A QUADRATIC LOWER BOUND FOR TWO-STAGE PREFERENTIAL G-OPTIMAL DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

In preference learning, G-optimal design, which minimizes the largest prediction variance over candidate responses, guides comparison allocation across prompts and response pairs. We study a two-stage strategy that first optimizes response distributions separately at each prompt, then allocates the prompt budget with those distributions fixed. For this strategy, a quadratic upper bound in the feature dimension was known, but whether prompt allocation could always recover linear scaling remained open. We establish a matching worst-case lower bound, showing that prompt allocation cannot always recover linear scaling. Joint allocation on the same instances attains the optimal linear value. A geometric decomposition explains this gap through within-prompt response concentration and cross-prompt covariance coverage, and yields linear scaling for uniformly comparable prompt covariances. The Bradley–Terry model assigns preference probabilities through a logistic function of response-score differences. With independent response pairs, these instances yield an analogous quadratic-versus-linear separation for uniform within-prompt score estimation at fixed parameter radius and sufficiently small target error.

1 INTRODUCTION

Efficient preference learning depends on how a comparison budget is divided across prompts and response pairs. In a linear reward model with feature dimension d, these choices determine how well comparisons cover the directions relevant for prediction; poorly covered directions require more data. A classical design principle minimizes the largest prediction variance over candidate responses; this is G-optimal design. Variants of this criterion appear in reward modeling, preference elicitation, and on-policy preference learning (Scheid et al., 2024; Mukherjee et al., 2024; Kim et al., 2026).

A common simplification separates the two allocation decisions. The locally G-optimal protocol of Kim et al. (2026) first chooses a response distribution at each prompt to minimize worst-case prediction variance after removing prompt-specific offsets, then freezes these distributions before allocating prompts. Their quadratic upper bound left open whether prompt allocation always reduces the global design value to linear scale, or whether locally optimal samplers can dilute coverage of complementary directions under a single budget. Figure 1 contrasts the budget shares assigned to strong response pairs.

We show that prompt allocation does not always recover linear scaling. In a signed-axis family, freezing locally G-optimal response distributions weakens coverage of complementary directions, at a cost proportional to the feature dimension. Built from positive and negative coordinate vectors, the family assigns a different strong coordinate to each prompt. Its local optima are unique, with full-rank covariances, yet its two-stage value is quadratic while joint allocation on those features attains the optimal linear value. Local balancing weakens the rotating strong directions before prompt allocation. Our decomposition separates response concentration within prompts from cross-prompt covariance coverage. Under logistic Bradley–Terry labels for independently sampled response pairs, the same poorly covered direction yields the corresponding comparison-complexity gap at fixed parameter radius and sufficiently small target error.

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

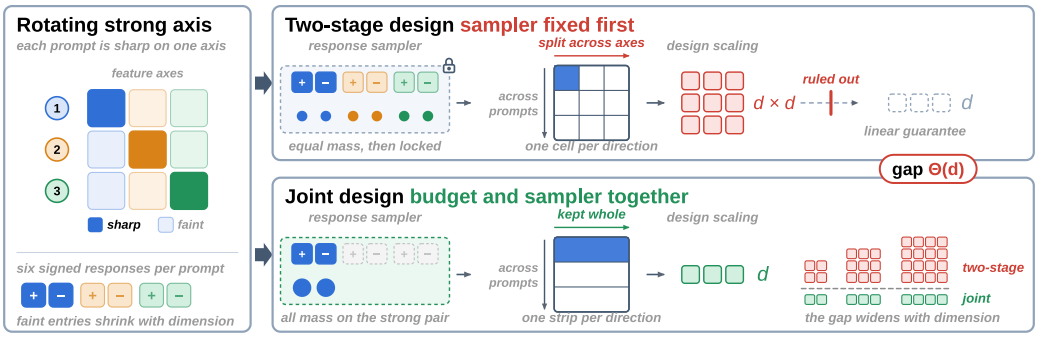


Figure 1: **A tight quadratic bound for two-stage design.** At the inverse-square-root weak-feature scale, our lower bound rules out the dashed route to a uniform linear guarantee. Joint allocation attains the linear benchmark on the same features. Dots indicate response probabilities; shaded budget regions show strong-pair shares. Counters show growth orders.

Our contributions are as follows:

(1) **Tight two-stage lower bound.** We construct a finite signed-axis family that attains the worst-case quadratic order with unique local optima and full-rank covariances. An exact robustness analysis quantifies the range of approximately optimal local designs for which quadratic scaling persists.

(2) **Geometric decomposition.** We develop a decomposition that separates response concentration within prompts from covariance coverage across prompts. An aligned comparison with matching local spectra shows that spectra alone do not determine the two-stage value, while linear scaling returns when all prompt covariances are uniformly comparable to a common reference.

(3) **Joint-design benchmark.** We show that joint prompt–response allocation reaches the optimal linear benchmark on the same feature family. This comparison isolates the cost of fixing local response distributions before allocating prompts.

(4) **Statistical consequence.** We translate the geometric gap into a quadratic-versus-linear separation in comparison complexity under the stated Bradley–Terry score-estimation model.

We relate the question to experimental design and preference learning in Section 2. Section 3 sets up the two protocols, constructs the signed-axis family, and computes its two-stage value exactly. Section 4 decomposes the gap into response concentration and covariance coverage, and Section 5 carries it into comparison complexity. We conclude in Section 6 with the conditions under which the separation persists.

2 Related Work

Pairwise preference data supports reward modeling, reinforcement learning from human feedback, and direct policy optimization (DPO) (Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023). Choosing which comparisons to label turns this pipeline into a two-level experimental-design problem: G-optimality controls the largest prediction variance, but comparisons must be allocated across prompts and response pairs. Fixing within-prompt distributions restricts the covariance mixtures available to prompt allocation, whereas joint design coordinates both choices.

Optimal experimental design. The covariance-normalized prediction variance of a response is its leverage, and G-optimality minimizes the largest leverage. Its classical roots include polynomial regression design and the Kiefer–Wolfowitz equivalence between G-optimality and determinant-maximizing D-optimality (Smith, 1918; Kiefer & Wolfowitz, 1960). The same criterion and related computational tools appear in linear-bandit pure exploration and minimum-volume-ellipsoid methods (Soare et al., 2014; Degenne et al., 2020; Lattimore & Szepesvári, 2020; Todd, 2016). This criterion supplies the local leverage and trace lower bound used in Section 3; whether prompt mixing preserves that worst-direction coverage is measured by covariance domination (Section 4).

Comparison selection and joint allocation. Optimal-design methods choose informative comparisons for linear reward estimation (Scheid et al., 2024; Mukherjee et al., 2024). Recent work extends this perspective to response sampling for reward modeling and to D-optimal selection of multiway comparison queries (Feng et al., 2025; Thekumparampil et al., 2025). More broadly, dataset coverage and exploration govern data efficiency in both offline and online preference learning (Agarwal et al., 2025; Xiong et al., 2024; Xie et al., 2025; Cen et al., 2025; Song et al., 2024). Sampler choice also affects the convergence of online DPO (Shi et al., 2025). Active collection (Bıyık et al., 2024; Metz et al., 2023; Das et al., 2024) and online reweighting (Chen et al., 2024) alter which comparisons are sampled or emphasized. Dataset diagnostics assess preference data (Shen et al., 2024), while dataset resets use offline examples to guide policy training (Chang et al., 2024). Preference-based RL likewise studies active, randomized, and reward-agnostic ways to collect pairwise feedback (Chen et al., 2022; Zhan et al., 2024; Wu & Sun, 2024; Du et al., 2024). These approaches motivate joint prompt–response allocation; we quantify the worst-direction coverage left after the within-prompt distributions have been fixed.

Pairwise-feedback models and bandits. The Bradley–Terry model turns the difference between two response scores into a pairwise preference probability through a logistic link (Bradley & Terry, 1952). Bandit methods that learn from pairwise comparisons, including dueling, contextual-dueling, and preference-based bandits, select comparisons adaptively to control regret or identify a best action (Yue et al., 2012; Dudík et al., 2015; Pacchiano et al., 2021; Novoseller et al., 2020; Bengs et al., 2022; Saha, 2021); linear Bradley–Terry scores also connect preference learning to generalized linear and logistic bandits (Filippi et al., 2010; Faury et al., 2020; Abeille et al., 2021; Lee & Oh, 2024). Our Bradley–Terry experiment draws two responses independently at the same prompt and measures the largest error in within-prompt score differences. The least-covered covariance direction then drives the score-estimation lower bound.

Two-stage preferential design. The closest prior result is the two-stage preferential G-optimal design of Kim et al. (2026). Their Proposition 4.2 gives a quadratic upper bound after the response distributions are fixed, and Remark 4.3 asks whether prompt optimization can always reduce the value to linear scale. Our signed-axis construction matches that order while joint allocation on those features remains linear.

The missing gap and our motivation. Beyond the worst-case scaling, the central issue is how locally optimal responses affect coverage after prompt mixing. Once response distributions are fixed, prompt allocation can only reweight their covariances, not change the response probabilities that produced them. Local spectra do not describe how strong directions align across prompts, so comparing spectra alone cannot explain the loss. Our signed-axis family attains quadratic order with unique local optima and full-rank covariances, while an aligned counterpart with matching spectra has linear two-stage value. The geometric decomposition explains this contrast and gives a sufficient condition for linear scaling. The Bradley–Terry analysis then connects the same geometry to the comparisons needed for within-prompt score estimation.

3 Two-stage and joint allocation on signed axes

The two protocols use the same prompts, responses, and feature map. Their difference is whether response distributions are optimized separately at each prompt or together with prompt weights. The signed-axis family makes the resulting difference in coverage explicit.

3.1 Setup and the two-stage objective

Fix a feature dimension $d \geq 1$. Let $\mathcal{X}$ be a finite prompt set, $\mathcal{A}$ a finite response set, and $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ a feature map. For a finite set S, write $\Delta(S)$ for the probability simplex on S. A conditional response sampler $\pi(\cdot \mid x)$ induces three objects. The local mean removes a prompt-specific offset, the centered feature records the identifiable score difference, and the covariance records how well the

sampler covers each feature direction:

$$\begin{aligned}\bar{\phi}_\pi(x) &:= \mathbb{E}_{b\sim\pi(\cdot|x)}[\phi(x,b)],\\ \phi_{g,\pi}(x,a) &:= \phi(x,a) - \bar{\phi}_\pi(x),\\ \Sigma_\pi(x) &:= \mathbb{E}_{a\sim\pi(\cdot|x)}[\phi_{g,\pi}(x,a)\phi_{g,\pi}(x,a)^\top].\end{aligned}$$

For symmetric matrices, $A \preceq B$ means that $B - A$ is positive semidefinite, and $A \succ 0$ means that A is positive definite. When $\Sigma_\pi(x)$ is positive definite, write its largest covariance-normalized centered prediction variance as

$$G_x(\pi) := \max_{a\in\mathcal{A}} \phi_{g,\pi}(x,a)^\top \Sigma_\pi(x)^{-1} \phi_{g,\pi}(x,a). \tag{1}$$

We call this quantity the centered conditional G-criterion and use the standard extended value $+\infty$ when the covariance is singular. A sampler is centered conditional G-optimal if it minimizes this criterion. Centering is essential for preference data: pairwise comparisons identify score differences at a prompt, not an arbitrary prompt-specific offset.

