LLM ALIGNMENT WITH BIASED PROXY PREFERENCES UNDER HETEROGENEOUS UTILITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Aligning a language model in a specialized domain such as medicine raises the question of whose judgment should count. Expert comparisons are on target but scarce, a general population's are plentiful but systematically biased, and pooling the two learns a utility between the two populations. We ask when the crowd's comparisons still help, which way of using them helps most, and why. Under heterogeneous utilities, we propose a two-headed estimator (THE) of the direction of the experts' utility center from a small gold and a large proxy sample, which fits one utility per population with the sign loss and couples the two through a penalty on their difference; re-weighting the pooled samples is its infinite-penalty limit. We prove that proxy comparisons speed up estimation while their bias is below the statistical error of pooling, and that once the bias exceeds the gold-only error the gain of any re-weighting over gold-only vanishes as the bias grows. For controlled comparisons, which vary one group of criteria at a time, with few contentious criteria and precisely estimated consensus criteria, a variant of THE has a smaller asymptotic error than every re-weighting. With a smoothed loss, a finite penalty improves on the best re-weighting whenever the bias lies, on average, in more curved directions of the gold loss than the error of the best re-weighting. For a bias sparse in s of d features, THE's two-step special case with an ℓ_1 penalty reduces the gold sample size that our sign-loss bounds require by a factor of order $\sqrt{d/s}$, under local conditions. In simulations with the exact sign loss, THE improves on gold-only by 15% at a bias twice the gold-only error, where the best re-weighting improves by 3%; on real answer embeddings with simulated preferences, it improves on the best re-weighting by up to 20% when the bias lies along high-variance features.

1 INTRODUCTION

Preference data used to align a language model (Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023) raise a basic question: whose judgment should count? In a specialized domain, a small group of qualified reviewers may define the desired preference, while comparisons from a broader population, or from AI labelers (Bai et al., 2022b; Lee et al., 2024), are easier to collect. For a medical answer, for example, a reassuring tone and a correct recommendation need not receive the same relative weight from the two groups. The broader population's comparisons can still reveal shared aspects of quality. We ask how to use them without moving the learned preference away from the qualified reviewers' target.

A single Bradley–Terry fit with unrestricted utilities aggregates heterogeneous annotators implicitly, by a Borda count (Siththaranjan et al., 2024), and cannot represent every subpopulation (Chakraborty et al., 2024); personalized and pluralistic approaches instead account for the differences (Park et al., 2024; Poddar et al., 2024; Sorensen et al., 2024), and Zhu et al. (2025) give conditions under which proxy preferences reduce the sample complexity of learning a target policy. Closest to our setting, the sign estimator of Aouad et al. (2025), a maximum score estimator (Manski, 1975), recovers the direction of a population's utility center under symmetric heterogeneity, where maximum likelihood for a single Bradley–Terry model can fail. It handles variation among users within one population but does not remove a shift between two population centers: a proxy fit can estimate the proxy direction accurately while remaining separated from the target.

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.

Borrowing from a biased source has been studied for regression. Data-enriched regression couples the fits to two samples through a quadratic penalty (Chen et al., 2015); Bastani (2021) corrects a proxy fit with an ℓ_1 penalty and proves that averaging and weighted losses improve on the better single source by at most a constant factor, and the correction extends to several sources and to generalized linear models (Li et al., 2022; Tian & Feng, 2023). These analyses rely on smooth losses. The sign loss is discontinuous: its estimators converge at the cube-root rate to non-Gaussian limits (Kim & Pollard, 1990), smoothing yields faster rates under additional smoothness (Horowitz, 1992), and sparse versions exist (Chen & Lee, 2018; Feng et al., 2022), all for a single population. This leaves open the question: *When can comparisons from a biased population improve the estimation of a target population's preference direction, and which way of using them helps most?*

1.1 Problem definition

Choice model. In the Bradley–Terry–Luce model with linear utilities (Bradley & Terry, 1952; Luce, 1959), a comparison from population $k \in \{1, 2\}$ has feature difference $X_{k,i} \in \mathbb{R}^d$ between two alternatives and response $Y_{k,i} \in \{0, 1\}$. Each comparison is answered by a different user, drawn independently from the population, with latent utility vector $\beta_k + \varepsilon_{k,i}$: $\beta_k \in \mathbb{R}^d$ is the population's utility center, and $\varepsilon_{k,i} \in \mathbb{R}^d$ is the user's deviation from it, a random vector symmetric about zero, independent of $X_{k,i}$ and independent across comparisons. Thus

$$\Pr(Y_{k,i} = 1 \mid X_{k,i}, \varepsilon_{k,i}) = \sigma\big(X_{k,i}^\top(\beta_k + \varepsilon_{k,i})\big)\,, \tag{1}$$

where $\sigma(t) = (1+e^{-t})^{-1}$. For language-model preference data, $X_{k,i}$ can be the difference between fixed features of two answers. Population 1 is gold: its center $\beta_1 = \beta^\star$ is the target and $\|\beta^\star\| = 1$. Population 2 is proxy, with center $\beta_2 = \beta^\star + \delta_2$. We observe N_1 gold and N_2 proxy comparisons, with $N = N_1 + N_2$ and often $N_2 \gg N_1$. The populations share the law of X and the law of ε; their centers may differ. Symmetry does not require a finite mean, so we use *center* rather than mean unless a moment assumption is added.

Target and loss. Ranking alternatives by the gold center's linear utility $x^\top\beta^\star$ is unchanged by a positive rescaling. We thus estimate its direction and evaluate any nonzero directional estimate $\widehat{\beta}$ by $\angle(\widehat{\beta}, \beta^\star) = \arccos(\widehat{\beta}^\top\beta^\star/(\|\widehat{\beta}\|\|\beta^\star\|)) \in [0, \pi]$, which vanishes when the estimate induces the same linear ranking. The proxy direction is $q_2 = \beta_2/\|\beta_2\|$ when $\beta_2 \neq 0$; more proxy data can estimate q_2 accurately without making it equal to $\beta^\star$.

The sign estimator. The *sign estimator* of Aouad et al. (2025) minimizes, over unit vectors, the empirical loss

$$\widehat{L}(\beta) = -1 + \frac{2}{N}\sum_{i=1}^{N} \mathbf{1}\{(2Y_i - 1)X_i^\top\beta < 0\}. \tag{2}$$

The loss rewards agreement between the observed choice and the ranking induced by β; zero empirical margins count as correct, as in Appendix B. Under the following conditions, its angular error for one population has a cube-root sample-size rate. The first condition controls individual heterogeneity; the second ensures that comparison features reveal every direction.

Assumption 1 (Heterogeneity)**.** *ε is independent of X, symmetric about zero, has the same law in both populations, and puts positive mass near the origin in every direction:* $\inf_{\|u\|=1} \Pr(|u^\top\varepsilon| \le 1/(2\sqrt{2})) > 0$.

Assumption 2 (Design)**.** *X has a bounded density with bounded support, bounded below by $f_{\min} > 0$ on a ball around the origin, and the same law in both populations.*

Systematic shift and two settings. The sign-based results below specialize to $\|\beta_2\| = 1$ and write $\delta = \beta_2 - \beta^\star = q_2 - \beta^\star$. The first setting places no restriction on the coordinate support of δ and studies how to weight the proxy by direction. The second assumes $\|\delta\|_0 \le s$ in a fixed feature basis and studies how many gold comparisons are needed to correct the differing coordinates. If those coordinates are predefined answer-quality criteria, sparsity means the populations' *normalized* evaluation weights agree outside a few criteria. This is a stronger model condition that depends on the basis; learned embeddings need not satisfy it. Documented differences between labeler populations are of this kind: crowd judgments reward assertiveness and underweight factual errors (Hosking et al., 2024), preference data and AI judges reward length (Singhal et al., 2024; Zheng et al., 2023),

and removing one length term from a model of an AI judge's preferences brings its rankings closer to human ones (Dubois et al., 2024). A proxy that adds weight w to a criterion the gold population ignores, such as length, has a shift δ that is 1-sparse up to a term of order w^2.

Local-asymptotic convention. The limit comparisons keep d fixed and $N_k/N \to \nu_k \in (0,1)$. Choose an orthonormal basis E of $\beta^{\star\perp}$ and set $\phi(v) = \sqrt{1-\|v\|^2}\beta^\star + Ev$. A uniformly directed local shift means $\beta_{2,N} = \phi(rN^{-a}\Theta)$, where Θ is uniform on $\mathbb{S}^{d-2}$, independent of the samples, and $a = 1/3$ for the exact loss or $a = 2/5$ for smoothing. Thus $\|\delta_N\| = rN^{-a} + O(N^{-3a})$; the random direction is tangent. Oracle tuning is deterministic and independent of the realized Θ. Limits in r are taken after the sample-size limit.

1.2 Results and techniques

Our main estimator is the *two-headed estimator* (THE): it fits one direction per population with the sign loss and couples the two fits through a quadratic penalty (Section 2). Its limits of zero and infinite penalty are gold-only estimation and *re-weighting* (WE), which minimizes one weighted loss of the pooled samples, so at its best tuning THE is never worse than either; the question is when the penalty makes it strictly better.

Theorem 1 (Informal; Theorem 4 and Proposition 1)**.** *Use the exact sign loss and a bias of the order of the noise, in a uniformly random direction. (a) While the bias is below the noise of the pooled sample, the proxy sample speeds up estimation from the gold-only scale $N_1^{-1/3}$ to the pooled scale $N^{-1/3}$, up to a logarithmic factor and constants made explicit in Section 3. (b) Once the bias is $m > 1$ times the gold-only error, every re-weighting has at least $(1-1/m)^{2/3}$ times that error, so its gain vanishes as the bias grows. (c) For a variant of THE on controlled comparisons (Section 3), which vary one group of criteria at a time, with few contentious criteria, consensus directions that are curved enough and have little noise, and a bias above the gold-only error, THE has a smaller error than every re-weighting.*

Parts (a)–(c) compare upper bounds for a use of the proxy with lower bounds for the baselines, the form of Bastani (2021) for regression, here for the sign estimator of Aouad et al. (2025), with lower bounds that we derive from the cube-root limit theory of Kim & Pollard (1990). Re-weighting fails because it shrinks the gold fit toward the proxy fit by the same fraction in every direction; the penalty shrinks less where the gold data are precise. Beyond controlled comparisons the estimators share the rate $N^{-1/3}$ and the constants that decide between them have no closed form; smoothing the sign loss, as in the implementation of Aouad et al. (2025), makes them exact.

Theorem 2 (Informal; Theorems 5, 6 and 28)**.** *With a smoothed sign loss and a bias of order $N^{-2/5}$, all estimators converge at this rate with exact constants; define C_{gold}, C_{WE} and C_{THE} as the infimum of the constant over each family. A finite penalty makes THE strictly better than the best re-weighting, $C_{\mathrm{THE}} < C_{\mathrm{WE}} < C_{\mathrm{gold}}$, whenever the gold loss is more curved, on average, along the bias than along the error of the best re-weighting. When the most curved directions are the least noisy, this holds for small bias, and with a common bandwidth the gain of THE over gold-only remains a fixed multiple, larger than one, of that of re-weighting as the bias grows.*

In linear regression with Gaussian noise, five or more coefficients and ten or more error degrees of freedom, a data-driven scalar average of the two least-squares fits (Chen et al., 2015) makes the small sample alone inadmissible at every bias size, and the correction of Bastani (2021) is compared with the best weighting by rates; here the penalty improves the constant of the best re-weighting, which at a fixed rescaled bias and fixed sample shares divides the sample size needed for a given accuracy by $(C_{\mathrm{WE}}/C_{\mathrm{THE}})^{5/2}$.

Our third result concerns sparse shifts, in the sense of Section 1.1. For them we use a special case of THE, its corner $\rho \to 1$ with an ℓ_1 penalty (Section 2).

Theorem 3 (Informal; Theorem 7)**.** *If the shift has s nonzero coordinates, then under local conditions this estimator has angular error $\widetilde{O}((\sqrt{sd}/N_1)^{1/3} + (d/N_2)^{1/3})$ with the exact sign loss, against $\widetilde{O}((d/N_1)^{1/3})$ for gold-only. With a known Gaussian heterogeneity law and a sparse difference between the utility centers, its error is at most a constant times $\sqrt{s \log d/d}$ times the lower bounds of gold-only, proxy-only, averaging and re-weighting at their best tuning, for fixed d, when the proxy sample is large, the shift is not too small and the sample sizes meet the thresholds of Theorem 7.*

For regression, Bastani (2021) replaces d by s, up to a logarithmic factor, in the gold term. With the sign loss the correction replaces it by $\sqrt{sd}$; with smoothing, and constants that depend on the design, the gold term is $(s \log d/N_1)^{2/5}$, the single-population rate of Feng et al. (2022). Appendix L discusses further related work.

Numerical findings. Experiments (Section 6) address what the analysis leaves open. With smoothing, asymptotic constants over random designs (Appendix K.6) indicate that low-noise curved directions help at every bias size we computed, that no finite penalty helps whenever a large one does not, and that the gain persists with separate bandwidths, which we prove only in limits.

2 Transfer estimators and benchmarks

Let $\widehat{L}_1$ and $\widehat{L}_2$ be the empirical sign losses (2) on the gold and proxy samples. The procedures below fit unit vectors. Section 4 also studies their smoothed versions; part (c) of Theorem 7 instead uses a marginal likelihood on raw centers.

Two-headed estimation without coordinate sparsity. We fit one direction per population,

$$(\widehat{\beta}_1, \widehat{\beta}_2) \in \arg \min_{\|\beta_1\|=\|\beta_2\|=1} (1-\rho)\widehat{L}_1(\beta_1) + \rho\widehat{L}_2(\beta_2) + \lambda\|\beta_1 - \beta_2\|^2, \tag{3}$$

with a weight $\rho \in (0,1)$ and a penalty $\lambda \geq 0$, and output the gold head $\widehat{\beta}_{\mathrm{THE}}(\rho,\lambda) = \widehat{\beta}_1$. At $\lambda = 0$ the output is the sign estimator on the gold sample; as $\lambda \to \infty$ the output minimizes the weighted loss $(1-\rho)\widehat{L}_1 + \rho\widehat{L}_2$.

Why a finite penalty can help. Write H for the Hessian of the gold population loss in tangent coordinates at $\beta^\star$. Under the regularity conditions of Section 4, the matrix weights on the proxy fit are

$$W_{\mathrm{WE}} = \rho I, \qquad W_{\mathrm{THE}} = \rho\Big(I + \frac{\rho(1-\rho)}{2\lambda} H\Big)^{-1} \tag{4}$$

for re-weighting and THE. Weighting moves the same fraction in every direction, while THE moves least along the most curved eigenvectors of H; this helps when gold noise is also small there, which curvature alone does not imply.

Sparse correction in a fixed feature basis. For sparse shifts we use a special case of THE. At its corner $\rho \to 1$, with λ proportional to $1-\rho$, the proxy head is the proxy fit $\widehat{\beta}_{\mathrm{proxy}} \in \arg\min_{\|\beta\|=1} \widehat{L}_2(\beta)$ and the gold head corrects it on the gold sample. With an ℓ_1 penalty on the difference, and within a neighborhood of the proxy fit, this gives the ℓ_1-corrected estimate

$$\widehat{\beta}_{\ell_1} \in \arg \min_{\substack{\|\beta\|=1 \\ \|\beta - \widehat{\beta}_{\mathrm{proxy}}\|_2 \leq r_{\mathrm{loc}}}} \Big\{ \widehat{L}_1(\beta) + \lambda\|\beta - \widehat{\beta}_{\mathrm{proxy}}\|_1 \Big\}. \tag{5}$$

As in THE, the large proxy sample estimates the common direction, and the penalty on the difference of the heads, which in the ℓ_1 norm depends on the coordinate basis, lets the gold comparisons correct only the few coordinates on which the normalized population directions differ.

Benchmarks. (i) *Gold-only* is the sign estimator on the gold sample, $\widehat{\beta}_{\mathrm{gold}} \in \arg\min \widehat{L}_1$. (ii) *Re-weighting* minimizes a single weighted loss, $\widehat{\beta}_{\mathrm{WE}}(\rho) \in \arg\min_{\|\beta\|=1}(1-\rho)\widehat{L}_1(\beta) + \rho\widehat{L}_2(\beta)$, and is the $\lambda = \infty$ boundary of THE. Ordinary pooling is the weight $\rho = N_2/N$; a fixed shift can leave it with a nonvanishing bias (Theorem 14). (iii) The *quadratic two-step correction* (TSE) first fits the proxy sample and then corrects its fit with the gold sample, in the spirit of Bastani (2021):

$$\widehat{\beta}_{\mathrm{TSE}}(\lambda) \in \arg \min_{\|\beta\|=1} \widehat{L}_1(\beta) + \lambda\|\beta - \widehat{\beta}_{\mathrm{proxy}}\|^2, \tag{6}$$

a distinct finite-sample objective; with smoothing, its asymptotic matrix weight is the limit $\rho \uparrow 1$ of those of THE (Appendix E).

3 Unrestricted shifts: when does the proxy help?

This section keeps the exact sign loss and adds to Assumptions 1–2 a condition on population losses.

Assumption 3 (Quadratic bracket). *In the tangent coordinates v at $\beta^\star$ of Section 1.1, so that $\beta = \phi(v)$ and the proxy center is $\beta_2 = \phi(\zeta)$, on a fixed neighborhood containing both centers, the gold excess loss is between $Q(v)$ and $Q'(v)$ and the proxy excess loss between $Q(v-\zeta)$ and $Q'(v-\zeta)$, where $Q(u) = \sum_j c_j \|u_j\|^2$ and $Q'(u) = \sum_j c'_j \|u_j\|^2$ for one orthogonal block decomposition, $0 < c_j \le c'_j$, and $\gamma = \max_j (c'_j - c_j)/\min_j c_j$.*

The upper bracket and the blocks let the losses of the two populations be compared direction by direction. Let $\ell = d \log N + \log(1/\eta)$; the cube-root noise scales of the two samples are $A(\ell/N_1)^{1/3}$ and $A(\ell/N_2)^{1/3}$, where A depends only on the design and on $\min_j c_j$ (Appendix B). All upper bounds hold with probability at least $1-\eta$ on one event, for all estimators at once, so that they can be compared directly; since the angle is at most π, the choice $\eta = 1/N$ turns each into a bound on the expected angle at the cost of a term π/N.

Theorem 4 (When the proxy helps). *Let Assumptions 1–2 hold with a continuous design density and $\mathbb{E}\|\varepsilon\| < \infty$, let the bias be of the order of the noise, $\|\delta\| = rN^{-1/3}$ with $r \ge 0$ fixed in the local-asymptotic convention of Section 1.1, and let $N_k/N \to \nu_k \in (0,1)$ for both samples. There is a constant $K > 0$, depending only on the design and the law of ε, such that*

$$\begin{aligned} \mathbb{E}\angle(\widehat{\beta}_{\mathrm{gold}}, \beta^\star) &\ge \frac{K + o(1)}{N_1^{1/3}}, \\ \inf_\rho \mathbb{E}\angle(\widehat{\beta}_{\mathrm{WE}}(\rho), \beta^\star) &\ge (1 + o(1)) \min\Big\{ \frac{K}{2^{2/3} N_1^{1/3}},\ \frac{\|\delta\|}{4} + \frac{K}{2^{5/3} N_2^{1/3}} \Big\}, \end{aligned} \tag{7}$$

and if $\|\delta\| \ge mKN_1^{-1/3}$ with $m > 1$, no weight brings the error below $(1 - 1/m)^{2/3}$ times that of gold-only. Under Assumption 3 in addition, if $\|\delta\| \le A(\ell/N)^{1/3}$, then with probability at least $1-\eta$ and up to lower-order terms the pooled sign estimator has angular error at most $(2 + \sqrt{2\gamma})A(\ell/N)^{1/3}$, and the two-headed and the two-step estimator at their best tuning have upper bounds of the same order.

Below the pooled noise scale, the ratio of the pooled upper bound to the gold-only lower bound is at most $[(2 + \sqrt{2\gamma})A/K](\ell N_1/N)^{1/3}$ up to lower-order terms; with $\eta = 1/N$, as above, it bounds the ratio of expected errors. A strict comparison requires this full expression to be below one. With fixed positive shares both procedures have the same cube-root exponent. At the level of constants no logarithmic factor enters: pooling is strictly better than gold-only while the bias is below a multiple of the gold-only error, by the factor $(N_1/N)^{1/3}$ at $\delta = 0$ (Corollary 2). When the bias exceeds the gold-only error by a factor $m > 1$, re-weighting can remove at most the fraction $1 - (1 - 1/m)^{2/3}$ of that error. Like the constant of the Chernoff distribution, K is the mean norm of the minimizer of a Gaussian process with quadratic drift and is known only numerically (Appendix C).

What the upper bounds show about the penalty. Under Assumption 3, in every design the bias part of the bound on the two-headed estimator is $\|\delta\|$ times the root-mean-square of the eigenvalues of the matrix weight (4) at the bracket constants, below the bias part $\rho\|\delta\|$ of re-weighting (Theorems 10 and 8): the penalty reduces the proxy weight most in the blocks of large curvature. The coupling adds the error of the proxy head, so the noise part can grow. These bounds cannot rank the estimators, which share the rate $N^{-1/3}$; the next result ranks them for a natural class of designs, and Section 4 in general under smoothing.

Controlled comparisons. Call comparisons *controlled* if the evaluation criteria split into a contentious and a consensus group, spanning complementary subspaces of the tangent space of dimensions p_L and $d-1-p_L$, and each comparison differs along one group only, with a continuous bounded density of its feature difference on the span of $\beta^\star$ and that group; pairs that change one aspect of an answer, such as its length, tone or facts, are of this kind. This replaces Assumption 2, and the curvature becomes block diagonal. For this singular design only, use the tie-neutral loss $-(2Y-1)\operatorname{sign}(X^\top\beta)$, with $\operatorname{sign}(0) = 0$, and minimize over a deterministic finite N^{-3}-net of the unit sphere, using the same net for both heads. This theoretical variant has an attained minimum and the limit laws proved in Appendix C.7. The zero-margin convention in (2) would invalidate global identification here. Let $K_L N_1^{-1/3}$ be the gold-only noise scale on the contentious directions, $\|\delta\| = mK_L N_1^{-1/3}$ with a uniformly random direction, h_S the smallest curvature of the gold loss on the consensus directions and $\bar{c}_S$ the noise constant of the gold head there at the pooling weight (Appendix C.7).

Proposition 1 (THE below every re-weighting). *Let Assumption 1 hold with $\mathbb{E}\|\varepsilon\| < \infty$ for controlled comparisons, $m \geq 1$ and $m\sqrt{p_L/(d-1)} + (N_1/N)^{1/3} < (1-1/m)^{2/3}$. For every $\epsilon > 0$ there is λ_0 such that, for each $\lambda \geq \lambda_0$, if the consensus directions are curved enough and have little noise, $h_S \geq 8\lambda mN/(\epsilon N_1)$ and $\bar{c}_S \leq \epsilon/4$, the limit of $N_1^{1/3}\,\mathbb{E}\angle(\widehat{\beta}_{\mathrm{THE}}(N_2/N, \lambda), \beta^\star)/K_L$ is at most $m\sqrt{p_L/(d-1)} + (N_1/N)^{1/3} + \epsilon$, while that of re-weighting at every weight is at least $(1-1/m)^{2/3}$, the lower bound of Theorem 4, proved for this design in Appendix C.7; for small ϵ, THE lies below every re-weighting.*

With the exact sign loss, the two conditions on the consensus group replace the condition of Section 4 that curved directions have low noise. Re-weighting has bias $\rho\|\delta\|$ at every weight and must keep ρ below $1/m$; the penalty removes most of the weight on the proxy fit in the consensus directions, where the gold data are precise, so THE keeps the pooling weight on the contentious ones and has bias at most $\sqrt{p_L/(d-1)}\,\|\delta\|$.

4 Unrestricted shifts: exact constants under smoothing

Assumption 4 (Smooth design). *The density of X is twice continuously differentiable with compact support, and $d \geq 3$, so that the tangent space has dimension at least two.*

Assumption 5 (Second moment). $\mathbb{E}\|\varepsilon\|^2 < \infty$.

Under them the sign can be replaced by $\tanh(u/h)$ with a bandwidth h. Smoothing shifts the population minimizer by an offset of order h^2 and reduces the noise to order $(Nh)^{-1/2}$; bandwidths $h \asymp N^{-1/5}$, one per population, balance the two, and the error is of order $N^{-2/5}$. We assume that the smoothing offset is nonzero (otherwise the constants at this rate vanish), let N_1/N converge to a share in $(0, 1/2)$, so that the proxy sample is the larger one, and measure the bias in units of the error, $\|\delta\| = rN^{-2/5}$ with $r > 0$ fixed.

Proposition 2 (Asymptotic error). *Under Assumptions 1, 2, 4 and 5, the rescaled error $N^{2/5}\,\mathbb{E}\angle(\widehat{\beta}, \beta^\star)$ converges, uniformly over hyperparameters in compact sets and over bias directions, to*

$$\mathbb{E}\left\|U_1 + W\,(U_2 - U_1)\right\|, \tag{8}$$

where U_1 and U_2 are the asymptotic errors of the gold-only and the proxy-only estimator, Gaussian vectors centered at smoothing offsets proportional to the squared bandwidths, the proxy one shifted in addition by the bias, and W is the matrix weight of the estimator.

Define C_{gold}, C_{WE}, C_{TSE} and C_{THE} as the infima of the limiting risks (8) of the four families over their hyperparameters and positive bandwidth constants. For every $r > 0$ the best weight ρ_{WE} lies strictly between 0 and the pooling weight ν_2, and $C_{\mathrm{WE}} < C_{\mathrm{gold}}$ (Lemma 42). Since re-weighting and the two-step estimator are boundaries of the two-headed family, $C_{\mathrm{THE}} \leq \min\{C_{\mathrm{WE}}, C_{\mathrm{TSE}}\}$, with equality throughout when H is a multiple of the identity (Proposition 6).

The criterion. Let $U = U_1 + \rho_{\mathrm{WE}}(U_2 - U_1)$ be the asymptotic error of the best weighted estimator and Θ the random direction of the bias. Two averages of the eigenvalues of H decide whether a penalty helps,

$$h_{\mathrm{bias}} = \frac{\mathbb{E}[U^\top H\Theta/\|U\|]}{\mathbb{E}[U^\top\Theta/\|U\|]}, \qquad h_{\mathrm{noise}} = \frac{\mathbb{E}[U^\top HU/\|U\|]}{\mathbb{E}\|U\|}\,: \tag{9}$$

the curvature along the bias, which weights each direction by how much of the bias remains in the error of the best weighted estimate, and the curvature along the full weighted-estimation error, including its smoothing offset and residual proxy bias.

Theorem 5 (Penalty versus best weight; proved in Appendix E). *Under the assumptions above, let ρ_{WE} and its two bandwidths minimize the limiting risk of re-weighting, and let $C_{\mathrm{THE}}(\rho_{\mathrm{WE}}, \lambda)$ be the constant of THE with this weight, these bandwidths and penalty λ. If*

$$h_{\mathrm{bias}} > h_{\mathrm{noise}}, \tag{10}$$

then $C_{\mathrm{THE}}(\rho_{\mathrm{WE}}, \lambda) < C_{\mathrm{WE}}$ for all large enough λ, the minimum over λ is attained at a finite $\lambda_\star > 0$ with $0 < C_{\mathrm{THE}} \leq C_{\mathrm{THE}}(\rho_{\mathrm{WE}}, \lambda_\star) < C_{\mathrm{WE}} < C_{\mathrm{gold}}$, and

$$\mathbb{E}\angle\big(\widehat{\beta}_{\mathrm{THE}}(\rho_{\mathrm{WE}}, \lambda_\star), \beta^\star\big) = \frac{C_{\mathrm{THE}}(\rho_{\mathrm{WE}}, \lambda_\star) + o(1)}{N^{2/5}}. \tag{11}$$

When $h_{\mathrm{bias}} < h_{\mathrm{noise}}$, instead, $C_{\mathrm{THE}}(\rho_{\mathrm{WE}}, \lambda) > C_{\mathrm{WE}}$ for all large enough λ.

The improving estimator has fixed hyperparameters, so (11) follows from Proposition 2 without interchanging tuning and limits. The penalty helps when the bias lies, on average, in more curved directions than the error of the best re-weighting; the two averages coincide when the curvature is isotropic, where no penalty can help. The proof rests on two cancellations, of the smoothing offset at the optimal bandwidths and of the scalar part of the curvature at the optimal weight (Appendix E).

Theorem 6 (When the criterion holds, and the best penalty; proved in Appendix E)**.** *Assume in addition that H and the covariance V of the gold noise have the same eigenvectors with eigenvalues in reverse order (curved directions have low noise), neither of them a multiple of the identity, and that the nonzero smoothing offset points along the least curved direction. Centered comparisons with fixed bandwidths are treated separately in Appendix E. Then* (10) *holds for all sufficiently small r. Keeping the WE-optimal weight and bandwidths, the best penalty along this path scales as $\lambda \asymp r^{-2}$, and $C_{\mathrm{WE}} - \min_\lambda C_{\mathrm{THE}}(\rho_{\mathrm{WE}}, \lambda) \asymp r^4$ as $r \to 0$. If moreover V is proportional to H^{-1}, the best penalty is, up to a factor $1 + o(1)$,*

$$\lambda = F \, \frac{\mathrm{tr}(HV)}{2hN\|\delta\|^2}, \tag{12}$$

where h is the small-bias bandwidth of the best weighted estimator and $F > 1$ is an explicit factor that accounts for the angular loss.

The condition is an association between curvature and precision: H and the covariance of the gold score are both second moments of the design on the hyperplane $\{x : x^\top \beta^\star = 0\}$, weighted differently by the heterogeneity, so when that weight varies little V is close to proportional to H^{-1} and curved directions have low noise. For a common bandwidth the relevant derivative uses the different criterion Δ_{c} in (142). If $\Delta_{\mathrm{c}} < 0$, no penalty improves on the common-bandwidth WE optimum for sufficiently small r; the equality case is not decided by this argument (Theorem 27).

Large bias and the two-step estimator. With one bandwidth for both samples, the advantage over the best weight also holds for large bias: under the reverse-order condition of Theorem 6 and a nonzero smoothing offset along a common eigenvector, the gains over gold-only of THE and of re-weighting both vanish at the rate r^{-2}, and the former exceeds the latter by a fixed factor (Theorem 28). The two-step estimator, the corner $\rho \to 1$ of the family, is never better than the tuned two-headed estimator, and for small bias with a common bandwidth it is strictly worse (Theorem 27).

5 Sparse shifts: correcting localized differences

Here the question is how many gold comparisons are needed when the populations' normalized evaluation weights differ on only s predefined criteria (Section 1.1). For the exact sign and smoothed results both centers are unit; the formal statements specify the radius of the neighborhood of the proxy fit and local bounds on $\|\delta\|$ and on the proxy error $e = \widehat{\beta}_{\mathrm{proxy}} - \beta_2$ (Appendices F–G).

Theorem 7 (Sparse correction; proved in Appendices F–H)**.** *For parts (a)–(b), let $\delta = q_2 - \beta^\star$ have at most s nonzero coordinates; for part (c), let $\delta_2 = \beta_2 - \beta^\star$ have at most s nonzero coordinates. Under the conditions below, the corresponding ℓ_1-corrected proxy fits obey these bounds; parts (a)–(b) hold conditionally on the proxy sample, and part (c) holds with high probability over both samples.*

- *(a) Exact sign loss, Assumptions 1–3, the localization conditions of Theorem 29, and the penalty level $\lambda \asymp (d/s^2)^{1/6}\sqrt{\log d}\, N_1^{-1/3}$: the angular error is $\tilde{O}\big((\sqrt{sd}/N_1)^{1/3} + (d/N_2)^{1/3}\big)$, against the upper bound $\tilde{O}((d/N_1)^{1/3})$ of gold-only.*
- *(b) Smoothed loss, Assumptions 1, 2, 4 and 5, a fixed design, bandwidth $h_1 \asymp (s\log(d/\eta)/N_1)^{1/5}$, $\lambda \asymp \sqrt{\log(d/\eta)/(N_1 h_1)}$ and a trust region of radius $h_1/(8R)$ around the proxy fit, R the radius of the design support: under the localization and sample conditions of Theorem 30, given the proxy sample the Euclidean error is $O\big((s\log(d/\eta)/N_1)^{2/5} + \|e\|\big)$, where $\|e\| = O_p(N_2^{-2/5})$ and the constants depend on the design, hence on d.*
- *(c) Known Gaussian heterogeneity law and a design uniform on a ball (Assumption 6, in place of Assumptions 3–5), with marginal likelihoods on raw centers in place of the sign*

loss and $\lambda \asymp \sqrt{\log d/N_1}$: under all sample-size conditions of Theorem 31, including the additional condition for the correction, the Euclidean error for the gold center $\beta^\star$ is $\tilde{O}\big(\sqrt{s/N_1} + \sqrt{d/N_2}\big)$. For fixed d, a shift of order $N^{-1/2}$ with $\|\delta_2\| \geq 4\sqrt{d/N_1}$, and $N_2/N_1 \gtrsim d/(s \log d)$, its expected value is at most of order $\sqrt{s \log d/d}$ times the asymptotic lower bounds of the unpenalized procedures, gold-only, proxy-only, averaging and re-weighting, each at its best tuning, with constants depending on Assumption 6 (Theorem 33).

Part (a) replaces d in the gold term by $\sqrt{sd}$; its comparison is between high-probability upper bounds under the same assumptions. The unit-norm constraint adds at most the proxy error: some unit vector that differs from the proxy fit in at most $2s$ coordinates lies within $\|e\|$ of $\beta^\star$ (Lemma 49). The factor $\sqrt{sd}$ comes from balancing the ℓ_1 geometry of the corrections at coarse scales against halfspace complexity at fine scales (Lemma 51). The sparsity in (c) concerns δ_2; neither this condition nor sparsity of $\delta = q_2 - \beta^\star$ implies the other. The likelihood result uses full empirical Hessian positivity and a localization radius of order $d^{-3/2}$; its sufficient sample thresholds are of order d^4 up to logarithms and design constants. It does not establish this bound for $d > N_1$, and the comparison in (c) is informative once d is large relative to $s \log d$.

6 Numerical Experiments

We compare THE with gold-only estimation, which measures the value of the proxy sample, with pooling, and with re-weighting at its best weight (WE), which measures the value of the penalty. The simulations use the model (1) with Gaussian perturbations and a bias of uniformly random direction and size r times the gold-only error, rather than rN^{-a} as in Section 1.1. Hyperparameters are chosen by a cross-fitted oracle or by an implementable rule. We report the reduction $1 - E/E_{\text{gold}}$ of the gold-only mean error, with paired 95% intervals (Appendix K).

