these experimental parameters, not the context distribution or the confidence level of the main text.

Contexts, actions, and features. Let $d_x = 5$ and $d_a = 3$. The context space is the unit sphere $\mathcal{X} = \mathbb{S}^{d_x-1} = \{x \in \mathbb{R}^{d_x} : \|x\|_2 = 1\}$, and contexts are drawn from the uniform distribution $\rho = \mathrm{Unif}(\mathbb{S}^{d_x-1})$. The context space is therefore continuous, and each data set contains N_x independent draws from it; we sample $X = Z/\|Z\|_2$ with $Z \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d_x})$. The action space consists of the eight normalized vertices of the cube in $\mathbb{R}^{d_a}$,

$$\mathcal{A} = \left\{ a = \tfrac{1}{\sqrt{3}}(s_1, s_2, s_3)^\top : s_1, s_2, s_3 \in \{-1, +1\} \right\},$$

and we identify each action with its vector, writing $s_k(a) \coloneqq \mathrm{sign}(a_k)$. We use the bilinear feature map and the corresponding linear class

$$\boldsymbol{\phi}(x, a) = \mathrm{vec}(xa^\top) \in \mathbb{R}^d, \qquad f_{\boldsymbol{W}}(x, a) = \langle \boldsymbol{\phi}(x, a), \mathrm{vec}(\boldsymbol{W}) \rangle = x^\top \boldsymbol{W} a, \qquad \boldsymbol{W} \in \mathbb{R}^{d_x \times d_a},$$

so that $d = d_x d_a = 15$.

Reward model. In each independently generated problem instance, we draw $\boldsymbol{u}^\star \sim \mathrm{Unif}(\mathbb{S}^{d_x-1})$, independently of all contexts, and keep it fixed within the instance. For $\rho \in [0, 1]$, let

$$\boldsymbol{c}_\rho = \sqrt{\rho}\, \boldsymbol{e}_1 + \sqrt{1-\rho}\, \boldsymbol{e}_2 \in \mathbb{R}^{d_a}, \qquad \boldsymbol{W}^\star_\rho = \sqrt{d_x d_a}\, \boldsymbol{u}^\star \boldsymbol{c}_\rho^\top, \qquad f^\star \coloneqq f_{\boldsymbol{W}^\star_\rho}.$$

Thus $f^\star(x, a) = \sqrt{15}\, (x^\top \boldsymbol{u}^\star)(\boldsymbol{c}_\rho^\top a)$ depends on the first action coordinate when $\rho = 1$ and on the second when $\rho = 0$. The factor $\sqrt{d_x d_a}$ normalizes the signal: $\mathbb{E}[f^\star(X, A)^2] = 1$ when $X \sim \mathrm{Unif}(\mathbb{S}^{d_x-1})$ and A is uniform on $\mathcal{A}$, and also under the behavior policy below. The reward is $r^\star = f^\star + b^\star$, where the nuisance $b^\star(x) = \beta h(x)$ with $\beta \geq 0$ does not depend on the action and

$$h(x) = \frac{x_1^2 - 1/d_x}{\sqrt{2(d_x - 1)/(d_x^2(d_x + 2))}}.$$

The constants are chosen so that $\mathbb{E}[h(X)] = 0$ and $\mathbb{E}[h(X)^2] = 1$ for $X \sim \mathrm{Unif}(\mathbb{S}^{d_x-1})$, so β is the root mean square of $b^\star$ under the context distribution. This reward satisfies Assumption 1 with $f^\star \in \mathcal{F}_{\mathrm{lin}}$ for every β. However, $\mathcal{F}_{\mathrm{lin}}$ contains no nonzero action-independent function (because $x^\top \boldsymbol{W} a = -x^\top \boldsymbol{W}(-a)$ and $-a \in \mathcal{A}$), so $r^\star \in \mathcal{F}_{\mathrm{lin}}$ if and only if $\beta = 0$. In other words, the model is well specified for VR exactly when $\beta = 0$.