A trace identity fixes the dimension-scale benchmark for both allocation protocols.

Lemma 1. *Let q be a distribution on finitely many vectors $y_a \in \mathbb{R}^d$, let $m = \mathbb{E}_q[y_a]$, and let $M_q = \mathbb{E}_q[(y_a - m)(y_a - m)^\top] \succ 0$. Then*

$$\mathbb{E}_q\big[(y_a - m)^\top M_q^{-1}(y_a - m)\big] = d, \qquad \max_a (y_a - m)^\top M_q^{-1}(y_a - m) \geq d.$$

Proof sketch. The equality is the trace calculation $\mathrm{tr}(M_q^{-1}M_q) = d$, and the maximum dominates its q-average; Appendix B treats the singular case. The identity sets d as the benchmark for both protocols, which differ only in when they require full directional coverage: separately at every prompt or only after mixing prompts.

Fix a centered conditional G-optimal sampler $\pi_{\mathrm{cg}}(\cdot \mid x)$ at each prompt and abbreviate $y_{x,a} := \phi_g(x,a) := \phi_{g,\pi_{\mathrm{cg}}}(x,a)$ and $\Sigma_x := \Sigma_{\pi_{\mathrm{cg}}}(x)$. After these objects have been fixed, the remaining choice is a distribution $\mu \in \Delta(\mathcal{X})$ over prompts, called the prompt marginal. Its mixture covariance determines how small the worst response leverage can be:

$$\begin{aligned}V_\mu &:= \sum_{x\in\mathcal{X}} \mu(x)\Sigma_x,\\ Q(\phi,\pi_{\mathrm{cg}}) &:= \min_{\mu\in\Delta(\mathcal{X})} \max_{x\in\mathcal{X},\, a\in\mathcal{A}} y_{x,a}^\top V_\mu^{-1} y_{x,a}.\end{aligned} \tag{2}$$

More generally, $Q(\phi,\pi)$ denotes the same criterion when the fixed response conditionals are π; thus $Q(\phi,\pi_{\mathrm{cg}})$ is its specialization to π_{cg}. We call (2) the exact two-stage design value. This is Eq. (7) of Kim et al. (2026): centering and local response design precede the choice of prompt marginal. Once the conditionals are fixed, changing μ can reweight the local covariance matrices but cannot change their centers, spectra, or relative response masses. To cover a direction that is weak in most fixed covariances, μ must concentrate on the few prompts where it is strong.

Figure 2 illustrates the alternative: choosing prompt weights and response samplers together lets complementary prompts specialize. The following construction evaluates both protocols on the same features; Appendix A summarizes their design variables.

3.2 A SIGNED-AXIS FAMILY

In the signed-axis family, the tension between local balance and global coverage is explicit. At each prompt, one coordinate is especially informative, yet the unique local optimizer spreads mass across every direction. Pairing each coordinate vector with its negative keeps centering transparent, while diagonal covariances expose how the strong direction changes across prompts. Once the local samplers are fixed, prompt allocation can only reweight these covariances, so the resulting loss of coverage is exactly calculable.

Fix $d \geq 2$, write $[d] := \{1,\dots,d\}$, and choose a cross-prompt scale $\rho \in (0,1)$. Let $\mathcal{X} = [d]$ and $\mathcal{A} = [d] \times \{+1,-1\}$, and write $e_1,\dots,e_d$ for the standard basis of $\mathbb{R}^d$. For prompt i and response

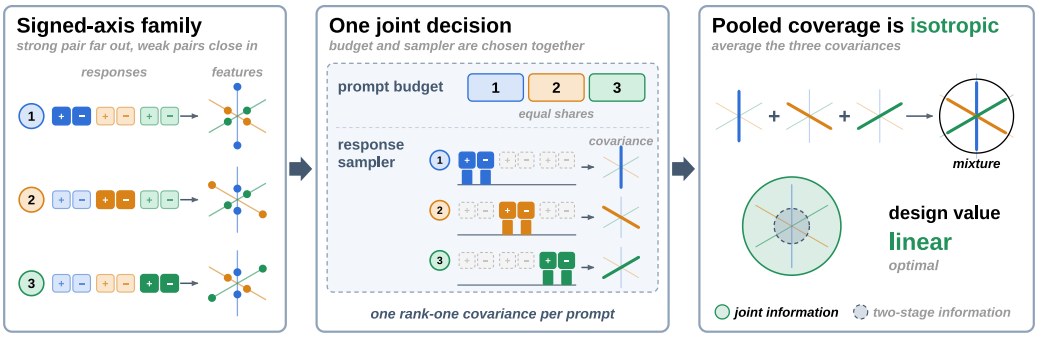


Figure 2: **Joint allocation across complementary directions.** Equal weights within strong response pairs and across prompts give an isotropic covariance with optimal design value d. Segments show rank-one local covariances; circles indicate relative information scales.

(j, s), define

$$\phi_\rho(i, (j, s)) = s\, c_{ij} e_j, \qquad c_{ij} = \begin{cases} 1, & i = j, \\ \rho, & i \neq j. \end{cases} \tag{3}$$

Prompt i has unit-scale access to coordinate i, but sees every other coordinate only at scale ρ. Thus every prompt contains all directions but specializes in a different one. As ρ decreases, the fixed local samplers provide less coverage away from their specialized prompts.

At every prompt, consider the uniform response sampler $\pi_{\mathrm{cg}}((j, s) \mid i) = 1/(2d)$. The centered conditional G-criterion uniquely selects this sampler.

Theorem 2. *Let $b_1, \ldots, b_d > 0$ and let the available vectors be $\{\pm b_j e_j : j \in [d]\}$. Under the full-rank centered criterion* (1)*, the unique G-optimal distribution assigns mass $1/(2d)$ to every signed vector.*

Proof sketch. The trace identity first fixes d as the optimal value. If a design attains this value, averaging the two sign constraints on each axis forces zero sign imbalance; the remaining leverage inequalities then force total mass $1/d$ on every axis. Hence each signed response has mass $1/(2d)$. The full calculation appears in Appendix C.

Because centering makes the design problem nonlinear in the response probabilities, changing sign balance moves both the mean and the covariance. Uniqueness therefore fixes the local mean and response masses before prompt allocation.

Proposition 3. *For every prompt i, the uniform sampler above is the unique centered conditional G-optimal sampler for* (3)*. Its mean is zero and*

$$\Sigma_i = \frac{1}{d}\,\mathrm{diag}(c_{i1}^2, \ldots, c_{id}^2) = \frac{1}{d} e_i e_i^\top + \frac{\rho^2}{d} \sum_{j \neq i} e_j e_j^\top. \tag{4}$$

Every response has local leverage d. The largest and smallest eigenvalues are $1/d$ and ρ^2/d; their ratio, called the local condition number, is ρ^{-2}.

Proof sketch. The signed pair on coordinate j contributes $c_{ij}^2 e_j e_j^\top / d$ to the covariance, so the jth diagonal entry of the inverse is d/c_{ij}^2 and every response has leverage d.

Every feature is bounded, every response has positive local mass, and every conditional covariance is full rank. Yet coordinate e_j is strong only at prompt j. A single prompt marginal must therefore divide its budget among d complementary directions, while every off-prompt contribution is weakened by ρ.

3.3 EXACT VALUE OF THE TWO-STAGE DESIGN

The unique local optimizers fix diagonal covariances, and the unit-scale response at prompt j is the worst action on coordinate j. The two-stage objective is governed by the least prompt mass and reduces to an exactly solvable one-dimensional worst-case optimization problem.

Theorem 4. *For every $d \geq 2$ and $\rho \in (0, 1)$, the signed-axis family in* (3) *satisfies*

$$\boxed{Q_\rho := Q(\phi_\rho, \pi_{\mathrm{cg}}) = \frac{d^2}{1 + (d-1)\rho^2}.} \tag{5}$$

The unique minimizing prompt marginal is uniform: $\mu_{\mathrm{unif}}(i) = 1/d$ for all i.

Proof sketch. In (5), the d^2 numerator comes from balancing d directions locally and then balancing prompts globally. The denominator credits the weak coverage that each coordinate receives at the other prompts. Under a prompt marginal μ, coordinate j has variance $[\rho^2 + (1-\rho^2)\mu(j)]/d$, and the unit response at prompt j is worst on that coordinate. The objective therefore depends only on $\min_j \mu(j)$. Since $\min_j \mu(j) \leq 1/d$, with equality only for the uniform marginal, substitution gives (5).

Corollary 5. *For every $c > 0$, if $0 < \rho \leq c/\sqrt{d-1} < 1$, then $Q_\rho \geq d^2/(1+c^2)$. In terms of the local condition number $\chi = \rho^{-2}$, the exact value is $d^2/[1 + (d-1)/\chi]$, so quadratic order requires $\chi = \Omega(d)$, $\chi = \Theta(d)$ suffices, and bounded χ gives linear order. Consequently, no dimension-independent constant C satisfies $Q(\phi, \pi_{\mathrm{cg}}) \leq Cd$ on every finite, bounded instance whose response features do not lie in a lower-dimensional affine subspace at any prompt, even when the first-stage algorithm may coordinate its choice among local G-optima. Combined with the upper bound of Kim et al. (2026, Proposition 4.2), the worst-case order on the same class is $\Theta(d^2)$.*

Proof sketch. Substituting $\rho^2 \leq c^2/(d-1)$ into (5) gives the first bound, and rewriting its denominator as $1 + (d-1)/\chi$ gives the condition-number form and the three regimes. At $\rho = d^{-1/2}$, the condition number is exactly d and $Q_{d^{-1/2}} = d^3/(2d-1) \geq d^2/2$; at $\rho = 1/d$, $Q_{1/d} = (1 - o(1))d^2$. Taking $d > 2C$ at $\rho = d^{-1/2}$ then rules out any dimension-independent linear bound C. Because the family satisfies the conditions of the known upper bound (Appendix J), the worst-case order is $\Theta(d^2)$.

The exact formula identifies a dimension-growing local condition number as one ingredient of the quadratic regime; an aligned construction with the same local spectra nevertheless has linear value (Section 4), so cross-prompt orientation is the complementary ingredient. Appendix E treats approximate local designs and related variants, and Appendix J verifies the standard preferential-design assumptions for the signed-axis family.