Simulated preference labels (Figure 1(a)). With the exact sign loss (Figure 2(b)), where every estimator is a global minimizer over the points of a grid, THE reduces the gold-only error by 44%, 32% and 15% at $r = 0.5$, 1 and 2, whereas WE reduces it by 37%, 18% and 3%, with significant paired differences (Table 5); under isotropic curvature the two perform alike. As N_2/N_1 grows from 1 to 100 (Figure 3(a)), THE's reduction at $r = 1$ grows from 16% to 40%, while WE's stays between 11% and 27%. With the smoothed loss (Figure 2(a)), THE reduces the gold-only error by 24% at $r = 2$, where WE reduces it by 11%, and asymptotic predictions agree with the simulations

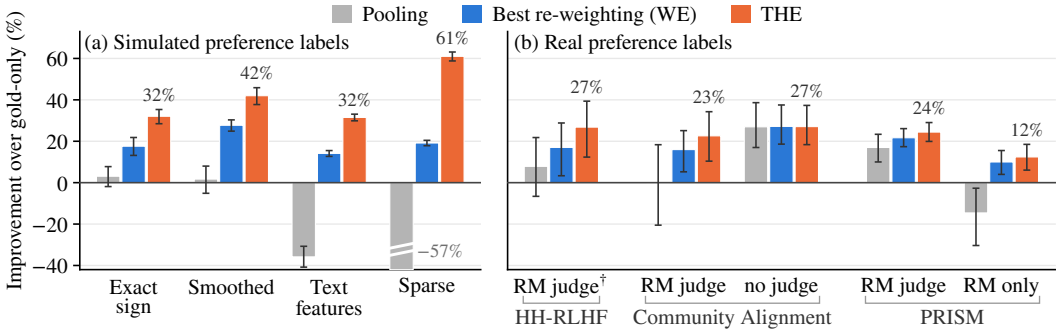


Figure 1: Improvement over gold-only, $1 - E/E_{\text{gold}}$ for the mean error E, of pooling, WE and THE. (a) Simulated preference labels: the exact sign ($d = 3$) and smoothed ($d = 8$) losses and answer embeddings from HH-RLHF with the bias in the top four principal components and implementable tuning, all at a bias of one gold-only error; and a sparse shift in $d = 80$ (Euclidean error), where WE is re-weighting at its oracle weight and THE its ℓ_1 special case with a cross-validated penalty. (b) Real preference labels, with human gold and reference labels, oracle tuning: HH-RLHF (Bai et al., 2022a), Community Alignment (Zhang et al., 2026) and PRISM (Kirk et al., 2024). A reward model (Köpf et al., 2023) relabels half of the proxy (RM judge; †in HH-RLHF the rest of the proxy comes from the gold population), or all of it (RM only). Bars: 95% intervals over replications and, in (b), over the held-out human reference.

(dashed lines). As the bias grows, both methods approach gold-only, but the improvement of THE over gold-only stays a growing multiple of that of WE, 1.5, 2.1 and 4.0 at $r = 1$, 2 and 4, toward the constant factor of Theorem 28 (asymptotic values 1.5, 2.3 and 3.1). On answer pairs from HH-RLHF (Bai et al., 2022a), embedded by a sentence encoder (Reimers & Gurevych, 2019) and reduced to 128 principal components, with the implementable rule, THE and WE reduce the gold-only error at $r = 1$ by 20% and 18% for a uniform bias, and by 32% and 14% within the top four principal components ($N_2 = 30N_1$; Figure 3(b)); with $d = 384$, THE reduces it by 11% at $r = 1$, where WE does not. With a known Gaussian heterogeneity law (Figure 4), the ℓ_1 correction reduces the gold-only error by 34–61% for d from 20 to 80, where the theorem's comparison applies, against 3–19% for re-weighting at its oracle weight, and by 71–88% against 34–67% for d from 160 to 1280, including $d > N_1$, where unconstrained gold-only diverges. With a heterogeneity law of the wrong shape or scale, its gain over re-weighting at its oracle weight stays at 30–57% (Appendix K.3). With the smoothed loss and a difference on two coordinates, the ℓ_1 correction reduces the error of re-weighting at its oracle tuning by 30–58% at $r = 1$ for d from 40 to 160, and its error grows with d at a log-log slope of 0.14, against 0.59 for gold-only (Appendix K.4). With tuned hyperparameters, THE also improves slightly on the two-step correction for $r \leq 2$ with the smoothed loss (Table 3).

Real preference labels (Figure 1(b)). Estimators here use a ridge-penalized logistic likelihood in place of the sign loss. With the gold sample from one population and the proxy from another, the proxy reduces the gold-only error by 27–54% with either method, since these shifts lie in flat directions. When a reward model relabels half of the proxy and the features are its own representation, where its bias lies in curved directions, THE reduces the gold-only error by 24% on PRISM and 23% on Community Alignment, against 22% and 16% for re-weighting, and on Community Alignment the gap agrees with its asymptotic prediction; across the designs with another population's proxy, predictions from the training data order the gains (rank correlations 0.69 to 0.83). Under the implementable rule this gap stays significant but smaller. Rerun with the smoothed sign loss and the reward model on 10–25% of the proxy, THE improves on WE by 4–6% on Community Alignment and on average by at most 1.5% on PRISM and HH-RLHF. When a rule preferring the shorter answer relabels a fifth of the HH-RLHF proxy (other labels from the gold population), the shift lies mostly in the answer-length feature, and the ℓ_1 correction reduces the error of WE by 32–48% for d from 16 to 64; for judges whose shift is not sparse it gives no consistent gain (Appendix K.5).

7 Conclusion

We found that proxy comparisons speed up estimation while their bias is below the pooled noise scale, and re-weighting loses its gain once the bias exceeds the gold-only error, because it shrinks the gold fit toward the proxy fit by the same fraction in every direction. THE shrinks less where the gold data are precise: with a smoothed loss it improves the constant of the best re-weighting when the bias lies, on average, in more curved directions than the error, and for controlled comparisons the same ordering holds for the exact sign loss. When the shift between the populations is sparse, its ℓ_1 special case lets a gold sample smaller by a factor of order $\sqrt{d/s}$ correct the shift. With several proxies, one head and one penalty per proxy can reach the Bayes-optimal combination of the single-population fits (Appendix I). In experiments with simulated preferences and real labels from PRISM, Community Alignment and HH-RLHF, a proxy from another population reduces the gold-only error by 27–54% with either method, and THE reduces it by 32–42% in the simulated designs at a bias of one gold-only error and by 23–24% on PRISM and Community Alignment when a reward model relabels half of the proxy, against 14–28% and 16–22% for WE (27% against 17% on HH-RLHF, whose proxy includes gold-population labels).

Several directions remain open. The design-level sufficient condition for the criterion is proved for small bias, and for large bias only with a common bandwidth; whether a data-driven penalty attains the oracle constant is not known. Outside controlled comparisons, the exact-sign constants can only be compared numerically. The shared laws of the features and of the heterogeneity exclude shifts in the prompt distribution or in the noise level between populations, and the sparse result needs a known feature basis. Natural next steps are finite-sample results for several proxies and weighting the proxy by direction for reward models on learned representations.

AI USE STATEMENT

In this work, we used generative AI tools for grammar and proofread, cleaning and reformatting the dataset. We have not used generative AI tools for other tasks that required disclosure, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to modify figures, edit code, identify relevant literature, and format references. We have reviewed all AI-assisted work. For example, for Chen et al. (2015) we confirmed that their result concerns a data-driven average of the two least-squares fits rather than the penalized coupling, and needs at least five coefficients and ten error degrees of freedom, and we corrected our description accordingly. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The appendix states and proves every result of the paper; Appendix A collects the assumptions and the asymptotic quantities, and Appendix K specifies every design, sample size, number of replications, grid, tuning rule and interval. The real-label experiments use public data (PRISM, Community Alignment and HH-RLHF) and public models (an OpenAssistant reward model, Qwen2.5-1.5B-Instruct and the all-MiniLM-L6-v2 sentence encoder). The supplementary material contains anonymized code that reproduces the figures and tables.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000.

Ali Aouad, Aymane El Gadarri, and Vivek F. Farias. The sign estimator: LLM alignment in the face of choice heterogeneity. *arXiv preprint arXiv:2510.23965*, 2025.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455. PMLR, 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a. URL https://arxiv.org/abs/2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022b. URL https://arxiv.org/abs/2212.08073.

Hamsa Bastani. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984, 2021. doi: 10.1287/mnsc.2020.3729.

Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathématique*, 334(6):495–500, 2002. doi: 10.1016/S1631-073X(02)02292-6.

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.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4971–5012. PMLR, 2024. URL https://proceedings.mlr.press/v235/burns24b.html.

T. Tony Cai and Hongji Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100–128, 2021. doi: 10.1214/20-AOS1949.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=bx24KpJ4Eb.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. MaxMin-RLHF: Alignment with diverse human preferences. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6116–6135. PMLR, 2024. URL https://proceedings.mlr.press/v235/chakraborty24b.html.

Aiyou Chen, Art B. Owen, and Minghui Shi. Data enriched linear regression. *Electronic Journal of Statistics*, 9(1):1078–1112, 2015. doi: 10.1214/15-EJS1027.

Le-Yu Chen and Sokbae Lee. Best subset binary prediction. *Journal of Econometrics*, 206(1):39–56, 2018. doi: 10.1016/j.jeconom.2018.05.001.

Ming-Hui Chen and Joseph G. Ibrahim. The relationship between the power prior and hierarchical models. *Bayesian Analysis*, 1(3):551–574, 2006. doi: 10.1214/06-BA118.

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. 4299–4307, 2017.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mosse, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. Position: Social choice should guide AI alignment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9346–9360. PMLR, 2024. URL https://proceedings.mlr.press/v235/conitzer24a.html.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *First Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2404.04475.

Huijie Feng, Yang Ning, and Jiwei Zhao. Nonregular and minimax estimation of individualized thresholds in high dimension with binary responses. *The Annals of Statistics*, 50(4):2284–2305, 2022. doi: 10.1214/22-AOS2188.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Joel L. Horowitz. A smoothed maximum score estimator for the binary response model. *Econometrica*, 60(3):505–531, 1992. doi: 10.2307/2951582.

Tom Hosking, Phil Blunsom, and Max Bartolo. Human feedback is not gold standard. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7W3GLNImfS.

Jeankyung Kim and David Pollard. Cube root asymptotics. *The Annals of Statistics*, 18(1):191–219, 1990. doi: 10.1214/aos/1176347498.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pp. 105236–105344, 2024.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. OpenAssistant conversations – Democratizing large language model alignment. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, pp. 47669–47681, 2023.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, Berlin, 1991.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 26874–26901. PMLR, 2024. URL https://proceedings.mlr.press/v235/lee24t.html.

Sai Li, T. Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173, 2022. doi: 10.1111/rssb.12479.

Dennis V. Lindley and Adrian F. M. Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(1):1–18, 1972. doi: 10.1111/j.2517-6161.1972.tb00885.x.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York, 1959.

Charles F. Manski. Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, 3(3):205–228, 1975. doi: 10.1016/0304-4076(75)90032-9.

Daniel McFadden and Kenneth Train. Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15(5):447–470, 2000.

Debarghya Mukherjee, Moulinath Banerjee, and Ya'acov Ritov. Optimal linear discriminators for the discrete choice model in growing dimensions. *The Annals of Statistics*, 49(6):3324–3357, 2021. doi: 10.1214/21-AOS2085.

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

Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman Ozdaglar. RLHF from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*, 2024. URL https://arxiv.org/abs/2405.00254.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 52516–52544, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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, 2023.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019. URL https://aclanthology.org/D19-1410/.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 1998.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in RLHF. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=G8LaO1P0xv.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=0tWTxYYPnW.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46280–46302. PMLR, 2024. URL https://proceedings.mlr.press/v235/sorensen24a.html.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981. doi: 10.1214/aos/1176345632.

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, pp. 3008–3021, 2020.

Henri Theil and Arthur S. Goldberger. On pure and mixed statistical estimation in economics. *International Economic Review*, 2(1):65–78, 1961. doi: 10.2307/2525589.

Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544):2684–2697, 2023. doi: 10.1080/01621459.2022.2071278.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

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

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020.

Lily Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Dwivedi-Yu, Vidya Sarma, Kristopher Rose, and Maximilian Nickel. Cultivating pluralism in algorithmic monoculture: The Community Alignment dataset. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ee208bfc04b1bf6125a6a34baa1c28d3-Abstract-Conference.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, pp. 46595–46623, 2023.

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, 2023. URL https://proceedings.mlr.press/v202/zhu23f.html.

Yuchen Zhu, Daniel Augusto De Souza, Zhengyan Shi, Mengyue Yang, Pasquale Minervini, Matt Kusner, and Alexander D'Amour. When can proxies improve the sample complexity of preference learning? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79790–79814. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhu25f.html.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. URL https://arxiv.org/abs/1909.08593.

The appendix states and proves every result of the paper, including the extension of Proposition 2 to several proxy populations.

A NOTATION, ASSUMPTIONS AND THE ASYMPTOTIC QUANTITIES

Table 1: Two settings under the same gold–proxy choice model. The result for a known Gaussian heterogeneity law estimates the gold center itself, in Euclidean loss, and uses stronger distributional information.

Shift structure	Transfer method	Result and comparison
No support restriction on δ	Quadratic THE puts a direction-dependent weight on the proxy fit	Rate bounds for the exact sign loss; with smoothing, a finite penalty can give a smaller asymptotic constant than the best re-weighting under (10).
At most s nonzeros in a fixed feature basis	Proxy fit followed by ℓ_1 gold correction	Gold term $\widetilde{O}((\sqrt{sd}/N_1)^{1/3})$ for the exact sign loss, with a proxy-error term; fixed-design bound for the smoothed loss; extension to a known Gaussian heterogeneity law and a sparse shift of the unnormalized centers (Theorem 7).

The main text uses the smallest possible notation; the appendices use a more detailed notation. Sample sizes N_1 (gold), N_2 (proxy) and $N = N_1 + N_2$ are the same in both; the appendices write n for the total N and ν_k for the limits of the shares N_k/n. Source $k = 1$ is gold and $k = 2$ is proxy. The smoothed losses use source-specific bandwidths $h_k = c_k n^{-1/5}$, which the main text optimizes within the $n^{-1/5}$ scale; a common bandwidth is the case $c_1 = c_2$, and the bandwidth h of Theorem 6 is $c_0 n^{-1/5}$, where c_0 is the small-bias limit of the best weighted bandwidth constant. The main text's cube-root scales $A(\ell/N_1)^{1/3}, A(\ell/N_2)^{1/3}$ are the appendices' κ_1, κ_2, its logarithmic factor ℓ is $\Lambda/C_{\mathrm{univ}}$, its bracket constants c_j, c_j' are $c_j^{\min}, c_j^{\max}$ with ratio γ, and the root-mean-square in its bias term is $\overline{w}(W^{\min})$. The main text's constants C_{gold}, C_{WE}, C_{TSE}, C_{THE} are the appendices' $\bar{C}^{\star}_{\mathrm{gold}}$, $\bar{C}^{\star}_{\mathrm{WE}}(r)$, $\bar{C}^{\star}_{\mathrm{TSE}}(r)$ and $\bar{C}^{\star}_{\mathrm{THE}}(r)$; its $C_{\mathrm{THE}}(\rho_{\mathrm{WE}}, \lambda)$ is the risk $g(t)$ of the two-headed estimator with the weight and bandwidths of the weighted optimum, with $C_{\mathrm{THE}}(\rho_{\mathrm{WE}}, \lambda_\star) = g(t_\star)$; and its curvature averages $h_{\mathrm{bias}}, h_{\mathrm{noise}}$ are $\bar{h}_{\mathrm{bias}}, \bar{h}_{\mathrm{noise}}$.

Conventions. The quadratic THE and two-step procedures use the loss (2) and penalty $\lambda\|\cdot\|^2$; the sparse correction (5) instead uses an ℓ_1 penalty. The main text's H is the Hessian of the population sign loss in tangent coordinates, and its bracket constants c_j are coefficients of $\|v_j\|^2$, so that the bracket matrix C corresponds to $H/2$. Appendix B keeps the loss (2) and writes the penalty as $\frac{\lambda}{2}\|\cdot\|^2$ with its own λ, twice the main text's. The appendices on the smoothed sign loss use the disagreement probability, which is $(1 + \widehat{L})/2$ in terms of (2), with the penalty $\lambda\|\cdot\|^2$ for a λ that is half the main text's, and their curvature H is half the main text's; the noise covariance V, the offset b, the asymptotic errors U_k, the matrix weights W, the criterion (10) and the penalty (12) are unchanged by this rescaling, because λ and H rescale together. Population quantities displayed in this appendix are in the appendix normalization.

Model and assumptions. The common choice model of Section 1.1 allows a nonunit proxy center $\beta_2 = \beta^\star + \delta_2$. The results for the exact and the smoothed sign loss in this appendix specialize to unit centers $\beta_1 = \beta^\star$ and $\beta_2 = \beta^\star + \delta$. Source k supplies N_k independent observations $(X_{k,j}, Y_{k,j})$ with a common covariate law and, conditionally on a fresh perturbation $\varepsilon_{k,j}$ independent of the covariates and of each other, $\Pr(Y_{k,j} = 1 \mid X_{k,j}, \varepsilon_{k,j}) = \sigma(X_{k,j}^\top(\beta_k + \varepsilon_{k,j}))$ as in (1), with $\varepsilon \overset{d}{=} -\varepsilon$ and one law across sources. The common assumptions are: X has a bounded Lebesgue density with bounded support and a positive lower bound on a ball around the origin; the projected central-mass condition $\inf_{\|u\|=1} \Pr(|u^\top \varepsilon| \le 1/(2\sqrt{2})) > 0$. Since both centers are unit vectors, $\beta^{\star\top}\delta = -\|\delta\|^2/2$ and the angular difference is $2\arcsin(\|\delta\|/2)$. The results for the smoothed sign loss add $f_X \in C_c^2(\mathbb{R}^d)$,

$\mathbb{E}\|\varepsilon\|^2 < \infty$ and $d \geq 3$. The smoothed empirical loss is

$$\widehat{L}_{k,h}(v) = \frac{1}{N_k} \sum_{j=1}^{N_k} \frac{1 - (2Y_{k,j} - 1)\tanh(X_{k,j}^\top v/h)}{2}, \qquad \|v\| = 1, \tag{13}$$

and the exact sign loss is (2) with the lower-semicontinuous convention that zero margins count as correct. Prediction errors are angular, $\angle(\widehat{\beta}, \beta^\star) = \arccos(\widehat{\beta}^\top \beta^\star)$ for a unit $\widehat{\beta}$, as in the main text.

Chart and population quantities. Write $x = t\beta^\star + Ez$ with E an orthonormal basis of $\beta^{\star\perp}$ and $\phi(v) = \sqrt{1 - \|v\|^2}\,\beta^\star + Ev$ the chart of the sphere at $\beta^\star$. With $m_1(x) = \mathbb{E}[2Y - 1 \mid X = x]$ for the gold source, the three population quantities for the smoothed sign loss are the tangential curvature of the gold sign risk, the score covariance shape and the direction of the smoothing offset,

$$\begin{aligned} &H = 2\int zz^\top \mathbb{E}_\varepsilon \sigma'(z^\top E^\top \varepsilon) f_X(Ez)\, dz, \qquad \Sigma = \tfrac{1}{3}\int zz^\top f_X(Ez)\, dz, \qquad V = H^{-1}\Sigma H^{-1}, \\ &b = \tfrac{\pi^2}{24} H^{-1}\Big[\tfrac{d^2}{dt^2}\int z\, m_1 f_X(t\beta^\star + Ez)\, dz\Big]_{t=0}. \end{aligned} \tag{14}$$

H and Σ are second moments of the design on the hyperplane $\beta^{\star\perp}$, differing only by the weight $\mathbb{E}_\varepsilon \sigma'(\cdot) \in (0, 1/4]$; when that weight is a constant $\bar{\sigma}'$ across directions, $V = \frac{1}{6\bar{\sigma}'} H^{-1}$. The vector b is the direction of the second-order smoothing offset; the main text calls V the gold noise covariance and b the smoothing offset.

Asymptotic errors. Let $N_k/n \to \nu_k > 0$, $h_k = c_k n^{-1/5}$, and let the bias be scaled as $\beta^\star + \delta_n = \phi(n^{-2/5}\zeta)$ with tangential coordinate $\zeta = r\Theta$, Θ uniform on the unit sphere of $\beta^{\star\perp}$ (so $\|\delta_n\| \asymp n^{-2/5} r$). With G_1, G_2 independent standard Gaussian vectors in $\mathbb{R}^{d-1}$,

$$U_1 = c_1^2 b + \frac{V^{1/2} G_1}{\sqrt{\nu_1 c_1}}, \qquad U_2 = \zeta + c_2^2 b + \frac{V^{1/2} G_2}{\sqrt{\nu_2 c_2}} \tag{15}$$

are the main text's U_1 and U_2, and the asymptotic risk of an estimator with matrix weight W is $\bar{C}(W, c_1, c_2; r) = \mathbb{E}\|(I - W)U_1 + WU_2\|$; the infimum of $\bar{C}$ over a family of matrix weights and over $c_1, c_2 > 0$ is written $\bar{C}^\star_{\mathcal{M}}(r)$. In (53) the matrix weight of the two-step estimator is written $W_{\mathrm{TSE}} = (H + 2\rho_{\mathrm{TSE}} I)^{-1} 2\rho_{\mathrm{TSE}}$ with ρ_{TSE} the two-step penalty, which is the penalty λ in (6). The criterion quantities (9) are evaluated at a direction-averaged minimizer $(\rho^\star, c_1^\star, c_2^\star)$ of the weighted family, where the main text's U is $(1 - \rho^\star)U_1 + \rho^\star U_2$; by Theorem 21, the derivative at $t = 0$ of the risk $g(t)$ with the THE matrix weight $\rho^\star(I + tH)^{-1}$ and the bandwidths of the weighted optimum has the sign of $\bar{h}_{\mathrm{noise}} - \bar{h}_{\mathrm{bias}}$. The identity $W_{\mathrm{THE}}(\rho, \lambda) = \rho\, W_{\mathrm{TSE}}(\lambda/(\rho(1 - \rho)))$ is (140); it concerns the families of matrix weights for the smoothed sign loss.

Where the proofs are. The lower bound due to the bias and the upper bounds described in Section 3 are Theorems 8, 14 and 10; Theorem 4 collects Theorems 14, 16, 17 and Corollaries 1 and 2, and the limit laws for the exact sign loss are Theorems 15 and 18 (Appendix C); Proposition 1 follows from Theorem 19. Proposition 2 is proved as Theorem 20 (Appendix D); the structure of the families of matrix weights (4) is Proposition 6; the statements on re-weighting after Proposition 2 are Lemma 42 and Proposition 6; Theorem 5 collects Theorems 21 and 22; Theorem 6 collects Theorems 23, 24 and Corollary 5; the paragraph on large bias and the two-step estimator in Section 4 rests on Theorems 28 and 27; Theorem 7(a) is proved as Theorem 29 (Appendix F), (b) as Theorem 30 (Appendix G), and (c) as Theorem 31 with the comparison of Theorem 33 (Appendix H); Proposition 14 collects Propositions 15–16 and Theorem 34.

Exact-loss normalization and controlled designs. Appendix C uses half the main-text loss, so its Hessian and penalty coefficient are $H_{\mathrm{main}}/2$ and $\lambda_{\mathrm{main}}/2$. Its controlled-design results use the tie-neutral finite-net definition in Appendix C.7. Full-density results retain (2); sample ties affect their local limit processes by $o(1)$. Appendix B works with the ± 1 loss and the penalty $\frac{\lambda}{2}\|\cdot\|^2$ (see its $\lambda' = \lambda/4$), Appendix C with the half loss and $\lambda\|\cdot\|^2$, $\lambda = \lambda_{\mathrm{main}}/2$.

B UPPER BOUNDS FOR THE EXACT SIGN LOSS

This appendix treats the lower-semicontinuous exact sign loss without smoothness. It compares high-probability angular-error upper bounds derived by one common analysis on one common event; it does not compare actual risks. The cube-root limit of the empirical minimizer is non-Gaussian, so the linear structure of Appendix D is not available; the weight matrices below appear only in the population bias and in the structure of the upper bounds.

B.1 SETTING, TOOLS AND THE QUADRATIC BRACKET ASSUMPTION

Empirical loss and its expectation:

$$\widehat{L}_k(\beta) = -1 + \frac{2}{N_k}\sum_{j=1}^{N_k} \mathbf{1}\{(2Y_{k,j}-1)X_{k,j}^\top \beta < 0\}, \qquad L_k(\beta) = -\mathbb{E}[(2Y-1)\operatorname{sign}(X^\top\beta)], \tag{16}$$

ties counted as correct. $\widehat{L}_k$ is lower semicontinuous, so all penalized objectives attain their minima on the compact spheres; every statement below holds for every global minimizer and no selection rule is needed. The estimators are as in Section 1.1 with penalty $\frac{\lambda}{2}\|\cdot\|_2^2$. Write $m_k(x) = \mathbb{E}[2Y-1 \mid X = x] = \mathbb{E}_\varepsilon \tanh(x^\top(\beta_k+\varepsilon)/2)$; symmetry of ε alone gives $\operatorname{sign} m_k(x) = \operatorname{sign}(x^\top\beta_k)$.

Lemma 1 (Population facts)**.** *For unit β, β' with $\alpha = \angle(\beta,\beta')$ and $\Gamma := \bar{f}_2 R^2$ ($\bar{f}_2$ bounds the density of any two-dimensional projection of X): (a) $L_k(\beta) - L_k(\beta_k) = 2\mathbb{E}[|m_k(X)|\mathbf{1}\{\operatorname{sign}(X^\top\beta) \neq \operatorname{sign}(X^\top\beta_k)\}] \geq 0$, with equality iff $\beta = \beta_k$; (b) $\Pr\{\operatorname{sign}(X^\top\beta) \neq \operatorname{sign}(X^\top\beta')\} \leq \Gamma\alpha$ and $|L_k(\beta) - L_k(\beta')| \leq 2\Gamma\alpha$; (c) $|(L_2 - L_1)(\beta) - (L_2 - L_1)(\beta')| \leq R\Gamma\|\delta\|_2\alpha$; (d) there is $c_{\mathrm{q}} > 0$, depending only on the design and noise constants, with $L_k(\beta) - L_k(\beta_k) \geq c_{\mathrm{q}}\angle(\beta,\beta_k)^2$ for all unit β.*

Proof. (a)–(b) as in Appendix D (the disagreement set is a double wedge). (c) $|m_2 - m_1| \leq \frac{1}{2}|x^\top\delta|$ since $\tanh$ is 1-Lipschitz. (d) Aouad et al. (2025), Theorem 3, gives $L_k(\beta) - L_k(\beta_k) \geq C_0\alpha^2$ for $\alpha \leq \pi/4$ under the central-mass condition with the same normalization of L_k; along any great circle through β_k the disagreement set is monotone in the angle, so the excess loss is nondecreasing, and $c_{\mathrm{q}} = C_0/16$ works globally. □

Fix the tangent frame E at $\beta^\star$, $\phi(v) = \sqrt{1-\|v\|^2}\beta^\star + Ev$, so that $\beta_2 = \phi(\zeta)$ with $\zeta = E^\top\delta$ and $r := \|\zeta\|$.

> **Assumption (S) (blockwise quadratic bracket).** There are $\eta_0 \in (0, 1/2]$, an orthogonal decomposition $\mathbb{R}^p = \bigoplus_{j=1}^J \mathcal{V}_j$ ($\dim\mathcal{V}_j = p_j$) and constants $0 < c_j^{\min} \leq c_j^{\max} < \infty$ such that, when $r \leq \eta_0/2$, for all $\|v\| \leq \eta_0$,
>
> $$Q^{\min}(v) \leq L_1(\phi(v)) - L_1(\beta^\star) \leq Q^{\max}(v),$$
> $$Q^{\min}(v-\zeta) \leq L_2(\phi(v)) - L_2(\beta_2) \leq Q^{\max}(v-\zeta),$$
>
> $Q^{\min}(u) = u^\top C^{\min}u = \sum_j c_j^{\min}\|u_j\|^2$, $Q^{\max}(u) = u^\top C^{\max}u$. Write $c_{\min} = \min_j c_j^{\min}$, $c_{\max} = \max_j c_j^{\max}$ and $\gamma := \max_j(c_j^{\max} - c_j^{\min})/c_{\min}$ (bracket ratio). Anisotropy means $c_j^{\max} < c_{j'}^{\min}$ for some $j \neq j'$.

Under smoothness (S) holds with $C^{\min} \approx C^{\max} \approx H$ and $\gamma \to 0$ as $\eta_0 \to 0$; here it is assumed with fixed constants on a fixed radius, and γ, η_0 enter every upper bound explicitly.

Chart geometry: for $\|v\|, \|v'\| \leq \eta_0 \leq 1/2$, $\|v - v'\| \leq \angle(\phi(v),\phi(v')) \leq \frac{\pi}{\sqrt{2}}\|v - v'\|$ and $\|v-v'\|^2 \leq \|\phi(v) - \phi(v')\|^2 \leq (1+4\eta_0^2)\|v-v'\|^2$; write $\lambda' := \lambda(1+4\eta_0^2)$.

Empirical-process tools. Let $\Lambda := C_{\mathrm{univ}}[d\log(e(N_1+N_2)) + \log(1/\eta)]$. The class $\{\mathbf{1}\{(2y-1)x^\top\beta < 0\}\}$ has VC dimension at most d and the difference class has variance $\leq \Gamma\angle(\beta,\beta')$ by Lemma 1(b).

- (E1) With probability $\geq 1-\eta/4$: $\sup_\beta |\widehat{L}_k(\beta) - L_k(\beta)| \leq \epsilon_k^{\mathrm{glob}} := 2\sqrt{\Lambda/N_k}$, $k = 1, 2$.
- (E2) With probability $\geq 1-\eta/4$, simultaneously for all weights $(\omega_1, \omega_2) \in [0,1]^2$ and all $(\beta_1, \beta_1', \beta_2, \beta_2')$,

$$\begin{aligned} &\Big|\sum_k \omega_k[(\widehat{L}_k - L_k)(\beta_k) - (\widehat{L}_k - L_k)(\beta_k')]\Big| \\ &\qquad \leq \mathsf{Inc}_\omega(\alpha_1, \alpha_2) := \sqrt{\Gamma\Lambda\Big[\frac{\omega_1^2\alpha_1}{N_1} + \frac{\omega_2^2\alpha_2}{N_2}\Big]} + \Lambda \max_k \frac{\omega_k}{N_k}. \end{aligned} \tag{17}$$

(E1) is the Vapnik–Chervonenkis uniform deviation bound; (E2) is a ratio-type inequality obtained by peeling the difference class over dyadic shells of (α_1, α_2) with Talagrand–Bousquet concentration on each shell (Bousquet, 2002). To obtain the asserted weight uniformity, first use homogeneity to normalize $\max_k \omega_k = 1$ (zero weights are trivial). Cover the two segments $(1, u)$ and $(u, 1)$, $u \in [0,1]$, by a mesh of size N^{-3}, $N = N_1 + N_2$, and take a union bound over their $O(N^3)$ grid points together with the variance shells. The extra $O(\log N)$ is absorbed in Λ. Moving from a grid point to its nearest weight changes the centered empirical increment by at most CN^{-3}. The square root of its variance bound changes by at most $CN^{-3}\sqrt{\Lambda/N_{\min}}$, by the triangle inequality for the weighted Euclidean norm. Both errors are absorbed, after increasing C_{univ}, by $\Lambda \max_k \omega_k/N_k \geq \Lambda/N$ in the normalized parameterization. Rescale to arbitrary weights. Thus the single (E2) event is uniform in the weights. All statements below hold on the intersection $\mathcal{E}$ of (E1) for both sources and the (E2) events used, $\Pr(\mathcal{E}) \geq 1-\eta$.

Lemma 2 (An elementary inequality). *If $m > 0$, $a, g, x_0 \geq 0$ and $x \geq 0$ satisfies $mx^2 \leq g + a\sqrt{x + x_0}$, then $x \leq (2a/m)^{2/3} + \sqrt{2(g + a\sqrt{x_0})/m}$.*

Proof. $\sqrt{x + x_0} \leq \sqrt{x} + \sqrt{x_0}$; then either $a\sqrt{x} \geq g'$ and $x^{3/2} \leq 2a/m$, or $mx^2 \leq 2g'$. □

B.2 Population level

Lemma 3 (WE population minimizer). *Let $F_\rho(v) = (1-\rho)[L_1(\phi(v)) - L_1(\beta^\star)] + \rho[L_2(\phi(v)) - L_2(\beta_2)]$ on $\|v\| \leq \eta_0$. Any minimizer v_ρ satisfies $Q^{\min}(v_\rho - \rho\zeta) \leq \rho(1-\rho)[Q^{\max}(\zeta) - Q^{\min}(\zeta)]$, hence $\|v_\rho - \rho\zeta\| \leq r\sqrt{\rho(1-\rho)\gamma}$. If both excess losses are exactly the same quadratic form, $v_\rho = \rho\zeta$ regardless of the curvature.*

Proof. Completing the square, $(1-\rho)Q(v) + \rho Q(v-\zeta) = Q(v - \rho\zeta) + \rho(1-\rho)Q(\zeta)$ for any block-diagonal quadratic Q; combine $F_\rho(v_\rho) \leq F_\rho(\rho\zeta)$ with the two brackets. □

Lemma 4 (Exact model for THE). *For $Q(u) = u^\top C u$ with block-diagonal $C \succ 0$, the unique minimizer of $F^Q(v_1, v_2) = (1-\rho)Q(v_1) + \rho Q(v_2 - \zeta) + \frac{\lambda}{2}\|v_1 - v_2\|^2$ is $v_1^Q = W_Q\zeta$, $v_2^Q = \zeta + K_2(v_1^Q - \zeta)$, $v_1^Q = K_1 v_2^Q$, with*

$$\begin{aligned} &K_1 = \tfrac{\lambda}{2}\big((1-\rho)C + \tfrac{\lambda}{2}I\big)^{-1}, \qquad K_2 = \tfrac{\lambda}{2}\big(\rho C + \tfrac{\lambda}{2}I\big)^{-1}, \\ &W_Q = (I - K_1K_2)^{-1}K_1(I - K_2) = \rho\Big(I + \frac{2\rho(1-\rho)}{\lambda}C\Big)^{-1}, \end{aligned} \tag{18}$$

blockwise $k_{1j} = \frac{\lambda/2}{(1-\rho)c_j + \lambda/2}$, $k_{2j} = \frac{\lambda/2}{\rho c_j + \lambda/2}$, $w_j = \frac{\rho\lambda/2}{\rho(1-\rho)c_j + \lambda/2}$, and $\min F^Q = \sum_j (1-\rho)c_j w_j \|\zeta_j\|^2$.

Proof. First-order conditions and the Moreau envelope $\min_{v_2}[\rho Q(v_2 - \zeta) + \frac{\lambda}{2}\|v_1 - v_2\|^2] = (v_1 - \zeta)^\top P(v_1 - \zeta)$, $P = \frac{\lambda\rho}{2}C(\rho C + \frac{\lambda}{2})^{-1}$; blockwise algebra. □

With $C = H$ the weight matrix for the exact sign loss $\rho(I + \frac{2\rho(1-\rho)}{\lambda}C)^{-1}$ coincides with the weight matrix for the smoothed loss $\rho(I + \frac{\rho(1-\rho)}{2\lambda'}H)^{-1}$ of Appendix D, whose penalty we write λ' here, exactly when $\lambda' = \lambda/4$. The two factors of two are the loss normalization $L_k = 2L_{k,0} - 1$ and the penalty convention $\frac{\lambda}{2}\|\cdot\|^2$ used here versus $\lambda\|\cdot\|^2$ used there; the family is the same either way.