Behavior policy and data. A *face* of the cube is the set of four actions sharing the same first sign s_1. Each context favors the face $\tau(x) \coloneqq \mathrm{sign}(x_1) \in \{-1, +1\}$ (with $\mathrm{sign}(0) \coloneqq 1$, a probability-zero event). For $\delta \in [0, 1/2]$, the behavior policy is

$$\pi_{\mathrm{ref}}(a \mid x) = \begin{cases} (1-\delta)/4, & s_1(a) = \tau(x), \\ \delta/4, & s_1(a) = -\tau(x). \end{cases}$$

That is, $A \sim \pi_{\mathrm{ref}}(\cdot \mid x)$ lies on the favored face with probability $1-\delta$ and on the opposite face with probability δ, and, given its face, $s_2(A)$ and $s_3(A)$ are independent uniform signs. Every action has positive probability whenever $\delta > 0$; the case $\delta = 0$, in which the opposite face is never observed, is included only as a limiting case. Following Section 1, each data set consists of $x_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathbb{S}^{d_x-1})$ for $i \in [N_x]$, $a_{i,j} \overset{\text{i.i.d.}}{\sim} \pi_{\mathrm{ref}}(\cdot \mid x_i)$ for $j \in [N_a]$, and

$$r_{i,j} = f^\star(x_i, a_{i,j}) + \beta h(x_i) + \eta_{i,j}, \qquad \eta_{i,j} = \sigma \xi_{i,j},$$

where the $\xi_{i,j}$ are independent Rademacher signs, independent of all contexts and actions. We use $(N_x, N_a) = (200, 8)$ and $\sigma = 0.5$.

We now explain δ, ρ, and β. Below, $X \sim \mathrm{Unif}(\mathbb{S}^{d_x-1})$ and $A, A' \overset{\text{i.i.d.}}{\sim} \pi_{\mathrm{ref}}(\cdot \mid X)$.

Role of δ: design constants. The conditional action moments are

$$\mathbb{E}[A \mid X = x] = \frac{(1-2\delta)\tau(x)}{\sqrt{3}} \boldsymbol{e}_1, \qquad \mathrm{Cov}(A \mid X = x) = \frac{1}{3}\,\mathrm{diag}\big(4\delta(1-\delta), 1, 1\big) =: \boldsymbol{D}_\delta. \quad (65)$$

Hence $\mathbb{E}[AA^\top \mid X] = \boldsymbol{I}_{d_a}/d_a$ for every δ, so that $\boldsymbol{\Sigma} = \boldsymbol{I}_d/d$ and $\|\boldsymbol{z}(X,A)\|_2 = \sqrt{d}$: the VR design constant is $\alpha_0 = 1$ in Assumption 2, regardless of δ. The centered covariance, in contrast, is $\boldsymbol{\Sigma}_{\mathrm{C}} = \boldsymbol{D}_\delta \otimes (\boldsymbol{I}_{d_x}/d_x)$ (in the ordering of vec), and its eigenvalue along the first action coordinate is proportional to $4\delta(1-\delta)$. For $\delta > 0$, we have $d_{\mathrm{C}} = d$, and the smallest constant admissible in Assumption 3 is

$$\alpha_{\mathrm{C}} = \sqrt{\frac{8 + 1/(\delta(1-\delta))}{6}} = \Theta(\delta^{-1/2}) \quad \text{as } \delta \to 0.$$

At $\delta = 0$, the first action coordinate lies entirely in $\ker(\boldsymbol{\Sigma}_{\mathrm{C}})$ and $d_{\mathrm{C}} = 10$. This is a contextual version of the first example in Section 4.3. Note that small δ does not mean little within-context variation: the total variation $\mathrm{tr}\,\mathrm{Cov}(A \mid X) = \frac{2}{3} + \frac{4}{3}\delta(1-\delta)$ is at least $2/3$ for every δ, but almost all of it lies in the second and third action coordinates.

Role of ρ: alignment with within-context variation. Whether the first coordinate matters depends on the direction of the true value. Let $m(x) := \mathbb{E}[f^\star(x, A) \mid X = x]$, so that the centered value is $\mathcal{C}f^\star(x,a) = f^\star(x,a) - m(x)$. Since $\mathbb{E}[f^\star(X,A)^2] = 1$, the conditional mean and centered component account for the following fractions of the target's second moment:

$$\kappa_{\mathrm{mean}}(\rho, \delta) := \mathbb{E}[m(X)^2] = \rho(1-2\delta)^2,$$
$$\kappa_{\mathrm{diff}}(\rho, \delta) := \mathbb{E}\big[(\mathcal{C}f^\star(X,A))^2\big] = \frac{1}{2}\mathbb{E}\big[(f^\star(X,A) - f^\star(X,A'))^2\big] = 1 - \rho(1-2\delta)^2, \quad (66)$$

which satisfy $\kappa_{\mathrm{mean}} + \kappa_{\mathrm{diff}} = 1$. Because the VDR loss depends on the data only through within-context reward differences, κ_{diff} is the fraction of the signal that VDR observes; the remaining fraction κ_{mean} is visible only through absolute reward levels, which VR also uses. When $\rho = 1$, the optimal face at x is $\mathrm{sign}(x^\top \boldsymbol{u}^\star)$, which differs from the favored face $\tau(x)$ for a constant fraction of contexts, yet only $\kappa_{\mathrm{diff}} = 4\delta(1-\delta)$ of the signal distinguishes the two faces within a context. At $(\rho, \delta) = (1, 0)$, $\kappa_{\mathrm{diff}} = 0$: all observed actions at a context have the same true value, and VDR sees only noise.