4 COVARIANCE GEOMETRY OF PROMPT ALLOCATION

The closed form for Q_ρ does not by itself reveal the mechanism behind the quadratic order; two geometric effects are in play. A local condition number records directional imbalance within a prompt, but not whether the best-covered directions align across prompts.

4.1 LOCAL LEVERAGE AND COVARIANCE COVERAGE

Within a prompt, we compare the largest squared response projection with its local average. Across prompts, we compare each local covariance with the mixture formed by prompt allocation.

Fix any collection of full-rank conditional G-optimal designs with local value d, and write $y_{x,a} = \phi_g(x, a)$. For a prompt x and nonzero direction $u \in \mathbb{R}^d$, define the within-prompt max-to-average ratio

$$s_x(u) := \frac{\max_{a \in \mathcal{A}} (u^\top y_{x,a})^2}{u^\top \Sigma_x u}.$$

This ratio measures how sharply one action represents u relative to the local covariance. A value near one means that the local second moment is representative of its available actions in direction u; a large value means that one response is much more extreme than the local average.

This ratio is a within-prompt property. Across prompts, we ask how much one mixture covariance must be enlarged to cover every local covariance. The resulting covariance domination factor is

$$\kappa_{\mathrm{dom}} := \min_{\mu \in \Delta(\mathcal{X})} \max_{\substack{x \in \mathcal{X} \\ u \neq 0}} \frac{u^\top \Sigma_x u}{u^\top V_\mu u}. \tag{6}$$

Equivalently, κ_{dom} is the smallest k for which some V_μ satisfies $\Sigma_x \preceq k V_\mu$ for every prompt.

Theorem 6. *Suppose the fixed conditional designs are full rank and G-optimal, with local value d at every prompt. Then*

$$Q(\phi, \pi_{\mathrm{cg}}) = \min_{\mu \in \Delta(\mathcal{X})} \max_{\substack{x \in \mathcal{X} \\ u \neq 0}} s_x(u) \, \frac{u^\top \Sigma_x u}{u^\top V_\mu u}. \tag{7}$$

Moreover, $1 \leq s_x(u) \leq d$ for every x and $u \neq 0$, and consequently

$$\boxed{\kappa_{\mathrm{dom}} \leq Q(\phi, \pi_{\mathrm{cg}}) \leq d \, \kappa_{\mathrm{dom}}.} \tag{8}$$

The identity gives the two factors distinct roles. The covariance ratio measures how much of a local covariance survives prompt mixing, while $s_x(u)$ measures how extreme the worst response is relative to its local average. Reweighting prompts can change only the covariance ratio, so the first-stage commitment persists in the global objective.

Proof sketch. For $V \succ 0$ and $y \in \mathbb{R}^d$, the variational identity $y^\top V^{-1} y = \max_{u \neq 0} (u^\top y)^2 / (u^\top V u)$ turns leverage into a directional supremum. Maximizing over actions produces $s_x(u)$, and multiplying and dividing by $u^\top \Sigma_x u$ produces the covariance ratio. Under the local sampler, the normalized squared projection has mean one; generalized Cauchy–Schwarz and local G-optimality bound it by d. This proves both the representation and the sandwich.

A mixture maximizing $\log\det V_\mu$ satisfies $\Sigma_x \preceq d V_\mu$ for every prompt, giving $1 \leq \kappa_{\mathrm{dom}} \leq \min\{d, |\mathcal{X}|\}$ (Proposition 19). Thus (8) recovers $Q(\phi, \pi_{\mathrm{cg}}) \leq d \min\{d, |\mathcal{X}|\} \leq d^2$. Bounded κ_{dom} gives linear leverage, whereas quadratic value requires $\kappa_{\mathrm{dom}} = \Omega(d)$, as attained by the signed-axis family.

4.2 Consequences of the decomposition

The covariance bound gives a sufficient condition for linear scaling. An aligned comparison shows why local spectra alone cannot determine the two-stage value.

Corollary 7. *Suppose there are $H \succ 0$ and constants $0 < \alpha \leq \beta$ such that $\alpha H \preceq \Sigma_x \preceq \beta H$ for every prompt x. Then $\kappa_{\mathrm{dom}} \leq \beta/\alpha$ and $Q(\phi, \pi_{\mathrm{cg}}) \leq (\beta/\alpha) d$. A constant covariance-comparability ratio therefore recovers linear leverage.*

Proof sketch. The condition allows nonidentical covariances and different eigenvectors, and requires only that every Σ_x lie between fixed multiples of a common reference matrix. Every mixture satisfies $V_\mu \succeq \alpha H$ while $\Sigma_x \preceq \beta H$, so $\Sigma_x \preceq (\beta/\alpha) V_\mu$ and $\kappa_{\mathrm{dom}} \leq \beta/\alpha$; Theorem 6 then gives the two-stage value at most $(\beta/\alpha) d$.

Proposition 8. *For every $d \geq 2$ and $\rho \in (0, 1)$, there is an aligned signed-axis feature map ϕ_ρ^{al} on the same prompt and response sets whose unique local covariances have the same spectrum and condition number as the local covariances of the witness* (3)*, yet $\kappa_{\mathrm{dom}} = 1$ and $Q(\phi_\rho^{\mathrm{al}}, \pi_{\mathrm{cg}}) = d$.*

Proof sketch. The comparison instance aligns the same strong direction at every prompt, so every local covariance equals the same diagonal matrix with the same spectrum and condition number as the witness; every prompt mixture then equals it, every response has leverage d, and $\kappa_{\mathrm{dom}} = 1$. The witness instead distributes those directions across prompts. Cross-prompt orientation therefore supplies information that the local spectra omit; the explicit instance appears in Appendix F.

Proposition 9. *For the family* (3)*, $\kappa_{\mathrm{dom}} = d/[1 + (d-1)\rho^2]$ and $Q_\rho = d\kappa_{\mathrm{dom}}$. Thus the factor d in* (8) *is sharp.*

Proof sketch. For the witness, the covariance ratio in the strong direction of prompt i is $1/[\rho^2 + (1-\rho^2)\mu(i)]$, minimized by the uniform marginal, and the same direction maximizes $s_i(e_i) = d$;

this gives the stated κ_{dom} and $Q_\rho = d\kappa_{\mathrm{dom}}$. Changing only the orientation of the local covariances therefore changes the two-stage value by up to a factor d.

Covariance domination diagnoses the geometry after the response conditionals have been frozen. Choosing those conditionals together with the prompt marginal avoids the mismatch on the same signed-axis features.

4.3 Joint prompt–response allocation

A joint design chooses response conditionals and the prompt marginal together.

Let $\pi = (\pi_x)_{x\in\mathcal{X}}$, with $\pi_x \in \Delta(\mathcal{A})$. For each prompt, define the centered feature $y^\pi_{x,a} := \phi_{g,\pi_x}(x,a)$, and let $\Sigma_{\pi_x}(x)$ denote its covariance. Given a prompt marginal μ, set $W(\mu,\pi) := \sum_x \mu(x)\Sigma_{\pi_x}(x)$. The joint value is

$$J_{\mathrm{joint}}(\phi) := \min_{\substack{\mu\in\Delta(\mathcal{X}),\ \{\pi_x\in\Delta(\mathcal{A})\} \\ W(\mu,\pi)\succ 0}} \max_{x,a} (y^\pi_{x,a})^\top W(\mu,\pi)^{-1} y^\pi_{x,a}.$$

Here the conditional means and covariances vary with the response distributions, which are optimized together with the prompt marginal.

Theorem 10. *For the signed-axis family* (3)*,*

$$\boxed{J_{\mathrm{joint}}(\phi_\rho) = d, \qquad \frac{Q_\rho}{J_{\mathrm{joint}}(\phi_\rho)} = \frac{d}{1+(d-1)\rho^2}.} \tag{9}$$

An optimum is attained by the uniform prompt marginal and by assigning probability $1/2$ *to each strong signed response at every prompt (Figure 2). At* $\rho = d^{-1/2}$*, the ratio is* $d^2/(2d-1) = \Theta(d)$*.*

Proof sketch. For the lower bound, write $W = W(\mu,\pi)$ for any feasible joint design. Averaging action leverage under its sampling law gives $\mathrm{tr}(W^{-1}W) = d$, so the maximum is at least d. For the matching upper bound, prompt i uses only its strong signed pair, and uniform prompt mixing gives $W = I_d/d$, where I_d is the $d\times d$ identity matrix. Strong actions then have leverage d and weak actions have leverage $d\rho^2$, matching the lower bound. Dividing (5) by d gives the ratio.

Appendix G shows that conditionals assigning positive probability to every response, while remaining full rank, retain linear joint scaling at the inverse-square-root scale. The Bradley–Terry experiment in Section 5 translates the gap between the frozen local samplers and the explicit joint sampler into comparison complexity.

5 Comparison complexity for Bradley–Terry score estimation

The comparison experiment below uses the least-covered covariance direction to distinguish two nearby parameters. The frozen-conditional design therefore yields a worst-case lower bound, whereas the explicit joint sampler yields a coordinatewise upper bound.

Fix a prompt marginal μ and retain the unique response conditionals of Proposition 3. The logistic link $\sigma(t) := (1+e^{-t})^{-1}$ converts a score difference into a preference probability. For a parameter $\theta\in\mathbb{R}^d$, one labeled comparison first draws a prompt and then two responses independently from its conditional sampler:

$$I\sim\mu, \qquad A,B \overset{\mathrm{iid}}{\sim} \pi_{\mathrm{cg}}(\cdot\mid I), \qquad Y\mid I,A,B \sim \mathrm{Bernoulli}\big(\sigma\big(\theta^\top[\phi_\rho(I,A)-\phi_\rho(I,B)]\big)\big). \tag{10}$$

The lower bound is uniform over the prompt marginal μ, while the response conditionals are the unique local G-optima analyzed above.

This is a Bradley–Terry model with linear reward $r_\theta(i,a) = \theta^\top\phi_\rho(i,a)$. Let $\widehat{\theta}$ be any estimator from n independent observations. Here $\mathbb{E}_\theta$ denotes expectation under (10) at parameter θ. Write $\mathcal{R}_\mu(\widehat{\theta},\theta)$ for the largest mean-squared centered-score error over all prompt–response pairs:

$$\mathcal{R}_\mu(\widehat{\theta},\theta) := \max_{i\in[d],\ a\in\mathcal{A}} \mathbb{E}_\theta\Big[\big(\phi_g(i,a)^\top(\widehat{\theta}-\theta)\big)^2\Big]. \tag{11}$$

For $R > 0$, write $B_2(0, R) := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le R\}$. Centered scores remove the unidentifiable prompt-specific offset. Their uniform mean-squared error is the statistical target throughout.