Lemma 5 (THE population pair). *Let $F(v_1, v_2)$ be the population THE objective on $\{\|v_1\|, \|v_2\| \le \eta_0\}$ and $F_\lambda^{\min} \le F \le F_{\lambda'}^{\max}$ the models built from $Q^{\min}$ with penalty $\frac{\lambda}{2}\|v_1 - v_2\|^2$ and from $Q^{\max}$ with $\frac{\lambda'}{2}\|v_1 - v_2\|^2$. Any minimizer (v_1°, v_2°) satisfies $\mathcal{Q}_\lambda^{\min}(v^\circ - v^{\min}) \le \mathfrak{g}_{\mathrm{THE}} := \min F_{\lambda'}^{\max} - \min F_\lambda^{\min}$, with $\mathcal{Q}_\lambda^{\min}$ the quadratic form of $F_\lambda^{\min}$ and $v^{\min} = (W^{\min}\zeta, \cdot)$ its minimizer; in particular $\|v_1^\circ - W^{\min}\zeta\| \le \sqrt{\mathfrak{g}_{\mathrm{THE}}/((1-\rho)c_{\min})}$.*

Proof. Assumption (S) brackets each population excess loss between $Q^{\min}$ and $Q^{\max}$ in the chart, and the chart geometry brackets $\frac{\lambda}{2}\|\phi(v_1) - \phi(v_2)\|^2$ between $\frac{\lambda}{2}\|v_1 - v_2\|^2$ and $\frac{\lambda'}{2}\|v_1 - v_2\|^2$; together these give $F_\lambda^{\min} \le F \le F_{\lambda'}^{\max}$ on $\{\|v_1\|, \|v_2\| \le \eta_0\}$. Hence $F_\lambda^{\min}(v^\circ) \le F(v^\circ) = \min F \le \min F_{\lambda'}^{\max}$, and subtracting $\min F_\lambda^{\min}$ and using the exact identity $F^Q(v) - \min F^Q = \mathcal{Q}^Q(v - v^Q)$ of Lemma 4 gives $\mathcal{Q}_\lambda^{\min}(v^\circ - v^{\min}) \le \mathfrak{g}_{\mathrm{THE}}$. For the gold head, minimizing $\mathcal{Q}_\lambda^{\min}(u_1, u_2)$ over u_2 at fixed u_1 leaves, by the same Moreau envelope as in Lemma 4, the profile $u_1^\top[(1-\rho)C^{\min} + P^{\min}]u_1$ with $P^{\min} = \frac{\lambda\rho}{2}C^{\min}(\rho C^{\min} + \frac{\lambda}{2}I)^{-1}$, i.e. blockwise the coefficient $c_{\mathrm{eff},j} = (1-\rho)c_j^{\min} + \frac{\lambda\rho c_j^{\min}/2}{\rho c_j^{\min} + \lambda/2}$. With $c_{\mathrm{eff}} = \min_j c_{\mathrm{eff},j} \ge (1-\rho)c_{\min}$ this yields $c_{\mathrm{eff}}\|v_1^\circ - W^{\min}\zeta\|^2 \le \mathfrak{g}_{\mathrm{THE}}$. □

B.3 Global localization

Lemma 6. *On $\mathcal{E}$, with $\epsilon_\rho^{\mathrm{glob}} = (1-\rho)\epsilon_1^{\mathrm{glob}} + \rho\epsilon_2^{\mathrm{glob}}$: (a) THE, $\rho \in (0,1)$, any $\lambda \ge 0$: $\angle(\widehat{\beta}_1, \beta^\star) \le \eta_1^{\mathrm{THE}} := \sqrt{\frac{\rho c_{\max} r^2 + 2\epsilon_\rho^{\mathrm{glob}}}{(1-\rho)c_{\mathrm{q}}}}$, $\angle(\widehat{\beta}_2, \beta^\star) \le \eta_2^{\mathrm{THE}} := \sqrt{\frac{\rho c_{\max} r^2 + 2\epsilon_\rho^{\mathrm{glob}}}{\rho c_{\mathrm{q}}}} + \arcsin r$; (b) WE: the minimum of the analogous two bounds; (c) gold-only $\alpha \le \sqrt{2\epsilon_1^{\mathrm{glob}}/c_{\mathrm{q}}}$, proxy-only $\angle(\widehat{\beta}_{\mathrm{proxy}}, \beta_2) \le \sqrt{2\epsilon_2^{\mathrm{glob}}/c_{\mathrm{q}}}$; (d) TSE second stage $\alpha \le \sqrt{(\lambda\angle(\widehat{\beta}_{\mathrm{proxy}}, \beta^\star)^2/2 + 2\epsilon_1^{\mathrm{glob}})/c_{\mathrm{q}}}$.*

Proof. Compare with the common candidate $(\beta^\star, \beta^\star)$, whose penalty vanishes; use $L_2(\widehat{\beta}_2) - L_2(\beta^\star) \ge c_{\mathrm{q}}\angle(\widehat{\beta}_2, \beta_2)^2 - Q^{\max}(\zeta)$ and drop the nonnegative penalty. (a) does not use the penalty and is therefore uniform in λ, but needs ρ bounded away from 0 and 1. □

Condition (Loc). All heads have chart coordinates of norm $\le \eta_0/2$ and $r \le \eta_0/2$; this is an explicit condition on (r, N_1, N_2, ρ) through Lemma 6.

B.4 The unified upper bound

Notation: $\sigma_\rho^2 = \frac{(1-\rho)^2}{N_1} + \frac{\rho^2}{N_2}$, $\mu_\rho = \max\{\frac{1-\rho}{N_1}, \frac{\rho}{N_2}\}$,

$$\kappa(s) = \Big(\frac{2\sqrt{\frac{\pi}{\sqrt{2}}\Gamma\Lambda s}}{c_{\min}}\Big)^{2/3}, \qquad \kappa_k = \kappa(1/N_k),$$
$$\kappa(\sigma_\rho^2) = [(1-\rho)^2\kappa_1^3 + \rho^2\kappa_2^3]^{1/3}, \qquad \tau_\rho = \sqrt{2\Lambda\mu_\rho/c_{\min}}. \tag{19}$$

Theorem 8 (WE, gold-only, proxy-only). *On $\mathcal{E}$ under (Loc), $\widehat{\beta}_{\mathrm{WE}} = \phi(\widehat{v})$ satisfies*

$$\|\widehat{v}\| \le \mathcal{B}_{\mathrm{WE}}(\rho; \zeta) := \rho r + \kappa(\sigma_\rho^2) + r\sqrt{2\rho(1-\rho)\gamma} + \tau_\rho. \tag{20}$$

Gold-only: $\|\widehat{v}\| \le \mathcal{B}_{\mathrm{gold}} := \kappa_1 + \tau_1$; proxy-only: $\|\widehat{v}_{\mathrm{p}} - \zeta\| \le \mathsf{N}_{\mathrm{p}} := \kappa_2 + \tau_2$.

Proof. Compare with $\phi(\rho\zeta)$: by (E2), $F_\rho(\widehat{v}) - F_\rho(\rho\zeta) \le \mathsf{Inc}_{(1-\rho,\rho)}(\alpha, \alpha)$ with $\alpha \le \frac{\pi}{\sqrt{2}}\|\widehat{v} - \rho\zeta\|$; by Lemma 3, $c_{\min}\|\widehat{v} - \rho\zeta\|^2 \le \rho(1-\rho)\gamma c_{\min} r^2 + \Lambda\mu_\rho + \sqrt{\frac{\pi}{\sqrt{2}}\Gamma\Lambda\sigma_\rho^2\|\widehat{v} - \rho\zeta\|}$; Lemma 2. □

Theorem 9 (TSE). *Let $K^{\mathrm{T}} = \frac{\lambda}{2}(C^{\min} + \frac{\lambda}{2}I)^{-1}$, $\|K^{\mathrm{T}}\| = \frac{\lambda/2}{c_{\min} + \lambda/2}$. On $\mathcal{E}$ under (Loc),*

$$\|\widehat{v}_{\mathrm{TSE}}\| \le \min\Big\{\|K^{\mathrm{T}}\zeta\| + \|K^{\mathrm{T}}\|\mathsf{N}_{\mathrm{p}} + \kappa_1\Big(\frac{c_{\min}}{c_{\min} + \lambda/2}\Big)^{2/3} + \sqrt{\frac{2\mathfrak{g}^{\mathrm{T}} + 2\Lambda/N_1}{c_{\min} + \lambda/2}},\ r + \mathsf{N}_{\mathrm{p}} + \frac{2}{\sqrt{\lambda}}\Big\}, \tag{21}$$

$\mathfrak{g}^{\mathrm{T}} = [\gamma c_{\min}\|K^{\mathrm{T}}\|^2 + 2\lambda\eta_0^2](r + \mathsf{N}_{\mathrm{p}})^2$. *At $\lambda = 0$ this is $\mathcal{B}_{\mathrm{gold}}$; as $\lambda \to \infty$ it tends to the proxy-only upper bound.*

Proof. Second stage: compare the penalized objective at $\widehat{v}_{\mathrm{TSE}}$ and at $K^{\mathrm{T}}\widehat{v}_{\mathrm{p}}$; the bracket models $G^{\min}(v;a) = Q^{\min}(v) + \frac{\lambda}{2}\|v-a\|^2 \le G(v;a) \le Q^{\max}(v) + \frac{\lambda'}{2}\|v-a\|^2$ have minimizer $K^{\mathrm{T}}a$ and excess $\|v - K^{\mathrm{T}}a\|^2_{C^{\min}+\lambda I/2}$; the gap between the models at $K^{\mathrm{T}}a$ is at most $[\gamma c_{\min}\|K^{\mathrm{T}}\|^2 + 2\lambda\eta_0^2]\|a\|^2$. Apply Lemma 2. The second branch: comparing with $\widehat{\beta}_{\mathrm{proxy}}$ itself gives $\frac{\lambda}{2}\|\widehat{\beta}_{\mathrm{TSE}} - \widehat{\beta}_{\mathrm{proxy}}\|^2 \le 2$. □

Theorem 10 (THE, analysis with the other head fixed)**.** *Let $K_1, K_2, W^{\min}$ be as in Lemma 4 with $C = C^{\min}$, and blockwise*

$$\Phi_j = \frac{1}{1 - k_{1j}k_{2j}} = \frac{[(1-\rho)c_j^{\min} + \frac{\lambda}{2}][\rho c_j^{\min} + \frac{\lambda}{2}]}{c_j^{\min}[\rho(1-\rho)c_j^{\min} + \frac{\lambda}{2}]}, \tag{22}$$

$$\Psi_j = \frac{k_{1j}}{1 - k_{1j}k_{2j}} = \frac{\frac{\lambda}{2}[\rho c_j^{\min} + \frac{\lambda}{2}]}{c_j^{\min}[\rho(1-\rho)c_j^{\min} + \frac{\lambda}{2}]}. \tag{23}$$

On $\mathcal{E}$ (with (E2) for $\omega = (1-\rho, 0)$ and $(0, \rho)$) under (Loc), every joint minimizer satisfies

$$\|\widehat{v}_1 - K_1\widehat{v}_2\| \le \mathsf{N}_1 := \kappa_1\Big(\frac{(1-\rho)c_{\min}}{(1-\rho)c_{\min} + \lambda/2}\Big)^{2/3} + \sqrt{\frac{2\mathfrak{g}_1 + 2(1-\rho)\Lambda/N_1}{(1-\rho)c_{\min} + \lambda/2}}, \tag{24}$$

$$\|\widehat{v}_2 - \zeta - K_2(\widehat{v}_1 - \zeta)\| \le \mathsf{N}_2 := \kappa_2\Big(\frac{\rho c_{\min}}{\rho c_{\min} + \lambda/2}\Big)^{2/3} + \sqrt{\frac{2\mathfrak{g}_2 + 2\rho\Lambda/N_2}{\rho c_{\min} + \lambda/2}}, \tag{25}$$

with $\mathfrak{g}_1 = [(1-\rho)\gamma c_{\min}\|K_1\|^2 + 2\lambda\eta_0^2](\eta_2^{\mathrm{THE}})^2$, $\mathfrak{g}_2 = [\rho\gamma c_{\min}\|K_2\|^2 + 2\lambda\eta_0^2](\eta_1^{\mathrm{THE}} + r)^2$, and, with ξ_1 and ξ_2 the two vectors bounded above, the gold head satisfies

$$\widehat{v}_1 = W^{\min}\zeta + (I - K_1K_2)^{-1}(\xi_1 + K_1\xi_2), \qquad \|\widehat{v}_1\| \le \mathcal{B}^{\mathrm{anch}}_{\mathrm{THE}} := \|W^{\min}\zeta\| + \Phi_{\max}\mathsf{N}_1 + \Psi_{\max}\mathsf{N}_2. \tag{26}$$

Proof. A joint minimizer is a partial minimizer in each head. Keeping $\widehat{v}_2$ fixed and comparing the objective of the gold head at $\widehat{v}_1$ with the point $K_1\widehat{v}_2$, the bracket models $(1-\rho)Q^{\min}(v) + \frac{\lambda}{2}\|v-a\|^2$ and $(1-\rho)Q^{\max}(v) + \frac{\lambda'}{2}\|v-a\|^2$ have minimizer K_1a and excess $\|v - K_1a\|^2_{(1-\rho)C^{\min}+\lambda I/2}$; the model gap at K_1a is bounded by $\mathfrak{g}_1$ using $\|\widehat{v}_2\| \le \eta_2^{\mathrm{THE}}$. Symmetrically for the proxy head, with the penalty centered at $\widehat{v}_1$. Both inequalities hold for the same actual pair. Eliminating $\widehat{v}_2$ from $\widehat{v}_1 = K_1(\zeta + e_2) + \xi_1$, $e_2 = K_2(\widehat{v}_1 - \zeta) + \xi_2$ gives the displayed representation, and $\|(I - K_1K_2)^{-1}\| = \Phi_{\max}$, $\|(I - K_1K_2)^{-1}K_1\| = \Psi_{\max}$ in the common block basis. □

The first term of N_1 is the cube-root noise term of the gold head, with the curvature increased by the penalty from $(1-\rho)c_{\min}$ to $(1-\rho)c_{\min} + \lambda/2$; this is the only effect of the penalty that such an upper bound can reflect. As $\lambda \to 0$ the bound reduces exactly to $\mathcal{B}_{\mathrm{gold}}$. As $\lambda \to \infty$, $\Phi_j \sim \lambda/(2c_j^{\min})$ while the chart term in $\mathfrak{g}_1$ does not vanish, so this upper bound diverges, because it treats the fluctuations of the two heads as independent worst cases, whereas under strong coupling they are highly correlated.

Theorem 11 (THE, comparison with equal heads)**.** *On $\mathcal{E}$ under (Loc),*

$$\|\widehat{v}_1\| \le \mathcal{B}^{\mathrm{ch}}_{\mathrm{THE}} := \rho r + \kappa(\sigma_\rho^2) + \sqrt{\frac{2\rho(1-\rho)\gamma c_{\min} r^2 + 2\Lambda\mu_\rho + \frac{5\rho\Lambda}{2N_2}}{c_{\min}} + \frac{9\pi^2\rho^2\Gamma^2}{4\lambda c_{\min}}}, \tag{27}$$

which tends to $\mathcal{B}_{\mathrm{WE}}(\rho)$ (up to the lower-order N_2 term) as $\lambda \to \infty$.

Proof. Compare with $(\phi(\rho\zeta), \phi(\rho\zeta))$; rewrite $\rho\widehat{L}_2(\widehat{\beta}_2) = \rho\widehat{L}_2(\widehat{\beta}_1) - \rho[\widehat{L}_2(\widehat{\beta}_1) - \widehat{L}_2(\widehat{\beta}_2)]$; bound the bracket by $3\rho\Gamma\alpha_{12} + \frac{5\rho\Lambda}{4N_2}$ and the penalty from below by $\frac{2\lambda}{\pi^2}\alpha_{12}^2$; maximize over α_{12} to get $\frac{9\pi^2\rho^2\Gamma^2}{8\lambda}$; conclude as in Theorem 8. □

Definition 1. *$\mathcal{B}_{\mathrm{THE}}(\rho, \lambda) := \min\{\mathcal{B}^{\mathrm{anch}}_{\mathrm{THE}}, \mathcal{B}^{\mathrm{ch}}_{\mathrm{THE}}\}$, with $\mathcal{B}_{\mathrm{THE}}(\rho, 0) = \mathcal{B}_{\mathrm{gold}}$ and $\mathcal{B}_{\mathrm{THE}}(\rho, \infty) = \mathcal{B}_{\mathrm{WE}}(\rho)$. Both branches hold on the same event, so the minimum is a valid upper bound that equals the gold-only bound at $\lambda = 0$ and the WE bound at $\lambda = \infty$.*

B.5 Direction averaging and structure

For $\zeta = r\Theta$ only the bias terms depend on Θ, and for block-diagonal W $\mathbb{E}_\Theta\|W\Theta\| \le \overline{w}(W) := \sqrt{\sum_j p_j w_j^2/p}$, $\mathbb{E}_\Theta\|\rho\Theta\| = \rho$.

Proposition 3. *(a) The bias terms of WE, THE and TSE are ρr, $\overline{w}(W^{\min})r$ and $\overline{w}(K^{\mathrm{T}})r$; the family $\{W^{\min}\}$ coincides with the family of Appendix D under $H \leftrightarrow C^{\min}$, with $\{\rho I\}$ its $\lambda = \infty$ boundary and $\{K^{\mathrm{T}}\}$ its $\rho \to 1$ closure; $\overline{w}(W^{\min}) \le \rho$ with equality iff $\lambda = \infty$. (b) In the isotropic case ($c_j^{\min} \equiv c$), $W^{\min} = wI$ with $w = \frac{\rho\tilde\lambda}{\rho(1-\rho)+\tilde\lambda}$, $\tilde\lambda = \lambda/(2c)$, and*

$$\Phi\,\kappa_1\Big(\frac{(1-\rho)c}{(1-\rho)c+\lambda/2}\Big)^{2/3} = \Phi^{1/3}(1-w)^{2/3}\kappa_1 \ge (1-w)^{2/3}\kappa_1, \tag{28}$$

$$\Psi\,\kappa_2\Big(\frac{\rho c}{\rho c+\lambda/2}\Big)^{2/3} = \Psi^{1/3}w^{2/3}\kappa_2, \tag{29}$$

with $\Phi \ge 1$ and $\Psi \ge 1 \iff \tilde\lambda \ge \tilde\lambda_\rho := \frac{(1-\rho)+\sqrt{(1-\rho)(1+3\rho)}}{2}$. Hence for $\tilde\lambda \ge \tilde\lambda_\rho$ the THE upper bound is at least $wr + \kappa(\sigma_w^2)$, the WE upper bound at weight w, so in the isotropic case the analysis gives THE no advantage over WE.

Proof. (a) The identification of $\{W^{\min}\}$ with the family of Appendix D and of its two boundaries is Lemma 4: blockwise $w_j = \frac{\rho\lambda/2}{\rho(1-\rho)c_j^{\min}+\lambda/2}$, which tends to ρ as $\lambda \to \infty$, while $W^{\min}$ tends to the matrix K^{T} of Theorem 9 along $\rho \to 1$, $\lambda \to 0$ at fixed $\rho(1-\rho)/\lambda$. Since $w_j \le \rho$ for every block, $\overline{w}(W^{\min}) \le \rho\sqrt{\sum_j p_j/p} = \rho$, and equality requires $w_j = \rho$ for every j, i.e. the endpoint $\lambda = \infty$, where $W^{\min} = \rho I$ is scalar. (b) With $c_j^{\min} \equiv c$ and $\tilde\lambda = \lambda/(2c)$ one gets $w = \frac{\rho\tilde\lambda}{\rho(1-\rho)+\tilde\lambda}$, $1-w = \frac{(1-\rho)(\rho+\tilde\lambda)}{\rho(1-\rho)+\tilde\lambda}$, $\Phi = \frac{(1-\rho+\tilde\lambda)(\rho+\tilde\lambda)}{\rho(1-\rho)+\tilde\lambda}$ and $\Psi = \frac{\tilde\lambda(\rho+\tilde\lambda)}{\rho(1-\rho)+\tilde\lambda}$, so that $\Phi\frac{1-\rho}{1-\rho+\tilde\lambda} = 1-w$ and $\Psi\frac{\rho}{\rho+\tilde\lambda} = w$; raising these to the power $2/3$ gives the two displayed identities. Moreover $\Phi \ge 1$ because $(\rho+\tilde\lambda)(1-\rho+\tilde\lambda) = \rho(1-\rho)+\tilde\lambda+\tilde\lambda^2$, and $\Psi \ge 1$ exactly when $\tilde\lambda(\rho+\tilde\lambda) \ge \rho(1-\rho)+\tilde\lambda$. When $\Phi \ge 1$ and $\Psi \ge 1$ the two noise terms are at least $(1-w)^{2/3}\kappa_1$ and $w^{2/3}\kappa_2$, and $x+y \ge (x^3+y^3)^{1/3}$ applied to $x = (1-w)^{2/3}\kappa_1$, $y = w^{2/3}\kappa_2$ turns their sum into $\kappa(\sigma_w^2)$, while the bias term is $\overline{w}(wI)r = wr$. □

Lemma 7. $\bar{\mathcal{B}}^\star_{\mathrm{WE}} := \inf_\rho \bar{\mathcal{B}}_{\mathrm{WE}}(\rho) \ge \inf_\rho[\rho r + \kappa_1(1-\rho)^{2/3}] = \min\{r, \kappa_1\}$ *and* $\bar{\mathcal{B}}_{\mathrm{gold}} \ge \kappa_1$.

Proof. $\kappa(\sigma_\rho^2) = [(1-\rho)^2\kappa_1^3 + \rho^2\kappa_2^3]^{1/3} \ge (1-\rho)^{2/3}\kappa_1$, and the bracket term $r\sqrt{2\rho(1-\rho)\gamma}$ and τ_ρ are nonnegative, so $\bar{\mathcal{B}}_{\mathrm{WE}}(\rho) \ge \rho r + \kappa_1(1-\rho)^{2/3}$. The right-hand side is concave in ρ on $[0,1]$ (a linear term plus a concave one), hence attains its minimum at an endpoint, with values κ_1 at $\rho = 0$ and r at $\rho = 1$. Finally $\bar{\mathcal{B}}_{\mathrm{gold}} = \kappa_1 + \tau_1 \ge \kappa_1$. □

B.6 Region of strict improvement

Theorem 12 (Comparison inequality)**.** *Under (S) and (Loc), if*

$$r\,\overline{w}(W^{\min}) + \Phi_{\max}\mathsf{N}_1(\rho,\lambda) + \Psi_{\max}\mathsf{N}_2(\rho,\lambda) < \min\{r, \kappa_1\}, \tag{30}$$

then on $\mathcal{E}$ the direction-averaged THE upper bound at (ρ,λ) is strictly below the WE upper bound optimized over $\rho \in [0,1]$ and below the gold-only upper bound. The same holds for TSE with its upper bound on the left.

Proof. Direction averaging of Theorem 10, using $\mathbb{E}_\Theta\|W^{\min}\Theta\| \le \overline{w}(W^{\min})$, gives on $\mathcal{E}$ $\bar{\mathcal{B}}_{\mathrm{THE}}(\rho,\lambda) \le \bar{\mathcal{B}}^{\mathrm{anch}}_{\mathrm{THE}}(\rho,\lambda) \le r\,\overline{w}(W^{\min}) + \Phi_{\max}\mathsf{N}_1 + \Psi_{\max}\mathsf{N}_2$, which by (30) is $< \min\{r,\kappa_1\}$. By Lemma 7 that quantity is at most $\bar{\mathcal{B}}^\star_{\mathrm{WE}}$ and at most $\kappa_1 \le \bar{\mathcal{B}}_{\mathrm{gold}}$. All the upper bounds involved were derived on the same event $\mathcal{E}$, with the same failure probability and the same constants, so the two strict inequalities hold simultaneously with probability at least $1-\eta$. Replacing the left side of (30) by the direction-averaged bound of Theorem 9 gives the corresponding statement for TSE. □

Theorem 13 (Nonemptiness, limit form). *Consider a high-curvature group ($c_j^{\min} \geq c_{\mathrm{hi}}$) and a low-curvature group ($c_j^{\min} = c_{\min}$) of total dimension p_L. Fix $\rho \in (0,1)$, $\tilde{\lambda} = \lambda/(2c_{\min})$, and set*

$$\overline{w}_\infty = \sqrt{\frac{p_L}{p}}\,\frac{\rho\tilde{\lambda}}{\rho(1-\rho)+\tilde{\lambda}}, \qquad \phi_\infty = \frac{(1-\rho+\tilde{\lambda})(\rho+\tilde{\lambda})}{\rho(1-\rho)+\tilde{\lambda}}\Big(\frac{1-\rho}{1-\rho+\tilde{\lambda}}\Big)^{2/3}. \tag{31}$$

As $c_{\mathrm{hi}}/c_{\min} \to \infty$, $\gamma, \eta_0 \to 0$ at rate $\gamma + \eta_0^2 \ll (\Lambda/N_1)^{1/6}$, $N_2/N_1 \to \infty$ and $N_1 \to \infty$, the left side of (30) *divided by κ_1 converges to $x\overline{w}_\infty + \phi_\infty$ with $x = r/\kappa_1$, and the right side to* $\min\{x, 1\}$. *Hence for every $\epsilon > 0$ the interval*

$$\frac{r}{\kappa_1} \in \Big[\frac{\phi_\infty}{1-\overline{w}_\infty} + \epsilon,\ \frac{1-\phi_\infty}{\overline{w}_\infty} - \epsilon\Big] \tag{32}$$

is eventually contained in the region where (30) *holds; it is nonempty iff $\phi_\infty + \overline{w}_\infty < 1$. Writing $\overline{w}_\infty = q\,w$ with $q = \sqrt{p_L/p}$ and $w = \frac{\rho\tilde{\lambda}}{\rho(1-\rho)+\tilde{\lambda}} \in (0,1)$ (as in Proposition 3(b) with $c = c_{\min}$), one has $\phi_\infty + w > 1$ for every $\rho \in (0,1)$ and $\tilde{\lambda} > 0$. Hence the interval is empty when $p_L = p$, and when it is nonempty both of its endpoints lie in the open interval $(0, 1/q)$.*

Proof. High-curvature blocks: $w_j \to 0$, $\Phi_j \to 1$, $\Psi_j \to 0$; low-curvature blocks give $\Phi_{\mathrm{lo}}, \Psi_{\mathrm{lo}}$. The gap terms are $O((\gamma + \eta_0^2)(\eta^{\mathrm{THE}})^2)$ with $\eta^{\mathrm{THE}} \asymp (\Lambda/N_1)^{1/4}$, which is below $\kappa_1^2 \asymp (\Lambda/N_1)^{2/3}$ under the stated rate; the lower-order Λ/N_k terms and $\Psi_{\mathrm{lo}}\mathsf{N}_2/\kappa_1 \propto (N_1/N_2)^{1/3}$ vanish. For the last claims, $1 - w = \frac{(1-\rho)(\rho+\tilde{\lambda})}{\rho(1-\rho)+\tilde{\lambda}}$ and $\phi_\infty = \frac{\rho+\tilde{\lambda}}{\rho(1-\rho)+\tilde{\lambda}}(1-\rho)^{2/3}(1-\rho+\tilde{\lambda})^{1/3}$, so $\phi_\infty/(1-w) = \big((1-\rho+\tilde{\lambda})/(1-\rho)\big)^{1/3} > 1$, i.e. $\phi_\infty + w > 1$. With $q = 1$ this is $\phi_\infty + \overline{w}_\infty > 1$, so the interval is empty. In general the lower end of the interval is positive because $\phi_\infty > 0$, and the upper end satisfies $\frac{1-\phi_\infty}{qw} < \frac{1}{q}$ because $1 - \phi_\infty < w$. □

The lower end of the interval comes from the requirement that the reduction of the bias term exceed the noise term of THE, and the upper end from the requirement that the THE upper bound be below the gold-only bound. The rate condition on γ, η_0 is needed because the loss is not smooth; a one-step refinement of the localization used inside $\mathfrak{g}_1, \mathfrak{g}_2$ would replace it by a fixed smallness condition; we do not pursue this.

Proposition 4 (TSE interval). *In the same limit with $\tilde{\lambda}^{\mathrm{T}} = \lambda/(2c_{\min})$ and $q = \sqrt{p_L/p} < 1$, the TSE interval is*

$$\frac{r}{\kappa_1} \in \big(x_-(\tilde{\lambda}^{\mathrm{T}}), x_+(\tilde{\lambda}^{\mathrm{T}})\big), \quad x_-(\tilde{\lambda}^{\mathrm{T}}) = \frac{(1+\tilde{\lambda}^{\mathrm{T}})^{1/3}}{1+(1-q)\tilde{\lambda}^{\mathrm{T}}}, \quad x_+(\tilde{\lambda}^{\mathrm{T}}) = \frac{(1+\tilde{\lambda}^{\mathrm{T}}) - (1+\tilde{\lambda}^{\mathrm{T}})^{1/3}}{q\,\tilde{\lambda}^{\mathrm{T}}}. \tag{33}$$

x_+ increases from $\frac{2}{3q}$ to $\frac{1}{q}$ and x_- tends to 0 as $\tilde{\lambda}^{\mathrm{T}} \to \infty$, so the union of the (possibly empty) TSE intervals over $\tilde{\lambda}^{\mathrm{T}} > 0$ is $(0, 1/q)$. Consequently every THE interval of Theorem 13, closure included, is contained in the TSE interval for all sufficiently large $\tilde{\lambda}^{\mathrm{T}}$. At equal penalties $\tilde{\lambda}^{\mathrm{T}} = \tilde{\lambda}$ neither interval contains the other in general.

Proof. In the limit of Theorem 13 the first branch of Theorem 9 has $k_j^{\mathrm{T}} = \frac{\lambda/2}{c_j^{\min}+\lambda/2} \to 0$ on the high-curvature blocks and $k_j^{\mathrm{T}} = \frac{\tilde{\lambda}^{\mathrm{T}}}{1+\tilde{\lambda}^{\mathrm{T}}}$ on the low-curvature ones, so $\overline{w}(K^{\mathrm{T}}) \to q\,\frac{\tilde{\lambda}^{\mathrm{T}}}{1+\tilde{\lambda}^{\mathrm{T}}}$, while its cube-root term is $\kappa_1(1+\tilde{\lambda}^{\mathrm{T}})^{-2/3}$. The remaining terms are negligible relative to κ_1: $\|K^{\mathrm{T}}\|\mathsf{N}_{\mathrm{p}}/\kappa_1 \to 0$ because $\kappa_2/\kappa_1 = (N_1/N_2)^{1/3} \to 0$, and the $\mathfrak{g}^{\mathrm{T}}$ and Λ/N_1 terms are suppressed by the same rate condition $\gamma + \eta_0^2 \ll (\Lambda/N_1)^{1/6}$. Hence the left side of (30) with the TSE upper bound, divided by κ_1, converges to $x\,\overline{w}^{\mathrm{T}} + \phi^{\mathrm{T}}$ with $\overline{w}^{\mathrm{T}} = q\,\tilde{\lambda}^{\mathrm{T}}/(1+\tilde{\lambda}^{\mathrm{T}})$ and $\phi^{\mathrm{T}} = (1+\tilde{\lambda}^{\mathrm{T}})^{-2/3}$, and solving $x\,\overline{w}^{\mathrm{T}} + \phi^{\mathrm{T}} < \min\{x, 1\}$ gives $x \in \big(\frac{\phi^{\mathrm{T}}}{1-\overline{w}^{\mathrm{T}}}, \frac{1-\phi^{\mathrm{T}}}{\overline{w}^{\mathrm{T}}}\big)$, which is (33) after clearing the fractions. With $t = 1+\tilde{\lambda}^{\mathrm{T}}$, $x_+ = \frac{t-t^{1/3}}{q(t-1)}$, and the derivative of $\frac{t-t^{1/3}}{t-1}$ has the sign of $\psi(t) = -1 + \frac{2}{3}t^{1/3} + \frac{1}{3}t^{-2/3}$; since $\psi(1) = 0$ and $\psi'(t) = \frac{2}{9}t^{-5/3}(t-1) > 0$ for $t > 1$, x_+ is increasing, with $x_+ \to \frac{2}{3q}$ as $t \downarrow 1$ and $x_+ \to \frac{1}{q}$ as $t \to \infty$. Since $q < 1$, $x_-(\tilde{\lambda}^{\mathrm{T}}) \asymp (\tilde{\lambda}^{\mathrm{T}})^{-2/3} \to 0$.

Every interval lies in $(0, 1/q)$, and for $x \in (0, 1/q)$ one has $x_-(\tilde{\lambda}^{\mathrm{T}}) < x < x_+(\tilde{\lambda}^{\mathrm{T}})$ for all large $\tilde{\lambda}^{\mathrm{T}}$, which gives the union. The closure of a THE interval is a compact subinterval of $(0, 1/q)$ by Theorem 13, so it lies in $(x_-(\tilde{\lambda}^{\mathrm{T}}), x_+(\tilde{\lambda}^{\mathrm{T}}))$ for all large $\tilde{\lambda}^{\mathrm{T}}$. For the last claim fix $\tilde{\lambda}$ large enough that $x_-(\tilde{\lambda}) < x_+(\tilde{\lambda})$. As $\rho \to 1$ one has $w \to 1$ and $\phi_\infty \to 0$, so the THE interval tends to $(0, 1/q)$ and strictly contains the TSE interval at $\tilde{\lambda}^{\mathrm{T}} = \tilde{\lambda}$; as $\rho \to 0$ one has $w \to 0$ and $\phi_\infty \to (1 + \tilde{\lambda})^{1/3} > 1$, so the THE interval is empty while the TSE interval at $\tilde{\lambda}^{\mathrm{T}} = \tilde{\lambda}$ is not. □

B.7 Scope

Upper bound orderings are not risk orderings, and the decomposition of Theorem 10 is conservative for THE, so the region in which THE is shown to have the smaller upper bound is a subset of the actual one. Assumption (S) is additional to the assumptions of the sign estimator, and its constants γ, η_0 must decrease as N_1 grows for a comparison at the cube-root scale. The localization requires ρ bounded away from 0 and 1; the corner $\rho \to 1$, $\lambda \to 0$ and a comparison of the upper bounds of THE and TSE are not treated. For a worst-case bias direction, $\overline{w}$ is replaced by $\max_j w_j$ and the interval changes accordingly.

C Lower bounds for the baselines and limit laws for the exact sign loss

This appendix proves the lower bounds of Theorem 4. It also establishes the cube-root limit laws of the three estimators with the exact sign loss, which show that exact asymptotic constants exist without smoothing and why they are not compared analytically. The notation is that of Appendices A and B: chart $\phi(v) = \sqrt{1 - \|v\|^2}\beta^\star + Ev$, $v \in \mathbb{R}^p$, $p = d - 1$; proxy center $\beta_2 = \phi(\zeta)$, $r = \|\zeta\|$; $N = N_1 + N_2$, $N_k/N \to \nu_k > 0$; $\sigma_\rho^2 = (1-\rho)^2/\nu_1 + \rho^2/\nu_2$ (in Appendix B the same symbol includes the factor $1/N$). We work with the 0-1 loss $\ell_v(x, y) = \frac{1}{2}[1 - (2y - 1)\operatorname{sign}(x^\top \phi(v))]$, with $\operatorname{sign}(0) = 0$. Its population version $L_{k,0}$ and tangential Hessian H are those of Appendix D (Lemma 17). For the original full-density estimator, the actual half-scaled empirical loss is $\ell_v^- = \mathbf{1}\{(2y - 1)x^\top \phi(v) < 0\}$. The two losses agree away from ties. For the full-density design of Assumption 2, general position implies that any nonzero direction is orthogonal to at most $d - 1$ observations almost surely, so $\sup_v |\mathbb{P}_{N_k}(\ell_v - \ell_v^-)| \le (d - 1)/(2N_k)$. Consequently their local processes differ by $o(1)$ on the $N^{2/3}$ scale for fixed d and positive sample shares. Limit statements for minimizers below include minimizers of the original loss as approximate minimizers of the neutral criterion. Penalties throughout this appendix use the half-loss convention: $\lambda = \lambda_{\mathrm{main}}/2$ and $H = H_{\mathrm{main}}/2$. Thus a displayed penalty λ in THE or TSE corresponds to penalty 2λ in the main-text objective; matrix weights are unchanged by this simultaneous rescaling. The weighted empirical criterion is $\widehat{M}_N(v) = (1 - \rho)\mathbb{P}_{N_1}\ell_v + \rho\,\mathbb{P}_{N_2}\ell_v$ and $\mathbb{G}_{N_k} = \sqrt{N_k}(\mathbb{P}_{N_k} - P_k)$.