Role of β: misspecification of VR. The nuisance affects VR only through its correlation with the features. By (65) and symmetry,

$$\mathbb{E}[b^\star(X)\boldsymbol{\phi}(X,A)] = \frac{\beta(1-2\delta)}{\sqrt{3}}\,\mathbb{E}\big[h(X)\,|X_1|\big]\ \mathrm{vec}(\boldsymbol{e}_1\boldsymbol{e}_1^\top),$$

where $\mathbb{E}[h(X)|X_1|] > 0$ because h is centered and increasing in $|X_1|$. Hence the population VR fit is $\boldsymbol{W}_\rho^\star$ plus a multiple of $\boldsymbol{e}_1\boldsymbol{e}_1^\top$ proportional to $\beta(1-2\delta)$. This shift acts on the first action coordinate, which is exactly the coordinate that separates the two faces when $\rho = 1$, and it vanishes only when $\beta = 0$ or $\delta = 1/2$. The same nuisance cancels from every VDR target.

Together, these calculations suggest two regimes at $\rho = 1$. When $\beta = 0$ and δ is small, VR is well specified with $\alpha_0 = 1$, whereas VDR has a large α_{C} and observes little of the signal, so VR should be favored. As β grows, the bias of VR grows in proportion while VDR is unaffected, so VDR should eventually be favored.

Estimators. Let $L_{\mathrm{VR}}(\boldsymbol{W}) \coloneqq \widehat{\mathcal{L}}(f_{\boldsymbol{W}})$ with $\mathcal{T}^{\mathrm{VR}}$ as in Section 2, and let $L_{\mathrm{VDR}}(\boldsymbol{W}) \coloneqq \frac{1}{2}\widehat{\mathcal{L}}(f_{\boldsymbol{W}})$ with $\mathcal{T}^{\mathrm{VDR}}$. Writing $e_{i,j}(\boldsymbol{W}) \coloneqq f_{\boldsymbol{W}}(x_i, a_{i,j}) - r_{i,j}$ and $\bar{e}_i(\boldsymbol{W}) \coloneqq N_a^{-1}\sum_j e_{i,j}(\boldsymbol{W})$,

$$L_{\mathrm{VR}}(\boldsymbol{W}) = \frac{1}{N_x N_a}\sum_{i=1}^{N_x}\sum_{j=1}^{N_a} e_{i,j}(\boldsymbol{W})^2, \qquad L_{\mathrm{VDR}}(\boldsymbol{W}) = \frac{1}{N_x(N_a-1)}\sum_{i=1}^{N_x}\sum_{j=1}^{N_a}\big(e_{i,j}(\boldsymbol{W}) - \bar{e}_i(\boldsymbol{W})\big)^2, \tag{67}$$

where the second expression follows from $\sum_{j<k}(e_{i,j} - e_{i,k})^2 = N_a\sum_j(e_{i,j} - \bar{e}_i)^2$. The extra factor $1/2$ in L_{VDR} does not change its minimizer; it gives the two losses the same expected value, σ^2, when evaluated at the true reward with $\beta = 0$, so that they are on a common scale when mixed below. The VR and VDR estimators of Sections 4.1 and 4.2 minimize these losses over $\boldsymbol{W}$; both are computed in closed form from the corresponding sufficient statistics.

Data-driven selection between VR and VDR. Since neither objective dominates, a practitioner who does not know β or δ would want to choose between them from the data. We therefore also consider the mixed objective

$$\widehat{\boldsymbol{W}}_\alpha \in \arg\min_{\boldsymbol{W}}\left\{(1-\alpha)L_{\mathrm{VR}}(\boldsymbol{W}) + \alpha L_{\mathrm{VDR}}(\boldsymbol{W}) + \lambda_\alpha\|\boldsymbol{W}\|_F^2\right\}, \qquad \alpha \in \{0, 0.05, \ldots, 0.95, 1\},$$

which interpolates between VR ($\alpha = 0$) and VDR ($\alpha = 1$). Here

$$\lambda_\alpha = 10^{-8}\max\left\{\frac{\mathrm{tr}(\boldsymbol{G}_\alpha)}{d}, 1\right\}, \qquad \boldsymbol{G}_\alpha \coloneqq (1-\alpha)\boldsymbol{G}_{\mathrm{VR}} + \alpha\boldsymbol{G}_{\mathrm{VDR}},$$