Theorem 11. *For every $R > 0$, $0 < \varepsilon \le R/2$, prompt marginal μ, and sample size $n \le Q_\rho/(8\varepsilon^2)$, the experiment* (10) *satisfies*

$$\boxed{\inf_{\widehat{\theta}} \sup_{\theta \in B_2(0,R)} \mathcal{R}_\mu(\widehat{\theta}, \theta) \ge \frac{3}{8}\varepsilon^2.}$$

Consequently, uniform centered-score mean-squared error below $3\varepsilon^2/8$ requires $\Omega(Q_\rho/\varepsilon^2)$ comparisons for every prompt marginal. At $\rho = d^{-1/2}$, this is $\Omega(d^2/\varepsilon^2)$.

Proof sketch. For the lower bound, choose the least-covered coordinate j and test $\theta = 0$ against $\theta = 2\varepsilon e_j$. Their target scores differ by 2ε. Since the response-difference covariance is $2V_\mu$, the n-sample Kullback–Leibler (KL) divergence is at most $n\varepsilon^2/Q_\rho$. Under the stated sample-size condition this is at most $1/8$; Pinsker's inequality and thresholding the estimated target score then give the stated risk.

For comparison, let $\mu^\star(i) = 1/d$ and let $\pi_i^\star$ put probability $1/2$ on each strong signed response at prompt i. Draw $I \sim \mu^\star$, $A, B \overset{\text{iid}}{\sim} \pi_I^\star$, and use the label law in (10). Proposition 3 gives $\phi_g = \phi_\rho$ for the fixed local sampler, and the joint sampler is likewise centered at zero by construction. Let $\mathbb{E}_\theta^\star$ denote expectation under n independent joint-design comparisons and set $m_R := \sigma(2R)[1 - \sigma(2R)]$, the smallest derivative of σ on the relevant score range $[-2R, 2R]$.

Theorem 12. *For every known $R > 0$ and $n \ge 1$, there is an estimator such that*

$$\sup_{\theta \in B_2(0,R)} \max_{i \in [d],\, a \in \mathcal{A}} \mathbb{E}_\theta^\star\Big[\big(\phi_g(i,a)^\top(\widehat{\theta} - \theta)\big)^2\Big] \le \frac{d}{4m_R^2 n}. \tag{12}$$

Proof sketch. For the upper bound, an unequal signed pair for coordinate i appears with probability $1/(2d)$. After reversing labels to a common orientation, the retained observations are Bernoulli with mean $\sigma(2\theta_i)$. A clipped empirical mean is then mapped back through the inverse logistic link to estimate each coordinate. The inverse-link derivative on the radius-R ball is at most $1/(2m_R)$, and averaging over the binomial number of informative samples gives $d/(4m_R^2 n)$. Since every feature coefficient has magnitude at most one, this coordinate bound controls the worst centered-score error. Consequently, for $0 < \delta \le 3R^2/32$, two-stage risk below δ requires $n > 3Q_\rho/(64\delta)$, while the joint sampler attains risk at most δ for $n \ge d/(4m_R^2\delta)$. At $\rho = d^{-1/2}$, the dimension bounds are $\Omega(d^2/\delta)$ and $O(d/\delta)$, where the latter constant depends only on R.

The statistical lower bound therefore follows the same least-covered direction that determines Q_ρ, while the joint sampler covers every coordinate at the trace scale. Low variance of the sampled feature differences in this direction makes nearby reward parameters difficult to distinguish under the two-stage design. Under the joint sampler, the label probability for an opposite-sign comparison depends on only one coordinate of the reward parameter, so each coordinate can be estimated separately.

6 Conclusion

We establish a quadratic-versus-linear separation for locally G-optimal two-stage allocation in preferential design. In the signed-axis family, each prompt has its own strong direction, yet the unique local optimum spreads mass over all coordinates. With weak features of magnitude $d^{-1/2}$, the frozen designs have condition number d and a quadratic two-stage value, whereas joint allocation on the same features reaches the linear trace lower bound. Allocation order therefore costs a factor d when local imbalance grows with the dimension and strong directions fail to align across prompts. We use the geometric decomposition to identify when this separation can occur. The within-prompt max-to-average ratio measures response concentration, while covariance domination measures whether a single prompt mixture preserves every local covariance. Uniform covariance comparability controls the latter and recovers linear scaling; local spectra alone do not. The robustness analysis characterizes the approximate local designs for which the separation persists, and the Bradley–Terry comparison model transfers the same geometry to a factor-d gap in uniform centered-score estimation.

AI USE STATEMENT

Generative AI tools assisted language editing and formatting.

REFERENCES

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pp. 3691–3699. PMLR, 2021. URL https://proceedings.mlr.press/v130/abeille21a.html.

Alekh Agarwal, Christoph Dann, and Teodor Vanislavov Marinov. Design considerations in offline preference-based RL. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 499–512. PMLR, 2025. URL https://proceedings.mlr.press/v267/agarwal25a.html.

Viktor Bengs, Aadirupa Saha, and Eyke Hüllermeier. Stochastic contextual dueling bandits under linear stochastic transitivity models. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 1764–1786. PMLR, 2022. URL https://proceedings.mlr.press/v162/bengs22a.html.

Erdem Bıyık, Nicolas Huynh, Mykel J. Kochenderfer, and Dorsa Sadigh. Active preference-based gaussian process regression for reward learning and optimization. *The International Journal of Robotics Research*, 43(5):665–684, 2024.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324. URL https://academic.oup.com/biomet/article-abstract/39/3-4/324/326091.

Shicong Cen, Jincheng Mei, Katayoon Goshvadi, Hanjun Dai, Tong Yang, Sherry Yang, Dale Schuurmans, Yuejie Chi, and Bo Dai. Value-incentivized preference optimization: A unified approach to online and offline RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SQnitDuow6.

Jonathan D. Chang, Wenhao Zhan, Owen Oertell, Kianté Brantley, Dipendra Misra, Jason D. Lee, and Wen Sun. Dataset reset policy optimization for RLHF. *arXiv preprint arXiv:2404.08495*, 2024. URL https://arxiv.org/abs/2404.08495.

Lichang Chen, Jiuhai Chen, Chenxi Liu, John Kirchenbauer, Davit Soselia, Chen Zhu, Tom Goldstein, Tianyi Zhou, and Heng Huang. OPTune: Efficient online preference tuning. *arXiv preprint arXiv:2406.07657*, 2024. URL https://arxiv.org/abs/2406.07657.

Xiaoyu Chen, Han Zhong, Zhuoran Yang, Zhaoran Wang, and Liwei Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. *arXiv preprint arXiv:2205.11140*, 2022. URL https://arxiv.org/abs/2205.11140.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Provably sample efficient RLHF via active preference optimization. *arXiv preprint arXiv:2402.10500*, 2024. URL https://arxiv.org/abs/2402.10500.

Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 2432–2442. PMLR, 2020. URL https://proceedings.mlr.press/v119/degenne20a.html.

Yihan Du, Anna Winnicki, Gal Dalal, Shie Mannor, and R. Srikant. Exploration-driven policy optimization in RLHF: Theoretical insights on efficient data utilization. *arXiv preprint arXiv:2402.10342*, 2024. URL https://arxiv.org/abs/2402.10342.

Miroslav Dudík, Katja Hofmann, Robert E. Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Proceedings of the Conference on Learning Theory*, pp. 563–587. PMLR, 2015. URL https://proceedings.mlr.press/v40/Dudik15.html.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3052–3060. PMLR, 2020. URL https://proceedings.mlr.press/v119/faury20a.html.

Yunzhen Feng, Ariel Kwiatkowski, Kunhao Zheng, Julia Kempe, and Yaqi Duan. PILAF: Optimal human preference sampling for reward modeling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16744–16776. PMLR, 2025. URL https://proceedings.mlr.press/v267/feng25g.html.

Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 23, 2010.

Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960. doi: 10.4153/CJM-1960-030-4.

Juno Kim, Jihun Yun, Jason D. Lee, and Kwang-Sung Jun. Coverage improvement and fast convergence of on-policy preference learning. *arXiv preprint arXiv:2601.08421*, 2026. URL https://arxiv.org/abs/2601.08421.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401. URL https://www.cambridge.org/core/books/bandit-algorithms/B600E1B842C4B00D3E6F1B7A44464792.

Joongkyu Lee and Min-Hwan Oh. Nearly minimax optimal regret for multinomial logistic bandit. *arXiv preprint arXiv:2405.09831*, 2024. URL https://arxiv.org/abs/2405.09831.

Yannick Metz, David Lindner, Raphaël Baur, Daniel Keim, and Mennatallah El-Assady. RLHF-blender: A configurable interactive interface for learning from diverse human feedback. *arXiv preprint arXiv:2308.04332*, 2023. URL https://arxiv.org/abs/2308.04332.

Subhojyoti Mukherjee, Anusha Lalitha, Kousha Kalantari, Aniket Deshmukh, Ge Liu, Yifei Ma, and Branislav Kveton. Optimal design for human preference elicitation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2861.

Ellen Novoseller, Yibing Wei, Yanan Sui, Yisong Yue, and Joel Burdick. Dueling posterior sampling for preference-based reinforcement learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pp. 1029–1038. PMLR, 2020. URL https://proceedings.mlr.press/v124/novoseller20a.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Aldo Pacchiano, Aadirupa Saha, and Jonathan Lee. Dueling RL: Reinforcement learning with trajectory preferences. *arXiv preprint arXiv:2111.04850*, 2021. URL https://arxiv.org/abs/2111.04850.

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, 2023.

Aadirupa Saha. Optimal algorithms for stochastic contextual preference bandits. *Advances in Neural Information Processing Systems*, 34:30050–30062, 2021.

Antoine Scheid, Etienne Boursier, Alain Durmus, Michael I. Jordan, Pierre Ménard, Eric Moulines, and Michal Valko. Optimal design for reward modeling in RLHF. *arXiv preprint arXiv:2410.17055*, 2024. URL https://arxiv.org/abs/2410.17055.

Judy Hanwen Shen, Archit Sharma, and Jun Qin. Towards data-centric RLHF: Simple metrics for preference dataset comparison. *arXiv preprint arXiv:2409.09603*, 2024. URL https://arxiv.org/abs/2409.09603.

Ruizhe Shi, Runlong Zhou, and Simon S. Du. The crucial role of samplers in online direct preference optimization. *arXiv preprint arXiv:2409.19605*, 2025. URL https://arxiv.org/abs/2409.19605.

Kirstine Smith. On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, 12(1–2):1–85, 1918.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

Yuda Song, Gokul Swamy, Aarti Singh, J. Andrew Bagnell, and Wen Sun. The importance of online data: Understanding preference fine-tuning via coverage. *Advances in Neural Information Processing Systems*, 37:12243–12270, 2024.