C.1 Lower bound due to the bias

Theorem 14 (Lower bound for re-weighting due to the bias)**.** *Under Assumptions 1–3, fix $\rho \in [0, 1]$. On the event $\mathcal{E}$ of Appendix B, which has probability at least $1 - \eta$, and under (Loc),*

$$\|\widehat{v}_{\mathrm{WE}}(\rho)\| \;\ge\; \rho\, r - \kappa_\rho - \tau_\rho - r\sqrt{2\rho(1-\rho)\gamma}, \tag{34}$$

where κ_ρ and τ_ρ are the noise term and the lower-order term of Theorem 8. Since $\angle(\phi(v), \beta^\star) = \arcsin\|v\| \ge \|v\|$, the same lower bound holds for the angular error.

Proof. The proof of Theorem 8 compares $\widehat{v}$ with $\rho\zeta$ and shows, on $\mathcal{E}$, that $\|\widehat{v} - \rho\zeta\| \le \kappa_\rho + \tau_\rho + r\sqrt{2\rho(1-\rho)\gamma}$; the upper bound there is $\|\widehat{v}\| \le \rho r + \|\widehat{v} - \rho\zeta\|$. The reverse triangle inequality $\|\widehat{v}\| \ge \rho r - \|\widehat{v} - \rho\zeta\|$ gives (34). □

For fixed nonzero shift, if (Loc) continues to hold and $\rho > 2(1 - \rho)\gamma$, the limiting lower bound $\rho r(1 - \sqrt{2(1-\rho)\gamma/\rho})$ is positive and excludes consistency at that weight. When this condition fails, the displayed bound alone is inconclusive about consistency.

C.2 THE INCREMENT KERNEL

For $s, t \in \mathbb{R}^p$ let

$$\mathcal{K}(s,t) = \int_{\mathbb{R}^p} f_X(Ez)\, \frac{|z^\top s| + |z^\top t| - |z^\top (s-t)|}{2}\, dz. \tag{35}$$

Lemma 8 (Increment kernel)**.** *Assume the common assumptions and that f_X is continuous. Let $\beta_N \to \beta^\star$ be unit vectors, E_N orthonormal bases of $\beta_N^\perp$ with $E_N \to E$, $A_N \to I$ ($p \times p$ matrices), $\alpha_N \downarrow 0$ and T compact, and put $\beta_N(s) = \beta_N + \alpha_N E_N A_N s + o(\alpha_N)$ uniformly over $s \in T$. Then for $s, t \in T$,*

$$\alpha_N^{-1} \Pr\big\{ \operatorname{sign}(X^\top \beta_N) \neq \operatorname{sign}(X^\top \beta_N(s)),\ \operatorname{sign}(X^\top \beta_N) \neq \operatorname{sign}(X^\top \beta_N(t)) \big\} \ \to\ \mathcal{K}(s,t).$$

Consequently $\mathcal{K}$ is a covariance kernel; it is homogeneous of degree one, $\mathcal{K}(cs, ct) = c\,\mathcal{K}(s,t)$ for $c > 0$; even, $\mathcal{K}(-s,-t) = \mathcal{K}(s,t)$; and its increments are stationary, $\mathcal{K}(s,s) + \mathcal{K}(t,t) - 2\mathcal{K}(s,t) = \|s-t\|_f$ with $\|u\|_f := \int f_X(Ez)|z^\top u|\, dz$, a norm on $\mathbb{R}^p$ with $\|u\|_f \le \bar{f}\, \mathrm{Vol}(B_p(R))R\|u\|$. The centered Gaussian process B with covariance $\mathcal{K}$ therefore has a continuous version; when $f_X(E\,\cdot)$ is radial, $\mathcal{K}(s,t)$ is a constant multiple of $\|s\| + \|t\| - \|s-t\|$ and B is Lévy's Brownian field.

Proof. Write $x = t'\beta_N + E_N z$. Then $x^\top \beta_N(s) = t'(1 + o(1)) + \alpha_N[z^\top A_N s + o(1)]$ uniformly over $\|z\| \le R$ and $s \in T$, so the two signs differ iff t' lies in an interval $I_N(s,z)$ with one end at 0, on the side $-\operatorname{sign}(z^\top A_N s)$, of length $\alpha_N |z^\top A_N s| + o(\alpha_N)$. Both signs differ iff $t' \in I_N(s,z) \cap I_N(t,z)$, which is empty when $z^\top A_N s$ and $z^\top A_N t$ have opposite signs and has length $\alpha_N \min(|z^\top A_N s|, |z^\top A_N t|) + o(\alpha_N)$ otherwise. By Fubini the probability is $\int_{\|z\| \le R} \int_{I_N(s,z) \cap I_N(t,z)} f_X(t'\beta_N + E_N z)\, dt'\, dz$; f_X is uniformly continuous, so $f_X(t'\beta_N + E_N z) = f_X(Ez) + o(1)$ uniformly for $|t'| \le C\alpha_N$; dividing by α_N and using dominated convergence gives $\int f_X(Ez) \min(|z^\top s|, |z^\top t|) \mathbf{1}\{(z^\top s)(z^\top t) > 0\} dz$, which equals (35) because $\min(|a|,|b|) = \frac{1}{2}(|a| + |b| - |a-b|)$ for $ab > 0$ and the right side vanishes for $ab < 0$. Taking $\beta_N = \beta^\star$ and $\alpha_N = \alpha$, the limit is that of $\alpha^{-1}\mathbb{E}[(\ell_{\alpha s} - \ell_0)(\ell_{\alpha t} - \ell_0)]$ (each factor is ± 1 exactly on its disagreement set, with a common sign), so $\mathcal{K}$ is nonnegative definite. The remaining properties are read off (35); the continuity of B follows from $\mathbb{E}(B(s) - B(t))^2 = \|s-t\|_f \lesssim \|s-t\|$ and Kolmogorov's criterion. □

C.3 CUBE-ROOT LIMIT LAW OF THE WEIGHTED ESTIMATOR

Theorem 15 (Limit law of re-weighting)**.** *Let Assumptions 1–2 hold, let f_X be continuous and $\mathbb{E}\|\varepsilon\| < \infty$ (both are implied by Assumptions 4–5). Fix a unit vector $\Theta \in \mathbb{R}^p$ and $r \ge 0$, let $\zeta_N = N^{-1/3} r\Theta$, $N_k/N \to \nu_k > 0$ and $\rho_N \to \rho \in [0,1]$. Let $\widehat{v}$ be the chart coordinate of any minimizer of $\widehat{M}_N$ with weight ρ_N. Then*

$$N^{1/3}\widehat{v} \ \Rightarrow\ \rho r\Theta + \sigma_\rho^{2/3} Z_1, \qquad Z_1 := \arg\min_{u \in \mathbb{R}^p} \Big\{ B(u) + \tfrac{1}{2} u^\top H u \Big\}, \tag{36}$$

where B is the centered Gaussian process with covariance (35)*. The argmin is almost surely unique, $Z_1 \overset{d}{=} -Z_1$, $Z_1 \neq 0$ almost surely, and $\mathbb{E}\|Z_1\|^q < \infty$ for every q. We write $K := \mathbb{E}\|Z_1\| \in (0, \infty)$.*

Proof. We cite van der Vaart & Wellner (1996): Theorem 2.14.1 (maximal inequality for classes with a uniform entropy bound), Theorem 3.2.5 (rate of convergence), Theorem 2.11.23 (functional central limit theorem for triangular arrays) and Theorem 3.2.2 (argmax continuous mapping); and Kim & Pollard (1990), Lemma 2.6 (a Gaussian process with continuous paths and nondegenerate increments, plus a deterministic drift, has an almost surely unique extremum).

Step 0: population expansion. Let $M_N(v) = (1 - \rho_N) L_{1,0}(\phi(v)) + \rho_N L_{2,0}^{(N)}(\phi(v))$, where $L_{2,0}^{(N)}$ is the proxy population loss with center $\beta_2^{(N)} = \phi(\zeta_N)$. Lemma 17, which uses only the continuity of f_X and $\mathbb{E}\|\varepsilon\| < \infty$, expands each loss in its own chart with a remainder $C_{\mathrm{R}} \|\cdot\|^2 \bar{\omega}_0(2R\|\cdot\|)$ whose modulus does not depend on the center. Choose orthonormal bases $E_2^{(N)} \to E$ of $\beta_2^{(N)\perp}$; then $H_2^{(N)} = \int z z^\top a_2^{(N)}(z) dz \to H$ by dominated convergence, and the chart change $\psi_N =$

$(\phi_2^{(N)})^{-1} \circ \phi$ satisfies $\psi_N(\zeta_N) = 0$ and $D\psi_N \to I$ uniformly on $\|v\| \le \eta_0$. Completing the square, there is a modulus $\omega(t) \downarrow 0$ such that, uniformly in N and $\|v\| \le \eta_0$,

$$\begin{aligned} F_N(v) &:= M_N(v) - (1-\rho_N)L_{1,0}(\beta^\star) - \rho_N L_{2,0}^{(N)}(\beta_2^{(N)}) \\ &= \tfrac{1}{2}(v - v_N^\circ)^\top H (v - v_N^\circ) + \tfrac{\rho_N(1-\rho_N)}{2}\zeta_N^\top H \zeta_N + \mathrm{Rem}_N(v), \end{aligned}$$

with $v_N^\circ := \rho_N \zeta_N$ and $|\mathrm{Rem}_N(v)| \le \omega(\|v\| + \|\zeta_N\|)(\|v\|^2 + \|v - \zeta_N\|^2)$. Hence $|\mathrm{Rem}_N(v_N^\circ)| = o(N^{-2/3})$ and, for $\|v - v_N^\circ\| \le \eta_1$ with $\omega(2\eta_1) \le \underline{h}/8$, where $\underline{h} = \lambda_{\min}(H) > 0$ by Lemma 18,

$$F_N(v) - F_N(v_N^\circ) \;\ge\; \tfrac{\underline{h}}{4}\|v - v_N^\circ\|^2 - \epsilon_N, \qquad \epsilon_N = o(N^{-2/3}). \tag{37}$$

Step 1: consistency. By Lemma 19, $L_{k,0}(\beta) - L_{k,0}(\beta_k) \ge c_{\mathrm{q}} \angle(\beta, \beta_k)^2$ for all unit β with constants independent of the center, and the Vapnik–Chervonenkis bound gives $\sup_\beta |\widehat{L}_k(\beta) - L_{k,0}(\beta)| = O_p(\sqrt{d/N_k})$. With $\widehat{\beta} = \phi(\widehat{v})$, $\widehat{M}_N(\widehat{\beta}) \le \widehat{M}_N(\beta^\star)$ and Lemma 20 ($L_{2,0}^{(N)}(\beta^\star) - L_{2,0}^{(N)}(\beta_2^{(N)}) \le \Gamma \angle(\beta^\star, \beta_2^{(N)}) = O(N^{-1/3})$),

$$c_{\mathrm{q}}\big[(1-\rho_N)\angle(\widehat{\beta}, \beta^\star)^2 + \rho_N \angle(\widehat{\beta}, \beta_2^{(N)})^2\big] \le O_p(\sqrt{d/N}) + O(N^{-1/3}),$$

and the bracket is at least $\frac{1}{2}\min\{\angle(\widehat{\beta}, \beta^\star), \angle(\widehat{\beta}, \beta_2^{(N)})\}^2$ with both centers converging to $\beta^\star$. So $\angle(\widehat{\beta}, \beta^\star) \to 0$ in probability, uniformly in ρ_N; with probability tending to one $\widehat{\beta}$ lies in the chart and $\widehat{v} - v_N^\circ \to 0$.

Step 2: rate. Apply Theorem 3.2.5 of van der Vaart & Wellner (1996) to $\mathbb{M}_N(v) = -\widehat{M}_N(v)$, centered at v_N°, with the Euclidean distance. The population condition is (37); the proof of that theorem is a peeling argument, which absorbs the slack $\epsilon_N = o(N^{-2/3})$ (equivalent to $\widehat{v}$ being an $o_p(r_N^{-2})$-approximate maximizer) and allows the center to depend on N as long as both conditions hold uniformly in N. For the modulus condition, let $\mathcal{F}_\alpha = \{\ell_v - \ell_{v_N^\circ} : \|v - v_N^\circ\| \le \alpha\}$. By Lemma 14(c), $\|\phi(v) - \phi(v_N^\circ)\| \le 2\|v - v_N^\circ\|$, so a change of sign requires $|x^\top \phi(v_N^\circ)| \le 2R\alpha$: the envelope $F_\alpha(x) = \mathbf{1}\{|x^\top \phi(v_N^\circ)| \le 2R\alpha\}$ has $PF_\alpha^2 \le 4R\bar{f}_1\alpha$. The class $\mathcal{F}_\alpha$ is contained in the difference class of the VC class $\{\mathbf{1}\{(2y-1)x^\top\beta < 0\}\}$ of index at most $d+1$, so its uniform entropy integral is at most $C\sqrt{d}$, and Theorem 2.14.1 gives $\mathbb{E}\|\mathbb{G}_{N_k}\|_{\mathcal{F}_\alpha} \le C\sqrt{d}\,\|F_\alpha\|_{P,2} \le C'\sqrt{d\alpha}$. The two sources being independent,

$$\begin{aligned} &\mathbb{E} \sup_{\|v - v_N^\circ\| \le \alpha} \sqrt{N}\big|(\widehat{M}_N - M_N)(v) - (\widehat{M}_N - M_N)(v_N^\circ)\big| \\ &\qquad \le \Big[(1-\rho_N)\sqrt{N/N_1} + \rho_N\sqrt{N/N_2}\Big] C'\sqrt{d\alpha} =: \varphi_N(\alpha), \end{aligned}$$

and $\varphi_N(\alpha)/\alpha^{1/2}$ is bounded in N and nonincreasing in α. With $r_N = cN^{1/3}$, $r_N^2 \varphi_N(1/r_N) \le \sqrt{N}$. Since $\widehat{v}$ minimizes $\widehat{M}_N$ exactly inside the chart and is consistent, $N^{1/3}\|\widehat{v} - v_N^\circ\| = O_p(1)$.

Step 3: local process. For a compact $T \subset \mathbb{R}^p$ and $t \in T$ let $v_N(t) = v_N^\circ + N^{-1/3}t$ and

$$\mathbb{Z}_N(t) := N^{2/3}\big[\widehat{M}_N(v_N(t)) - \widehat{M}_N(v_N^\circ)\big] = N^{2/3}\big[F_N(v_N(t)) - F_N(v_N^\circ)\big] + \mathbb{W}_N(t).$$

By Step 0 the deterministic part equals $\frac{1}{2}t^\top H t$ plus $N^{2/3}$ times $\mathrm{Rem}_N(v_N(t)) - \mathrm{Rem}_N(v_N^\circ)$, and the latter is $\omega(O(N^{-1/3}))\,O(N^{-2/3})$ uniformly on T; so the deterministic part converges to $\frac{1}{2}t^\top H t$ in $\ell^\infty(T)$. The stochastic part is

$$\mathbb{W}_N(t) = (1-\rho_N)\sqrt{N/N_1}\,\mathbb{G}_{N_1}(f_{N,t}^{(1)}) + \rho_N\sqrt{N/N_2}\,\mathbb{G}_{N_2}(f_{N,t}^{(2)}), \qquad f_{N,t}^{(k)} := N^{1/6}\big[\ell_{v_N(t)} - \ell_{v_N^\circ}\big].$$

For each source we verify the conditions of Theorem 2.11.23 for $\mathcal{F}_N^{(k)} = \{f_{N,t}^{(k)} : t \in T\}$. (i) Envelope: $\bar{F}_N = N^{1/6}\mathbf{1}\{|x^\top\phi(v_N^\circ)| \le 2RN^{-1/3}\sup_T\|t\|\}$ has $P_k\bar{F}_N^2 \le 4R\bar{f}_1 \sup_T\|t\| = O(1)$ and $\bar{F}_N \le N^{1/6}$, so $P_k\bar{F}_N^2\mathbf{1}\{\bar{F}_N > \epsilon\sqrt{N_k}\} = 0$ for large N. (ii) By Lemma 13 and the chart geometry, $\sup_{\|s-t\|<\delta_N} P_k(f_{N,s} - f_{N,t})^2 \le N^{1/3}\Gamma\frac{\pi}{\sqrt{2}}N^{-1/3}\delta_N \to 0$. (iii) $\mathcal{F}_N^{(k)}$ is a constant multiple of a subset of the difference class of a VC class, so $\sup_Q N(\epsilon\|\bar{F}_N\|_{Q,2}, \mathcal{F}_N^{(k)}, L_2(Q)) \le (C/\epsilon)^{2(d+1)}$

and the entropy integral over $(0, \delta_N)$ tends to zero. (iv) Covariances: $P_k f_{N,t} = N^{1/6} O(N^{-1/3}) \to 0$ by Lemma 20, and $P_k(f_{N,s} f_{N,t}) = N^{1/3} \Pr_k\{\text{both signs differ}\}$. Put $\beta_N = \phi(v_N^\circ) \to \beta^\star$ and $\alpha_N = N^{-1/3}$; by Lemma 14(b), $\phi(v_N(t)) = \beta_N + N^{-1/3} D\phi(v_N^\circ) t + O(N^{-2/3})$ with $\beta_N^\top D\phi(v_N^\circ) = 0$, so $D\phi(v_N^\circ) t = E_N A_N t$ for an orthonormal basis $E_N \to E$ of $\beta_N^\perp$ and $A_N = E_N^\top D\phi(v_N^\circ) \to I$. Lemma 8 gives $P_k(f_{N,s} f_{N,t}) \to \mathcal{K}(s,t)$ for both sources, which share the design. Hence $\mathbb{G}_{N_k}(f_{N,\cdot}^{(k)}) \Rightarrow B_k$ in $\ell^\infty(T)$, with B_1, B_2 independent centered Gaussian processes with covariance $\mathcal{K}$, and

$$\mathbb{Z}_N \;\Rightarrow\; \mathbb{Z} := (1-\rho)\nu_1^{-1/2} B_1 + \rho\, \nu_2^{-1/2} B_2 + \tfrac{1}{2} t^\top H t \;\overset{d}{=}\; \sigma_\rho B + \tfrac{1}{2} t^\top H t \qquad \text{in } \ell^\infty(T).$$

Step 4: argmin. $\mathbb{Z}$ has continuous paths (Lemma 8) and its Gaussian part has nondegenerate increments, $\mathrm{Var}(\sigma_\rho B(s) - \sigma_\rho B(t)) = \sigma_\rho^2 \|s-t\|_f > 0$ for $s \neq t$ (note $\sigma_\rho^2 \geq 1$), so by Lemma 2.6 of Kim & Pollard (1990) its minimizer is almost surely unique. It exists: by Dudley's bound and the Borell–TIS inequality, $\Pr\{\sup_{\|u\| \leq 2^{j+1}} |B(u)| > 2^{2j} \underline{h}/8\} \leq \exp(-c\, 2^{3j})$ is summable, so $\mathbb{Z}(u) \to \infty$ as $\|u\| \to \infty$ almost surely. Step 2 shows that $\hat{t}_N := N^{1/3}(\widehat{v} - v_N^\circ)$ is uniformly tight, and it minimizes $\mathbb{Z}_N$ exactly over a set containing any given compact for large N. Theorem 3.2.2 (applied to $-\mathbb{Z}_N$) gives $\hat{t}_N \Rightarrow \arg\min \mathbb{Z}$.

Step 5: scaling, symmetry, moments. Homogeneity of $\mathcal{K}$ gives $B(a\,\cdot) \overset{d}{=} a^{1/2} B(\cdot)$; substituting $u = \sigma_\rho^{2/3} u'$, $\sigma_\rho B(\sigma_\rho^{2/3} u') + \frac{1}{2}\sigma_\rho^{4/3} u'^\top H u' \overset{d}{=} \sigma_\rho^{4/3}[B(u') + \frac{1}{2} u'^\top H u']$, so $\arg\min \mathbb{Z} \overset{d}{=} \sigma_\rho^{2/3} Z_1$. Evenness of $\mathcal{K}$ gives $B(-\cdot) \overset{d}{=} B(\cdot)$, hence $Z_1 \overset{d}{=} -Z_1$. Along a ray $u = s u_0$, $u_0 \neq 0$, $s \mapsto B(s u_0)$ is a Brownian motion with variance rate $\|u_0\|_f > 0$; by the law of the iterated logarithm $B(s u_0) + \frac{s^2}{2} u_0^\top H u_0 < 0 = \mathbb{Z}(0)$ for some small s, so $Z_1 \neq 0$ almost surely. If $\|Z_1\| > 2^j$ then $\sup_{\|u\| \leq 2^{j'+1}} |B(u)| \geq 2^{2j'} \underline{h}/8$ for some $j' \geq j$, an event of probability at most $\sum_{j' \geq j} e^{-c 2^{3j'}}$; all moments are finite. Finally $N^{1/3}\widehat{v} = N^{1/3} v_N^\circ + \hat{t}_N$ and $N^{1/3} v_N^\circ = \rho_N r \Theta \to \rho r \Theta$. □

C.4 Lower bounds for gold-only and re-weighting

Theorem 16 (Lower bounds due to the noise)**.** *Under the conditions of Theorem 15, with Θ uniform on the unit sphere of $\beta^{\star\perp}$ and independent of the data,*

$$\liminf_{N\to\infty} N^{1/3}\, \mathbb{E}\, \alpha\big(\widehat{\beta}_{\mathrm{WE}}(\rho_N), \beta^\star\big) \;\geq\; \mathbb{E}\big\|\rho r \Theta + \sigma_\rho^{2/3} Z_1\big\| \;\geq\; \max\big\{\rho r,\; \sigma_\rho^{2/3} K\big\}. \tag{38}$$

In particular $\liminf N_1^{1/3} \mathbb{E}\angle(\widehat{\beta}_{\mathrm{gold}}, \beta^\star) \geq K$ *and* $\liminf N^{1/3} \mathbb{E}\angle(\widehat{\beta}_{\mathrm{proxy}}, \beta^\star) \geq \max\{r, \nu_2^{-1/3} K\}$. *For every* $x < \max\{\rho r, \sigma_\rho^{2/3} K\}$, $\liminf_N \Pr\{N^{1/3}\alpha \geq x\} \geq \Pr\{\|\rho r \Theta + \sigma_\rho^{2/3} Z_1\| > x\} > 0$.

Proof. With probability tending to one $\widehat{\beta}$ is in the chart and $\alpha = \arcsin \|\widehat{v}\| \geq \|\widehat{v}\|$; on the complement $\alpha \geq 0$. Given Θ, $N^{1/3}\|\widehat{v}\| \mathbf{1}\{\text{chart}\} \Rightarrow \|\rho r \Theta + \sigma_\rho^{2/3} Z_1\|$, and the portmanteau theorem for the nonnegative continuous function $\|\cdot\|$ gives $\liminf \mathbb{E}[N^{1/3}\alpha \mid \Theta] \geq \mathbb{E}[\|\rho r \Theta + \sigma_\rho^{2/3} Z_1\| \mid \Theta]$ without any uniform integrability; Fatou's lemma in Θ gives the first inequality. For the second, Z_1 and Θ are independent and symmetric, so Jensen's inequality conditionally on Z_1 gives $\mathbb{E}\|\rho r \Theta + \sigma_\rho^{2/3} Z_1\| \geq \mathbb{E}\|\rho r\, \mathbb{E}\Theta + \sigma_\rho^{2/3} Z_1\| = \sigma_\rho^{2/3} K$, and conditionally on Θ gives $\geq \rho r$. The probability statement is the portmanteau theorem for open sets. □

Corollary 1 (Re-weighting at its best weight)**.** *Under the conditions of Theorem 15, write $\Phi_{\mathrm{gold}} = \nu_1^{-1/3} K$ and $\underline{\Phi}_{\mathrm{WE}}(r) = \min_{\rho \in [0,1]} \max\{\rho r, \sigma_\rho^{2/3} K\}$.*

(a) $\liminf_N \inf_{\rho \in [0,1]} N^{1/3} \mathbb{E}\angle(\widehat{\beta}_{\mathrm{WE}}(\rho), \beta^\star) \geq \underline{\Phi}_{\mathrm{WE}}(r) \geq K$.

(b) $\underline{\Phi}_{\mathrm{WE}}(r) \geq \min\{2^{-2/3}\Phi_{\mathrm{gold}},\; \frac{r}{4} + 2^{-5/3}\nu_2^{-1/3} K\}$*; in the original units,*

$$\liminf_N \inf_\rho \; \mathbb{E}\angle(\widehat{\beta}_{\mathrm{WE}}(\rho), \beta^\star) \Big/ \min\Big\{2^{-2/3} K N_1^{-1/3},\; \tfrac{1}{4}\|\delta\| + 2^{-5/3} K N_2^{-1/3}\Big\} \;\geq\; 1.$$

(c) If $\|\delta\| \geq mKN_1^{-1/3}$ *with* $m > 1$*, then* $\liminf_N \inf_\rho N_1^{1/3}\mathbb{E}\angle(\widehat{\beta}_{\mathrm{WE}}(\rho), \beta^\star) \geq (1 - 1/m)^{2/3}K$*, while gold-only has the limit constant* K*.*

Proof. (a) Let ρ_N be a $1/N$-approximate minimizer of $\rho \mapsto N^{1/3}\mathbb{E}\angle(\widehat{\beta}_{\mathrm{WE}}(\rho), \beta^\star)$; along any subsequence with $\rho_N \to \rho_\infty$, Theorem 16 gives the liminf $\geq \max\{\rho_\infty r, \sigma_{\rho_\infty}^{2/3}K\} \geq \underline{\Phi}_{\mathrm{WE}}(r)$; and $\sigma_\rho^2 \geq 1$ with equality at $\rho = \nu_2$. (b) For $\rho \leq \frac{1}{2}$, $\sigma_\rho^2 \geq (1-\rho)^2/\nu_1 \geq 1/(4\nu_1)$, so $\sigma_\rho^{2/3}K \geq 2^{-2/3}\Phi_{\mathrm{gold}}$; for $\rho \geq \frac{1}{2}$, $\rho r \geq r/2$ and $\sigma_\rho^{2/3}K \geq 2^{-2/3}\nu_2^{-1/3}K$, and a maximum is at least the average. (c) Since $\sigma_\rho^{2/3} \geq (1-\rho)^{2/3}\nu_1^{-1/3}$, $\underline{\Phi}_{\mathrm{WE}}(r) \geq \min_\rho g(\rho)$ with $g(\rho) = \max\{\rho r, (1-\rho)^{2/3}\Phi_{\mathrm{gold}}\}$; ρr increases from 0 and $(1-\rho)^{2/3}\Phi_{\mathrm{gold}}$ decreases from Φ_{gold} to 0, so $\min g = g(\rho_\times) = (1-\rho_\times)^{2/3}\Phi_{\mathrm{gold}}$ at the crossing, where $\rho_\times r = (1-\rho_\times)^{2/3}\Phi_{\mathrm{gold}} \leq \Phi_{\mathrm{gold}}$ gives $\rho_\times \leq \Phi_{\mathrm{gold}}/r \leq 1/m$. Multiplying by $\nu_1^{1/3}$ converts $N^{1/3}$ into $N_1^{1/3}$ units. □

Corollary 2 (Pooling against gold-only at the level of constants)**.** *Under the conditions of Theorem 15, with* Θ *uniform on the unit sphere of* $\beta^{\star\perp}$ *and independent of the data, let* $\widehat{\beta}_{\mathrm{gold}}$ *and* $\widehat{\beta}_{\mathrm{WE}}(N_2/N)$ *be any minimizers and put* $\Phi_{\mathrm{pool}}(r) = \mathbb{E}\|\nu_2 r\Theta + Z_1\|$*.*

(a) $N_1^{1/3}\mathbb{E}\angle(\widehat{\beta}_{\mathrm{gold}}, \beta^\star) \to K$*, which needs only* $N_1 \to \infty$*, so that* $N^{1/3}\mathbb{E}\angle(\widehat{\beta}_{\mathrm{gold}}, \beta^\star) \to \Phi_{\mathrm{gold}} = \nu_1^{-1/3}K$ *under the shares; and* $N^{1/3}\mathbb{E}\angle(\widehat{\beta}_{\mathrm{WE}}(N_2/N), \beta^\star) \to \Phi_{\mathrm{pool}}(r)$*.*

(b) $\max\{\nu_2 r, K\} \leq \Phi_{\mathrm{pool}}(r) \leq \nu_2 r + K$*;* Φ_{pool} *is convex and even, hence nondecreasing on* $[0, \infty)$*, and strictly increasing wherever it exceeds* $\Phi_{\mathrm{pool}}(0) = K$*.*

(c) There is a unique r^**, with*

$$\frac{\nu_1^{-1/3} - 1}{\nu_2}\,K \;\leq\; r^* \;\leq\; \frac{\nu_1^{-1/3}}{\nu_2}\,K,$$

such that $\Phi_{\mathrm{pool}}(r) < \Phi_{\mathrm{gold}}$ *for* $r < r^*$ *and* $\Phi_{\mathrm{pool}}(r) > \Phi_{\mathrm{gold}}$ *for* $r > r^*$*. In the units of the gold-only error,* $m := \nu_1^{1/3}r/K = \lim_N N_1^{1/3}\|\delta_N\|/K$*, pooling has a strictly smaller limit than gold-only if* $m < (1 - \nu_1^{1/3})/\nu_2$ *and a strictly larger one if* $m > 1/\nu_2$*; at* $m = 0$ *the ratio of the two limits is* $\nu_1^{1/3}$*.*

For every fixed direction $\Theta = \theta$ *the bounds in (b), and the bounds on the threshold in (c), hold as well, with a threshold that may depend on* θ*, because* Z_1 *is symmetric.*

Proof. (a) Gold-only is re-weighting at $\rho = 0$. Theorem 15 with $\rho_N \equiv 0$ gives $N^{1/3}\widehat{v}_{\mathrm{gold}} \Rightarrow \sigma_0^{2/3}Z_1 = \nu_1^{-1/3}Z_1$, and with $\rho_N = N_2/N \to \nu_2$, where $\sigma_{\nu_2}^2 = \nu_1 + \nu_2 = 1$, it gives $N^{1/3}\widehat{v} \Rightarrow \nu_2 r\Theta + Z_1$ conditionally on Θ. On the chart $\angle = \arcsin\|\widehat{v}\| = \|\widehat{v}\|\{1 + O(\|\widehat{v}\|^2)\}$, so $N^{1/3}\angle$ has the same limits by Slutsky's lemma; off the chart $N^{1/3}\angle \leq \pi N^{1/3}$ on an event of probability $O(e^{-cN})$ (proof of Lemma 9). Lemma 9, whose bounds are uniform in the weight $\rho \in [0,1]$ and in the direction, gives uniform integrability, so the conditional expectations converge to $\nu_1^{-1/3}K$ and to $\mathbb{E}[\|\nu_2 r\Theta + Z_1\| \mid \Theta]$ and are bounded uniformly in Θ; dominated convergence in Θ gives the limits. For gold-only the proxy sample does not enter, and the single-population case of Theorem 15 and Lemma 9 with N_1 in place of N, as in Theorem 17(a), gives $N_1^{1/3}\mathbb{E}\angle(\widehat{\beta}_{\mathrm{gold}}, \beta^\star) \to K$ from $N_1 \to \infty$ alone.

(b) The upper bound is the triangle inequality. Conditionally on Z_1, Jensen's inequality with $\mathbb{E}\Theta = 0$ gives $\Phi_{\mathrm{pool}}(r) \geq \mathbb{E}\|Z_1\| = K$; conditionally on Θ, the symmetry of Z_1 gives $\mathbb{E}\|\nu_2 r\Theta + Z_1\| = \frac{1}{2}\mathbb{E}[\|\nu_2 r\Theta + Z_1\| + \|\nu_2 r\Theta - Z_1\|] \geq \nu_2 r$. Convexity holds because $r \mapsto \|\nu_2 r\Theta + Z_1\|$ is convex for every outcome, and evenness because $\Theta \overset{d}{=} -\Theta$; a convex even function is nondecreasing on $[0,\infty)$, and if $\Phi_{\mathrm{pool}}(r) > K$ and $r' > r > 0$ then $\Phi_{\mathrm{pool}}(r') \geq \Phi_{\mathrm{pool}}(r) + (r' - r)\{\Phi_{\mathrm{pool}}(r) - K\}/r > \Phi_{\mathrm{pool}}(r)$.

(c) Φ_{pool} is continuous, $\Phi_{\mathrm{pool}}(0) = K < \Phi_{\mathrm{gold}}$ and $\Phi_{\mathrm{pool}}(r) \geq \nu_2 r \to \infty$, so, by the monotonicity in (b), $\{\Phi_{\mathrm{pool}} < \Phi_{\mathrm{gold}}\} = [0, r^*)$ for some r^*, and by the strict increase in (b) $\Phi_{\mathrm{pool}} > \Phi_{\mathrm{gold}}$ on

(r^*, ∞). The bounds on r^* follow from (b): $\nu_2 r + K < \Phi_{\text{gold}}$ when $r < (\nu_1^{-1/3} - 1)K/\nu_2$, and $\nu_2 r > \Phi_{\text{gold}}$ when $r > \nu_1^{-1/3} K/\nu_2$. Multiplying by $\nu_1^{1/3}$ converts $N^{1/3}$ into $N_1^{1/3}$ units and r into mK. For a fixed direction θ, $r \mapsto \mathbb{E}\|\nu_2 r\theta + Z_1\|$ is even because $Z_1 \overset{d}{=} -Z_1$, Jensen's inequality with $\mathbb{E}Z_1 = 0$ bounds it below by $\nu_2 r$, and $\frac{1}{2}\mathbb{E}[\|\nu_2 r\theta + Z_1\| + \|\nu_2 r\theta - Z_1\|] \geq \mathbb{E}\|Z_1\|$ bounds it below by K; the rest of (b) and (c) is unchanged, with a threshold that satisfies the same bounds. □

C.5 Upper bound against lower bound: when the proxy helps

Theorem 17 (Dichotomy). *Let Assumptions 1–3 hold, let f_X be continuous and $\mathbb{E}\|\varepsilon\| < \infty$, and let ℓ, A and γ be as in Section 3. Parts (b)–(c)'s upper bounds require (Loc), uniform bracket constants, and negligible displayed lower-order terms. The gold-only limit in (a) requires only $N_1 \to \infty$ in the fixed gold design; the pooled upper bound can therefore be compared with it even when N_2/N_1 diverges. The large-rescaled-bias lower bound in (c), in contrast, uses the local alternatives and positive limiting shares of Theorem 15.*

(a) $\liminf_{N_1} N_1^{1/3} \mathbb{E}\angle(\widehat{\beta}_{\text{gold}}, \beta^\star) \geq K$.

(b) With probability at least $1-\eta$, up to a term of order $\sqrt{\ell/N}$ at the pooling weight, the pooled sign estimator, which is re-weighting with $\rho = N_2/N$, satisfies $\angle(\widehat{\beta}_{\text{WE}}(N_2/N), \beta^\star) \leq \big(\frac{N_2}{N} + \frac{\sqrt{2N_1 N_2 \gamma}}{N}\big)\|\delta\| + A(\ell/N)^{1/3}$, because $A[(1-\rho)^2\ell/N_1 + \rho^2\ell/N_2]^{1/3} = A(\ell/N)^{1/3}$ at $\rho = N_2/N$. The two-headed estimator at $\lambda = \infty$ is re-weighting, and the two-step upper bound at its best penalty is at most the smaller of the gold-only and proxy-only upper bounds (Theorem 9); so when $\|\delta\| \leq A(\ell/N)^{1/3}$ each of the three estimators that use the proxy, at its best tuning, has an upper bound of order $(\ell/N)^{1/3}$.