where $\boldsymbol{G}_{\mathrm{VR}}$ and $\boldsymbol{G}_{\mathrm{VDR}}$ are the empirical Gram matrices of the two quadratic losses in (67). The ℓ_2-regularization serves only to stabilize the linear solve; in our design $\mathrm{tr}(\boldsymbol{G}_\alpha)/d < 1$, so $\lambda_\alpha = 10^{-8}$ in practice. The same term is used at the endpoints $\alpha \in \{0, 1\}$, so the reported VR and VDR fits are unregularized up to this negligible ridge. We choose α by five-fold cross-validation (CV) over contexts. Specifically, we partition the $N_x = 200$ context blocks into five folds of 40 contexts each; for each fold, we fit $\widehat{\boldsymbol{W}}_\alpha$ on the other four folds and compute L_{VDR} on the held-out fold. The selected α minimizes the average of the five held-out losses, and the reported estimator is $\widehat{\boldsymbol{W}}_\alpha$ at the selected α, fitted on all N_x contexts. Folds are formed over whole context blocks so that held-out actions never share a context with training actions. We validate with the VDR loss rather than the VR loss because the held-out VR loss contains the nuisance $b^\star$ and would reward fits that absorb it, whereas the held-out VDR loss is unaffected by $b^\star$ and, up to scaling and a noise constant, measures the centered prediction error that controls regret in Proposition B.1. We call this rule DIFFERENCE-CV (CV-best in Figure 3). As a reference that uses information unavailable in practice, ORACLE chooses α to minimize the true regret, computed with $\boldsymbol{W}_\rho^\star$, on 2,000 additional simulated contexts.

Evaluation and protocol. Each fitted matrix $\widehat{\boldsymbol{W}}$ defines the greedy policy $x \mapsto \arg\max_{a\in\mathcal{A}} f_{\widehat{\boldsymbol{W}}}(x, a)$. We estimate its regret (1) on 10,000 fresh contexts by enumerating all eight actions; since $b^\star$ does not depend on the action, it does not affect the regret. We run two experiments.

- **Crossover grid.** With $\rho = 1$, we vary $\delta \in \{0, 0.02, 0.05, 0.1, 0.2, 0.5\}$ and $\beta \in \{0, 0.25, 0.5, 1, 2\}$ to test the two regimes predicted above.
- **Alignment sweep.** With $(\delta, \beta) = (0.02, 0)$ fixed, we vary $\rho \in \{0, 0.25, 0.5, 0.75, 0.9, 1\}$. The behavior policy, and hence the total within-context variation, is unchanged; only the direction of the true value moves toward the weak first coordinate. This tests whether what matters is the within-context variation *in the direction of the true value* (measured by κ_{diff}) rather than its total amount.

For each condition, we generate eight independent problem instances and 20 training data sets per instance. Repetitions are first averaged within each problem instance, and the reported 95% intervals are computed from the eight independent instance means rather than by treating repetitions within an instance as independent.

Crossover results. Let $\Delta_{\mathrm{Reg}} = \mathrm{Reg}_{\mathrm{VR}} - \mathrm{Reg}_{\mathrm{VDR}}$, so that negative values favor VR and positive values favor VDR. Table 2 reports its mean over the grid, and the top-left panel of Figure 2 shows the same values as a heat map.

Both predicted regimes appear. When $\beta = 0$, VR has lower regret in all eight problem instances for every tested $\delta > 0$, so the VR-favored regime is not an artifact of the degenerate case $\delta = 0$; the advantage shrinks as δ grows, consistent with α_{C} decreasing and κ_{diff} increasing. For each $\delta \in \{0.02, 0.05, 0.1, 0.2\}$, the sign of Δ_{Reg} flips as β grows. At $\delta = 0.02$, for example, the paired mean difference and its 95% interval are -0.0040 ± 0.0005 at $\beta = 0$, $+0.0014 \pm 0.0014$ at $\beta = 0.25$, and $+0.0188 \pm 0.0050$ at $\beta = 0.5$. Over the same range, the population bias of VR grows as 0, 0.129, 0.259, 0.518, and 1.036 for $\beta = 0, 0.25, 0.5, 1, 2$, i.e., proportional to β as predicted. Here the population bias is the root-mean-square centered error of the population VR fit (the minimizer of the expected VR loss under π_{ref}, with the same negligible ridge), computed on the 2,000 validation contexts, where predictions are centered by their mean over all eight actions;[8] averaged over the eight problem instances. The row $\delta = 0.5$ is nearly flat because the VR bias vanishes at $\delta = 1/2$.