Kiran Koshy Thekumparampil, Gaurush Hiranandani, Kousha Kalantari, Shoham Sabach, and Branislav Kveton. Comparing few to rank many: Active human preference learning using randomized frank-Wolfe method. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59355–59376. PMLR, 2025. URL https://proceedings.mlr.press/v267/thekumparampil25a.html.

Michael J. Todd. *Minimum-Volume Ellipsoids: Theory and Algorithms*. SIAM, 2016. ISBN 9781611974379. doi: 10.1137/1.9781611974386.

Runzhe Wu and Wen Sun. Making RL with preference-based feedback efficient via randomization. *arXiv preprint arXiv:2310.14554*, 2024. URL https://arxiv.org/abs/2310.14554.

Tengyang Xie, Dylan J. Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Hassan Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit $Q^\star$-approximation for sample-efficient RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QYigQ6gXNw.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/xiong24a.html.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012. doi: 10.1016/j.jcss.2011.12.028.

Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D. Lee. Provable reward-agnostic preference-based reinforcement learning. *arXiv preprint arXiv:2305.18505*, 2024. URL https://arxiv.org/abs/2305.18505.

A ALLOCATION PROTOCOLS

The protocols in Figures 1 and 2 use the same prompts, responses, and features. They differ in which quantities are fixed before the prompt budget is chosen. The table records the corresponding design variables.

Design	Variables	Centering and covariance
Local response design	one $\pi(\cdot \mid x)$ for each fixed x	center and invert separately at each prompt
Prompt marginal	μ, with every $\pi_{\mathrm{cg}}(\cdot \mid x)$ fixed	center locally, then mix the fixed Σ_x's
Joint design	μ and all $\pi(\cdot \mid x)$	recenter and recompute every conditional covariance jointly

Theorem 2 makes the first-stage optimizer in the signed-axis family unique, so the prompt-marginal step is independent of tie-breaking. Theorem 10 evaluates the joint design on that family.

B PROOF OF THE TRACE LOWER BOUND

This trace calculation gives the common dimension-d lower bound for the conditional and joint criteria.

Proof of Lemma 1. For any vector z, $z^\top M_q^{-1} z = \mathrm{tr}(M_q^{-1} z z^\top)$. Linearity and cyclicity of the trace give

$$\begin{aligned}
&\sum_a q(a)(y_a - m)^\top M_q^{-1}(y_a - m) \\
&\quad = \mathrm{tr}\left(M_q^{-1} \sum_a q(a)(y_a - m)(y_a - m)^\top \right) \\
&\quad = \mathrm{tr}(M_q^{-1} M_q) = d.
\end{aligned}$$

The maximum over all available actions is at least this average.

The criterion maximizes over every available action, including actions outside the support of q, so the same lower bound applies to the all-action maximum. Under the inverse-based convention, a singular covariance has value $+\infty$. An alternative convention uses the Moore–Penrose pseudoinverse, which acts like an inverse on the directions covered by the design. The two conventions are equivalent on the signed-axis family, as established in Appendix J.

C LOCAL G-OPTIMALITY ON SIGNED AXES

The trace identity gives the lower bound d, which the uniform sampler attains. Equality forces zero sign imbalance and mass $1/d$ on each axis; the resulting sampler has the covariance spectrum used by the prompt-marginal analysis.

C.1 UNIQUENESS OF THE SIGNED-AXIS DESIGN

Leverage is invariant under invertible coordinate rescaling, so it is enough to analyze the signed standard basis vectors.

Proof of Theorem 2. The uniform distribution has mean zero, covariance $d^{-1}\,\mathrm{diag}(b_1^2, \ldots, b_d^2)$, and leverage d at every available vector. Lemma 1 therefore makes d the optimal value.

Let q be any optimizer. Write $q_{j,+}$ and $q_{j,-}$ for its masses on the two signs and define

$$p_j = q_{j,+} + q_{j,-}, \qquad \delta_j = q_{j,+} - q_{j,-}, \qquad D_p = \mathrm{diag}(p_1, \ldots, p_d).$$

Write $\delta = (\delta_1, \ldots, \delta_d)^\top$. With $B = \mathrm{diag}(b_1, \ldots, b_d)$, the mean is $B\delta$, and the covariance is BMB, where $M = D_p - \delta\delta^\top$. A finite criterion value implies $M \succ 0$. Leverage is invariant under the invertible map B, so optimality gives, for every j,

$$(\pm e_j - \delta)^\top M^{-1}(\pm e_j - \delta) \le d. \tag{13}$$

Set $k = \delta^\top M^{-1}\delta$. Averaging the two sign inequalities cancels the cross term and yields $(M^{-1})_{jj} + k \le d$. Since $D_p = M + \delta\delta^\top$ and $\sum_j p_j = 1$,

$$d + k = \mathrm{tr}(M^{-1}D_p) = \sum_j p_j (M^{-1})_{jj} \le d - k.$$

Since $M^{-1} \succ 0$, we have $k \ge 0$; the preceding inequality therefore forces $k = 0$, and positive definiteness then forces $\delta = 0$. Now $M = D_p$, so (13) becomes $1/p_j \le d$. Every p_j is at least $1/d$, and their sum is one; hence $p_j = 1/d$ for all j. Therefore $q_{j,\pm} = (p_j \pm \delta_j)/2 = 1/(2d)$.

C.2 Centering, leverage, and spectrum

Prompt-dependent scaling preserves sign cancellation but changes the covariance eigenvalues direction by direction.

Proof of Proposition 3. At prompt i, the two signs on every coordinate cancel, so

$$\bar{\phi}_{\pi_{\mathrm{cg}}}(i) = 0, \qquad \phi_g(i, (j, s)) = s\, c_{ij} e_j. \tag{14}$$

The signed pair on coordinate j contributes $c_{ij}^2 e_j e_j^\top / d$ to the covariance. Summing over coordinates gives (4). The inverse has jth diagonal entry d/c_{ij}^2, and hence

$$\phi_\rho(i, (j, s))^\top \Sigma_i^{-1} \phi_\rho(i, (j, s)) = c_{ij}^2 \frac{d}{c_{ij}^2} = d.$$

Theorem 2, applied with $b_j = c_{ij}$, proves uniqueness. The diagonal covariance also gives

$$\lambda_{\max}(\Sigma_i) = \frac{1}{d}, \qquad \lambda_{\min}(\Sigma_i) = \frac{\rho^2}{d}, \qquad \mathrm{cond}(\Sigma_i) = \rho^{-2}. \tag{15}$$

D Exact optimization over prompts

Because the fixed local covariances are diagonal, the objective depends on μ through its least prompt mass. The uniform marginal maximizes this quantity and yields $Q_\rho = d^2/[1 + (d-1)\rho^2]$.

Proof of Theorem 4. For $\mu \in \Delta([d])$, the covariance is diagonal. Its jth entry is

$$\begin{aligned}(V_\mu)_{jj} &= \frac{1}{d}\sum_{i=1}^{d} \mu(i) c_{ij}^2 \\ &= \frac{1}{d}\big(\mu(j) + \rho^2(1 - \mu(j))\big) \\ &= \frac{1}{d}\big(\rho^2 + (1 - \rho^2)\mu(j)\big).\end{aligned} \tag{16}$$

Every entry is at least ρ^2/d, including on the boundary of the prompt simplex.

Response (j, s) at prompt i has leverage

$$\frac{c_{ij}^2}{(V_\mu)_{jj}}.$$

For fixed j, the numerator is largest at prompt j, where it equals one. Thus the full action-level objective reduces to

$$F_\rho(\mu) = \max_{j \in [d]} \frac{d}{\rho^2 + (1 - \rho^2)\mu(j)}. \tag{17}$$

The scalar map $t \mapsto d/[\rho^2 + (1-\rho^2)t]$ is strictly decreasing, so $F_\rho(\mu)$ is determined by $\min_j \mu(j)$. The simplex inequality $\min_j \mu(j) \le 1/d$ is an equality only for the uniform marginal. Substitution gives

$$Q_\rho = \frac{d}{\rho^2 + (1-\rho^2)/d} = \frac{d^2}{1 + (d-1)\rho^2}.$$

Proof of Corollary 5. For the stated scale interval, $(d-1)\rho^2 \le c^2$; inserting this inequality into (5) gives $Q_\rho \ge d^2/(1+c^2)$. The condition-number form follows from $\chi = \rho^{-2}$. By contrast, for any fixed $K > 0$, if $\chi \le K$, then $Q_\rho \le Kd^2/(d-1) \le 2Kd$ for $d \ge 2$.

At $\rho = d^{-1/2}$, Theorem 4 gives

$$Q_\rho = \frac{d^3}{2d-1} \ge \frac{d^2}{2},$$

and Proposition 3 gives the local condition number $\rho^{-2} = d$. Local and prompt uniqueness follow from Theorems 2 and 4. At the smaller scale $\rho = 1/d$, the formula gives $Q_\rho = d^4/(d^2+d-1) = (1-o(1))d^2$.

Suppose a dimension-independent constant C satisfied $Q(\phi, \pi_{\mathrm{cg}}) \le Cd$ for every instance in the stated class. Choose an integer $d > 2C$ and take the signed-axis family with $\rho = d^{-1/2}$. Then $Q(\phi_{d^{-1/2}}, \pi_{\mathrm{cg}}) \ge d^2/2 > Cd$, a contradiction. Proposition 4.2 of Kim et al. (2026) supplies the matching d^2 upper bound under the assumptions verified in Appendix J.

E Robustness and approximate local design

The quadratic effect persists under heterogeneous off-diagonal scales of order $d^{-1/2}$ or smaller. Fix $\bar\rho \in (0, 1]$ and let

$$\widetilde\phi(i,(j,s)) = s\,\widetilde c_{ij} e_j, \qquad \widetilde c_{ii} = 1, \qquad 0 < \widetilde c_{ij} \le \bar\rho \quad (i \ne j), \tag{18}$$

where the weak scales may vary arbitrarily with both i and j. By Theorem 2, the uniform response distribution remains the unique conditional G-optimal design at every prompt. For $\mu \in \Delta([d])$, its local and mixture covariances are

$$\widetilde\Sigma_i := \frac{1}{d}\,\mathrm{diag}(\widetilde c_{i1}^2, \dots, \widetilde c_{id}^2), \qquad \widetilde V_\mu := \sum_{i=1}^d \mu(i)\widetilde\Sigma_i.$$

Proposition 13. *For every prompt marginal μ, the worst leverage under the fixed response conditionals of* (18) *is at least*

$$\frac{d^2}{1 + d\bar\rho^2}. \tag{19}$$

Consequently, for any $c > 0$, $\bar\rho \le c/\sqrt{d}$ gives the uniform lower bound $d^2/(1+c^2)$ over all weak scales satisfying (18)*.*

Proof. Choose j with $\mu(j) \le 1/d$. Under the unique local sampler, the variance on coordinate j obeys

$$(\widetilde V_\mu)_{jj} = \frac{1}{d}\sum_i \mu(i)\widetilde c_{ij}^2 \le \frac{1}{d}\left(\frac{1}{d} + \bar\rho^2\right).$$

The unit-scale response at prompt j therefore has leverage at least

$$\frac{1}{(\widetilde V_\mu)_{jj}} \ge \frac{d}{d^{-1} + \bar\rho^2} = \frac{d^2}{1 + d\bar\rho^2}.$$

This holds for every μ, proving the claim.