(c) If $\|\delta\| \leq A(\ell/N)^{1/3}$, then with probability at least $1-\eta$,

$$\frac{\angle(\widehat{\beta}_{\text{WE}}(N_2/N), \beta^\star)}{K N_1^{-1/3}} \leq \frac{(2+\sqrt{2\gamma})A}{K}\Big(\frac{\ell N_1}{N}\Big)^{1/3}(1+o(1)),$$

which tends to zero when $N_2/N_1 \gg \ell$. If $\|\delta\| \geq mKN_1^{-1/3}$ with $m > 1$, no weight improves on gold-only by more than the factor $(1-1/m)^{2/3}$ (Corollary 1(c)), and for fixed δ and fixed $\rho > 0$ Theorem 14 gives a finite-sample lower bound due to the bias.

Proof. (a) follows from the single-population proof of Theorem 15 with its sample size set to N_1, and portmanteau. It does not require a joint limit for the proxy sample. (b) is Theorem 8 at $\rho = N_2/N$ together with the identity $W_{\text{THE}}(\rho, \infty) = \rho I$ and the monotonicity of the two-step upper bound; (c) divides (b) by (a) and uses $(\ell/N)^{1/3} N_1^{1/3} = (\ell N_1/N)^{1/3}$. □

Remark 1 (Scope of the comparisons). *For general full-dimensional designs, the upper bounds here have logarithmic and empirical-process constants and do not resolve the constant-factor ordering of THE, WE, and TSE. The controlled-design variant below admits a separate comparison of exact constants. This is a limitation of the bounds established here, not an impossibility of a sharper analysis.*

C.6 Limit laws of the two-headed and two-step estimators for the exact sign loss

Theorem 18 (Limit laws of THE and TSE without smoothing). *Under the conditions of Theorem 15, fix $\rho \in (0,1)$ and $\lambda \in (0, \infty)$, and let B_1, B_2 be independent centered Gaussian processes with covariance* (35).

(a) For the two-headed estimator (3) *with the exact sign loss, $N^{1/3}(\widehat{v}_1, \widehat{v}_2) \Rightarrow (u_1^\circ, u_2^\circ) + \mathcal{Z}$, where (u_1°, u_2°) minimizes $\frac{1-\rho}{2} u_1^\top H u_1 + \frac{\rho}{2}(u_2 - r\Theta)^\top H(u_2 - r\Theta) + \lambda\|u_1 - u_2\|^2$, so that*

$u_1^\circ = W_{\mathrm{THE}} r\Theta$ *with* $W_{\mathrm{THE}} = \rho(I + \frac{\rho(1-\rho)}{2\lambda} H)^{-1}$, *and*

$$\mathcal{Z} = (\mathcal{Z}_1, \mathcal{Z}_2) := \arg \min_{(a,b)\in\mathbb{R}^{2p}} \Big\{ (1-\rho)\nu_1^{-1/2} B_1(a) + \rho\, \nu_2^{-1/2} B_2(b) + \tfrac{1-\rho}{2} a^\top H a + \tfrac{\rho}{2} b^\top H b + \lambda \|a-b\|^2 \Big\}$$

is almost surely unique, symmetric, independent of Θ*, and its law does not depend on* r*. The gold head has the exact constant* $\Phi_{\mathrm{THE}}(\rho, \lambda; r) = \mathbb{E}\|W_{\mathrm{THE}} r\Theta + \mathcal{Z}_1\|$.

(b) For the two-step estimator (6) *with the exact sign loss, with* $W_{\mathrm{TSE}} = (I + \frac{1}{2\lambda} H)^{-1}$ *and* $Z^{(2)}$ *a copy of* Z_1 *built from the proxy sample and independent of* B_1*,*

$$N^{1/3}\widehat{v}_{\mathrm{TSE}} \;\Rightarrow\; W_{\mathrm{TSE}}\big(r\Theta + \nu_2^{-1/3} Z^{(2)}\big) + \arg\min_a \Big\{ \nu_1^{-1/2} B_1(a) + \tfrac{1}{2} a^\top (H + 2\lambda I) a \Big\}.$$

Proof. The steps are those of Theorem 15; we list the differences. (a) The parameter is $(v_1, v_2) \in \mathbb{R}^{2p}$. The population criterion is $(1-\rho)L_{1,0}(\phi(v_1)) + \rho L_{2,0}^{(N)}(\phi(v_2)) + \lambda\|\phi(v_1) - \phi(v_2)\|^2$, and by Lemma 14(c) the penalty equals $\lambda\|v_1 - v_2\|^2(1 + O(\eta_0^2))$, a correction of order $N^{-4/3}$ at scale $N^{-1/3}$. Its quadratic part has the Hessian $\begin{pmatrix} (1-\rho)H + 2\lambda I & -2\lambda I \\ -2\lambda I & \rho H + 2\lambda I \end{pmatrix}$, positive definite since $(1-\rho)a^\top H a + \rho b^\top H b + 2\lambda\|a-b\|^2 > 0$ unless $a = b = 0$; the first-order conditions $(1-\rho)Hu_1 + 2\lambda(u_1 - u_2) = 0$, $\rho H(u_2 - \zeta) - 2\lambda(u_1 - u_2) = 0$ give $u_1^\circ = \rho(I + \frac{\rho(1-\rho)}{2\lambda} H)^{-1}\zeta$. Consistency uses the global growth of each loss and the nonnegativity of the penalty; the modulus is again $C\sqrt{d\alpha}$ since the empirical process is a sum of two independent processes each depending on one coordinate; the local process converges to $(1-\rho)\nu_1^{-1/2} B_1(a) + \rho\nu_2^{-1/2} B_2(b)$ plus the centered quadratic. The increments of that Gaussian process on $\mathbb{R}^{2p}$ have variance $(1-\rho)^2\nu_1^{-1}\|a - a'\|_f + \rho^2\nu_2^{-1}\|b - b'\|_f$, positive for $(a, b) \neq (a', b')$ when $\rho \in (0, 1)$, so Lemma 2.6 of Kim & Pollard (1990) applies. Stationarity of the increments makes the law of the noise part invariant under the shift of the center, hence independent of r and symmetric. (b) Step 1 is Theorem 15 at $\rho = 1$: $N^{1/3}\widehat{v}_2 \Rightarrow r\Theta + \nu_2^{-1/3} Z^{(2)}$. Given $\widehat{v}_2$, which is independent of the gold sample, the quadratic part of Step 2 is $\frac{1}{2} u^\top H u + \lambda\|u - \hat{u}_2\|^2$ with $\hat{u}_2 = N^{1/3}\widehat{v}_2$, minimized at $(H/2 + \lambda I)^{-1}\lambda\hat{u}_2 = W_{\mathrm{TSE}}\hat{u}_2$, and the centered quadratic is $\frac{1}{2} a^\top (H + 2\lambda I) a$; the gold local process is as in Theorem 15, and joint convergence follows from independence. □

Lemma 9 (Expected-risk convergence for the exact loss)**.** *In the fixed-dimensional local regime of Theorem 15, with* r *bounded and sample shares bounded away from zero, the rescaled angular errors* $N^{1/3}\angle(\widehat{\beta}, \beta^\star)$ *are uniformly integrable for WE, for THE at fixed* $\rho \in (0, 1)$ *and finite penalty, and for TSE at fixed finite penalty. The claim is uniform in the bias direction and in WE weights. It also holds for the controlled-design net estimators defined below. Hence their expected rescaled errors converge to the expectations of the displayed limit norms.*

Proof. Here and below constants may depend on the fixed design, dimension, bound on r, and fixed penalty parameters. We give the tail argument, including the finite-sample step needed in addition to the tails of the limit field. On a sufficiently small fixed chart, the population expansion yields

$$M_N(v) - M_N(v_N^\circ) \geq c\|v - v_N^\circ\|^2 - o(N^{-2/3}), \qquad \|v_N^\circ\| \leq CN^{-1/3}.$$

For THE use the product chart and its positive definite block Hessian; the same inequality holds for the pair of heads. Outside the chart, Lemma 6(a) gives the fixed population gap. The deterministic penalty contributes no empirical-process fluctuation. For a ball of radius a about the comparison point, the increment class has a bounded envelope supported on a slab of probability at most Ca, fixed VC entropy integral, variance at most Ca, and expected empirical supremum at most $C\sqrt{a/N}$. These are the envelope and maximal-inequality bounds in Step 2 of Theorem 15, applied to each source. Talagrand–Bousquet concentration therefore gives, for a within the fixed chart and $a \geq t_0 N^{-1/3}$, with $t_0 = (4C/c)^{2/3}$ so that $C\sqrt{a/N} \leq ca^2/4$, and for N large enough that the $o(N^{-2/3})$ remainder is at most $ca^2/2$ at $a = t_0 N^{-1/3}$,

$$\Pr\left\{ \sup_{\|v - v_N^\circ\| \leq 2a} |(\widehat{M}_N - M_N)(v) - (\widehat{M}_N - M_N)(v_N^\circ)| > ca^2/2 \right\} \leq Ce^{-c' N a^3}.$$

Indeed $\sqrt{a/N} = o(a^2)$ when Na^3 is sufficiently large, and the Bernstein denominator is $O(a+a^2)$ at threshold a^2. Peel the annuli $a, 2a, 4a, \dots$ and sum the resulting exponentials. Outside the chart, global identification gives a fixed positive population gap, so the bounded VC deviation inequality gives probability at most $Ce^{-c'N}$. The $O(d/N)$ tie discrepancy is absorbed by ca^2. Consequently, for large N and $t_0 \le t \le \pi N^{1/3}$ (for $t < t_0$ the trivial bound suffices),

$$\Pr\{N^{1/3}\angle(\widehat{\beta}, \beta^\star) > t\} \le Ce^{-c't^3} + Ce^{-c'N}.$$

For TSE first apply this bound to the independent proxy fit. On the event that its error is at most ϵa, the population gold criterion plus the quadratic penalty has growth at scale a about a point of size $O(N^{-1/3}) + O(\epsilon a)$; choose the fixed ϵ small enough and repeat the gold-sample argument. The complement has the same e^{-cNa^3} bound. Independence permits this conditioning.

For the controlled design, the global identification and local slab bounds are established below. Compare the empirical minimum with a deterministic net point within N^{-3} of the population comparison point (one per head). Its loss discrepancy is bounded by the number of observations in a slab of probability $O(N^{-3})$, divided by N, and its penalty discrepancy is $O(N^{-3})$. Bernstein's inequality for the binomial count of observations in that slab makes a discrepancy larger than $ca^2/4$ have probability at most $Ce^{-c'Na^2} \le Ce^{-c'Na^3}$ throughout the same range. This supplies the same peeling bound for the finite-net minimum.

Finally, integrating the tail bound proves uniform integrability. The bad-chart contribution is bounded by $C\pi N^{1/3}e^{-c'N}$; in the tail above M it is needed only for $\pi N^{1/3} > M$ and its supremum tends to zero. Finitely many small N are bounded variables. Convergence in distribution and uniform integrability give the claimed expected-risk limits, including direction averaging. □

Corollary 3 (The two-step estimator against gold-only at the level of constants)**.** *Under the conditions of Theorem 15, fix $\lambda \in (0,\infty)$, let $\mathcal{Z}_\lambda = \arg\min_a\{\nu_1^{-1/2}B_1(a) + \frac{1}{2}a^\top Ha + \lambda\|a\|^2\}$ be the gold noise of Theorem 18(b), so that $\mathcal{Z}_0 \overset{d}{=} \nu_1^{-1/3}Z_1$ is the gold-only noise, let $\bar{u}_2 = r\Theta + \nu_2^{-1/3}Z^{(2)}$, $w = \|W_{\mathrm{TSE}}\|_{\mathrm{op}} = 2\lambda/(2\lambda + \underline{h})$, and write $\Phi_{\mathrm{TSE}}(\lambda; r) = \mathbb{E}\|W_{\mathrm{TSE}}\bar{u}_2 + \mathcal{Z}_\lambda\|$ and $\Phi_{\mathrm{gold}} = \nu_1^{-1/3}K$ for the constants of the limit laws.*

(a) $\|\mathcal{Z}_\lambda\| \le \|\mathcal{Z}_{\lambda'}\|$ *almost surely whenever* $\lambda \ge \lambda' \ge 0$*; hence* $\theta(\lambda) := \mathbb{E}\|\mathcal{Z}_\lambda\|/\Phi_{\mathrm{gold}} \le 1$ *is nonincreasing.*

(b) $\max\{\theta(\lambda)\Phi_{\mathrm{gold}},\ \mathbb{E}\|W_{\mathrm{TSE}}r\Theta\|,\ \nu_2^{-1/3}\mathbb{E}\|W_{\mathrm{TSE}}Z^{(2)}\|\} \le \Phi_{\mathrm{TSE}}(\lambda; r) \le \theta(\lambda)\Phi_{\mathrm{gold}} + w\,(r + \nu_2^{-1/3}K)$.

(c) If $H = \underline{h}I$, then $\mathcal{Z}_\lambda \overset{d}{=} (1-w)^{2/3}\mathcal{Z}_0$, $\frac{d}{dw}\Phi_{\mathrm{TSE}}|_{w=0} = -\frac{2}{3}\Phi_{\mathrm{gold}} < 0$, so that $\Phi_{\mathrm{TSE}} < \Phi_{\mathrm{gold}}$ for all small $\lambda > 0$ whatever the bias; and with $m = \|\delta\|N_1^{1/3}/K$,

$$\frac{\Phi_{\mathrm{TSE}}(\lambda; r)}{\Phi_{\mathrm{gold}}} \le (1-w)^{2/3} + w\big(m + (\nu_1/\nu_2)^{1/3}\big),$$

whose right side is below one for some $w \in (0,1)$ if and only if $m + (\nu_1/\nu_2)^{1/3} < 1$.

Proof. (a) With $f(a) = \nu_1^{-1/2}B_1(a) + \frac{1}{2}a^\top Ha$, minimality gives $f(\mathcal{Z}_\lambda) + \lambda\|\mathcal{Z}_\lambda\|^2 \le f(\mathcal{Z}_{\lambda'}) + \lambda\|\mathcal{Z}_{\lambda'}\|^2$ and the same with the roles exchanged; adding the two, $(\lambda - \lambda')(\|\mathcal{Z}_\lambda\|^2 - \|\mathcal{Z}_{\lambda'}\|^2) \le 0$. No convexity of f is used. (b) The upper bound is the triangle inequality; for the lower bound the three terms are independent and symmetric, and Jensen's inequality conditionally on the other two applies as in Theorem 16. (c) Substituting $a = tu'$, $\sigma B(tu') + \frac{c}{2}t^2\|u'\|^2 \overset{d}{=} t^2 c[\sigma c^{-1}t^{-3/2}B(u') + \frac{1}{2}\|u'\|^2]$, and $t = (\sigma/c)^{2/3}$ shows $\arg\min\{\sigma B + \frac{c}{2}\|\cdot\|^2\} \overset{d}{=} (\sigma/c)^{2/3}\arg\min\{B + \frac{1}{2}\|\cdot\|^2\}$; comparing $c = \underline{h} + 2\lambda$ with $c = \underline{h}$ gives the factor $(1-w)^{2/3}$. Writing $\Phi_{\mathrm{TSE}} = \mathbb{E}\|(1-w)^{2/3}\mathcal{Z}_0 + w\bar{u}_2\|$ with $\mathcal{Z}_0$ independent of $\bar{u}_2$, the difference quotient is dominated by $C\|\mathcal{Z}_0\| + \|\bar{u}_2\|$, and dominated convergence gives the derivative $\mathbb{E}[\|\mathcal{Z}_0\|^{-1}\mathcal{Z}_0^\top(-\frac{2}{3}\mathcal{Z}_0 + \bar{u}_2)] = -\frac{2}{3}\mathbb{E}\|\mathcal{Z}_0\|$ since $\mathbb{E}\bar{u}_2 = 0$. The quantitative bound is (b) with $r + \nu_2^{-1/3}K = \nu_1^{-1/3}K(m + (\nu_1/\nu_2)^{1/3})$; with $u = 1 - w$ the condition reads $m + (\nu_1/\nu_2)^{1/3} < (1 - u^{2/3})/(1-u)$, and the right side decreases on $(0,1)$ with range $(\frac{2}{3}, 1)$. □

The corollary is the counterpart, for the exact sign loss, of the comparison of Bastani (2021) between the two-step estimator and gold-only, with the same constant K on both sides and no logarithm, because the error of the proxy fit and the bias enter linearly through W_{TSE} and the gold noise is the argmin of the gold-only limit process plus an isotropic quadratic, which (a) controls pathwise. Turning the constants of the limit laws into risks uses Lemma 9; the lower bounds alone require only portmanteau. The same structure is not available for the two-headed estimator, whose gold noise is one component of a coupled argmin; and it does not order the two-step against re-weighting, whose constant $\mathbb{E}\|\rho r\Theta + \sigma_\rho^{2/3} Z_1\|$ has bounds of the same form.

C.7 Controlled comparisons: a design in which the constants can be ranked

Split the tangent space $\beta^{\star\perp}$ into orthogonal subspaces of dimensions $p_L \geq 1$ and $p_S \geq 1$, with orthonormal bases E_L, E_S, and let $\mathcal{V}_L = \mathrm{span}(\beta^\star, E_L)$, $\mathcal{V}_S = \mathrm{span}(\beta^\star, E_S)$. Comparisons are *controlled*: with probability π_L the feature difference is $X = X_0\beta^\star + E_L X_L$, and with probability $\pi_S = 1 - \pi_L$ it is $X = X_0\beta^\star + E_S X_S$, where (X_0, X_L) and (X_0, X_S) have continuous bounded densities with bounded support, positive near the origin, and the heterogeneity is symmetric and independent of X. Each comparison thus varies along one group of criteria, the contentious group (subscript L) or the consensus group (subscript S). The design replaces Assumption 2: it has no density on $\mathbb{R}^d$, but is supported on the two subspaces $\mathcal{V}_L, \mathcal{V}_S$ containing $\beta^\star$, with a continuous density on each, with $\pi_L, \pi_S > 0$. Define its estimators using the neutral half loss $[1 - (2y-1)\,\mathrm{sign}(x^\top\beta)]/2$ and a deterministic finite N^{-3}-net $\mathcal{B}_N$ of $\mathbb{S}^{d-1}$, fixed independently of the observations and the unknown centers. Use the same net for both heads, with deterministic tie-breaking. Finite minimization ensures existence despite the neutral loss's discontinuity. All controlled-design claims below concern these estimators; the original tie-rewarding loss need not identify the truth on this singular design. In the main-text normalization multiply both this loss and its penalty by two.

Since each comparison involves one group, the kernel (35) and the curvature split, $\mathcal{K}(s,t) = \mathcal{K}_L(s_L, t_L) + \mathcal{K}_S(s_S, t_S)$ and $H = \mathrm{diag}(H_L, H_S)$, where $\mathcal{K}_L, H_L$ and $\mathcal{K}_S, H_S$ are computed from the two comparison types, mixture weights included. Let B_L, B_S be independent centered Gaussian processes with these covariances, $Z_L = \arg\min_u\{B_L(u) + \frac{1}{2}u^\top H_L u\}$, Z_S likewise, $K_L = \mathbb{E}\|Z_L\|$ and $K_S = \mathbb{E}\|Z_S\|$, and let $\mathcal{Z}^L(\mu) = (\mathcal{Z}_1^L(\mu), \mathcal{Z}_2^L(\mu))$ be the coupled argmin of Theorem 18(a) built from the L field with penalty μ, and $\mathcal{Z}^S(\mu)$ likewise. Two constants describe the consensus group at a weight ρ: its smallest curvature $h_S = \lambda_{\min}(H_S)$ and the noise constant $\bar{c}_S = \nu_1^{1/3} \sup_{\mu\in[0,\infty]} \mathbb{E}\|\mathcal{Z}_1^S(\mu)\|/K_L$ of the gold head on it. For the uniform direction $\Theta = (\Theta_L, \Theta_S)$, $\mathbb{E}\|\Theta_L\|^2 = p_L/p$, so

$$\theta_L = \mathbb{E}\|\Theta_L\| \leq \sqrt{\frac{p_L}{p}}, \qquad \theta_S = \mathbb{E}\|\Theta_S\| \leq 1. \tag{39}$$

Measure the bias in units of the contentious gold noise, $\|\delta\| = mK_L N_1^{-1/3}$, note that $\nu_1^{1/3}\sigma_\rho^{2/3} = [(1-\rho)^2 + \rho^2\nu_1/\nu_2]^{1/3}$ is minimized at $\rho = \nu_2$, where it equals $\nu_1^{1/3}$, and let $\psi(m) \in (0,1)$ solve $\psi/m + \psi^{3/2} = 1$, so that $\psi(m) = \min_\rho \max\{\rho m, (1-\rho)^{2/3}\}$, $\psi(m) \geq (1 - 1/m)^{2/3}$ for $m \geq 1$ and $\psi(m) \geq m/2$ for $m \leq 1$.

Theorem 19 (Ranking of the exact constants for controlled comparisons)**.** *Let Assumption 1 hold with $\mathbb{E}\|\varepsilon\| < \infty$ for controlled comparisons, with the bias direction uniform on the unit sphere of $\beta^{\star\perp}$ and independent of the data.*

(a) For re-weighting at every weight and for gold-only,

$$\liminf_N \inf_{\rho'\in[0,1]} \frac{N_1^{1/3}\,\mathbb{E}\angle(\widehat{\beta}_{\mathrm{WE}}(\rho'), \beta^\star)}{K_L} \geq \min_{\rho'} \max\{\rho' m, \nu_1^{1/3}\sigma_{\rho'}^{2/3}\} \geq \psi(m),$$

$$1 \leq \lim_N \frac{N_1^{1/3}\,\mathbb{E}\angle(\widehat{\beta}_{\mathrm{gold}}, \beta^\star)}{K_L} \leq 1 + \frac{K_S}{K_L}.$$

(b) For every $\rho \in (0,1)$ and $\lambda \in (0,\infty)$, with $t = \rho(1-\rho)/(2\lambda)$,

$$\lim_N \frac{N_1^{1/3}\,\mathbb{E}\angle(\widehat{\beta}_{\mathrm{THE}}(\rho,\lambda), \beta^\star)}{K_L} \;\leq\; \rho m\theta_L + \frac{\nu_1^{1/3}\,\mathbb{E}\|\mathcal{Z}_1^L(\lambda)\|}{K_L} + \frac{\rho m\theta_S}{1 + t\,h_S} + \bar{c}_S,$$

where the second term tends to $\nu_1^{1/3}\sigma_\rho^{2/3}$ as $\lambda \to \infty$.

(c) For every $\epsilon > 0$ there is λ_0 such that, for $\lambda \geq \lambda_0$, if $h_S \geq 4\rho m\theta_S/(\epsilon t)$ and $\bar{c}_S \leq \epsilon/4$, the limit risk of THE is at most $K_L[\rho m\theta_L + \nu_1^{1/3}\sigma_\rho^{2/3} + \epsilon]$; if $\rho m\theta_L + \nu_1^{1/3}\sigma_\rho^{2/3} < \psi(m)$ and ϵ is below the difference, this is below the re-weighting lower bound of (a) at every weight. At the pooling weight $\rho = \nu_2$ the ratio of the two sides is at most $[\nu_2 m\theta_L + (N_1/N)^{1/3} + \epsilon]/\psi(m)$, which for $m \geq 1$, with $\nu_2 \leq 1$, $\theta_L \leq \sqrt{p_L/p}$ and $\theta_S \leq 1$, gives Proposition 1.

Proof. Limit laws in the design. The proofs of Theorems 15 and 18 apply with four modifications. (i) The kernel is computed within each subspace: parametrizing $\mathcal{V}_\mathsf{T}$ by (t', z), the computation of Lemma 8 gives the kernel of type T. (ii) The thin-slab bound $P\mathbf{1}\{|X^\top\beta| \leq \delta\} \leq C\delta$ and the Lipschitz bound of the disagreement probability in the angle fail for directions orthogonal to one subspace, but the proofs use them only for $\beta^{\star\top}\beta \geq \frac{1}{2}$, where both hold within each type with its own constants. (iii) For global identification, first set both centers equal to $\beta^\star$. Write $a = \beta^\top\beta^\star$ and $v_\mathsf{T} = E_\mathsf{T}^\top\beta$. If the projection on $\mathcal{V}_\mathsf{T}$ is nonzero, its normalized direction has angle α_T to $\beta^\star$ and the one-population argument gives excess half loss at least $c_\mathsf{T}\alpha_\mathsf{T}^2$ (the projected design (X_0, X_T) and the projected heterogeneity satisfy the one-population assumptions, and c_T is the half-loss constant). If that projection is zero, the neutral half loss is $1/2$, whereas at the truth it is $[1 - \mathbb{E}_\mathsf{T}|2\Pr(Y = 1 \mid X) - 1|]/2$. The resulting excess is strictly positive. Thus a zero projection cannot be treated as a normalized direction, and does not receive a reward. When $a > 0$, both projections are nonzero and $\sin\alpha_\mathsf{T} = \|v_\mathsf{T}\|/\sqrt{a^2 + \|v_\mathsf{T}\|^2} \geq \|v_\mathsf{T}\|$. Summing gives excess at least $c\sin^2\angle(\beta, \beta^\star)$, which is at least $4c\angle(\beta, \beta^\star)^2/\pi^2$ on this hemisphere. When $a \leq 0$, at least one projection is nonzero and its angle is at least $\pi/2$; its mixture weight gives a uniform positive gap. Adjusting the constant proves global quadratic growth on the whole sphere. For centers $\phi(\zeta_N) \to \beta^\star$, the conditional label means change uniformly by at most $R\|\phi(\zeta_N) - \beta^\star\|/2$, since the logistic function is $1/4$-Lipschitz. This transfers consistency to the triangular array by the uniform VC law. Locally, expand within each support; the gold Hessian is $\mathrm{diag}(H_L, H_S)$ and the proxy quadratic has shift ζ_N with an $o(N^{-2/3})$ remainder.

(iv) The finite nets have mesh $o(N^{-1/3})$. On each fixed rescaled compact set, rounding a direction to the net changes a margin only in a slab of width $O(N^{-3})$. The VC bound for these slabs makes the uniform empirical discrepancy $O_p(\sqrt{N^{-3}\log N/N} + \log N/N) = o_p(N^{-2/3})$; the penalty discrepancy is $O(N^{-3})$ for fixed penalty. The rate argument uses the deterministic comparison point and its net approximation as in Lemma 9. Thus the dense rescaled nets have the same local argmin limits. On the local cap, general position within each support also bounds sample ties by $p_L + p_S$ per head, so the local neutral/indicator discrepancy is $O(1/N)$. The last two steps of Theorem 15 use only the kernel. Hence the limit fields are $B_L(u_L) + B_S(u_S)$, the argmins split into independent blocks, and

$$N^{1/3}\widehat{v}_{\mathrm{WE}}(\rho') \Rightarrow \big(\rho' r\Theta_L + \sigma_{\rho'}^{2/3}Z_L,\ \rho' r\Theta_S + \sigma_{\rho'}^{2/3}Z_S\big),$$
$$N^{1/3}\widehat{v}_{\mathrm{THE}}(\rho, \lambda) \Rightarrow \big(\rho(I + tH_L)^{-1}r\Theta_L + \mathcal{Z}_1^L(\lambda),\ \rho(I + tH_S)^{-1}r\Theta_S + \mathcal{Z}_1^S(\lambda)\big).$$

Limit constants equal limit risks by Lemma 9; the lower bounds need only portmanteau. The main-text h_S and λ are both twice their values here, so the ratio th_S and the consensus threshold are unchanged by the normalization.

(a) This is Theorem 16 and Corollary 1: Jensen's inequality conditionally on Θ gives $\mathbb{E}\|\rho' r\Theta + \sigma^{2/3}Z\| \geq \rho' r$, and conditionally on the noise it gives $\geq \sigma_{\rho'}^{2/3}\mathbb{E}\|(Z_L, Z_S)\| \geq \sigma_{\rho'}^{2/3}K_L$; in units of $\nu_1^{-1/3}K_L$, with $r = m\nu_1^{-1/3}K_L$, this is $\max\{\rho' m, \nu_1^{1/3}\sigma_{\rho'}^{2/3}\}$, and $\nu_1^{1/3}\sigma_{\rho'}^{2/3} \geq (1 - \rho')^{2/3}$ gives $\psi(m)$. The exchange of $\inf_{\rho'}$ and $\liminf$ is that of Corollary 1. The gold-only bounds are $\|(u, v)\| \geq \|u\|$ and the triangle inequality.

(b) By the triangle inequality across and within blocks, $\|(I + tH_L)^{-1}\|_{\mathrm{op}} \leq 1$, $\|(I + tH_S)^{-1}\|_{\mathrm{op}} = (1 + th_S)^{-1}$ and $\nu_1^{1/3}\mathbb{E}\|\mathcal{Z}_1^S(\lambda)\|/K_L \leq \bar{c}_S$. The supremum in $\bar{c}_S$ is finite because the tail bound of Step 5 in the proof of Theorem 15 holds uniformly in the penalty: the penalty is nonnegative and the objective vanishes at the origin, so at a minimizer z the quantity $\frac{1}{2}\min\{\rho, 1 - \rho\}h_S\|z\|^2$ is at most the sum of the suprema of the two fields over the ball of radius $\|z\|$. For the limit of the second term, write $J_\lambda(x, y)$ for the coupled objective and J_∞ for its restriction to $x = y$, which is

$\sigma_1 B_L^{(1)}(x) + \sigma_2 B_L^{(2)}(x) + \frac{1}{2} x^\top H_L x$ with $\sigma_1 = (1-\rho)\nu_1^{-1/2}$, $\sigma_2 = \rho\nu_2^{-1/2}$. Since $J_\lambda \geq J_0$ and $J_\lambda(z_\lambda) \leq J_\lambda(0,0) = 0$ at a minimizer z_λ, all minimizers lie in $\{J_0 \leq 0\}$, which is almost surely bounded because the tail argument of Theorem 15 holds for every positive threshold constant; J_∞ has a unique minimizer x^* by Lemma 2.6 of Kim & Pollard (1990), since its Gaussian part has increment variance $(\sigma_1^2 + \sigma_2^2)\|x - x'\|_f > 0$. Let $\lambda_n \to \infty$. For $\lambda' \leq \lambda_n$, monotonicity in the penalty gives $J_{\lambda'}(z_{\lambda_n}) \leq J_{\lambda_n}(z_{\lambda_n}) \leq J_{\lambda_n}(x^*, x^*) = J_\infty(x^*, x^*)$; along any subsequence with $z_{\lambda_n} \to \bar{z}$, continuity of $J_{\lambda'}$ gives $J_{\lambda'}(\bar{z}) \leq J_\infty(x^*, x^*)$, and letting $\lambda' \to \infty$ forces $\bar{z}$ onto the diagonal with $J_\infty(\bar{z}) \leq J_\infty(x^*, x^*)$, so $\bar{z} = (x^*, x^*)$. Hence $z_{\lambda_n} \to (x^*, x^*)$ almost surely. As $\sigma_1 B_L^{(1)} + \sigma_2 B_L^{(2)}$ has the law of $\sigma_\rho B_L$, self-similarity (Step 5 of Theorem 15) gives $x^* \overset{d}{=} \sigma_\rho^{2/3} Z_L$, and the uniform tail bound makes the convergence hold in L^1: $\mathbb{E}\|\mathcal{Z}_1^L(\lambda_n)\| \to \sigma_\rho^{2/3} K_L$.

(c) Take λ_0 such that for $\lambda \geq \lambda_0$ the second term of (b) is within $\epsilon/2$ of its limit; under the two conditions the two S terms in (b) are at most $\epsilon/4$ each, so (b) is at most $\rho m \theta_L + \nu_1^{1/3}\sigma_\rho^{2/3} + \epsilon$, and (a) gives the rest, with $\nu_1^{1/3}\sigma_{\nu_2}^{2/3} = \nu_1^{1/3}$ at the pooling weight. □

The conditions of (c) hold, for instance, when the consensus features of a controlled design are stretched by a factor a and the consensus part of the heterogeneity is shrunk by $1/a$. At zero shift this preserves the label law, multiplies h_S by a^2, and divides the block argmin scale by a (the effective penalty changes from μ to μ/a^2, leaving its supremum unchanged). Hence $\bar{c}_S$ is divided by a, so both conditions hold for all large a. Re-weighting has bias $\rho'\|\delta\|$ at every weight, because its matrix weight is the same in every direction, and must keep its weight below $1/m$; the matrix weight (4) of the two-headed estimator is small in the consensus directions and stays close to the pooling weight in the contentious ones, so the bias of THE is only the share $\theta_L\|\delta\|$ and THE keeps the pooled noise $(N_1/N)^{1/3}$. The threshold λ_0 is not explicit: a rate for the convergence of the coupled argmin to the constrained one would need a small-ball estimate for near-ties of the minimum of a Gaussian process with quadratic drift, available only in one dimension. When $N_2 \gg N_1$ the bound of (c) at the pooling weight reduces to $m\theta_L + \epsilon$ against $\psi(m)$, so the ratio is at most $2\theta_L \max\{m, 1\}$ up to ϵ: of order $\max\{m, 1\}\sqrt{p_L/p}$.

Remark 2 (Why the constants for the exact sign loss are not compared analytically)**.** *Without smoothing the three estimators have errors $N^{-1/3}$ times exact constants of the same form, a bias $Wr\Theta$ with the matrix weights* (4) *plus a noise term whose law does not depend on r: $\sigma_\rho^{2/3} Z_1$ for re-weighting, the first component $\mathcal{Z}_1$ of the coupled argmin for the two-headed estimator, and $W_{\mathrm{TSE}}\nu_2^{-1/3} Z^{(2)}$ plus an argmin with the larger curvature $H + 2\lambda I$ for the two-step estimator. In Section 4 the noise is Gaussian and linear in the matrix weight, and the comparison is made by differentiating the risk in the penalty. Here the noise is the argmin of a Gaussian process with quadratic drift, which depends on the parameters non-smoothly (in one dimension the argmin jumps), and no analytic tool for ordering $\mathbb{E}\|\cdot\|$ of such functionals is available; even the moments of the one-dimensional Chernoff distribution are known only numerically. Outside controlled comparisons whose consensus group is curved and has low noise (Theorem 19), the constants for the exact sign loss can therefore be compared only by simulating the field* (35)*, which is Lévy's Brownian field for a radial design. For this reason Section 4 uses the smoothed sign loss.*

D Asymptotic risk of the smoothed estimators: complete proof

D.1 Scope, normalization, and notation

Scope. This appendix proves the first-order expansion of the risk of the three estimators at fixed hyperparameters, uniformly over compact sets of hyperparameters and of rescaled biases ζ. For WE, the first head of THE and TSE, in the regime $h_k = c_k n^{-1/5}$, $E^\top \delta_n / r_n \to \zeta$, we prove

$$\mathbb{E}\angle(\widehat{\beta}_{\mathcal{M}}, \beta^\star) = n^{-2/5}\,\mathbb{E}\|U_{\mathcal{M}}\| + o(n^{-2/5}), \tag{40}$$

with the $o(\cdot)$ uniform over ζ in a compact set, over $\Theta \in \mathbb{S}^{p-1}$, and over hyperparameters in compact sets. Optimization over the hyperparameters and the comparison between estimators are the subject of Appendix E. Standard results are quoted with references.