The row $\delta = 0$ is qualitatively different. There, the opposite face is never observed, and at $\rho = 1$ the centered target $\mathcal{C}f^\star$ is identically zero because $s_1(A) = \tau(X)$ almost surely. VDR therefore recovers $\mathcal{C}f^\star$ but obtains no information about the cross-face value gaps, which lie outside the support of π_{ref} and are exactly what the greedy policy needs; accordingly, the regret guarantee of Proposition B.1 is vacuous in this case. VR, in contrast, still identifies $\boldsymbol{W}_\rho^\star$ because $\boldsymbol{\Sigma} \succ \mathbf{0}$: the favored face varies with the context, so absolute reward levels across contexts reveal the first action coordinate. We include this row to verify the construction, but center our discussions in the regime $\delta \geq 0.02$.

Data-driven selection results. The remaining panels of Figure 2 evaluate DIFFERENCE-CV. For $\delta > 0$, its selected α follows the same pattern as the ORACLE choice: it leans toward VR when $\beta = 0$ and moves to $\alpha \approx 1$ as β grows. Its regret is then close to that of the better of VR and VDR throughout the grid. At $\delta = 0$, however, DIFFERENCE-CV falls behind the better endpoint, increasingly so for larger β. This is expected: its validation loss is itself a VDR loss and cannot assess a direction that no observed difference reveals.

Alignment sweep results. In this experiment, the behavior policy and the nuisance are fixed, and only the direction of the true value changes. At $\rho = 0$, the true value varies only within each face, $\kappa_{\mathrm{diff}} = 1$, and both methods have mean regret 0.0003. For every $\rho > 0$, part of the true value moves into the weak first coordinate, and VR has lower mean regret than VDR (Figure 3, left). As ρ increases from 0 to 1, κ_{diff} decreases from 1 to 0.0784 and κ_{mean} increases from 0 to 0.9216, matching (66) (right panel). The results are consistent with a growing estimation disadvantage for VDR as the true value becomes more aligned with the weakly sampled first action direction.

Summary. The crossover and alignment experiments illustrate the geometric bias–variance tradeoff discussed in Section 4.3. In this construction, VR benefits from absolute reward levels under correct specification when cross-face comparisons are rare and the true value is aligned with the first action direction. For fixed $\delta < 1/2$, its population coefficient bias grows linearly with the nuisance magnitude β, whereas the VDR objective is unchanged. The observed regret crossover shows how increasing misspecification can outweigh the estimation advantage of using absolute levels. Across the tested grid with $\delta > 0$, cross-validation with the VDR loss achieves regret close to that of the better objective without knowing which regime applies.

δ	$\beta = 0$	$\beta = 0.25$	$\beta = 0.5$	$\beta = 1$	$\beta = 2$
0.00	−0.8390	−0.8330	−0.8154	−0.7472	−0.5686
0.02	−0.0040	+0.0014	+0.0188	+0.0813	+0.2485
0.05	−0.0013	+0.0036	+0.0189	+0.0737	+0.2298
0.10	−0.0007	+0.0033	+0.0157	+0.0592	+0.1978
0.20	−0.0002	+0.0019	+0.0090	+0.0340	+0.1250
0.50	−0.0001	+0.0000	+0.0002	+0.0012	+0.0045

Table 2: Mean greedy-regret difference Δ_{Reg} in the crossover grid ($\rho = 1$). Negative values favor VR and positive values favor VDR.

[8] This uniform centering differs from the centering operator $\mathcal{C}$ of Section 1, which subtracts the mean under $\pi_{\mathrm{ref}}(\cdot \mid x)$. Both remove action-independent terms and preserve all action gaps, so neither affects the regret. The population VR bias derived in the paragraph on the role of β is a multiple of $\boldsymbol{e}_1\boldsymbol{e}_1^\top$ proportional to $\beta(1 - 2\delta)$, so its centered magnitude is proportional to β under either centering; only the constant differs.