We next consider ridge regularization, which replaces V_μ by $V_\mu + \lambda I_d$. For $\lambda \ge 0$, define

$$Q_{\rho,\lambda} := \min_{\mu\in\Delta([d])} \max_{i,j,s} \phi_\rho(i,(j,s))^\top (V_\mu + \lambda I_d)^{-1} \phi_\rho(i,(j,s)). \tag{20}$$

Proposition 14. *For the homogeneous family* (3)*,*

$$Q_{\rho,\lambda} = \frac{d^2}{1 + (d-1)\rho^2 + \lambda d^2}, \tag{21}$$

and the uniform prompt marginal is the unique minimizer. Hence, for any $c, L > 0$, if $\rho \le c/\sqrt{d-1}$ and $\lambda \le L/d^2$, then

$$Q_{\rho,\lambda} \ge \frac{d^2}{1 + c^2 + L}. \tag{22}$$

Proof. The regularized covariance is diagonal, and the largest action on coordinate j remains the unit action at prompt j. Thus

$$\max_{i,j,s} \phi_\rho(i,(j,s))^\top (V_\mu + \lambda I_d)^{-1} \phi_\rho(i,(j,s)) = \max_j \frac{1}{\lambda + d^{-1}[\rho^2 + (1-\rho^2)\mu(j)]}.$$

This is strictly decreasing in $\min_j \mu(j)$, so uniformity is uniquely optimal. Evaluating at $\mu(j) = 1/d$ and multiplying numerator and denominator by d^2 gives (21). The two assumed inequalities bound its denominator by $1 + c^2 + L$.

Fix $0 < c_1 \le c_2$. Any sequence $\rho_d \in [c_1/d, c_2/d]$ satisfies $Q_{\rho_d}/d^2 \to 1$, so the near-saturating behavior at $\rho = 1/d$ is stable under constant-factor changes. Proposition 13 further permits the weak scales to vary independently up to order $d^{-1/2}$, while the ridge calculation shows that regularization at covariance scale d^{-2} preserves quadratic growth.

E.1 Approximate local G-optimality

To quantify how local approximation changes the two-stage value, fix $\varepsilon \ge 0$ and let $G_i(\pi_i)$ denote the centered conditional criterion (1) at prompt i. Define the best two-stage value attainable by $(1+\varepsilon)$-approximate local designs as

$$Q_\rho^{(\varepsilon)} := \inf_{\substack{\pi = (\pi_i)_{i=1}^d \\ G_i(\pi_i) \le (1+\varepsilon)d \ \forall i}} Q(\phi_\rho, \pi). \tag{23}$$

Proposition 15. *For every $d \ge 2$, $\rho \in (0,1)$, and $\varepsilon \ge 0$,*

$$\boxed{Q_\rho^{(\varepsilon)} = \frac{(1+\varepsilon)d^2}{1 + \varepsilon d + (d-1)\rho^2}.} \tag{24}$$

At $\rho = d^{-1/2}$, relative error $\varepsilon = O(1/d)$ preserves quadratic order, whereas every fixed $\varepsilon > 0$ gives linear order.

Proof. Set $L = (1+\varepsilon)d$. At prompt i, write p_{ij} for the total mass on the two signs of axis j, δ_{ij} for their mass difference, and $\delta_i = (\delta_{i1}, \dots, \delta_{id})^\top$. Define

$$M_i := \operatorname{diag}(p_{i1}, \dots, p_{id}) - \delta_i \delta_i^\top.$$

The finite local bound implies $M_i \succ 0$. After the invertible coordinate rescaling in (3), the two centered leverage constraints on axis j are at most L. Averaging them gives

$$(M_i^{-1})_{jj} + \delta_i^\top M_i^{-1} \delta_i \le L.$$

Since $M_i \preceq \operatorname{diag}(p_{i1}, \dots, p_{id})$, inverse monotonicity yields $1/p_{ij} \le (M_i^{-1})_{jj}$. Hence

$$p_{ij} \ge \frac{1}{L}, \qquad p_{ii} \le 1 - \frac{d-1}{L} =: a.$$

The local centered covariance therefore satisfies

$$\operatorname{tr}(\Sigma_{\pi_i}(i)) \le p_{ii} + \rho^2 \sum_{j \ne i} p_{ij} \le \rho^2 + (1-\rho^2)a =: T.$$

For any prompt marginal μ, the mixture V_μ has trace at most T, so some coordinate j obeys $(V_\mu)_{jj} \le T/d$. The two centered strong responses at prompt j differ by $2e_j$. The parallelogram identity and the matrix Cauchy–Schwarz inequality give

$$\max_{s\in\{+1,-1\}} y_{j,(j,s)}^\top V_\mu^{-1} y_{j,(j,s)} \ge e_j^\top V_\mu^{-1} e_j \ge \frac{1}{(V_\mu)_{jj}} \ge \frac{d}{T}.$$

Because

$$T = \frac{1+\varepsilon d + (d-1)\rho^2}{(1+\varepsilon)d},$$

this proves the lower bound in (24).

For equality, use sign-balanced conditionals whose axis masses at prompt i are

$$p_{ii} = a = \frac{1+\varepsilon d}{(1+\varepsilon)d}, \qquad p_{ij} = \frac{1}{(1+\varepsilon)d} \quad (j \ne i),$$

and split each axis mass equally between its two signs. Their local G-values are exactly L; all response masses are positive and the local covariances are full rank. Under the uniform prompt marginal, the mixture covariance is $(T/d)I_d$, the strong responses have leverage d/T, and the weak responses have leverage $\rho^2 d/T$. Thus the lower bound is attained. Substituting $\rho = d^{-1/2}$ gives

$$Q^{(\varepsilon)}_{d^{-1/2}} = \frac{(1+\varepsilon)d^2}{1+\varepsilon d + (d-1)/d},$$

which has the stated two regimes.

F Covariance domination: proofs and matrix formulations

The leverage decomposition follows from a directional variational identity. Schur complements and covariance ordering then give finite semidefinite formulations and bounds.

F.1 Leverage decomposition

Writing leverage as a directional supremum makes the action and covariance factors explicit.

Proof of Theorem 6. For $V \succ 0$ and $y, u \in \mathbb{R}^d$ with $u \ne 0$, Cauchy–Schwarz gives

$$y^\top V^{-1} y = \max_{u\ne 0} \frac{(u^\top y)^2}{u^\top V u},$$

with equality at $u = V^{-1}y$ when $y \ne 0$; when $y = 0$, both sides vanish. Maximizing over x, a, commuting the two maxima, and multiplying and dividing by $u^\top \Sigma_x u$ yield (7).

The $\pi_{\mathrm{cg}}(\cdot \mid x)$-average of $(u^\top y_{x,a})^2/(u^\top \Sigma_x u)$ is one, so $s_x(u) \ge 1$. Generalized Cauchy–Schwarz and local G-optimality give

$$(u^\top y_{x,a})^2 \le (u^\top \Sigma_x u)(y_{x,a}^\top \Sigma_x^{-1} y_{x,a}) \le d\, u^\top \Sigma_x u,$$

and hence $s_x(u) \le d$. Applying these bounds inside (7) proves (8).

Proof of Corollary 7. Every mixture satisfies $V_\mu \succeq \alpha H$, while $\Sigma_x \preceq \beta H$. Thus $\Sigma_x \preceq (\beta/\alpha)V_\mu$ for every x, so $\kappa_{\mathrm{dom}} \le \beta/\alpha$. Apply Theorem 6.

Proof of Proposition 8. Fix $\rho \in (0,1)$. At every prompt i, define

$$\phi^{\mathrm{al}}_\rho(i,(j,s)) := s\,\bar{c}_j e_j, \qquad \bar{c}_1 = 1, \quad \bar{c}_j = \rho \ \ (j \ne 1).$$

By Theorem 2, the uniform response distribution is the unique local optimum at every prompt. Its covariance is independent of the prompt:

$$\Sigma_x = \Sigma = \frac{1}{d}\,\mathrm{diag}(1, \rho^2, \dots, \rho^2).$$

This matrix has the same eigenvalues, and hence the same condition number, as every covariance in the signed-axis family. For any prompt marginal, $V_\mu = \Sigma$. Every available response has leverage d, so the two-stage design value is d. Finally, $\Sigma_x = V_\mu$ gives $\kappa_{\mathrm{dom}} \le 1$. Conversely, any domination $\Sigma_x \preceq kV_\mu$ averages to $V_\mu \preceq kV_\mu$, so $k \ge 1$.

Proof of Proposition 9. For an arbitrary μ, the covariance ratio in the strong direction of prompt i is

$$\frac{e_i^\top \Sigma_i e_i}{e_i^\top V_\mu e_i} = \frac{1}{\rho^2 + (1-\rho^2)\mu(i)},$$

whereas every weak-direction ratio is at most one. The maximum strong ratio is minimized uniquely by $\mu(i) = 1/d$, giving the stated value of κ_{dom}. In the same direction,

$$s_i(e_i) = \frac{\max_a (e_i^\top y_{i,a})^2}{e_i^\top \Sigma_i e_i} = d.$$

Equations (7) and (5) now give $Q_\rho = d\kappa_{\mathrm{dom}}$.

F.2 Finite matrix formulations

Because the prompt and action sets are finite, both geometric quantities admit finite optimization problems with positive-semidefinite matrix constraints.

Proposition 16. *For fixed full-rank conditional designs, the value in* (2) *is the optimum of*

$$\begin{aligned} &\textit{minimize} && t \\ &\textit{over} && \mu \in \Delta(\mathcal{X}),\ t \in \mathbb{R} \\ &\textit{subject to} && \begin{pmatrix} \sum_{z\in\mathcal{X}} \mu(z)\Sigma_z & y_{x,a} \\ y_{x,a}^\top & t \end{pmatrix} \succeq 0 \quad \textit{for every } (x,a) \in \mathcal{X}\times\mathcal{A}. \end{aligned} \tag{25}$$

The program has $|\mathcal{X}| + 1$ scalar variables and $|\mathcal{X}||\mathcal{A}|$ matrix inequalities of order $d + 1$.

Proof. Each V_μ is positive definite. The Schur complement gives

$$\begin{pmatrix} V_\mu & y \\ y^\top & t \end{pmatrix} \succeq 0 \quad\Longleftrightarrow\quad y^\top V_\mu^{-1} y \le t.$$

Imposing the inequality for every action and minimizing t gives (2).