Normalization relative to Aouad et al. (2025). Their loss is $L_k(\beta) = -\mathbb{E}[(2Y-1)\operatorname{sign}(X^\top\beta)]$, as in Appendix B, whereas we use $L_{k,0}(\beta) = \frac{1}{2} - \frac{1}{2}\mathbb{E}[(2Y-1)\operatorname{sign}(X^\top\beta)]$, so that

$$L_{k,0}(\beta) - L_{k,0}(\beta_k) = \tfrac{1}{2}\big[L_k(\beta) - L_k(\beta_k)\big]. \tag{41}$$

Their central-mass condition $\inf_u \Pr(|u^\top\varepsilon| \le \|\beta_k\|/(2\sqrt{2})) > 0$ coincides with our assumption when $\|\beta_k\| = 1$, and the constant C_0 of their Theorem 3 (local quadratic growth for $\alpha \le \pi/4$) must be multiplied by $1/2$ in our normalization. All statements needed below are proved in a self-contained way; the original paper is invoked only in remarks, to supply explicit constants.

Notation. We use $\beta^\star, \delta, \rho, N_1, N_2, d, h_1, h_2, \widehat{\beta}$ as in the main text and λ in the normalization of Appendix A (half the main text's), and b always denotes the direction of the smoothing offset. Throughout $p = d-1$, so tangent coordinates live in $\mathbb{R}^p$ and directions of the bias in $\mathbb{S}^{p-1} = \mathbb{S}^{d-2}$. The symbol $\|\cdot\|$ is the Euclidean norm of a vector or the operator norm of a matrix, and $\angle(\beta,\beta') = \arccos(\beta^\top\beta') \in [0,\pi]$. With $\sigma(t) = 1/(1+e^{-t})$ we have $\sigma'(t) = \sigma(t)(1-\sigma(t)) = \frac{1}{4}\operatorname{sech}^2(t/2)$ and $2\sigma(t) - 1 = \tanh(t/2)$. The kernels are $\kappa_0(u) = \operatorname{sech}^2 u$, $\kappa_1(u) = \operatorname{sech}^2 u \tanh u$, $\kappa_2(u) = \operatorname{sech}^4 u$. A generic constant C depends only on the DGP constants $(d, R, \bar{f}, \underline{f}, r_0, \mathcal{L}(\varepsilon))$ and on the compact parameter sets, may change from line to line, and its dependence is made explicit whenever this matters. G_1, G_2 denote independent standard Gaussian vectors in $\mathbb{R}^p$, and $\mathcal{G}_k(t)$ denotes the tangential first moment of $m_k f_X$, defined in (51) and Lemma 22.

D.2 Setting

Data-generating process. $\|\beta^\star\| = \|\beta^\star + \delta\| = 1$, $\beta_1 = \beta^\star$, $\beta_2 = \beta^\star + \delta$. The design X has a Lebesgue density f_X with $f_X \le \bar{f}$, $\|X\| \le R$ almost surely, and $f_X \ge \underline{f} > 0$ on $B(0, r_0)$. The noise $\varepsilon \overset{d}{=} -\varepsilon$ is independent of X and has the same law in both sources. Labels follow $\Pr(Y = 1 \mid X, \varepsilon) = \sigma(X^\top(\beta_k + \varepsilon))$. The additional conditions for the smoothed sign loss are $f_X \in C_c^2(\mathbb{R}^d)$ and $\mathbb{E}\|\varepsilon\|^2 < \infty$. Write

$$m_k(x) = \mathbb{E}[2Y - 1 \mid X = x] = \mathbb{E}_\varepsilon \tanh\big(x^\top(\beta_k + \varepsilon)/2\big). \tag{42}$$

Since $\|X\| \le R$ and f_X is continuous, $\operatorname{supp} f_X \subseteq \bar{B}(0, R)$.

Loss and estimators.

$$\begin{aligned} \ell_h(\beta; x, y) &= \frac{1 - (2y-1)\tanh(x^\top\beta/h)}{2}, \qquad L_{k,h}(\beta) = \mathbb{E}_k\, \ell_h(\beta; X, Y), \\ \widehat{L}_{k,h}(\beta) &= \frac{1}{N_k}\sum_{j=1}^{N_k} \ell_h(\beta; X_{k,j}, Y_{k,j}), \end{aligned} \tag{43}$$

$$L_{k,h}(\beta) = \tfrac{1}{2} - \tfrac{1}{2}\,\mathbb{E}\big[m_k(X)\tanh(X^\top\beta/h)\big], \qquad L_{k,0}(\beta) = \tfrac{1}{2} - \tfrac{1}{2}\,\mathbb{E}\big[m_k(X)\operatorname{sign}(X^\top\beta)\big]. \tag{44}$$

The three estimators are (there is no factor $1/2$ in front of the penalties)

$$\widehat{\beta}_{\mathrm{WE}} \in \operatorname*{argmin}_{\|\beta\|=1} \big\{(1-\rho)\widehat{L}_{1,h_1}(\beta) + \rho\widehat{L}_{2,h_2}(\beta)\big\}, \tag{45}$$

$$(\widehat{\beta}_1, \widehat{\beta}_2) \in \operatorname*{argmin}_{\|\gamma_1\|=\|\gamma_2\|=1} \big\{(1-\rho)\widehat{L}_{1,h_1}(\gamma_1) + \rho\widehat{L}_{2,h_2}(\gamma_2) + \lambda\|\gamma_1 - \gamma_2\|^2\big\}, \tag{46}$$

$$\widehat{\beta}_{\mathrm{proxy}} \in \operatorname*{argmin}_{\|\beta\|=1} \widehat{L}_{2,h_2}(\beta), \qquad \widehat{\beta}_{\mathrm{TSE}} \in \operatorname*{argmin}_{\|\beta\|=1} \big\{\widehat{L}_{1,h_1}(\beta) + \rho_{\mathrm{TSE}}\|\beta - \widehat{\beta}_{\mathrm{proxy}}\|^2\big\}. \tag{47}$$

THE outputs, and is evaluated through, its first head $\widehat{\beta}_1$ only. When a minimizer is not unique, a measurable selection rule fixed in advance and independent of the truth is applied.

Asymptotic framework. $n = N_1 + N_2$, $N_k = N_k(n)$, $N_k/n \to \nu_k > 0$; $r_n = n^{-2/5}$; $h_k = c_k n^{-1/5}$. The bias varies with n: $\delta = \delta_n$ and $E^\top\delta_n/r_n = \zeta_n \in \mathcal{Z}$ with $\mathcal{Z} \subset \mathbb{R}^p$ compact. By the identity for the two unit centers below, $\beta_2 = \phi(r_n\zeta_n)$, so δ_n is uniquely determined by ζ_n and

$\|\delta_n\| = O(r_n)$ uniformly over $\mathcal{Z}$.[1] The main-text choice $\zeta = r\Theta$, $\Theta \in \mathbb{S}^{p-1}$, is the special case $\mathcal{Z} = r\,\mathbb{S}^{p-1}$.

The hyperparameter compacts are: WE, $\rho \in [0,1]$, $c_k \in [\underline{c}, \overline{c}] \subset (0,\infty)$; THE, $\rho \in [\underline{\rho}, \overline{\rho}] \subset (0,1)$, $\lambda \in [\underline{\lambda}, \overline{\lambda}] \subset (0,\infty)$, $c_k \in [\underline{c}, \overline{c}]$; TSE, $\rho_{\mathrm{TSE}} \in [0, \overline{\rho}_{\mathrm{TSE}}]$, $c_k \in [\underline{c}, \overline{c}]$. We write θ for the hyperparameter vector and $\mathcal{K}_{\mathcal{M}}$ for its compact set.

Frame and chart. Fix $E \in \mathbb{R}^{d\times p}$ with $E^\top \beta^\star = 0$ and $E^\top E = I_p$, and set

$$\phi(v) = \sqrt{1 - \|v\|^2}\,\beta^\star + Ev, \qquad \|v\| < 1. \tag{48}$$

For β_2, let Q be the rotation in the plane $\mathrm{span}\{\beta^\star, \beta_2\}$ taking $\beta^\star$ to β_2 (the identity on the orthogonal complement), $E_2 := QE$, $\phi_2(v) = \sqrt{1-\|v\|^2}\,\beta_2 + E_2 v$; also $E_1 := E$, $\phi_1 := \phi$. Then $\|Q - I\| = 2\sin(\angle(\beta_2, \beta^\star)/2) = \|\delta_n\|$. For source k write $x = t\beta_k + E_k z$ ($t \in \mathbb{R}$, $z \in \mathbb{R}^p$; orthonormal coordinates with unit Jacobian) and

$$\mu_k(t,z) := m_k(t\beta_k + E_k z), \qquad f_k(t,z) := f_X(t\beta_k + E_k z), \qquad g_k(t,z) := |\mu_k(t,z)|\, f_k(t,z). \tag{49}$$

Projected density constants. $\bar{f}_1 := \bar{f}\,\mathrm{Vol}(B_p(R))$ bounds the density of any one-dimensional projection $X^\top\beta$ ($\|\beta\| = 1$), and $\bar{f}_2 := \bar{f}\,\mathrm{Vol}(B_{p-1}(R))$ bounds the density of any two-dimensional orthogonal projection (with $\mathrm{Vol}(B_0) = 1$ when $d = 2$). Indeed, a projected density is the integral of f_X over the orthogonal complement, whose intersection with $\bar{B}(0,R)$ has volume at most that of the ball of the corresponding dimension.

Coordinates of the two unit centers. From $\beta_2 = \beta^\star + \delta$ and $\|\beta_2\| = \|\beta^\star\| = 1$ we get $\beta^{\star\top}\delta = -\|\delta\|^2/2$, hence $\beta_2 = (1 - \|\delta\|^2/2)\beta^\star + EE^\top\delta$ and $\sqrt{1 - \|E^\top\delta\|^2} = \sqrt{1 - \|\delta\|^2 + \|\delta\|^4/4} = 1 - \|\delta\|^2/2$. Therefore $\beta_2 = \phi(E^\top\delta)$: the chart coordinate of β_2 in the $\beta^\star$-chart is exactly $E^\top\delta$.

The DGP quantities and the random limit.

$$H = 2\int zz^\top\, \mathbb{E}_\varepsilon \sigma'(z^\top E^\top \varepsilon)\, f_X(Ez)\, dz, \qquad \Sigma = \frac{1}{3}\int zz^\top f_X(Ez)\, dz, \qquad V = H^{-1}\Sigma H^{-1}, \tag{50}$$

$$\mathcal{G}_1(t) := \int z\, m_1(t\beta^\star + Ez)\, f_X(t\beta^\star + Ez)\, dz, \qquad b = \frac{\pi^2}{24} H^{-1}\mathcal{G}_1''(0). \tag{51}$$

The source-k versions $H_k, \Sigma_k, \mathcal{G}_k, b_k$ replace $(\beta^\star, E)$ by (β_k, E_k), so $H_1 = H$ and so on. With G_1, G_2 independent standard Gaussian vectors in $\mathbb{R}^p$,

$$U_1 = c_1^2 b + \frac{V^{1/2}G_1}{\sqrt{\nu_1 c_1}}, \qquad U_2 = \zeta + c_2^2 b + \frac{V^{1/2}G_2}{\sqrt{\nu_2 c_2}}, \qquad U_{\mathcal{M}} = (I - W_{\mathcal{M}})U_1 + W_{\mathcal{M}}U_2, \tag{52}$$

$$W_{\mathrm{WE}} = \rho I, \qquad W_{\mathrm{THE}} = \rho\Big[I + \frac{\rho(1-\rho)}{2\lambda}H\Big]^{-1}, \qquad W_{\mathrm{TSE}} = (H + 2\rho_{\mathrm{TSE}} I)^{-1} 2\rho_{\mathrm{TSE}}. \tag{53}$$

Each of these formulas is derived below (Lemmas 17, 22, 30 and 33–35).

D.3 Preliminary lemmas

Lemma 10 (Kernel integrals)**.** *With all integrals over $\mathbb{R}$ unless indicated otherwise,*

$$\begin{aligned} &\int \kappa_0 = 2, \quad \int u\kappa_0 = 0, \quad \int u^2\kappa_0 = \frac{\pi^2}{6}, \\ &\int u\kappa_1 = 1, \quad \int \kappa_2 = \frac{4}{3}, \quad \int_0^\infty (1 - \tanh u)\, du = \log 2, \end{aligned} \tag{54}$$

and moreover $|\kappa_1| \le \kappa_0$, $\kappa_2 \le \kappa_0$, $\int |u|^3\kappa_0 < \infty$ and $\kappa_0(u) \le 4e^{-2|u|}$.

[1] Since $\|\delta_n\|^2 = \|E^\top\delta_n\|^2 + \|\delta_n\|^4/4$, one has $\|\delta_n\| = \|E^\top\delta_n\|/\sqrt{1 - \|\delta_n\|^2/4}$, so $\|\delta_n\| \le \frac{2}{\sqrt{3}}\, r_n \max_{\mathcal{Z}} \|z\|$ once $\|\delta_n\| \le 1$. Only $\|\delta_n\| = O(r_n)$ is used below.

Proof. $\int \operatorname{sech}^2 = [\tanh]_{-\infty}^{\infty} = 2$, and odd integrands integrate to zero. The value $\int_0^\infty u^2 \operatorname{sech}^2 u\, du = \pi^2/12$ is a standard integral: integrating by parts against $\frac{d}{du}(\tanh u - 1) = \operatorname{sech}^2 u$ turns it into $\int_0^\infty 2u(1-\tanh u)\, du$, and since $1 - \tanh u = 2\sum_{m\geq 1}(-1)^{m-1}e^{-2mu}$ for $u > 0$,

$$\int_0^\infty 2u(1-\tanh u)\, du = 2\sum_{m\geq 1}(-1)^{m-1}\int_0^\infty 2u\, e^{-2mu}\, du = \sum_{m\geq 1}(-1)^{m-1}m^{-2} = \frac{\pi^2}{12}. \quad (55)$$

From $\frac{d}{du}(-\frac{1}{2}\operatorname{sech}^2 u) = \kappa_1(u)$, integration by parts gives $\int u\kappa_1 = [-\frac{u}{2}\operatorname{sech}^2 u]_{-\infty}^{\infty} + \frac{1}{2}\int \operatorname{sech}^2 = 1$. Next, $\int \operatorname{sech}^4 = \int(1-\tanh^2)\operatorname{sech}^2 = [\tanh - \tanh^3/3]_{-\infty}^{\infty} = 4/3$. Finally $\int_0^\infty (1-\tanh u)\, du = \int_0^\infty \frac{2e^{-2u}}{1+e^{-2u}}\, du = [-\log(1+e^{-2u})]_0^\infty = \log 2$. The remaining claims are immediate. □

Lemma 11 (Kernel expansions)**.** *Let ω be a nondecreasing modulus with $\omega(0^+) = 0$.*

(K1) If $\psi \in C^2(\mathbb{R};\mathbb{R}^m)$, $\psi(0) = 0$, and ψ'' is bounded with modulus of continuity ω, then

$$\int \psi(hu)\kappa_0(u)\, du = \frac{\pi^2 h^2}{12}\psi''(0) + \rho_1(h),$$
$$\|\rho_1(h)\| \leq \frac{h^2}{2}\int u^2\omega(h|u|)\kappa_0(u)\, du = o(h^2). \quad (56)$$

(K2) If $\psi \in C^1(\mathbb{R};\mathbb{R}^m)$ and ψ' is bounded with modulus of continuity ω, then

$$\frac{1}{h}\int \psi(hu)\kappa_1(u)\, du = \psi'(0) + \rho_2(h), \qquad \|\rho_2(h)\| \leq \int |u|\,\omega(h|u|)\kappa_0(u)\, du = o(1). \quad (57)$$

(K3) If ψ is bounded and Lipschitz with constant Lip_ψ, then $\big|\int \psi(hu)\kappa_j(u)\, du - \psi(0)\int \kappa_j\big| \leq \mathrm{Lip}_\psi h \int |u|\kappa_0$ for $j = 0, 2$.

Proof. (K1): Taylor's formula gives $\psi(hu) = hu\,\psi'(0) + \frac{h^2u^2}{2}\psi''(\vartheta hu)$ with $\vartheta \in (0,1)$; the odd term integrates to zero against κ_0; $\|\psi''(\vartheta hu) - \psi''(0)\| \leq \omega(h|u|)$; then use $\int u^2\kappa_0 = \pi^2/6$ from Lemma 10 and dominated convergence (the integrand $u^2\omega(h|u|)\kappa_0$ is dominated by $2\|\psi''\|_\infty u^2\kappa_0$, which is integrable, and tends to zero pointwise). (K2): $\psi(hu) = \psi(0) + hu\,\psi'(\vartheta hu)$, $\int \psi(0)\kappa_1 = 0$, $\int u\kappa_1 = 1$ and $|\kappa_1| \leq \kappa_0$. (K3) is immediate. □

Lemma 12 (Regularity of m_k and the use of the moment conditions)**.** *Write $a_k(x,\varepsilon) = x^\top(\beta_k + \varepsilon)/2$.*

(a) Using symmetry only: $|m_k| \leq 1$; $m_k(-x) = -m_k(x)$; $m_k(x) = \frac{1}{2}\mathbb{E}[\tanh(a_k(x,\varepsilon)) + \tanh(a_k(x,-\varepsilon))]$; $m_k(x) > 0 \iff x^\top\beta_k > 0$, and $m_k = 0$ on $\beta_k^\perp$.

(b) If $\mathbb{E}\|\varepsilon\| < \infty$, then $m_k \in C^1(\mathbb{R}^d)$,

$$\nabla m_k(x) = \tfrac{1}{2}\mathbb{E}\big[\operatorname{sech}^2(a_k)(\beta_k + \varepsilon)\big], \qquad \|\nabla m_k\|_\infty \leq \tfrac{1}{2}(1 + \mathbb{E}\|\varepsilon\|) =: \Lambda_m, \quad (58)$$

∇m_k is uniformly continuous on $\mathbb{R}^d$ with a modulus ω_1 not depending on β_k; $|m_k(x)| \leq \Lambda_m|x^\top\beta_k|$; and

$$\beta_k^\top \nabla m_k(E_k z) = 2\,\mathbb{E}_\varepsilon \sigma'(z^\top E_k^\top \varepsilon) > 0. \quad (59)$$

For β_1, β_2: $\|m_2 - m_1\|_{L^\infty(\bar B(0,R))} \leq \frac{R}{2}\|\delta\|$ and $\|\nabla m_2 - \nabla m_1\|_{L^\infty(\bar B(0,R))} \leq C\|\delta\|$ (all integrals below are over the support $\bar B(0,R)$, so bounds on the ball suffice).

(c) If $\mathbb{E}\|\varepsilon\|^2 < \infty$, then $m_k \in C^2(\mathbb{R}^d)$,

$$\nabla^2 m_k(x) = -\tfrac{1}{2}\mathbb{E}\big[\operatorname{sech}^2(a_k)\tanh(a_k)(\beta_k + \varepsilon)(\beta_k + \varepsilon)^\top\big], \quad (60)$$

which is bounded and uniformly continuous (with a modulus ω_2 not depending on β_k), and $\|\nabla^2 m_2 - \nabla^2 m_1\|_{L^\infty(\bar B(0,R))} \leq C\|\delta\|$.

Proof. (a) The symmetrized representation follows from $\varepsilon \overset{d}{=} -\varepsilon$. If $t = x^\top \beta_k > 0$ then, for every ε and with $s = x^\top \varepsilon$, $\tanh((t+s)/2) + \tanh((t-s)/2) > 0$ because $\tanh$ is strictly increasing and odd: $\tanh((t+s)/2) > \tanh((-t+s)/2) = -\tanh((t-s)/2)$. Hence $m_k(x) > 0$; the case $t < 0$ is symmetric; for $t = 0$ the integrand is a sum of odd functions and vanishes.

(b) For fixed ε the gradient of $x \mapsto \tanh(a_k)$ is $\frac{1}{2}\operatorname{sech}^2(a_k)(\beta_k + \varepsilon)$, of norm at most $\frac{1}{2}\|\beta_k + \varepsilon\|$, which is integrable iff the first moment is finite; differentiability and the formula follow by dominated convergence. Uniform continuity: $|\operatorname{sech}^2(a) - \operatorname{sech}^2(a')| \le \min(1, |a - a'|)$ and $|a_k(x,\varepsilon) - a_k(x',\varepsilon)| \le \frac{1}{2}\|x - x'\|\|\beta_k + \varepsilon\|$, so

$$\|\nabla m_k(x) - \nabla m_k(x')\| \le \tfrac{1}{2}\mathbb{E}\big[\min\big(1, \tfrac{1}{2}\|x - x'\|\|\beta_k + \varepsilon\|\big)\|\beta_k + \varepsilon\|\big] =: \omega_1(\|x - x'\|), \quad (61)$$

where the right-hand side tends to zero by dominated convergence and depends only on $\mathcal{L}(\varepsilon)$ (since $\|\beta_k + \varepsilon\| \le 1 + \|\varepsilon\|$). The bound $|m_k(x)| \le \Lambda_m |x^\top \beta_k|$ follows from $m_k(x - (x^\top \beta_k)\beta_k) = 0$ and the mean value inequality. At $x = E_k z$ we have $a_k = z^\top E_k^\top \varepsilon/2$ and $\beta_k^\top \nabla m_k(E_k z) = \frac{1}{2}\mathbb{E}[\operatorname{sech}^2(a_k)(1 + \beta_k^\top \varepsilon)]$; $\operatorname{sech}^2(a_k)$ is an even function of ε and $\beta_k^\top \varepsilon$ is odd, their product is integrable (first moment), so its expectation vanishes; what remains is $\frac{1}{2}\mathbb{E}\operatorname{sech}^2(z^\top E_k^\top \varepsilon/2) = 2\mathbb{E}\sigma'(z^\top E_k^\top \varepsilon) > 0$. Differences: $|m_2(x) - m_1(x)| \le \frac{1}{2}|x^\top \delta|$, and $\nabla m_2 - \nabla m_1 = \frac{1}{2}\mathbb{E}[(\operatorname{sech}^2(a_2) - \operatorname{sech}^2(a_1))(\beta_2 + \varepsilon)] + \frac{1}{2}\mathbb{E}[\operatorname{sech}^2(a_1)]\delta$, where $|a_2 - a_1| = |x^\top \delta|/2 \le R\|\delta\|/2$ does not involve ε, so the norm is at most $\frac{1}{2} \cdot \frac{R}{2}\|\delta\|(1 + \mathbb{E}\|\varepsilon\|) + \frac{1}{2}\|\delta\|$.

(c) Likewise: the integrand of ∇^2 has norm at most $\frac{1}{2}\|\operatorname{sech}^2 \tanh\|_\infty \|\beta_k + \varepsilon\|^2$, integrable under the second moment; uniform continuity follows from the Lipschitz property of $\operatorname{sech}^2 \tanh$ and dominated convergence; the difference bound is as in (b), using again that $|a_2 - a_1| \le R\|\delta\|/2$ does not involve ε, so the second moment suffices. □

Lemma 13 (Lipschitz bound for the disagreement probability)**.** *For unit vectors β, β', $\Pr\{\operatorname{sign}(X^\top \beta) \ne \operatorname{sign}(X^\top \beta')\} \le \Gamma \angle(\beta, \beta')$ with $\Gamma := \bar{f}_2 R^2$.*

Proof. The case $\beta = \beta'$ is trivial. Otherwise the disagreement set is the product of a double wedge of opening angle α in the plane $P = \operatorname{span}\{\beta, \beta'\}$ with the orthogonal complement; the orthogonal projection of X onto P has density at most $\bar{f}_2$ and lies in the disc of radius R; the area of the double wedge inside that disc is αR^2. □

Lemma 14 (Chart geometry)**.** *For $\|v\|, \|v_1\|, \|v_2\| \le 1/2$:*

(a) $\angle(\phi(v), \beta^\star) = \arcsin\|v\|$, hence $\|v\| \le \angle(\phi(v), \beta^\star) \le \frac{\pi}{2}\|v\|$ and $|\arcsin\|v\| - \|v\|| \le \|v\|^3$.

(b) $D\phi(v) = E - \dfrac{\beta^\star v^\top}{\sqrt{1 - \|v\|^2}}$, $\phi(v)^\top D\phi(v) = 0$, $D\phi(v)^\top D\phi(v) = I + \dfrac{vv^\top}{1 - \|v\|^2}$, $\|D\phi(v) - E\| \le 2\|v\|$, $\|D\phi(v)\| \le 2$; componentwise $\nabla^2 \phi_i(v) = \beta_i^\star \nabla^2 \sqrt{1 - \|v\|^2}$ and $\sup_{\|v\| \le 1/2} \|\nabla^2 \phi_i(v)\| \le C_\phi$.

(c) $\|\phi(v_1) - \phi(v_2)\|^2 = \|v_1 - v_2\|^2 + \dfrac{(\|v_1\|^2 - \|v_2\|^2)^2}{(\sqrt{1 - \|v_1\|^2} + \sqrt{1 - \|v_2\|^2})^2}$, hence

$$0 \le \|\phi(v_1) - \phi(v_2)\|^2 - \|v_1 - v_2\|^2 \le \|v_1 - v_2\|^2 (\|v_1\| + \|v_2\|)^2. \quad (62)$$

(d) Let $P(v_1, v_2) = \lambda\|\phi(v_1) - \phi(v_2)\|^2$. Then

$$\nabla_{v_1} P = 2\lambda(v_1 - v_2) + \epsilon_1, \qquad \|\epsilon_1\| \le 2\lambda\|v_1\|(\|v_1\|^2 + \|v_2\|^2), \quad (63)$$

symmetrically for $\nabla_{v_2} P$, and

$$\nabla^2 P = 2\lambda \begin{pmatrix} I & -I \\ -I & I \end{pmatrix} + \Delta_P, \qquad \|\Delta_P\| \le C\lambda(\|v_1\|^2 + \|v_2\|^2). \quad (64)$$

Proof. (a) $\cos\alpha = \phi(v)^\top\beta^\star = \sqrt{1-\|v\|^2}$ and $\sin\alpha = \|v\|$; on $s \in [0,1/2]$, $\arcsin s \le \frac{\pi}{2}s$ and $\arcsin s - s \le s^3$, since $\arcsin s - s = \sum_{m\ge1} \frac{(2m)!}{4^m(m!)^2(2m+1)} s^{2m+1} \le s^3 \sum_{m\ge1} \frac{1}{2m+1}(1/4)^{m-1} \le s^3$. (b) Direct differentiation; $\|\beta^\star v^\top\|/\sqrt{1-\|v\|^2} \le 2\|v\|$. (c) Expand $\|\phi(v_1)-\phi(v_2)\|^2 = (\sqrt{1-\|v_1\|^2}-\sqrt{1-\|v_2\|^2})^2 + \|v_1-v_2\|^2$; the denominator is at least 1 (each root is at least $\sqrt{3}/2$), and $|\|v_1\|-\|v_2\|| \le \|v_1-v_2\|$. (d) $\nabla_{v_1} P = 2\lambda D\phi(v_1)^\top(\phi(v_1)-\phi(v_2)) = -2\lambda D\phi(v_1)^\top\phi(v_2) = 2\lambda\Big[v_1\sqrt{\frac{1-\|v_2\|^2}{1-\|v_1\|^2}} - v_2\Big]$, and

$$\Big|\sqrt{\tfrac{1-\|v_2\|^2}{1-\|v_1\|^2}} - 1\Big| = \frac{\big|\|v_1\|^2-\|v_2\|^2\big|}{\sqrt{1-\|v_1\|^2}\big(\sqrt{1-\|v_1\|^2}+\sqrt{1-\|v_2\|^2}\big)} \le \|v_1\|^2+\|v_2\|^2. \tag{65}$$

For the Hessian, $P = 2\lambda\big[1-\sqrt{(1-\|v_1\|^2)(1-\|v_2\|^2)} - v_1^\top v_2\big]$,

$$\nabla^2_{v_1v_1} P = 2\lambda\sqrt{1-\|v_2\|^2}\Big[\frac{I}{\sqrt{1-\|v_1\|^2}} + \frac{v_1v_1^\top}{(1-\|v_1\|^2)^{3/2}}\Big], \tag{66}$$

$$\nabla^2_{v_1v_2} P = -2\lambda I - 2\lambda\frac{v_1v_2^\top}{\sqrt{(1-\|v_1\|^2)(1-\|v_2\|^2)}}, \tag{67}$$

and each block differs from $2\lambda I$, respectively $-2\lambda I$, by $O(\lambda\|v\|^2)$. □

Lemma 15 (Uniform deviation via a net and Bernstein's inequality)**.** *Let $\{\psi_\beta : \beta \in \mathbb{S}^{d-1}\}$ be a family of real functions on the sample space with $\sup_\beta \|\psi_\beta\|_\infty \le B$, $\sup_\beta \operatorname{Var}\psi_\beta \le \varsigma^2$, and $|\psi_\beta - \psi_{\beta'}| \le \operatorname{Lip}\|\beta-\beta'\|$ pointwise. Let $\mathbb{P}_N$ be the empirical measure of N i.i.d. samples. Then for all $\epsilon \in (0,1]$ and $t > 0$,*

$$\Pr\Big(\sup_\beta |\mathbb{P}_N\psi_\beta - \mathbb{P}\psi_\beta| > 2\operatorname{Lip}\epsilon + t\Big) \le 2\Big(\frac{3}{\epsilon}\Big)^d \exp\Big(-\frac{Nt^2}{2\varsigma^2+2Bt/3}\Big). \tag{68}$$

Proof. Take an ϵ-net $\mathcal{N}$ of $\mathbb{S}^{d-1}$ with $|\mathcal{N}| \le (3/\epsilon)^d$. For each β choose $\beta' \in \mathcal{N}$ with $\|\beta-\beta'\| \le \epsilon$; then $|\mathbb{P}_N\psi_\beta - \mathbb{P}\psi_\beta| \le |\mathbb{P}_N\psi_{\beta'} - \mathbb{P}\psi_{\beta'}| + 2\operatorname{Lip}\epsilon$. Apply Bernstein's inequality at each net point and a union bound. The supremum is measurable because the family is continuous in β, so it may be taken over a countable dense subset. □

D.4 The population sign loss

Lemma 16 (Disagreement representation of the excess loss and identification)**.** *For every unit β,*

$$L_{k,0}(\beta) - L_{k,0}(\beta_k) = \mathbb{E}\big[|m_k(X)|\,\mathbf{1}\{\operatorname{sign}(X^\top\beta) \ne \operatorname{sign}(X^\top\beta_k)\}\big] \ge 0, \tag{69}$$

with equality iff $\beta = \beta_k$. Hence β_k is the unique minimizer of $L_{k,0}$ on $\mathbb{S}^{d-1}$.

Proof. Since X has a density, $\Pr(X^\top\beta = 0) = 0$, so the convention at zero margin is immaterial. We have $L_{k,0}(\beta) - L_{k,0}(\beta_k) = \frac{1}{2}\mathbb{E}[m_k(X)(\operatorname{sign}(X^\top\beta_k) - \operatorname{sign}(X^\top\beta))]$; by Lemma 12(a), $\operatorname{sign} m_k(X) = \operatorname{sign}(X^\top\beta_k)$ a.s., and the bracket equals $2\operatorname{sign}(X^\top\beta_k)$ on the disagreement set and 0 elsewhere, which gives the representation. If $\beta \ne \beta_k$, the disagreement set is a nonempty open double wedge in $\mathbb{R}^d$ (up to a null set), whose intersection with $B(0,r_0)$ has positive Lebesgue measure; there $f_X \ge \underline{f} > 0$ and $|m_k| > 0$ off the hyperplane $\beta_k^\perp$, so the expectation is strictly positive. □

Lemma 17 (Local second-order expansion and the integral formula for H_k)**.** *Assume $\mathbb{E}\|\varepsilon\| < \infty$ and that f_X is continuous (the rest are the common assumptions). Define*

$$a_k(z) := \beta_k^\top\nabla m_k(E_kz)\, f_X(E_kz) = 2\,\mathbb{E}_\varepsilon\sigma'(z^\top E_k^\top\varepsilon)\, f_X(E_kz), \qquad H_k := \int_{\mathbb{R}^p} zz^\top a_k(z)\,dz. \tag{70}$$

Then for $\|v\| \le 1/2$,

$$L_{k,0}(\phi_k(v)) - L_{k,0}(\beta_k) = \frac{v^\top H_k v}{2(1-\|v\|^2)} + \operatorname{Rem}_k(v), \qquad |\operatorname{Rem}_k(v)| \le C_{\mathrm{R}}\,\|v\|^2\,\bar\omega_0\big(2R\|v\|\big), \tag{71}$$

where $\bar{\omega}_0(s) := \Lambda_m \omega_f(s) + \bar{f}\,\omega_1(s)$, ω_f *is the modulus of continuity of* f_X, ω_1 *that of* ∇m_k *(Lemma 12(b)), and* $C_{\mathrm{R}} = 2R^2 \operatorname{Vol}(B_p(R))$. *In particular*

$$L_{k,0}(\phi_k(v)) - L_{k,0}(\beta_k) = \tfrac{1}{2} v^\top H_k v + o(\|v\|^2), \tag{72}$$

where the $o(\cdot)$ *is uniform in the direction of* v, *over* $k = 1, 2$ *and over* $\beta_2 \in \mathbb{S}^{d-1}$. *For* $k = 1$, $H_1 = H$ *is the matrix in* (50).