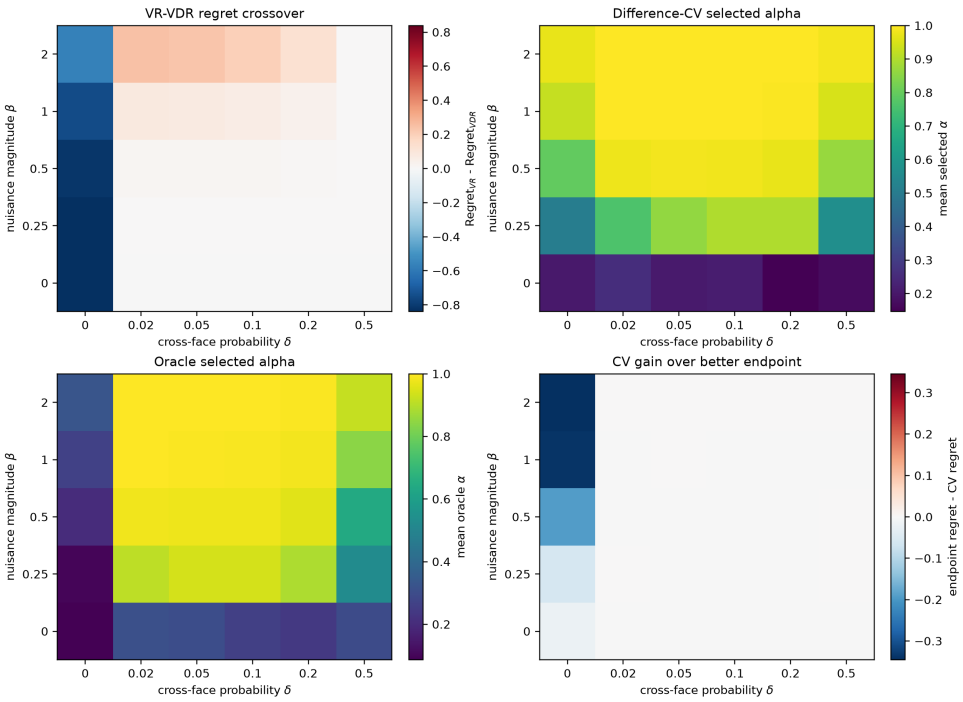

Figure 2: Crossover grid ($\rho = 1$). Top left: $\mathrm{Reg}_{\mathrm{VR}} - \mathrm{Reg}_{\mathrm{VDR}}$; negative values favor VR and positive values favor VDR. Top right and bottom left: mean α selected by DIFFERENCE-CV and by ORACLE. Bottom right: regret of the better of VR and VDR minus that of DIFFERENCE-CV; positive values mean that DIFFERENCE-CV improves on both.

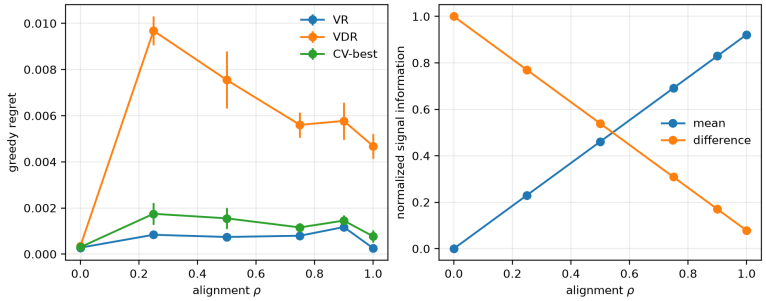

Figure 3: Alignment sweep at fixed $(\delta, \beta) = (0.02, 0)$. **Left:** greedy-policy regret as the true value direction moves toward the weak first action coordinate; CV-best denotes DIFFERENCE-CV. **Right:** the fractions κ_{mean} ("mean") and κ_{diff} ("difference") of the target's second moment contributed by its conditional mean and centered component, respectively; see (66).

H.2 LLM RESPONSE-SELECTION EXPERIMENT

Data and representations. We curate 1,000 English prompts from WildChat (Zhao et al., 2024), UltraFeedback (Cui et al., 2024), GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), MBPP (Austin et al., 2021), HelpSteer2 (Wang et al., 2024b), and TL;DR (Stiennon et al., 2020), and cache 1,000 responses per prompt generated by Llama-3.2-3B-Instruct (Meta, 2024). Each response is represented by a 3,072-dimensional feature vector obtained by mean-pooling the generator's final-layer hidden states over its response tokens. Keeping the prompts, responses, and representations fixed, we repeat the experiment with reward models Skywork-Reward-V2-Llama-3.1-8B and Skywork-Reward-V2-Qwen3-8B (Liu et al., 2026), and ArmoRM-Llama3-8B-v0.1 (Wang et al., 2024a).