Proposition 17. *The covariance-domination factor in* (6) *is the optimal value of the following program, where $\mathbb{R}_+^{\mathcal{X}}$ denotes the nonnegative weights indexed by prompts:*

$$\begin{aligned} \kappa_{\mathrm{dom}} = \min_{w\in\mathbb{R}_+^{\mathcal{X}}} \quad & \sum_{z\in\mathcal{X}} w_z \\ \textit{subject to} \quad & \sum_{z\in\mathcal{X}} w_z \Sigma_z \succeq \Sigma_x \quad \textit{for every } x \in \mathcal{X}. \end{aligned} \tag{26}$$

Proof. For fixed μ, the maximum in (6) is the least k such that $\Sigma_x \preceq kV_\mu$ for every x. Setting $w_z = k\mu(z)$ yields a feasible w with $\sum_z w_z = k$. Conversely, any feasible nonzero w gives $k = \sum_z w_z$ and $\mu(z) = w_z/k$, which reverses the construction.

F.3 Bounds from covariance domination

Rank-one domination converts leverage bounds into positive-semidefinite inequalities. This yields bounds in terms of d and $|\mathcal{X}|$.

Lemma 18. *For $V \succ 0$, $y \in \mathbb{R}^d$, and $t \ge 0$,*

$$y^\top V^{-1} y \le t \quad\Longleftrightarrow\quad yy^\top \preceq tV. \tag{27}$$

Proof. Premultiplying and postmultiplying by $V^{-1/2}$, with $z := V^{-1/2}y$, reduces the claim to $zz^\top \preceq tI_d$. The only nonzero eigenvalue of $zz^\top$ is $\|z\|_2^2 = y^\top V^{-1} y$.

Thus $Q(\phi, \pi_{\mathrm{cg}})$ is the smallest scale at which one covariance mixture dominates every centered action outer product, while κ_{dom} is the corresponding scale for every local covariance.

Proposition 19. *For fixed full-rank local covariances and $m = |\mathcal{X}|$,*

$$1 \le \kappa_{\mathrm{dom}} \le \min\{d, m\}. \tag{28}$$

For conditional G-optimal designs with local value d, the following implications hold for any $K, c > 0$:

1. *$\kappa_{\mathrm{dom}} \le K$ implies $Q(\phi, \pi_{\mathrm{cg}}) \le Kd$;*
2. *$Q(\phi, \pi_{\mathrm{cg}}) \ge cd^2$ implies $\kappa_{\mathrm{dom}} \ge cd$.*

Proof. If $\Sigma_x \preceq kV_\mu$ for every x, averaging with weights $\mu(x)$ gives $V_\mu \preceq kV_\mu$, so $k \ge 1$. Taking $w_x = 1$ in (26) gives $\kappa_{\mathrm{dom}} \le m$. For the dimension bound, choose μ_D maximizing $\log\det V_\mu$ and write $V_D = V_{\mu_D}$. A maximizer exists because the simplex is compact and every mixture is positive definite. Moving from V_D toward any Σ_x is feasible, so first-order optimality gives

$$0 \ge \frac{\mathrm{d}}{\mathrm{d}t} \log\det\big((1-t)V_D + t\Sigma_x\big)\Big|_{t=0^+} = \mathrm{tr}(V_D^{-1}\Sigma_x) - d.$$

The positive-semidefinite matrix $V_D^{-1/2}\Sigma_x V_D^{-1/2}$ has largest eigenvalue at most its trace, hence $\Sigma_x \preceq dV_D$ for every x. Thus $\kappa_{\mathrm{dom}} \le d$, and the two implications follow from (8).

G Full-support joint designs

Prompt-wise specialization attains the trace lower bound, and full-support conditionals approach the same value.

Proof of Theorem 10. For any feasible joint design, write $W := W(\mu, \pi)$ and average its action leverage under the joint sampling distribution:

$$\sum_x \mu(x)\mathbb{E}_{a\sim\pi_x}[(y^\pi_{x,a})^\top W^{-1} y^\pi_{x,a}] = \mathrm{tr}\left(W^{-1}\sum_x \mu(x)\Sigma_{\pi_x}(x)\right) = d.$$

The maximum over all prompt–action pairs is therefore at least d.

For the design in the theorem, every conditional mean is zero and $\Sigma_{\pi_i}(i) = e_i e_i^\top$. Hence

$$W(\mu_{\mathrm{unif}}, \pi) = \frac{1}{d}\sum_{i=1}^{d} e_i e_i^\top = \frac{1}{d} I_d.$$

The unit-scale actions have leverage d, and the weak actions have leverage $d\rho^2 \le d$. The maximum is d, which matches the trace lower bound. Dividing (5) by d proves (9).

Although the exact optimizer is sparse, full-support conditionals can approach its value. For $0 < \eta < 1$, put total probability $1 - \eta$ on the strong signed pair at prompt i and spread the remaining mass uniformly across weak signed pairs:

$$\pi_i^\eta((j,s)) = \begin{cases} (1-\eta)/2, & j = i, \\ \eta/[2(d-1)], & j \ne i. \end{cases} \tag{29}$$

With $\mu = \mu_{\mathrm{unif}}$, set $W_\eta := W(\mu_{\mathrm{unif}}, \pi^\eta)$. Direct averaging gives

$$W_\eta = \frac{1 - \eta + \eta\rho^2}{d} I_d, \qquad \max_{i,j,s} \phi_\rho(i,(j,s))^\top W_\eta^{-1} \phi_\rho(i,(j,s)) = \frac{d}{1 - \eta + \eta\rho^2}. \tag{30}$$

The value converges to d as $\eta \downarrow 0$, while every response has positive probability and every conditional covariance is full rank.

H BRADLEY–TERRY LOWER AND UPPER BOUNDS

The two-stage lower bound and joint upper bound use different reductions: a two-point test for the former and coordinatewise estimation for the latter.

H.1 A GENERAL FIXED-DESIGN TESTING LEMMA

The two-point argument of Theorem 11 extends to any fixed prompt marginal and fixed response conditionals.

Lemma 20. *For $t \in \mathbb{R}$, let $B_0 = \mathrm{Bernoulli}(1/2)$ and $B_t = \mathrm{Bernoulli}(\sigma(t))$. Then*

$$\mathrm{KL}(B_0 \| B_t) = \log\cosh(t/2) \le \frac{t^2}{8}. \tag{31}$$

Proof. Direct substitution gives the equality. The function $f(t) = \log\cosh(t/2)$ satisfies $f(0) = f'(0) = 0$ and $f''(t) = 1/[4\cosh^2(t/2)] \le 1/4$. Taylor's theorem therefore gives $f(t) \le t^2/8$.

Proposition 21. *For $R, \varepsilon > 0$ and an integer $n \ge 1$, fix response conditionals $\pi(\cdot \mid x)$ and a prompt marginal μ, and let $y_{x,a}$ denote the corresponding centered features. Define*

$$\Sigma_x := \mathbb{E}_{a\sim\pi(\cdot|x)}[y_{x,a}y_{x,a}^\top], \qquad V_\mu := \sum_x \mu(x)\Sigma_x \succ 0.$$

Let

$$F_\mu = \max_{x,a} y_{x,a}^\top V_\mu^{-1} y_{x,a},$$

and choose an attaining centered feature $y_\star$. Draw the n comparisons independently, each according to

$$X \sim \mu, \qquad A, B \overset{\text{iid}}{\sim} \pi(\cdot \mid X), \qquad Y \mid X, A, B \sim \mathrm{Bernoulli}\big(\sigma\big(\theta^\top[y_{X,A} - y_{X,B}]\big)\big).$$

Define

$$\Delta := \frac{2\varepsilon}{F_\mu} V_\mu^{-1} y_\star. \tag{32}$$

If $\|\Delta\|_2 \le R$ and $n \le F_\mu/(8\varepsilon^2)$, then every estimator satisfies

$$\sup_{\theta\in\{0,\Delta\}} \max_{x,a} \mathbb{E}_\theta[(y_{x,a}^\top(\widehat{\theta} - \theta))^2] \ge \frac{3}{8}\varepsilon^2. \tag{33}$$

Proof. The target score at $y_\star$ changes by

$$y_\star^\top \Delta = \frac{2\varepsilon}{F_\mu} y_\star^\top V_\mu^{-1} y_\star = 2\varepsilon.$$

For the response difference $Z := y_{X,A} - y_{X,B}$, conditional independence and centering give $\mathbb{E}[ZZ^\top] = 2V_\mu$. Let P_θ^n denote the law of n independent comparisons under parameter θ. Lemma 20 and the chain rule give

$$\begin{aligned}
\mathrm{KL}(P_0^n \| P_\Delta^n) &\le \frac{n}{8}\mathbb{E}[(Z^\top\Delta)^2] \\
&= \frac{n}{4}\Delta^\top V_\mu \Delta \\
&= \frac{n\varepsilon^2}{F_\mu} \le \frac{1}{8}.
\end{aligned}$$

Pinsker's inequality bounds the total variation distance by $1/4$. Thresholding the estimated score $y_\star^\top\widehat{\theta}$ at ε reduces estimation to testing; an error forces squared loss at least ε^2. With TV denoting this distance, the sum of the two risks is at least $\varepsilon^2(1 - \mathrm{TV}) \ge 3\varepsilon^2/4$, so their maximum is at least $3\varepsilon^2/8$.

To apply Proposition 21 to the signed-axis family, choose a least-covered coordinate j. Then $y_\star = e_j$, $F_\mu = 1/(V_\mu)_{jj}$, and $V_\mu^{-1} y_\star / F_\mu = e_j$. Thus $\Delta = 2\varepsilon e_j$, and the radius condition is exactly $2\varepsilon \le R$. Since $F_\mu \ge Q_\rho$, the proposition recovers Theorem 11 uniformly over every prompt marginal.

H.2 COORDINATEWISE ESTIMATION UNDER THE JOINT SAMPLER

Under the joint sampler, informative comparisons decouple by coordinate and reduce to clipped Bernoulli estimation.

Proof of Theorem 12. At prompt i, write the two sampled strong responses as (i, S_A) and (i, S_B), with independent uniform signs $S_A, S_B \in \{-1, +1\}$. A comparison is informative for coordinate i when its prompt is i and $S_A \neq S_B$. Let T_i be the number of such comparisons. Then

$$T_i \sim \text{Binomial}\left(n, \frac{1}{2d}\right).$$

For an informative comparison, retain Y when $S_A = +1$ and replace it by $1 - Y$ when $S_A = -1$. The resulting label is $\text{Bernoulli}(p_i)$, where $p_i = \sigma(2\theta_i)$.