Proof. In the coordinates $x = t\beta_k + E_k z$ we have $x^\top \phi_k(v) = t\sqrt{1 - \|v\|^2} + z^\top v$ and $x^\top \beta_k = t$. The two have opposite signs iff t lies strictly between 0 and $-s_v(z) := -z^\top v / \sqrt{1 - \|v\|^2}$; call this open interval $I_v(z)$. By Lemma 16 and Fubini,

$$\Delta_k(v) := L_{k,0}(\phi_k(v)) - L_{k,0}(\beta_k) = \int_{\|z\| \le R} \int_{I_v(z)} g_k(t, z)\, dt\, dz \tag{73}$$

(the support of f_X lies in $\bar{B}(0, R)$, so $\|z\| \le R$ and $|t| \le R$). By Lemma 12(a)(b), $\mu_k(t, z) = \operatorname{sign}(t)|\mu_k(t, z)|$, $\mu_k(0, z) = 0$ and $\partial_t \mu_k(t, z) = \beta_k^\top \nabla m_k(t\beta_k + E_k z)$, so that

$$|\mu_k(t, z)| = |t| \big[\beta_k^\top \nabla m_k(E_k z) + \theta_1(t, z)\big], \qquad |\theta_1| \le \omega_1(|t|), \tag{74}$$

and $f_k(t, z) = f_X(E_k z) + \theta_f(t, z)$ with $|\theta_f| \le \omega_f(|t|)$. Hence $g_k(t, z) = |t|\, a_k(z) + |t|\, \theta_g(t, z)$ with $|\theta_g(t, z)| \le \Lambda_m \omega_f(|t|) + \bar{f}\omega_1(|t|) = \bar{\omega}_0(|t|)$ (using $0 < \beta_k^\top \nabla m_k \le \Lambda_m$ and $f_X \le \bar{f}$). Integrating for fixed z: $\int_{I_v(z)} |t| a_k(z)\, dt = \frac{1}{2} s_v(z)^2 a_k(z) = \dfrac{(z^\top v)^2 a_k(z)}{2(1 - \|v\|^2)}$, and the remainder is at most $\frac{1}{2} s_v(z)^2 \bar{\omega}_0(|s_v(z)|)$. Since $|s_v(z)| \le R\|v\| / \sqrt{1 - \|v\|^2} \le 2R\|v\|$,

$$|\mathrm{Rem}_k(v)| \le \int_{\|z\| \le R} \frac{(2R\|v\|)^2}{2} \bar{\omega}_0(2R\|v\|)\, dz = C_{\mathrm{R}} \|v\|^2 \bar{\omega}_0(2R\|v\|). \tag{75}$$

Integrating the main term over z gives $v^\top H_k v / (2(1 - \|v\|^2))$. The modulus $\bar{\omega}_0$ depends only on f_X and $\mathcal{L}(\varepsilon)$, not on β_k (Lemma 12(b)). The second expression for a_k is the last display of Lemma 12(b). □

Lemma 18 (Positive definiteness of H_k)**.** H_k *is symmetric positive definite, with*

$$\lambda_{\min}(H_k) \ge \underline{h} := \frac{2\,\underline{\sigma}'\,\underline{f}\, r_0^2 \operatorname{Vol}(B_p(r_0))}{p + 2}, \qquad \lambda_{\max}(H_k) \le \overline{h} := \tfrac{1}{2} R^2 \bar{f} \operatorname{Vol}(B_p(R)), \tag{76}$$

where $\underline{\sigma}' := \inf_{\|w\| \le R} \mathbb{E}\sigma'(w^\top \varepsilon) > 0$ *depends only on* $\mathcal{L}(\varepsilon)$. *If in addition the central-mass condition* $\pi_\varepsilon := \inf_{\|u\|=1} \Pr(|u^\top \varepsilon| \le \frac{1}{2\sqrt{2}}) > 0$ *is used, then* $\underline{\sigma}' \ge \sigma'(R/(2\sqrt{2}))\pi_\varepsilon$ *is explicit.*

Proof. $v^\top H_k v = \int (z^\top v)^2 a_k(z)\, dz$ with $a_k \ge 0$. The map $w \mapsto \mathbb{E}\sigma'(w^\top \varepsilon)$ is continuous (dominated convergence, $\sigma' \le 1/4$) and everywhere positive ($\sigma' > 0$), hence bounded below by $\underline{\sigma}' > 0$ on the compact set $\|w\| \le R$. On $\|z\| < r_0$ we have $a_k(z) \ge 2\underline{\sigma}'\underline{f}$, while $\int_{\|z\| < r_0} (z^\top v)^2\, dz = \|v\|^2 r_0^2 \operatorname{Vol}(B_p(r_0))/(p + 2)$. Upper bound: $a_k \le 2 \cdot \frac{1}{4}\bar{f}$ and $\int_{\|z\| \le R} (z^\top v)^2\, dz \le R^2 \operatorname{Vol}(B_p(R))\|v\|^2$. Explicit lower bound: on the event $\{|u^\top \varepsilon| \le 1/(2\sqrt{2})\}$, with $u = w/\|w\|$, we have $|w^\top \varepsilon| \le R/(2\sqrt{2})$, and σ' is even and decreasing on $[0, \infty)$. □

Lemma 19 (Global quadratic growth)**.** *There exist* $c_{\mathrm{q}} > 0$ *and* $\alpha_1 \in (0, \pi/6]$, *depending only on the DGP, such that for all unit* β *and* $k = 1, 2$,

$$L_{k,0}(\beta) - L_{k,0}(\beta_k) \ge c_{\mathrm{q}}\, \angle(\beta, \beta_k)^2. \tag{77}$$

Concretely, choose α_1 *such that* $C_{\mathrm{R}} \bar{\omega}_0(2R \sin \alpha_1) \le \underline{h}/4$;[2] *then one may take* $c_{\mathrm{q}} = \dfrac{\underline{h} \sin^2 \alpha_1}{4\pi^2}$.

[2] In addition, $\alpha_1 \le \pi/6$ is required so that $\|v\| = \sin\alpha \le 1/2$ in Step 1 and Lemma 17 applies; the curvature bound of Aouad et al. (2025) (their Theorem 3) requires only $\alpha \le \pi/4$. Shrinking α_1 is harmless because $\bar{\omega}_0$ is nondecreasing, so the displayed condition remains satisfied.

Proof. Step 1 (local). For $\alpha = \angle(\beta, \beta_k) \le \alpha_1$ write $\beta = \phi_k(v)$ with $\|v\| = \sin\alpha \le \sin\alpha_1$; by Lemmas 17 and 18,

$$\Delta_k(v) \ge \tfrac{1}{2}\underline{h}\|v\|^2 - \tfrac{1}{4}\underline{h}\|v\|^2 = \tfrac{1}{4}\underline{h}\sin^2\alpha \ge \tfrac{1}{4}\underline{h}\Big(\frac{2\alpha}{\pi}\Big)^2 \ge c_{\mathrm{q}}\alpha^2. \tag{78}$$

Step 2 (monotonicity along a plane). Fix a unit $\omega \in \mathbb{R}^p$ and let $\beta(\alpha) = \cos\alpha\,\beta_k + \sin\alpha\,E_k\omega$, $\alpha \in (0,\pi)$. In the coordinates $t = x^\top\beta_k$, $s = \omega^\top E_k^\top x$ we have $x^\top\beta(\alpha) = t\cos\alpha + s\sin\alpha$; for $t > 0$ disagreement occurs iff $s < -t\cot\alpha$, and symmetrically for $t < 0$. Since $-\cot\alpha$ is strictly increasing on $(0,\pi)$, the disagreement set $D(\alpha)$ is nondecreasing in α (up to null sets). By the representation of Lemma 16, $\alpha \mapsto L_{k,0}(\beta(\alpha)) - L_{k,0}(\beta_k)$ is nondecreasing. Every $\beta \ne \pm\beta_k$ is of the form $\beta(\alpha)$ (and $\beta = -\beta_k$ is the limit $\alpha = \pi$, where the disagreement set is the whole space up to a null set, so the claim persists). Step 3. If $\alpha \ge \alpha_1$, then $L_{k,0}(\beta) - L_{k,0}(\beta_k) \ge L_{k,0}(\beta(\alpha_1)) - L_{k,0}(\beta_k) \ge \tfrac{1}{4}\underline{h}\sin^2\alpha_1 \ge \tfrac{1}{4}\underline{h}\sin^2\alpha_1\cdot\alpha^2/\pi^2 = c_{\mathrm{q}}\alpha^2$. □

Remark 3 (Relation to Theorem 3 of Aouad et al. (2025))**.** *For $\alpha \le \pi/4$ the original paper gives $L_k(\beta) - L_k(\beta_k) \ge C_0\alpha^2$ for the loss L_k of Appendix B, with C_0 of order $\sigma'(r_0/\sqrt{2})\,\pi_\varepsilon\,\underline{f}_2\,r_0^3$ (their statement has r_0^2, from a slip in the last display of their proof, where the integral equals $b^3\tan^2\alpha/6$ with $b = r_0/(2\sqrt{2})$), where we take $r_0 \le 1$ (shrinking the ball on which $f_X \ge \underline{f}$ if necessary), π_ε is the central-mass probability of Lemma 18 and $\underline{f}_2$ is a lower bound on the density of the projection of X onto a two-dimensional plane over the region used in their proof (with their region of radius $r_0/2$ one may take $\underline{f}_2 = \underline{f}\,\mathrm{Vol}(B_{p-1}(\sqrt{3}r_0/2))$). In our normalization this gives the constant $C_0/2$ for $\alpha \le \pi/4$; combined with the monotonicity of Step 2 above it extends to the whole sphere with a further factor $1/16$. This is an explicit alternative for c_{q}, but it uses the central-mass condition; the main argument only needs the qualitative constant of Lemma 19.*

Lemma 20 (Lipschitz upper bound for the sign loss)**.** $0 \le L_{k,0}(\beta) - L_{k,0}(\beta_k) \le \Gamma\,\angle(\beta,\beta_k)$ *with* $\Gamma = \bar{f}_2 R^2$.

Proof. Use $|m_k| \le 1$ in the representation of Lemma 16 and then Lemma 13. □

D.5 The smoothed population loss

Lemma 21 (Global distance between smoothed and sign loss)**.** *For all $h > 0$, k and unit β: $|L_{k,h}(\beta) - L_{k,0}(\beta)| \le \bar{f}_1 \log 2 \cdot h$.*

Proof. $|L_{k,h}(\beta) - L_{k,0}(\beta)| = \tfrac{1}{2}|\mathbb{E}[m_k(X)(\tanh(X^\top\beta/h) - \mathrm{sign}(X^\top\beta))]| \le \tfrac{1}{2}\mathbb{E}[1 - \tanh(|X^\top\beta|/h)] \le \tfrac{1}{2}\bar{f}_1\int_{\mathbb{R}}(1-\tanh(|t|/h))\,dt = \bar{f}_1 h\int_0^\infty(1-\tanh u)\,du = \bar{f}_1 h\log 2$ by Lemma 10. □

Lemma 22 (Tangential gradient expansion and the formula for b_k)**.** *Assume $f_X \in C_c^2$ and $\mathbb{E}\|\varepsilon\|^2 < \infty$. Let $\mathcal{G}_k(t) := \int_{\mathbb{R}^p} z\,\mu_k(t,z) f_k(t,z)\,dz$. Then $\mathcal{G}_k \in C_c^2(\mathbb{R};\mathbb{R}^p)$, $\mathcal{G}_k(0) = 0$,*

$$\mathcal{G}_k''(t) = \int z\,\beta_k^\top\nabla^2(m_k f_X)(t\beta_k + E_k z)\beta_k\,dz, \tag{79}$$

and

$$\nabla_v\big(L_{k,h}\circ\phi_k\big)(0) = -\tfrac{1}{2}\int \mathcal{G}_k(hu)\kappa_0(u)\,du = -\frac{\pi^2 h^2}{24}\mathcal{G}_k''(0) + \rho_b(h) = -h^2 H_k b_k + \rho_b(h), \tag{80}$$

$$b_k := \frac{\pi^2}{24}H_k^{-1}\mathcal{G}_k''(0), \qquad \|\rho_b(h)\| \le \frac{h^2}{4}R\,\mathrm{Vol}(B_p(R))\int u^2\,\bar{\omega}_2(h|u|)\kappa_0(u)\,du = o(h^2), \tag{81}$$

where $\bar{\omega}_2$ is the modulus of continuity of $\nabla^2(m_k f_X)$, which does not depend on β_k. For $k = 1$, $b_1 = b$ is the vector in (51)*.*

Proof. $\nabla_\beta L_{k,h}(\beta) = -\frac{1}{2h}\mathbb{E}[m_k(X)\kappa_0(X^\top\beta/h)X]$ (for $\|x\| \le R$ and fixed $h > 0$ the map $\beta \mapsto \tanh(x^\top\beta/h)$ is smooth with bounded derivative; dominated convergence). Since $D\phi_k(0) = E_k$, $\nabla_v(L_{k,h}\circ\phi_k)(0) = E_k^\top\nabla_\beta L_{k,h}(\beta_k) = -\frac{1}{2h}\iint \mu_k(t,z)f_k(t,z)\kappa_0(t/h)z\,dz\,dt =$

$-\frac{1}{2h}\int \mathcal{G}_k(t)\kappa_0(t/h)\,dt$, and the substitution $t = hu$ gives the first equality. $\mathcal{G}_k(0) = 0$ because $\mu_k(0,z) = m_k(E_k z) = 0$ (Lemma 12(a)). By Lemma 12(c) and $f_X \in C_c^2$, $m_k f_X \in C_c^2(\mathbb{R}^d)$; the integrand $z\,(m_k f_X)(t\beta_k + E_k z)$ is twice continuously differentiable in t, supported in $\|z\| \le R$, with bounded derivatives, so one may differentiate twice under the integral sign, which gives the formula for $\mathcal{G}_k''$; the modulus of continuity of $\mathcal{G}_k''$ is at most $R\,\mathrm{Vol}(B_p(R))\bar{\omega}_2$. Apply Lemma 11(K1) with $\psi = \mathcal{G}_k$, $\psi(0) = 0$: $\int \mathcal{G}_k(hu)\kappa_0 = \frac{\pi^2 h^2}{12}\mathcal{G}_k''(0) + \rho_1(h)$ with $\|\rho_1\| \le \frac{h^2}{2} R\,\mathrm{Vol}(B_p(R))\int u^2 \bar{\omega}_2(h|u|)\kappa_0$; multiply by $-\frac{1}{2}$. H_k is invertible by Lemma 18. □

Lemma 23 (The Hessian: formula, limit $\mathcal{H}_k(\beta)$, uniform convergence and a gradient bound)**.** *Assume $\mathbb{E}\|\varepsilon\| < \infty$ and $f_X \in C_c^1$. For unit β let $\Pi_\beta = I - \beta\beta^\top$ and*

$$\mathcal{H}_k(\beta) := \int_{\beta^\perp} w w^\top \beta^\top \nabla(m_k f_X)(w)\,dw \qquad (\textit{a symmetric } d\times d \textit{ matrix}). \tag{82}$$

(a) $\nabla_\beta^2 L_{k,h}(\beta) = \dfrac{1}{h^2}\mathbb{E}\big[m_k(X)\kappa_1(X^\top\beta/h)XX^\top\big]$*, and*

$$\begin{aligned} &\sup_{\beta\in\mathbb{S}^{d-1},\,k} \big\|\Pi_\beta \nabla_\beta^2 L_{k,h}(\beta)\Pi_\beta - \mathcal{H}_k(\beta)\big\| \le \omega_H(h),\\ &\omega_H(h) := R^2\,\mathrm{Vol}(B_p(R))\int |u|\,\bar{\omega}_1(h|u|)\kappa_0(u)\,du \xrightarrow[h\to 0]{} 0, \end{aligned} \tag{83}$$

where $\bar{\omega}_1$ is the modulus of continuity of $\nabla(m_k f_X)$, which does not depend on β_k.

(b) $\mathcal{H}_k(\beta_k) = E_k H_k E_k^\top$*;* $\|\mathcal{H}_k(\beta) - \mathcal{H}_k(\beta')\| \le \omega_{\mathcal{H}}(\|\beta-\beta'\|) := C\big(\|\beta-\beta'\| + \bar{\omega}_1(R\|\beta-\beta'\|)\big)$*;* $\sup_\beta \|\mathcal{H}_2(\beta) - \mathcal{H}_1(\beta)\| \le C\|\delta\|$*. Under $f_X \in C^2$ and $\mathbb{E}\|\varepsilon\|^2 < \infty$ one has $\bar{\omega}_1(s) \le Cs$, hence $\omega_{\mathcal{H}}(s) \le Cs$.*

(c) $\|\nabla_\beta L_{k,h}(\beta)\| \le C_{\mathrm{g}}\big(h + \|\beta - \beta_k\|\big)$ *with* $C_{\mathrm{g}} = \Lambda_m R \bar{f}_1 \max(\int |u|\kappa_0,\, 2R)$.

(d) (Uniform convergence of the tangential Hessian in the tangent coordinates at $\beta^\star$.) For $\eta \le 1/2$, $\|v\| \le \eta$ and $k = 1, 2$,

$$\begin{aligned} &\big\|\nabla_v^2\big(L_{k,h}\circ\phi\big)(v) - H\big\| \le \omega_{\mathrm{tot}}(h,\eta,\|\delta\|),\\ &\omega_{\mathrm{tot}}(h,\eta,\|\delta\|) := 4\omega_H(h) + \omega_{\mathcal{H}}(2\eta) + C\eta + C\|\delta\| + C(h + 2\eta + \|\delta\|), \end{aligned} \tag{84}$$

and in its own chart likewise $\|\nabla_v^2(L_{k,h}\circ\phi_k)(v) - H_k\| \le \omega_{\mathrm{tot}}(h,\eta,0)$. The curvatures of the two sources converge at rate $O(\|\delta\|)$.

Proof. (a) From $\nabla_\beta^2 \tanh(x^\top\beta/h) = h^{-2}\kappa_0'(x^\top\beta/h)xx^\top = -2h^{-2}\kappa_1(x^\top\beta/h)xx^\top$ and $L_{k,h} = \frac{1}{2} - \frac{1}{2}\mathbb{E}[m_k \tanh]$ we get the formula (dominated convergence). Write $x = \tau\beta + w$ with $w \in \beta^\perp$, so $\Pi_\beta x = w$:

$$\Pi_\beta \nabla_\beta^2 L_{k,h}(\beta)\Pi_\beta = \int_{\beta^\perp} w w^\top \Big[\frac{1}{h^2}\int_{\mathbb{R}} \psi_w(\tau)\kappa_1(\tau/h)\,d\tau\Big]dw, \qquad \psi_w(\tau) := (m_k f_X)(\tau\beta + w). \tag{85}$$

Here $\psi_w \in C^1$, $\psi_w'(\tau) = \beta^\top\nabla(m_k f_X)(\tau\beta + w)$ with modulus at most $\bar{\omega}_1$; after $\tau = hu$, $\frac{1}{h^2}\int \psi_w(\tau)\kappa_1(\tau/h)\,d\tau = \frac{1}{h}\int \psi_w(hu)\kappa_1(u)\,du$, which equals $\psi_w'(0) + \rho_2$ with $|\rho_2| \le \int |u|\bar{\omega}_1(h|u|)\kappa_0$ by Lemma 11(K2). Integrate over $\|w\| \le R$ (with $\|ww^\top\| \le R^2$). Finally $\nabla(m_k f_X) = f_X\nabla m_k + m_k\nabla f_X$ is uniformly continuous with a modulus not depending on β_k: $f_X \in C_c^1$ is uniformly continuous and bounded; ∇m_k is uniformly continuous (Lemma 12(b)) and m_k is Λ_m-Lipschitz.

(b) On $\beta_k^\perp$, $m_k = 0$, so $\beta_k^\top\nabla(m_k f_X)(w) = f_X(w)\beta_k^\top\nabla m_k(w)$; with $w = E_k z$ this is $a_k(z)$ of Lemma 17, whence $E_k H_k E_k^\top$. Continuity: let $\beta' = Q'\beta$ with Q' a plane rotation, $\|Q' - I\| = \|\beta - \beta'\|$; substituting $w = Q'\tilde{w}$,

$$\mathcal{H}_k(\beta') = Q'\Big[\int_{\beta^\perp} \tilde{w}\tilde{w}^\top (Q'\beta)^\top\nabla(m_k f_X)(Q'\tilde{w})\,d\tilde{w}\Big]Q'^\top, \tag{86}$$

and comparing term by term with $\mathcal{H}_k(\beta)$: the $\|Q'-I\|$ terms contribute $C\|\beta-\beta'\|$ and the shift of the argument, $\|Q'\tilde{w}-\tilde{w}\| \le R\|\beta-\beta'\|$, contributes $C\bar{\omega}_1(R\|\beta-\beta'\|)$. Source difference (only on the support $\bar{B}(0,R)$): $\|\nabla(m_2 f_X)-\nabla(m_1 f_X)\|_\infty \le \bar{f}\|\nabla m_2-\nabla m_1\|_{L^\infty(\bar{B}(0,R))} + \|\nabla f_X\|_\infty \|m_2-m_1\|_{L^\infty(\bar{B}(0,R))} \le C\|\delta\|$ (Lemma 12(b)). In the C^2 case $\nabla(m_k f_X)$ is Lipschitz.

(c) $|m_k(x)| \le \Lambda_m |x^\top \beta_k| \le \Lambda_m(|x^\top\beta| + R\|\beta-\beta_k\|)$, so

$$\begin{aligned}\|\nabla_\beta L_{k,h}(\beta)\| &\le \frac{\Lambda_m R}{2h}\mathbb{E}\big[(|X^\top\beta| + R\|\beta-\beta_k\|)\kappa_0(X^\top\beta/h)\big] \\ &\le \frac{\Lambda_m R \bar{f}_1}{2h}\Big[h^2\int |u|\kappa_0 + 2hR\|\beta-\beta_k\|\Big].\end{aligned} \tag{87}$$

(d) Chain rule: $\nabla_v^2(L_{k,h}\circ\phi)(v) = D\phi(v)^\top \nabla_\beta^2 L_{k,h}(\phi(v))D\phi(v) + \sum_i \partial_i L_{k,h}(\phi(v))\nabla^2\phi_i(v)$. Since $\phi(v)^\top D\phi(v) = 0$ (Lemma 14(b)), $D\phi(v) = \Pi_{\phi(v)}D\phi(v)$, so the first term equals $D\phi(v)^\top \Pi\nabla^2 L \Pi D\phi(v)$ and differs from $D\phi(v)^\top\mathcal{H}_k(\phi(v))D\phi(v)$ by at most $\|D\phi\|^2\omega_H(h) \le 4\omega_H(h)$ by (a). Next,

$$\begin{aligned}&\|D\phi(v)^\top\mathcal{H}_k(\phi(v))D\phi(v) - E^\top\mathcal{H}_1(\beta^\star)E\| \\ &\quad\le \|D\phi(v)-E\|(\|D\phi(v)\|+1)\|\mathcal{H}_k\| + \|\mathcal{H}_k(\phi(v))-\mathcal{H}_k(\beta^\star)\| + \|\mathcal{H}_k(\beta^\star)-\mathcal{H}_1(\beta^\star)\| \\ &\quad\le C\eta + \omega_{\mathcal{H}}(2\eta) + C\|\delta\|\end{aligned} \tag{88}$$

(using $\|\phi(v)-\beta^\star\| \le 2\|v\|$ and (b)), while $E^\top\mathcal{H}_1(\beta^\star)E = H$. The second term of the chain rule is at most $\sqrt{d}\,C_\phi\|\nabla_\beta L_{k,h}(\phi(v))\| \le C(h+2\eta+\|\delta\|)$ by (c) and $\|\phi(v)-\beta_k\| \le 2\|v\|+\|\delta\|$. The version in the chart of source k replaces $(\beta^\star, E, \delta)$ by $(\beta_k, E_k, 0)$. □

Lemma 24 (Smoothed population minimizer: existence, location, uniqueness, $O(h^2)$ bias)**.** *There is $h_0 > 0$, depending only on the DGP, such that for $0 < h \le h_0$ and $k = 1, 2$:*

- *(a) $L_{k,h}$ has a minimizer on $\mathbb{S}^{d-1}$, and every minimizer $\beta_{k,h}$ satisfies $\angle(\beta_{k,h},\beta_k) \le \eta_h^{\mathrm{pop}} := \sqrt{2\bar{f}_1 \log 2 \cdot h/c_{\mathrm{q}}}$.*
- *(b) The minimizer is unique, $\beta_{k,h} = \phi_k(v_{k,h})$, and the Hessian of $L_{k,h}\circ\phi_k$ on the ball $\{\|v\| \le \eta_h^{\mathrm{pop}}\}$ satisfies $\nabla^2(L_{k,h}\circ\phi_k) \succeq \frac{1}{2}\underline{h} I$.*
- *(c) Under the assumptions of Lemma 22,*

$$v_{k,h} = h^2 b_k + \rho_v(h), \qquad \|\rho_v(h)\| \le \frac{2}{\underline{h}}\big(\omega_{\mathrm{tot}}(h,\eta_h^{\mathrm{pop}},0)\,h^2\|b_k\| + \|\rho_b(h)\|\big) = o(h^2), \tag{89}$$

and $\|v_{k,h}\| \le \frac{2}{\underline{h}}\|\nabla_v(L_{k,h}\circ\phi_k)(0)\| = O(h^2)$. All bounds are uniform over $\beta_2 \in \mathbb{S}^{d-1}$.

Proof. (a) $L_{k,h}$ is continuous (dominated convergence) and the sphere is compact, so a minimizer exists. By Lemmas 19 and 21, $c_{\mathrm{q}}\angle(\beta_{k,h},\beta_k)^2 \le L_{k,0}(\beta_{k,h}) - L_{k,0}(\beta_k) \le L_{k,h}(\beta_{k,h}) - L_{k,h}(\beta_k) + 2\bar{f}_1\log 2\cdot h \le 2\bar{f}_1\log 2\cdot h$. (b) Choose h_0 such that $\eta_{h_0}^{\mathrm{pop}} \le 1/2$ and $\omega_{\mathrm{tot}}(h,\eta_h^{\mathrm{pop}},0) \le \underline{h}/2$ for $h \le h_0$ (Lemma 23(d), $\omega_{\mathrm{tot}} \to 0$). All minimizers lie in the chart ball $\{\|v\| \le \sin\eta_h^{\mathrm{pop}}\} \subseteq \{\|v\| \le \eta_h^{\mathrm{pop}}\}$ (the angle is below $\pi/2$, so they are in the domain of the chart); this ball is convex and on it the Hessian is $\succeq H_k - \frac{1}{2}\underline{h} I \succeq \frac{1}{2}\underline{h} I$, so $L_{k,h}\circ\phi_k$ is strictly convex there and has at most one critical point; a minimizer is a critical point, hence unique. (c) By the integral form of the mean value theorem, $0 = \nabla(L_{k,h}\circ\phi_k)(v_{k,h}) = \nabla(L_{k,h}\circ\phi_k)(0) + \bar{H}_{k,h}v_{k,h}$ with $\bar{H}_{k,h} := \int_0^1 \nabla^2(L_{k,h}\circ\phi_k)(s v_{k,h})\,ds$, $\|\bar{H}_{k,h} - H_k\| \le \omega_{\mathrm{tot}}(h,\eta_h^{\mathrm{pop}},0) \le \underline{h}/2$, hence $\|\bar{H}_{k,h}^{-1}\| \le 2/\underline{h}$. Inserting Lemma 22, $v_{k,h} = \bar{H}_{k,h}^{-1}(h^2 H_k b_k - \rho_b(h)) = h^2 b_k + \bar{H}_{k,h}^{-1}[(H_k - \bar{H}_{k,h})h^2 b_k - \rho_b(h)]$. □

Lemma 25 (Rates of convergence of the three quantities across sources)**.** *Under the assumptions of Lemma 22,*

$$\|H_2 - H\| \le C\|\delta\|, \qquad \|\Sigma_2 - \Sigma\| \le C\|\delta\|, \qquad \|b_2 - b\| \le C\|\delta\| + C\bar{\omega}_2(R\|\delta\|) = o(1), \tag{90}$$

where $\Sigma_k := \frac{1}{3}\int zz^\top f_X(E_k z)\,dz$. If $\nabla^2(m_k f_X)$ is Lipschitz, which holds when $f_X \in C^{2,1}$ and $\mathbb{E}\|\varepsilon\|^3 < \infty$, then $\|b_2 - b\| \le C\|\delta\|$. The first two bounds use only $f_X \in C^1$ and a first moment.

Proof. $E_2 = QE$ with $\|Q - I\| = \|\delta\|$. We have $a_2(z) - a_1(z) = 2[\mathbb{E}\sigma'(z^\top E^\top Q^\top \varepsilon) - \mathbb{E}\sigma'(z^\top E^\top \varepsilon)] f_X(QEz) + 2\mathbb{E}\sigma'(z^\top E^\top \varepsilon)[f_X(QEz) - f_X(Ez)]$; the first term is at most $2\|\sigma''\|_\infty R\|\delta\|\mathbb{E}\|\varepsilon\|\bar{f}$ and the second at most $\frac{1}{2}\|\nabla f_X\|_\infty R\|\delta\|$; integrate $zz^\top$ over $\|z\| \le R$. The same argument handles Σ. Next, $\mathcal{G}_2''(0) - \mathcal{G}_1''(0) = \int z\big[(Q\beta^\star)^\top \nabla^2(m_2 f_X)(QEz)(Q\beta^\star) - \beta^{\star\top}\nabla^2(m_1 f_X)(Ez)\beta^\star\big]dz$: replacing $Q\beta^\star$ by $\beta^\star$ contributes $C\|\delta\|$; replacing m_2 by m_1 contributes $C\|\nabla^2(m_2 f_X) - \nabla^2(m_1 f_X)\|_{L^\infty(\bar{B}(0,R))} \le C\|\delta\|$ (Lemma 12(b)(c), term by term in $\nabla^2(mf) = f\nabla^2 m + \nabla m \nabla f^\top + \nabla f \nabla m^\top + m\nabla^2 f$); replacing the argument QEz by Ez contributes $C\bar{\omega}_2(R\|\delta\|)$. Finally $b_2 - b = \frac{\pi^2}{24}[(H_2^{-1} - H^{-1})\mathcal{G}_2''(0) + H^{-1}(\mathcal{G}_2''(0) - \mathcal{G}_1''(0))]$ with $\|H_2^{-1} - H^{-1}\| \le \|H_2 - H\|/\underline{h}^2$. □

Lemma 26 (Coordinates of the proxy smoothed population minimizer in the tangent coordinates at $\beta^\star$)**.** *Let $h_2 \le h_0$ and $\|\delta_n\| \le 1/4$. Then $\beta_{2,h_2} = \phi(v_2^\circ)$ with $v_2^\circ := E^\top \beta_{2,h_2}$, and*

$$v_2^\circ = E^\top \delta_n + h_2^2 b + \rho_2^\circ, \qquad \|\rho_2^\circ\| \le \|\delta_n\|\big(\|v_{2,h_2}\| + \|v_{2,h_2}\|^2\big) + h_2^2\|b_2 - b\| + \|\rho_v(h_2)\|. \tag{91}$$

In the asymptotic framework ($\|\delta_n\| = O(r_n)$, $h_2^2 = c_2^2 r_n$), $\rho_2^\circ = o(r_n)$ uniformly over $c_2 \in [\underline{c}, \bar{c}]$ and $\zeta_n \in \mathcal{Z}$. Likewise set $v_1^\circ := v_{1,h_1} = h_1^2 b + \rho_v(h_1)$.

Proof. The angle between $\beta_{2,h_2} = \sqrt{1 - \|v_{2,h_2}\|^2}\beta_2 + E_2 v_{2,h_2}$ and $\beta^\star$ is at most $\angle(\beta_2, \beta^\star) + \angle(\beta_{2,h_2}, \beta_2) < \pi/2$, so β_{2,h_2} lies in the chart and $\beta_{2,h_2} = \phi(E^\top \beta_{2,h_2})$. By the identity for the two unit centers, $E^\top \beta_2 = E^\top \delta_n$; moreover $E^\top E_2 = E^\top QE$ with $\|E^\top QE - I\| \le \|\delta_n\|$, $|\sqrt{1 - \|v_{2,h_2}\|^2} - 1| \le \|v_{2,h_2}\|^2$ and $\|E^\top \delta_n\| \le \|\delta_n\|$. Hence $v_2^\circ = E^\top \delta_n + v_{2,h_2} + \rho'$ with $\|\rho'\| \le \|\delta_n\|\|v_{2,h_2}\|^2 + \|\delta_n\|\|v_{2,h_2}\|$; now insert Lemma 24(c) and Lemma 25. Orders: $\|v_{2,h_2}\| = O(h_2^2) = O(r_n)$, $\|\delta_n\| \cdot O(r_n) = O(r_n^2)$, $h_2^2\|b_2 - b\| = o(r_n)$, $\rho_v(h_2) = o(h_2^2) = o(r_n)$. □

D.6 Empirical process bounds

Throughout this subsection $h_n := n^{-1/5}$, $h_k = c_k h_n$ with $c_k \in [\underline{c}, \bar{c}]$, and

$$\Lambda_n := d\log\big(3(1+R)^3(1+\underline{c}^{-1})^3 n^2\big) + 3\log n = O(\log n), \tag{92}$$

which covers the logarithmic size of each of the three nets below plus $3\log n$: for $j = 1, 2, 3$, $3R^j n/h^j \le 3(1+R)^3(1+\underline{c}^{-1})^3 n^{8/5}$.

Lemma 27 (Uniform deviations of loss, gradient and Hessian over the sphere; the good event $\mathcal{E}_n$)**.** *For each k, with probability at least $1 - 2(1 + d + d^2)n^{-3}$ the following hold simultaneously:*

$$\sup_\beta |\widehat{L}_{k,h_k}(\beta) - L_{k,h_k}(\beta)| \le \epsilon_n^{\mathrm{L}} := \frac{1}{n} + \sqrt{\frac{4\Lambda_n}{N_k}} + \frac{4\Lambda_n}{3N_k}, \tag{93}$$

$$\sup_\beta \|\nabla_\beta \widehat{L}_{k,h_k}(\beta) - \nabla_\beta L_{k,h_k}(\beta)\| \le \epsilon_n^{\mathrm{g}} := \sqrt{d}\Big[\frac{1}{n} + \sqrt{\frac{4R^2 \bar{f}_1 \Lambda_n}{3h_k N_k}} + \frac{2R\Lambda_n}{3h_k N_k}\Big], \tag{94}$$

$$\sup_\beta \|\nabla_\beta^2 \widehat{L}_{k,h_k}(\beta) - \nabla_\beta^2 L_{k,h_k}(\beta)\| \le \epsilon_n^{\mathrm{H}} := d\Big[\frac{1}{n} + \sqrt{\frac{8R^4 \bar{f}_1 \Lambda_n}{h_k^3 N_k}} + \frac{2R^2\Lambda_n}{3h_k^2 N_k}\Big]. \tag{95}$$

In the asymptotic framework $\epsilon_n^{\mathrm{L}} = O(\sqrt{\log n/n})$, $\epsilon_n^{\mathrm{g}} = O(n^{-2/5}\sqrt{\log n})$ and $\epsilon_n^{\mathrm{H}} = O(n^{-1/5}\sqrt{\log n})$; all tend to zero uniformly over $c_k \in [\underline{c}, \bar{c}]$. Let $\mathcal{E}_n$ be the event that all six bounds (both sources) hold; then $\Pr(\mathcal{E}_n^c) \le 4(1 + d + d^2)n^{-3}$.

Proof. Apply Lemma 15 with the choice $t = \sqrt{4\varsigma^2\Lambda/N} + 4B\Lambda/(3N)$: then, writing $t = t_1 + t_2$, $Nt^2 \ge Nt_1^2 + Nt_2 t \ge 2\varsigma^2\Lambda + \frac{2}{3}B\Lambda t$, so the exponential factor is at most $e^{-\Lambda}$.

(i) Loss: $\psi_\beta = \ell_h(\beta; \cdot)$, $B = 1$, $\varsigma^2 \le 1$, Lipschitz constant $\mathrm{Lip} = R/(2h)$ (as $|\partial_\beta \tanh(x^\top\beta/h)| \le \|x\|/h$). Take $\epsilon = h/(Rn)$, so $2\mathrm{Lip}\,\epsilon = 1/n$ and the net has at most $(3Rn/h)^d$ points.

(ii) Gradient components: $\psi_\beta = -\frac{2y-1}{2h}\kappa_0(x^\top\beta/h)x_i$, $B = R/(2h)$; $\mathrm{Var} \le \frac{R^2}{4h^2}\mathbb{E}\kappa_0^2(X^\top\beta/h) \le \frac{R^2}{4h^2}\bar{f}_1 h\int\kappa_2 = \frac{R^2\bar{f}_1}{3h}$; $|\kappa_0'| = 2|\kappa_1| \le 1$, so the Lipschitz constant is at most $R^2/(2h^2)$; take $\epsilon = h^2/(R^2 n)$. A union bound over the d components and $\|\cdot\| \le \sqrt{d}\max_i|\cdot|$.

(iii) Hessian components: $\psi_\beta = \frac{2y-1}{h^2}\kappa_1(x^\top\beta/h)x_ix_j$, $B \le R^2/(2h^2)$ (since $\|\kappa_1\|_\infty < 1/2$); $\mathrm{Var} \le \frac{R^4}{h^4}\bar{f}_1 h\int\kappa_1^2 \le \frac{2R^4\bar{f}_1}{h^3}$; $|\kappa_1'| \le 1$, so the Lipschitz constant is at most R^3/h^3; take $\epsilon = h^3/(R^3n)$. A union bound over d^2 components and the Frobenius bound $\le d\max_{ij}|\cdot|$.