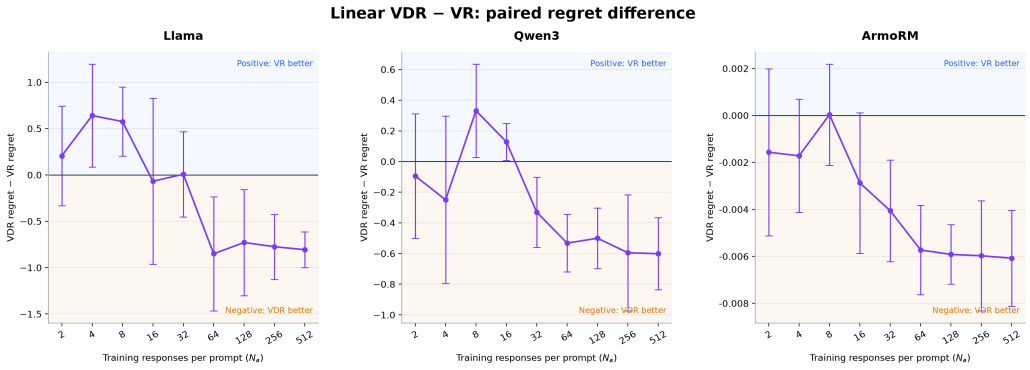

Figure 4: Paired regret differences on the 200 evaluation prompts, $\mathrm{Reg}_{\mathrm{VDR}} - \mathrm{Reg}_{\mathrm{VR}}$, with pointwise 95% Student-t confidence intervals over five seeds. Negative values favor VDR.

Training protocol. For each of five seeds, we split the prompts into $N_x = 800$ training prompts and 200 evaluation prompts that are not used for fitting or selecting the regularization strengths. We vary the number of labeled responses per training prompt over $N_a \in \{2, 4, 8, 16, 32, 64, 128, 256, 512\}$. Within each seed and response budget, VR and VDR use the same sampled responses. To stabilize estimation in the 3,072-dimensional feature space, we use ℓ_2-regularized versions of the objectives in Sections 4.1 and 4.2. Writing $\boldsymbol{\phi}_{i,j} = \boldsymbol{\phi}(x_i, a_{i,j})$, the estimators minimize

$$(\widehat{\boldsymbol{\theta}}_{\mathrm{VR}}, \widehat{c}) \in \arg\min_{\boldsymbol{\theta}, c} \left\{ \sum_{i=1}^{N_x} \sum_{j=1}^{N_a} \left(r_{i,j} - \boldsymbol{\theta}^\top \boldsymbol{\phi}_{i,j} - c\right)^2 + \lambda_{\mathrm{VR}} \|\boldsymbol{\theta}\|_2^2 \right\},$$

$$\widehat{\boldsymbol{\theta}}_{\mathrm{VDR}} \in \arg\min_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{N_x} \sum_{(j,k) \in \mathcal{P}_i} \left[(r_{i,j} - r_{i,k}) - \boldsymbol{\theta}^\top (\boldsymbol{\phi}_{i,j} - \boldsymbol{\phi}_{i,k})\right]^2 + \lambda_{\mathrm{VDR}} \|\boldsymbol{\theta}\|_2^2 \right\},$$

where $\mathcal{P}_i$ is the set of within-prompt pairs used for prompt x_i. VDR uses all $\binom{N_a}{2}$ pairs when $N_a \leq 32$ and samples $16N_a$ pairs without replacement otherwise. Features are standardized before fitting. We select λ_{VR} and λ_{VDR} separately for each reward model, method, and response budget using validation regret on a 640/160 split of the training prompts, and then refit on all 800 training prompts. The VR intercept c is not penalized; VDR requires no intercept because it cancels under differencing.

Evaluation and results. Each fitted scorer ranks all 1,000 cached responses for every evaluation prompt and selects the response with the highest predicted score. For an evaluation prompt x, offline regret is the difference between the reward of its best cached response and that of the selected response. We average this quantity over the 200 evaluation prompts. The primary figure reports means and pointwise 95% Student-t confidence intervals over the five seeds.

Because both methods use the same train–test split and sampled responses within each seed, we also compute the seed-wise paired difference $\mathrm{Reg}_{\mathrm{VDR}} - \mathrm{Reg}_{\mathrm{VR}}$. Figure 4 reports its mean and pointwise 95% Student-t confidence interval. For every $N_a \geq 64$, these intervals lie below zero under all three reward models.

H.3 MCTS-Based Mathematical Reasoning Experiment

We complement our theoretical results with an LLM-based mathematical reasoning experiment, examining whether relative supervision can improve downstream decision quality even when numerical rewards are available. We compare VR with a binary-preference variant of VDR (which we refer to VDR for simplicity) for guiding Monte Carlo tree search (MCTS). Following the process-preference approach of Guan et al. (2025), we train reward models using either search-derived Q-values or within-context preferences constructed from those values. Rather than evaluating selection from a

fixed candidate pool, we measure how these learned models guide a sequence of reasoning steps toward correct final answers.