Let $\widetilde{p}_i$ be the mean of these labels when $T_i > 0$, and set $\widetilde{p}_i = 1/2$ when $T_i = 0$. Project $\widetilde{p}_i$ onto $[a_R, 1 - a_R]$, where $a_R := \sigma(-2R)$, to obtain $\widehat{p}_i$, and define

$$\widehat{\theta}_i := \frac{1}{2} \log \frac{\widehat{p}_i}{1 - \widehat{p}_i}.$$

Because $|\theta_i| \leq R$, the true p_i belongs to the clipping interval. Projection is nonexpansive, and conditional on T_i,

$$\mathbb{E}_\theta[(\widehat{p}_i - p_i)^2 \mid T_i] \leq \frac{1}{4(T_i \vee 1)}.$$

The derivative of $p \mapsto \frac{1}{2} \log[p/(1-p)]$ is at most $1/(2m_R)$ on that interval. Consequently,

$$\mathbb{E}_\theta[(\widehat{\theta}_i - \theta_i)^2 \mid T_i] \leq \frac{1}{16 m_R^2 (T_i \vee 1)}. \tag{34}$$

To remove the conditioning on T_i, let $T \sim \text{Binomial}(n, q)$. Then

$$\mathbb{E}\left[\frac{1}{T \vee 1}\right] \leq 2\mathbb{E}\left[\frac{1}{T+1}\right] = \frac{2[1 - (1-q)^{n+1}]}{(n+1)q}.$$

Taking $q = 1/(2d)$ bounds this expectation by $4d/n$. Averaging (34) therefore gives

$$\sup_{\theta \in B_2(0,R)} \mathbb{E}_\theta^\star[(\widehat{\theta}_i - \theta_i)^2] \leq \frac{d}{4m_R^2 n}.$$

Finally, Proposition 3 gives $\phi_g = \phi_\rho$, and every action has the form $\phi_\rho(i, (j, s)) = s c_{ij} e_j$ with $c_{ij} \leq 1$. Its centered-score error is therefore at most the corresponding coordinate error, uniformly over all prompts and responses, which proves (12).

I PROPERTIES OF THE SIGNED-AXIS FAMILY

The full parameterized family has finite spaces, bounded features, positive conditional masses, and nonsingular mixture covariances. At $\rho = 1/d$, the local eigenvalues are d^{-3} and d^{-1}, and the optimal mixture covariance is isotropic.

I.1 FINITE SPACES, BOUNDED FEATURES, AND POSITIVE MASS

The prompt and response cardinalities are

$$|\mathcal{X}| = d, \qquad |\mathcal{A}| = 2d,$$

and the unique conditional G-optimal design satisfies $\pi_{\text{cg}}((j, s) \mid i) = 1/(2d) > 0$. Every feature is a scalar multiple of a standard basis vector with magnitude in $\{1, \rho\}$, so

$$\|\phi_\rho(i, (j, s))\|_2 = c_{ij} \leq 1. \tag{35}$$

At a fixed prompt, the response set contains a nonzero signed pair on every coordinate; it therefore affinely spans $\mathbb{R}^d$.

I.2 Conditional and mixture spectra

Equation (15) gives

$$\lambda_{\min}(\Sigma_i) = \frac{\rho^2}{d}, \qquad \lambda_{\max}(\Sigma_i) = \frac{1}{d}.$$

For every prompt marginal, (16) yields

$$(V_\mu)_{jj} = \frac{\rho^2 + (1-\rho^2)\mu(j)}{d} \geq \frac{\rho^2}{d}. \tag{36}$$

Thus

$$V_\mu \succeq \frac{\rho^2}{d} I_d$$

throughout the closed simplex, including its boundary. At μ_{unif}, every diagonal entry is the same:

$$V_{\mu_{\text{unif}}} = \frac{1 + (d-1)\rho^2}{d^2} I_d. \tag{37}$$

The optimal covariance is isotropic even though every local covariance has a distinct strong direction. For $\rho = 1/d$, these formulas specialize to

$$\lambda_{\min}(\Sigma_i) = d^{-3}, \qquad \lambda_{\max}(\Sigma_i) = d^{-1}, \tag{38}$$

$$V_\mu \succeq d^{-3} I_d, \qquad V_{\mu_{\text{unif}}} = \frac{d^2 + d - 1}{d^4} I_d. \tag{39}$$

I.3 Small dimensions and exact ratios

At $\rho = d^{-1/2}$, the local condition number grows linearly with dimension:

$$\text{cond}(\Sigma_i) = d, \qquad Q_{d^{-1/2}} = \frac{d^3}{2d-1}, \qquad \frac{Q_{d^{-1/2}}}{d^2} = \frac{d}{2d-1} \longrightarrow \frac{1}{2}.$$

At $\rho = 1/d$, Q_ρ/d^2 tends to one. Its exact small-dimensional values are:

d	$Q_{1/d}$	$Q_{1/d}/d^2$	Unique prompt marginal
2	$16/5$	$4/5$	$(1/2, 1/2)$
3	$81/11$	$9/11$	$(1/3, 1/3, 1/3)$
4	$256/19$	$16/19$	$(1/4, 1/4, 1/4, 1/4)$

More generally,

$$\frac{Q_\rho}{d^2} = \frac{1}{1 + (d-1)\rho^2}. \tag{40}$$

At $\rho = 1/d$, this equals $1/(1 + d^{-1} - d^{-2}) \to 1$.

I.4 Embedding in a linear-reward Bradley–Terry model

Choose a uniform base policy $\pi_0(\cdot \mid i)$, any parameter radius $R > 0$, and $\theta \in B_2(0, R)$. The score

$$r_\theta(i, (j, s)) = \theta^\top \phi_\rho(i, (j, s)) = s\, c_{ij} \theta_j$$

is finite. Bradley–Terry comparisons use

$$\Pr_\theta(a \succ b \mid i) = \frac{\exp(r_\theta(i, a))}{\exp(r_\theta(i, a)) + \exp(r_\theta(i, b))}, \tag{41}$$

and the corresponding softmax policy assigns positive probability to every response:

$$\pi_\theta(a \mid i) = \frac{\pi_0(a \mid i) \exp(\theta^\top \phi_\rho(i, a))}{\sum_b \pi_0(b \mid i) \exp(\theta^\top \phi_\rho(i, b))}.$$

For any $\theta^* \in B_2(0, R)$, the reward $r^*(i, a) = (\theta^*)^\top \phi_\rho(i, a)$ belongs to this bounded linear family and induces the full-support policy π_{θ^*}. The lower bound uses only $\theta_0 = 0$ and $\theta_1 = 2\varepsilon e_j$, both of which lie in the ball whenever $2\varepsilon \leq R$.

J PREFERENTIAL-DESIGN ASSUMPTIONS FOR THE SIGNED-AXIS FAMILY

The signed-axis construction satisfies the finite-dimensional hypotheses of Eq. (7), Lemma 4.1, and Proposition 4.2 of Kim et al. (2026): finite prompt and response spaces, bounded features, response features that affinely span $\mathbb{R}^d$ at every prompt, fixed centered features in the prompt-marginal step, and a global span of dimension d.

Property	Signed-axis family
Finite spaces	$\lvert\mathcal{X}\rvert = d$ and $\lvert\mathcal{A}\rvert = 2d$.
Feature bound	$\lVert\phi_\rho(i,(j,s))\rVert_2 = c_{ij} \le 1$.
Conditional affine span	At each i, differences between the two signed responses give a nonzero multiple of every e_j; hence the affine hull is $\mathbb{R}^d$, and the augmented vectors $(1,\phi_\rho)$ span a space of dimension $d+1$.
Centered conditional design	Theorem 2 gives the uniform sampler as the unique minimizer. Its mean is zero, its covariance is positive definite, and its maximum leverage is d.
Sequential allocation order	The unique π_{cg} determines $\phi_g = \phi_\rho$ and Σ_i before optimizing $\mu \in \Delta(\mathcal{X})$, matching Eq. (7) of Kim et al. (2026).
Global feature span	Every prompt already contains a nonzero signed pair on each of the d coordinates, so $\{\phi_g(i,a) : i,a\}$ spans $\mathbb{R}^d$.
Invertibility	Equation (16) gives $V_\mu \succeq (\rho^2/d) I_d$ for every prompt marginal, including every simplex boundary point.
Linear-reward realization	Choose the base policy $\pi_0(\cdot \mid i)$ uniform and $r^*(i,a) = (\theta^*)^\top \phi_\rho(i,a)$ for any θ^* in the prescribed parameter ball. Then $\pi_{\theta^*}(a \mid i) \propto \pi_0(a \mid i) e^{(\theta^*)^\top \phi_\rho(i,a)}$, and all policy probabilities are positive.
Bradley–Terry pairs	Score differences $(\theta^*)^\top[\phi_\rho(i,a) - \phi_\rho(i,b)]$ give the standard Bradley–Terry probabilities used in Section 5.

Proposition 4.2 of Kim et al. (2026) gives the $O(d^2)$ upper bound for this fixed-centered-feature criterion, while Corollary 5 supplies a matching $\Omega(d^2)$ signed-axis instance. Hence the worst-case order under the verified assumptions is $\Theta(d^2)$.

We also verify compatibility with the standard pseudoinverse convention. For a positive semidefinite matrix $M \in \mathbb{R}^{d\times d}$, let $M^\dagger$ denote the Moore–Penrose pseudoinverse and $\mathrm{range}(M)$ its column space. For $y \in \mathbb{R}^d$, the standard estimability convention is

$$\ell_M^\dagger(y) = \begin{cases} y^\top M^\dagger y, & y \in \mathrm{range}(M), \\ +\infty, & y \notin \mathrm{range}(M). \end{cases} \tag{42}$$

Proposition 22. *On the signed-axis family, replacing the extended inverse convention by* (42) *leaves the conditional and prompt-marginal optimization problems unchanged.*

Proof. Fix a prompt i, write $y_a := \phi_\rho(i,a)$, and let m_q and M_q be the mean and centered covariance under any response distribution q. The range of M_q is the span of the centered vectors receiving positive q-mass. If every available centered vector $y_a - m_q$ belonged to $\mathrm{range}(M_q)$, then their span would also be contained in that range. But the available signed axes have affine hull $\mathbb{R}^d$, so

$$\mathrm{span}\{y_a - m_q : a \in \mathcal{A}\} = \mathbb{R}^d.$$

Because a singular M_q has a proper range, at least one available centered action lies outside it. The conditional maximum is therefore $+\infty$ under (42), matching the extended-inverse convention. At the unique optimum, $M_q = \Sigma_i \succ 0$, so the inverse and pseudoinverse agree.

At the prompt-marginal stage, $V_\mu \succeq (\rho^2/d) I_d$ for every μ. It is never singular, and again $V_\mu^\dagger = V_\mu^{-1}$. Both optimization problems and all reported values are therefore identical.