Data, labels, and representations. We use 1,000 competition mathematics problems from the American Mathematics Competitions (Mathematical Association of America, 2024): 800 AMC12 problems from the 2002–2024 contests and 200 AIME problems from the 2000–2024 contests.[9] A context x_i contains a problem and its reasoning prefix; $a_{i,j}$ is a candidate next step with label $r_{i,j} \coloneqq Q(x_i, a_{i,j})$. Labels are taken from an eight-rollout data-generation snapshot.

During data generation, final-outcome rewards ($+1$ for a correct answer and -1 for a recorded unsuccessful outcome) are backed up through the search tree. For a visited node v, the Q-value estimates eventual success under the generating policy and budget, and is defined as

$$Q(v) = \frac{S(v)}{n(v)} = \frac{N_+(v) - N_-(v)}{N_+(v) + N_-(v)} = 2\widehat{p}(v) - 1,$$

where $S(v)$ and $n(v)$ are the reward sum and visit count; N_+ and N_- count successful and unsuccessful continuations, and $\widehat{p}$ is their success frequency.

For each candidate step a, we encode the parent context x (the problem, preceding reasoning steps, and prior execution observations) followed by a in the generation format, excluding a's own execution output. We extract the final-layer hidden state at the last nonpadding token of frozen Qwen2.5-Coder-7B-Instruct (Hui et al., 2024). The resulting 3,584-dimensional vectors are standardized to obtain $\phi(x, a)$, using statistics fitted on training candidates during model selection and on training-plus-validation candidates for the final refit.

Training objectives and protocol. A trainable linear or one-hidden-layer GELU head maps $\phi(x, a)$ to a scalar score $f_\theta(x, a)$, while the Qwen backbone remains frozen. For VDR, this encoder–head combination is the process preference model (PPM); paired candidates are scored separately by the same head. VR learns an unpenalized global intercept; VDR fixes it to zero.

For each context x_i, let P_i contain preferred–rejected pairs, with $a_{i,j}$ preferred to $a_{i,k}$ for $(j, k) \in P_i$. Let E_i denote their unique endpoint indices, $N_{a,i} = |E_i|$, and m_i^{pos} and m_i^{neg} denote the dataset-recorded positive and negative role counts. VR regresses the numerical labels $r_{i,j} = Q(x_i, a_{i,j})$, whereas VDR learns pairwise orderings using the ranking loss of Guan et al. (2025), with dataset-specific normalization: with $\sigma(z) = (1 + e^{-z})^{-1}$,

$$\widehat{L}_{\text{VR}}(f_\theta) = \frac{1}{N_x} \sum_{i=1}^{N_x} \frac{1}{N_{a,i}} \sum_{j \in E_i} \big(f_\theta(x_i, a_{i,j}) - r_{i,j}\big)^2,$$

$$\widehat{L}_{\text{VDR}}(f_\theta) = -\frac{1}{N_x} \sum_{i=1}^{N_x} \frac{1}{m_i^{\text{pos}} m_i^{\text{neg}}} \sum_{(j,k) \in P_i} \log \sigma\big(f_\theta(x_i, a_{i,j}) - f_\theta(x_i, a_{i,k})\big).$$

We train the linear heads using ridge/SVD with penalty $\lambda\|w\|_2^2$ for VR, and L-BFGS (Liu & Nocedal, 1989) with $(\lambda/2)\|w\|_2^2$ penalty for VDR; both select $\lambda = 1$. MLPs use AdamW (Loshchilov & Hutter, 2019), averaging context-wise errors for VR and summing role-count-weighted pair losses for VDR. We select the hyperparameters by validation loss, averaged over three training seeds for MLPs, then refit on training plus validation data.

We evaluate all 100 test problems with Qwen2.5-Coder-7B-Instruct using one MCTS tree, two iterations, four children per expansion, maximum depth 16, and token limits of 1,024 per step and 16,384 per context.

Results and discussion. VDR achieves higher observed accuracy than VR with both linear heads (16% versus 15%) and MLPs (13% versus 10%; Table 3).

The synthetic experiment in Appendix H.1 examines the bias–variance tradeoff by separately controlling nuisance magnitude and the geometry of within-context comparisons. The response-selection

[9] Problem statements are taken from the public archive at https://live.poshenloh.com/past-contests, and answer keys are cross-checked against published solution compilations.

Table 3: Final-answer accuracy on 100 test problems under a common MCTS budget. Each method uses one search seed; MLPs use one prespecified refit.

Reward head	Accuracy (%)
Linear VR	15
Linear VDR	16
MLP VR	10
MLP VDR	13

experiment in Appendix H.2 compares the objectives on real data without isolating the mechanism responsible for their relative performance. The present experiment asks whether preferences derived from numerical feedback can guide search toward correct final answers. Its observed gains provide additional evidence that relative supervision can be useful even when numerical rewards are available.