The logarithmic sizes of the three nets are at most $d\log(3R^jn/h^j)$, $j=1,2,3$, each at most $\Lambda_n - 3\log n$ (as $h \ge \underline{c}n^{-1/5}$). For the orders: $h_kN_k \asymp n^{4/5}$, $h_k^3N_k \asymp n^{2/5}$, $h_k^2N_k \asymp n^{3/5}$. □

Lemma 28 (Global localization of the empirical minimizer for one source)**.** *On $\mathcal{E}_n$, every minimizer $\widehat{\beta}$ of $\widehat{L}_{k,h_k}$ satisfies*

$$\angle(\widehat{\beta},\beta_k) \le \eta_n^{(1)} := \sqrt{\frac{2\epsilon_n^{\mathrm{L}} + 2\bar{f}_1\log 2\cdot\bar{c}\,h_n}{c_{\mathrm{q}}}} = O(n^{-1/10}). \tag{96}$$

Proof. By Lemmas 19 and 21, $c_{\mathrm{q}}\alpha^2 \le L_{k,0}(\widehat{\beta}) - L_{k,0}(\beta_k) \le [L_{k,h_k}(\widehat{\beta}) - L_{k,h_k}(\beta_k)] + 2\bar{f}_1\log 2\,h_k \le [\widehat{L}_{k,h_k}(\widehat{\beta}) - \widehat{L}_{k,h_k}(\beta_k)] + 2\epsilon_n^{\mathrm{L}} + 2\bar{f}_1\log 2\,h_k$, and the bracket is ≤ 0. □

Lemma 29 (Variance of local increments)**.** *For unit β,β' with $\alpha := \angle(\beta,\beta')$,*

$$\begin{aligned}&\mathrm{Var}\big(\ell_h(\beta;X,Y) - \ell_h(\beta';X,Y)\big) \le C_{\mathrm{inc}}\min\Big\{\frac{\alpha^2}{h},\ \alpha\Big\},\\ &C_{\mathrm{inc}} := \max\{4R^2\bar{f}_1(2R+\tfrac{1}{2}),\ 3(\bar{f}_1\log 2 + \Gamma)\}.\end{aligned} \tag{97}$$

Proof. $\ell_h(\beta) - \ell_h(\beta') = -\frac{2Y-1}{2}[\tanh(u/h) - \tanh(u'/h)]$ with $u = X^\top\beta$, $u' = X^\top\beta'$, $|u-u'| \le R\|\beta-\beta'\| \le R\alpha$.

Bound A ($\alpha \le h$; $\tanh(\cdot/h)$ varies mainly in a band of width $O(h)$): by the mean value theorem $|\tanh(u/h) - \tanh(u'/h)| \le \frac{|u-u'|}{h}\sup_{s\in[u\wedge u', u\vee u']}\kappa_0(s/h)$, where $\kappa_0(s/h) \le 4e^{-2|s|/h}$ and the distance from the interval to 0 is at least $(|u| - R\alpha)_+$. Hence

$$\begin{aligned}\mathbb{E}\big[(\tanh(u/h) - \tanh(u'/h))^2\big] &\le \frac{16R^2\alpha^2}{h^2}\mathbb{E}\,e^{-4(|X^\top\beta| - R\alpha)_+/h}\\ &\le \frac{16R^2\alpha^2}{h^2}\bar{f}_1\Big[2R\alpha + \int_{|u|>R\alpha} e^{-4(|u|-R\alpha)/h}du\Big]\\ &= \frac{16R^2\bar{f}_1\alpha^2}{h^2}\Big[2R\alpha + \frac{h}{2}\Big].\end{aligned} \tag{98}$$

Multiplying by $\frac{1}{4}$ and using $\alpha \le h$ gives $\le \frac{4R^2\bar{f}_1(2R+\frac{1}{2})}{h}\alpha^2$.

Bound B (any α): $|\tanh(u/h) - \tanh(u'/h)| \le (1 - \tanh(|u|/h)) + 2\cdot\mathbf{1}\{\mathrm{sign}\,u \ne \mathrm{sign}\,u'\} + (1 - \tanh(|u'|/h))$, and squaring, $\le 3[(\cdot)^2 + 4\cdot\mathbf{1} + (\cdot)^2]$; $\mathbb{E}(1 - \tanh(|X^\top\beta|/h))^2 \le \mathbb{E}(1 - \tanh(|X^\top\beta|/h)) \le 2\bar{f}_1 h\log 2$ (the computation of Lemma 21) and $\Pr(\mathrm{sign}\,u \ne \mathrm{sign}\,u') \le \Gamma\alpha$ (Lemma 13). Hence $\mathrm{Var} \le \frac{1}{4}\cdot 3[4\bar{f}_1 h\log 2 + 4\Gamma\alpha] = 3(\bar{f}_1\log 2\,h + \Gamma\alpha) \le 3(\bar{f}_1\log 2 + \Gamma)\alpha$ when $\alpha \ge h$. □

Lemma 30 (Score covariance at the smoothed population minimizer and CLT)**.** *Let v_k° be as in Lemma 26 (so $\beta_{k,h_k} = \phi(v_k^\circ)$), and*

$$\xi_{k,j} := D\phi(v_k^\circ)^\top\nabla_\beta\ell_{h_k}(\beta_{k,h_k};X_{k,j},Y_{k,j}) = -\frac{2Y_{k,j}-1}{2h_k}\kappa_0\Big(\frac{X_{k,j}^\top\beta_{k,h_k}}{h_k}\Big)D\phi(v_k^\circ)^\top X_{k,j}, \tag{99}$$

$$\widehat{S}_k := \frac{1}{N_k}\sum_j\xi_{k,j} = \nabla_v\big(\widehat{L}_{k,h_k}\circ\phi\big)(v_k^\circ). \tag{100}$$

(a) $\mathbb{E}\xi_{k,j} = 0$ *and* $\|\xi_{k,j}\| \le R/h_k$.

(b) $\big\|h_k\,\mathrm{Cov}(\xi_{k,j}) - \Sigma\big\| \le C\big(h_k + \|v_k^\circ\|\big)$*; in particular* $h_k\,\mathrm{Cov}(\xi_{k,j}) \to \Sigma$ *for* $k = 1,2$.

(c) $\mathbb{E}\|r_n^{-1}\widehat{S}_k\|^2 = \dfrac{\operatorname{tr}(h_k \operatorname{Cov}\xi_{k,j})}{r_n^2 N_k h_k} \le C$*, uniformly in the asymptotic framework and over* $c_k \in [\underline{c}, \overline{c}]$.

(d) Along any sequence $c_{k,n} \to c_k^\infty \in [\underline{c}, \overline{c}]$, $\zeta_n \in \mathcal{Z}$: $\sqrt{N_k h_k}\,\widehat{S}_k \Rightarrow \mathcal{N}(0,\Sigma)$, $r_n^{-1}\widehat{S}_k \Rightarrow \mathcal{N}\big(0, \Sigma/(\nu_k c_k^\infty)\big)$, *and* $(r_n^{-1}\widehat{S}_1, r_n^{-1}\widehat{S}_2)$ *converges jointly to an independent Gaussian pair.*

Proof. (a) β_{k,h_k} is the minimizer of $L_{k,h_k} \circ \phi$ at an interior point of the chart (Lemmas 24, 26), so $\mathbb{E}\xi = \nabla_v(L_{k,h_k} \circ \phi)(v_k^\circ) = 0$. The bound uses $\kappa_0 \le 1$ and $\|D\phi\| \le 2$.

(b) Since $(2Y-1)^2 = 1$, $\operatorname{Cov}\xi = \frac{1}{4h^2}\mathbb{E}[\kappa_2(X^\top\beta_{k,h}/h) D\phi^\top X X^\top D\phi]$. Write $x = \tau\beta_{k,h} + w$ with $w \perp \beta_{k,h}$; from $D\phi(v_k^\circ)^\top \phi(v_k^\circ) = 0$ we get $D\phi^\top x = D\phi^\top w$. Hence

$$h \operatorname{Cov}\xi = \frac{1}{4}\int_{\beta_{k,h}^\perp} D\phi^\top w w^\top D\phi \Big[\int \kappa_2(u) f_X(hu\beta_{k,h} + w)\, du\Big] dw. \tag{101}$$

By Lemma 11(K3) with $\psi(\tau) = f_X(\tau\beta_{k,h} + w)$ (Lipschitz constant $\|\nabla f_X\|_\infty$), the inner integral is $\frac{4}{3} f_X(w) + O(h)$. Thus $h \operatorname{Cov}\xi = D\phi(v_k^\circ)^\top \Sigma^{(\beta_{k,h})} D\phi(v_k^\circ) + O(h)$ with $\Sigma^{(\beta)} := \frac{1}{3}\int_{\beta^\perp} w w^\top f_X(w)\, dw$ and $E^\top \Sigma^{(\beta^\star)} E = \Sigma$. The map $\beta \mapsto \Sigma^{(\beta)}$ is Lipschitz (rotation substitution and Lipschitz f_X, as in Lemma 23(b)), $\|\beta_{k,h} - \beta^\star\| \le 2\|v_k^\circ\|$ and $\|D\phi(v_k^\circ) - E\| \le 2\|v_k^\circ\|$.

(c) Directly from (b): $r_n^2 N_k h_k = (N_k/n) c_k \to \nu_k c_k \ge \frac{1}{2}\min_k \nu_k\, \underline{c}$ for large n.

(d) Lyapunov's condition (the vector case reduces to scalars via Cramér–Wold): $\sum_j \mathbb{E}\|\sqrt{h_k/N_k}\,\xi_{k,j}\|^3 \le N_k (h_k/N_k)^{3/2} \frac{R}{h_k} \mathbb{E}\|\xi\|^2 \le \frac{CR}{\sqrt{N_k h_k}} \to 0$ (as $N_k h_k \asymp n^{4/5}$), and the covariance $h_k \operatorname{Cov}\xi \to \Sigma$ by (b) ($h_k \to 0$, $v_k^\circ \to 0$). Hence the Lindeberg–Feller CLT for triangular arrays gives $\sqrt{N_k h_k}\widehat{S}_k \Rightarrow \mathcal{N}(0,\Sigma)$; multiply by $(r_n\sqrt{N_k h_k})^{-1} \to (\nu_k c_k^\infty)^{-1/2}$. The two sources are independent, so the convergence is joint. □

Lemma 31 (Stability of the empirical tangential Hessian on shrinking balls)**.** *On* $\mathcal{E}_n$*, for* $\eta \le 1/2$*,* $\|v\| \le \eta$ *and* $k = 1, 2$*,*

$$\big\|\nabla_v^2\big(\widehat{L}_{k,h_k} \circ \phi\big)(v) - H\big\| \le \epsilon_n^{\mathcal{H}}(\eta) := \omega_{\mathrm{tot}}(h_k, \eta, \|\delta_n\|) + 4\epsilon_n^{\mathrm{H}} + \sqrt{d}\, C_\phi\, \epsilon_n^{\mathrm{g}}. \tag{102}$$

If $\eta = \eta_n \to 0$*, then* $\epsilon_n^{\mathcal{H}}(\eta_n) \to 0$ *uniformly over* $c_k \in [\underline{c}, \overline{c}]$ *and* $\zeta_n \in \mathcal{Z}$*.*

Proof. Chain rule as in Lemma 23(d): for the difference between the empirical and the population objects, the first term is at most $\|D\phi(v)\|^2 \sup_\beta \|\nabla^2\widehat{L} - \nabla^2 L\| \le 4\epsilon_n^{\mathrm{H}}$ and the second at most $\sqrt{d} C_\phi \sup_\beta \|\nabla\widehat{L} - \nabla L\| \le \sqrt{d} C_\phi \epsilon_n^{\mathrm{g}}$; the population term is Lemma 23(d). □

D.7 Linearization of the first-order conditions

Common notation and the threshold n_0**.** Set $\widehat{U}_k := v_k^\circ - H^{-1}\widehat{S}_k$ ($k = 1, 2$). By Lemmas 26 and 30(d), along any sequence $(c_{k,n}, \zeta_n) \to (c_k^\infty, \zeta^\infty)$,

$$r_n^{-1}(\widehat{U}_1, \widehat{U}_2) \Rightarrow (U_1, U_2), \qquad U_1 = (c_1^\infty)^2 b + \frac{V^{1/2} G_1}{\sqrt{\nu_1 c_1^\infty}}, \quad U_2 = \zeta^\infty + (c_2^\infty)^2 b + \frac{V^{1/2} G_2}{\sqrt{\nu_2 c_2^\infty}}, \tag{103}$$

because $-H^{-1}\mathcal{N}(0, \Sigma/(\nu c)) \overset{d}{=} \mathcal{N}(0, V/(\nu c))$. Moreover $\|\widehat{U}_k\| \le \|v_k^\circ\| + \underline{h}^{-1}\|\widehat{S}_k\|$.

Let $\alpha_n^\circ := \angle(\beta^\star + \delta_n, \beta^\star) = \arcsin\|E^\top\delta_n\| \le \frac{\pi}{2} r_n \max_{\mathcal{Z}} \|z\|$, and define the localization radius

$$\begin{aligned} \eta_n := &\sqrt{\frac{2\epsilon_n^{\mathrm{L}} + 2\bar{f}_1 \log 2 \cdot \overline{c}\, h_n + c_{\mathrm{q}}(\alpha_n^\circ)^2 + \Gamma\alpha_n^\circ}{c_{\mathrm{q}} \min(\frac{1}{2},\, 1 - \overline{\rho},\, \underline{\rho})}} + \alpha_n^\circ \\ &+ \sqrt{\frac{\overline{\rho}_{\mathrm{TSE}}}{c_{\mathrm{q}}}}\,\big(\eta_n^{(1)} + \alpha_n^\circ\big) = O(n^{-1/10}), \end{aligned} \tag{104}$$

which covers the localization needs of the three lemmas below at once (each uses only part of it; for WE and TSE the minimum is read as $\frac{1}{2}$, and the THE constants $\underline{\rho}, \overline{\rho}$ do not enter). Let

$\epsilon_n := \max_k \epsilon_n^{\mathcal{H}}(\eta_n) \to 0$ (Lemma 31). Choose n_0 (depending only on the DGP and the compact parameter sets) such that for $n \geq n_0$: $\overline{c}h_n \leq h_0$; $\|\delta_n\| \leq 1/4$; $\eta_n \leq 1/4$; $\|v_k^\circ\| \leq \eta_n$; and $\kappa(\epsilon_n + 4\max(\overline{\lambda}, \overline{\rho}_{\mathrm{TSE}})\eta_n^2) \leq 1/2$, where $\kappa := 1/(\min(\underline{\rho}, 1-\overline{\rho})\underline{h})$. From now on $n \geq n_0$ and we argue on $\mathcal{E}_n$.

Lemma 32 (Perturbed linear system)**.** *Let $\mathcal{A}$ be invertible with $\|\mathcal{A}^{-1}\| \leq \kappa$, and let a vector $\mathbf{v}$ satisfy $(\mathcal{A}+\Delta)\mathbf{v} + \boldsymbol{\epsilon} = \mathbf{r}$ with $\|\Delta\| \leq \epsilon_1$, $\|\boldsymbol{\epsilon}\| \leq \epsilon_2\|\mathbf{v}\|$ and $\kappa(\epsilon_1+\epsilon_2) \leq 1/2$. Then*

$$\|\mathbf{v}\| \leq 2\kappa\|\mathbf{r}\|, \qquad \|\mathbf{v} - \mathcal{A}^{-1}\mathbf{r}\| \leq 2\kappa^2(\epsilon_1+\epsilon_2)\|\mathbf{r}\|. \tag{105}$$

Proof. $\mathcal{A}\mathbf{v} = \mathbf{r} - \Delta\mathbf{v} - \boldsymbol{\epsilon}$ gives $\|\mathbf{v}\| \leq \kappa\|\mathbf{r}\| + \kappa(\epsilon_1+\epsilon_2)\|\mathbf{v}\| \leq \kappa\|\mathbf{r}\| + \frac{1}{2}\|\mathbf{v}\|$; and $\mathbf{v} - \mathcal{A}^{-1}\mathbf{r} = -\mathcal{A}^{-1}(\Delta\mathbf{v} + \boldsymbol{\epsilon})$. □

Lemma 33 (WE)**.** *Let $\rho \in [0,1]$ and $c_k \in [\underline{c}, \overline{c}]$. On $\mathcal{E}_n$, every minimizer $\widehat{\beta}_{\mathrm{WE}}$ of $\widehat{Q}_{\mathrm{WE}} := (1-\rho)\widehat{L}_{1,h_1} + \rho\widehat{L}_{2,h_2}$ satisfies*

(a) $\angle(\widehat{\beta}_{\mathrm{WE}}, \beta^\star) \leq \eta_n$;

(b) $\widehat{\beta}_{\mathrm{WE}} = \phi(\widehat{v})$ with

$$\widehat{v} = (1-\rho)\widehat{U}_1 + \rho\widehat{U}_2 + \rho_{\mathrm{WE}}, \qquad \|\rho_{\mathrm{WE}}\| \leq C\epsilon_n\big(\|\widehat{U}_1\| + \|\widehat{U}_2\| + \|v_1^\circ\| + \|v_2^\circ\|\big), \tag{106}$$

where C depends only on $\underline{h}$ and $\overline{h}$.[3]

Proof. (a) Let $Q_0 := (1-\rho)L_{1,0} + \rho L_{2,0}$ and $Q_h := (1-\rho)L_{1,h_1} + \rho L_{2,h_2}$. Then $\sup|\widehat{Q}_{\mathrm{WE}} - Q_h| \leq \epsilon_n^{\mathrm{L}}$ and $\sup|Q_h - Q_0| \leq \bar{f}_1 \log 2\,\overline{c}h_n$ (Lemmas 27, 21), so $Q_0(\widehat{\beta}) - Q_0(\beta^\star) \leq [\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta^\star)] + 2\epsilon_n^{\mathrm{L}} + 2\bar{f}_1\log 2\,\overline{c}h_n \leq 2\epsilon_n^{\mathrm{L}} + 2\bar{f}_1\log 2\,\overline{c}h_n$. Lower bound, by Lemmas 19 and 20:

$$\begin{aligned} Q_0(\beta) - Q_0(\beta^\star) &= (1-\rho)[L_{1,0}(\beta) - L_{1,0}(\beta^\star)] + \rho[L_{2,0}(\beta) - L_{2,0}(\beta^\star + \delta_n)] \\ &\quad - \rho[L_{2,0}(\beta^\star) - L_{2,0}(\beta^\star + \delta_n)] \\ &\geq (1-\rho)c_{\mathrm{q}}\angle(\beta,\beta^\star)^2 + \rho c_{\mathrm{q}}\angle(\beta,\beta^\star+\delta_n)^2 - \Gamma\alpha_n^\circ \\ &\geq \tfrac{1}{2}c_{\mathrm{q}}\angle(\beta,\beta^\star)^2 - c_{\mathrm{q}}(\alpha_n^\circ)^2 - \Gamma\alpha_n^\circ, \end{aligned} \tag{107}$$

where the last step uses $\angle(\beta,\beta^\star) \leq \angle(\beta,\beta^\star+\delta_n) + \alpha_n^\circ$ and $(x+y)^2 \leq 2x^2 + 2y^2$. Combining, $\angle(\widehat{\beta},\beta^\star) \leq \eta_n$, uniformly in $\rho \in [0,1]$.

(b) Since $\eta_n < \pi/2$, $\widehat{\beta} = \phi(\widehat{v})$ with $\|\widehat{v}\| = \sin\alpha \leq \eta_n$, and $\widehat{v}$ is a minimizer of the C^2 function $\widehat{Q}_{\mathrm{WE}} \circ \phi$ in an open ball, so the first-order condition $(1-\rho)\nabla_v(\widehat{L}_{1,h_1}\circ\phi)(\widehat{v}) + \rho\nabla_v(\widehat{L}_{2,h_2}\circ\phi)(\widehat{v}) = 0$ holds. By the integral form of the mean value theorem (the segment $[v_k^\circ, \widehat{v}]$ lies in the convex ball $\{\|v\| \leq \eta_n\}$),

$$\nabla_v(\widehat{L}_{k,h_k}\circ\phi)(\widehat{v}) = \widehat{S}_k + \widehat{\widehat{H}}_k(\widehat{v} - v_k^\circ), \qquad \widehat{\widehat{H}}_k := \int_0^1 \nabla_v^2(\widehat{L}_{k,h_k}\circ\phi)(v_k^\circ + s(\widehat{v} - v_k^\circ))\,ds = H + \Delta_k, \tag{108}$$

with $\|\Delta_k\| \leq \epsilon_n$ (Lemma 31). Substituting and rearranging,

$$\big[H + (1-\rho)\Delta_1 + \rho\Delta_2\big]\widehat{v} = H\big[(1-\rho)\widehat{U}_1 + \rho\widehat{U}_2\big] + (1-\rho)\Delta_1 v_1^\circ + \rho\Delta_2 v_2^\circ =: \mathbf{r}. \tag{109}$$

Lemma 32 with $\mathcal{A} = H$, $\kappa = 1/\underline{h}$, $\epsilon_1 = \epsilon_n$, $\epsilon_2 = 0$ gives $\widehat{v} = H^{-1}\mathbf{r} + \mathbf{e}$ with $\|\mathbf{e}\| \leq 2\underline{h}^{-2}\epsilon_n\|\mathbf{r}\|$; here $H^{-1}\mathbf{r} = (1-\rho)\widehat{U}_1 + \rho\widehat{U}_2 + H^{-1}[(1-\rho)\Delta_1 v_1^\circ + \rho\Delta_2 v_2^\circ]$ and $\|\mathbf{r}\| \leq \overline{h}(\|\widehat{U}_1\| + \|\widehat{U}_2\|) + \epsilon_n(\|v_1^\circ\| + \|v_2^\circ\|)$. □

Lemma 34 (THE: joint localization of both heads and linearization of the first head)**.** *Let $\rho \in [\underline{\rho}, \overline{\rho}] \subset (0,1)$, $\lambda \in [\underline{\lambda}, \overline{\lambda}] \subset (0,\infty)$, $c_k \in [\underline{c}, \overline{c}]$, and*

$$\widehat{Q}_{\mathrm{THE}}(\gamma_1,\gamma_2) := (1-\rho)\widehat{L}_{1,h_1}(\gamma_1) + \rho\widehat{L}_{2,h_2}(\gamma_2) + \lambda\|\gamma_1 - \gamma_2\|^2, \qquad (\gamma_1,\gamma_2) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}. \tag{110}$$

On $\mathcal{E}_n$, every joint minimizer $(\widehat{\beta}_1, \widehat{\beta}_2)$ satisfies

[3] The dependence on $\overline{h}$ comes from the bound $\|\mathbf{r}\| \leq \overline{h}(\|\widehat{U}_1\| + \|\widehat{U}_2\|) + \cdots$ in the proof.

(a) $\angle(\widehat{\beta}_1, \beta^\star) \le \eta_n$ *and* $\angle(\widehat{\beta}_2, \beta^\star) \le \eta_n$*;*

(b) $\widehat{\beta}_k = \phi(\widehat{v}_k)$*, and* $\widehat{\mathbf{v}} = (\widehat{v}_1, \widehat{v}_2)$ *satisfies*

$$\begin{aligned} &\widehat{v}_1 = (I - W_{\mathrm{THE}})\widehat{U}_1 + W_{\mathrm{THE}}\widehat{U}_2 + \rho^{\mathrm{rem}}_{\mathrm{THE}}, \\ &\|\rho^{\mathrm{rem}}_{\mathrm{THE}}\| \le C(\epsilon_n + \eta_n^2)\big(\|\widehat{U}_1\| + \|\widehat{U}_2\| + \|v_1^\circ\| + \|v_2^\circ\|\big), \end{aligned} \tag{111}$$

with $W_{\mathrm{THE}} = \rho[I + \frac{\rho(1-\rho)}{2\lambda}H]^{-1}$ *and* C *depending only on* $\underline{h}, \overline{h}, \underline{\rho}, \overline{\rho}, \overline{\lambda}$*; moreover* $\|\widehat{\mathbf{v}}\| \le C(\|\widehat{U}_1\| + \|\widehat{U}_2\| + \|v_1^\circ\| + \|v_2^\circ\|)$.

Proof. (a) Let $Q_0(\gamma_1, \gamma_2) := (1-\rho)L_{1,0}(\gamma_1) + \rho L_{2,0}(\gamma_2) + \lambda\|\gamma_1 - \gamma_2\|^2$ and let Q_h be the same with $L_{k,0}$ replaced by L_{k,h_k}. The penalty cancels in $\widehat{Q} - Q_h$ and in $Q_h - Q_0$, so $\sup|\widehat{Q}_{\mathrm{THE}} - Q_h| \le \epsilon_n^{\mathrm{L}}$ and $\sup|Q_h - Q_0| \le \bar{f}_1 \log 2\,\overline{c}h_n$. As in Lemma 33, $Q_0(\widehat{\beta}_1, \widehat{\beta}_2) - Q_0(\beta^\star, \beta^\star) \le 2\epsilon_n^{\mathrm{L}} + 2\bar{f}_1 \log 2\,\overline{c}h_n$ (the comparison point $(\beta^\star, \beta^\star)$ has zero penalty). Lower bound (dropping the nonnegative penalty):

$$Q_0(\gamma_1, \gamma_2) - Q_0(\beta^\star, \beta^\star) \ge (1 - \overline{\rho})c_{\mathrm{q}}\angle(\gamma_1, \beta^\star)^2 + \underline{\rho}\, c_{\mathrm{q}}\angle(\gamma_2, \beta^\star + \delta_n)^2 - \Gamma\alpha_n^\circ. \tag{112}$$

Both terms are nonnegative, so each is controlled separately: $\angle(\widehat{\beta}_1, \beta^\star)^2 \le \frac{2\epsilon_n^{\mathrm{L}} + 2\bar{f}_1 \log 2\,\overline{c}h_n + \Gamma\alpha_n^\circ}{(1-\overline{\rho})c_{\mathrm{q}}}$, $\angle(\widehat{\beta}_2, \beta^\star + \delta_n)^2 \le \frac{2\epsilon_n^{\mathrm{L}} + 2\bar{f}_1 \log 2\,\overline{c}h_n + \Gamma\alpha_n^\circ}{\underline{\rho}c_{\mathrm{q}}}$, and $\angle(\widehat{\beta}_2, \beta^\star) \le \angle(\widehat{\beta}_2, \beta^\star + \delta_n) + \alpha_n^\circ$. All are at most η_n. This step does not use the penalty, so it holds for all $\lambda \in [0, \overline{\lambda}]$; it does need ρ bounded away from 0 and 1.

(b) Both heads lie in the chart, and $(\widehat{v}_1, \widehat{v}_2)$ minimizes the C^2 function $\widehat{Q}_{\mathrm{THE}}(\phi(v_1), \phi(v_2))$ on the open set $\{\|v_1\|, \|v_2\| < 1\}$, so the first-order conditions hold. Expanding the penalty gradient by Lemma 14(d) and the two losses by the mean value expansion of the proof of Lemma 33,

$$(1-\rho)\big[\widehat{S}_1 + (H + \Delta_1)(\widehat{v}_1 - v_1^\circ)\big] + 2\lambda(\widehat{v}_1 - \widehat{v}_2) + \mathbf{e}_1 = 0, \tag{113}$$

$$\rho\big[\widehat{S}_2 + (H + \Delta_2)(\widehat{v}_2 - v_2^\circ)\big] + 2\lambda(\widehat{v}_2 - \widehat{v}_1) + \mathbf{e}_2 = 0, \tag{114}$$

with $\|\Delta_k\| \le \epsilon_n$ and $\|\mathbf{e}_k\| \le 2\lambda\|\widehat{v}_k\|(\|\widehat{v}_1\|^2 + \|\widehat{v}_2\|^2) \le 4\overline{\lambda}\eta_n^2\|\widehat{\mathbf{v}}\|$. Write this as

$$\begin{aligned} &(\mathcal{A} + \Delta)\widehat{\mathbf{v}} + \mathbf{e}^{\mathrm{foc}} = \mathcal{B}\widehat{\mathbf{U}} + \Delta\mathbf{v}^\circ, \qquad \mathbf{e}^{\mathrm{foc}} = (\mathbf{e}_1, \mathbf{e}_2), \\ &\mathcal{A} = \begin{pmatrix} (1-\rho)H + 2\lambda I & -2\lambda I \\ -2\lambda I & \rho H + 2\lambda I \end{pmatrix}, \quad \mathcal{B} = \begin{pmatrix} (1-\rho)H & 0 \\ 0 & \rho H \end{pmatrix}, \end{aligned} \tag{115}$$

with $\Delta = \mathrm{diag}((1-\rho)\Delta_1, \rho\Delta_2)$, $\widehat{\mathbf{U}} = (\widehat{U}_1, \widehat{U}_2)$, $\mathbf{v}^\circ = (v_1^\circ, v_2^\circ)$ (using $H\widehat{U}_k = Hv_k^\circ - \widehat{S}_k$). For $\mathbf{x} = (x, y)$, $\mathbf{x}^\top\mathcal{A}\mathbf{x} = (1-\rho)x^\top Hx + \rho y^\top Hy + 2\lambda\|x - y\|^2 \ge \min(\underline{\rho}, 1-\overline{\rho})\underline{h}\|\mathbf{x}\|^2$, so $\|\mathcal{A}^{-1}\| \le \kappa$. Lemma 32 with $\epsilon_1 = \epsilon_n$ and $\epsilon_2 = 4\overline{\lambda}\eta_n^2$ gives

$$\begin{aligned} &\widehat{\mathbf{v}} = \mathcal{A}^{-1}\mathcal{B}\widehat{\mathbf{U}} + \mathcal{A}^{-1}\Delta\mathbf{v}^\circ + \mathbf{e}, \qquad \|\mathbf{e}\| \le 2\kappa^2(\epsilon_n + 4\overline{\lambda}\eta_n^2)\big(\|\mathcal{B}\|\|\widehat{\mathbf{U}}\| + \epsilon_n\|\mathbf{v}^\circ\|\big), \\ &\|\widehat{\mathbf{v}}\| \le 2\kappa\big(\overline{h}\|\widehat{\mathbf{U}}\| + \epsilon_n\|\mathbf{v}^\circ\|\big). \end{aligned} \tag{116}$$

To compute $\mathcal{A}^{-1}\mathcal{B}\widehat{\mathbf{U}} =: (x, y)$, note that it solves $(1-\rho)H(x - \widehat{U}_1) + 2\lambda(x - y) = 0$ and $\rho H(y - \widehat{U}_2) + 2\lambda(y - x) = 0$. From the second equation, $x - y = (\rho H + 2\lambda I)^{-1}\rho H(x - \widehat{U}_2)$; substituting into the first and multiplying by H^{-1} on the left (note that $(\rho H + 2\lambda I)^{-1}$ commutes with H): $(1-\rho)(x - \widehat{U}_1) + M(x - \widehat{U}_2) = 0$ with $M := 2\lambda\rho(\rho H + 2\lambda I)^{-1} = \rho(I + \frac{\rho}{2\lambda}H)^{-1}$. Hence $x = (I - W)\widehat{U}_1 + W\widehat{U}_2$ with $W = [(1-\rho)I + M]^{-1}M$. Since $(1-\rho)I + M = (I + \frac{\rho}{2\lambda}H)^{-1}[I + \frac{\rho(1-\rho)}{2\lambda}H]$,

$$W = \Big[I + \frac{\rho(1-\rho)}{2\lambda}H\Big]^{-1}\Big(I + \frac{\rho}{2\lambda}H\Big)\rho\Big(I + \frac{\rho}{2\lambda}H\Big)^{-1} = \rho\Big[I + \frac{\rho(1-\rho)}{2\lambda}H\Big]^{-1} = W_{\mathrm{THE}}. \tag{117}$$

The remainder $\rho^{\mathrm{rem}}_{\mathrm{THE}}$ of the first block consists of the first block of $\mathcal{A}^{-1}\Delta\mathbf{v}^\circ$ (of norm at most $\kappa\epsilon_n\|\mathbf{v}^\circ\|$) and of $\mathbf{e}$. □

Remark 4 (The penalty and the loss increments are both of order r_n^2). *On $\mathcal{E}_n$, $\|\widehat{v}_k\| \le C(\|\widehat{U}_1\| + \|\widehat{U}_2\| + \|\mathbf{v}^\circ\|) = O_P(r_n)$, so $\lambda\|\widehat{\beta}_1 - \widehat{\beta}_2\|^2 = \lambda\|\widehat{v}_1 - \widehat{v}_2\|^2(1+O(\eta_n^2)) = O_P(r_n^2)$ (Lemma 14(c)). On the other hand, $\widehat{L}_{k,h_k}(\widehat{\beta}_k) - \widehat{L}_{k,h_k}(\beta_{k,h_k}) = \widehat{S}_k^\top(\widehat{v}_k - v_k^\circ) + \frac{1}{2}(\widehat{v}_k - v_k^\circ)^\top \widehat{H}_k'(\widehat{v}_k - v_k^\circ) = O_P(r_n \cdot r_n)$, and by Lemma 29 the stochastic fluctuation of the empirical loss increments on an $O(r_n)$ ball is also $\sqrt{r_n^2/(h_k N_k)} \asymp r_n^2$. The three are of the same order, which is why a fixed finite λ produces a nondegenerate matrix weight W_{THE} in the limit. As $\lambda \to \infty$, $W_{\mathrm{THE}} \to \rho I$ (WE); as $\lambda \to 0$, $W_{\mathrm{THE}} \to 0$ (gold only). These two limits lie outside the compact set considered here.*

Lemma 35 (TSE: retention of the first-stage error and plug-in). *Let $\rho_{\mathrm{TSE}} \in [0, \overline{\rho}_{\mathrm{TSE}}]$ and $c_k \in [\underline{c}, \overline{c}]$. On $\mathcal{E}_n$:*

(a) First stage: every minimizer $\widehat{\beta}_{\mathrm{proxy}}$ of $\widehat{L}_{2,h_2}$ satisfies $\angle(\widehat{\beta}_{\mathrm{proxy}}, \beta^\star) \le \eta_n$, $\widehat{\beta}_{\mathrm{proxy}} = \phi(\widehat{v}_{\mathrm{p}})$, and $\widehat{v}_{\mathrm{p}} = \widehat{U}_2 + \rho_{\mathrm{p}}$ with $\|\rho_{\mathrm{p}}\| \le C\epsilon_n(\|\widehat{U}_2\| + \|v_2^\circ\|)$.

(b) Second stage: every minimizer $\widehat{\beta}_{\mathrm{TSE}}$ of $\widehat{Q}_{\mathrm{TSE}}(\gamma) := \widehat{L}_{1,h_1}(\gamma) + \rho_{\mathrm{TSE}}\|\gamma - \widehat{\beta}_{\mathrm{proxy}}\|^2$ satisfies $\angle(\widehat{\beta}_{\mathrm{TSE}}, \beta^\star) \le \eta_n$, $\widehat{\beta}_{\mathrm{TSE}} = \phi(\widehat{v})$, and

$$\begin{aligned} &\widehat{v} = (I - W_{\mathrm{TSE}})\widehat{U}_1 + W_{\mathrm{TSE}}\widehat{U}_2 + \rho_{\mathrm{TSE}}^{\mathrm{rem}}, \\ &\|\rho_{\mathrm{TSE}}^{\mathrm{rem}}\| \le C(\epsilon_n + \eta_n^2)\big(\|\widehat{U}_1\| + \|\widehat{U}_2\| + \|v_1^\circ\| + \|v_2^\circ\|\big), \end{aligned} \tag{118}$$

with $W_{\mathrm{TSE}} = (H + 2\rho_{\mathrm{TSE}} I)^{-1} 2\rho_{\mathrm{TSE}}$ and $I - W_{\mathrm{TSE}} = (H + 2\rho_{\mathrm{TSE}} I)^{-1} H$. The first-stage stochastic error $-H^{-1}\widehat{S}_2$ is retained in full through $\widehat{U}_2$.

Proof. (a) Lemma 28 gives $\angle(\widehat{\beta}_{\mathrm{proxy}}, \beta^\star + \delta_n) \le \eta_n^{(1)}$, hence $\angle(\widehat{\beta}_{\mathrm{proxy}}, \beta^\star) \le \eta_n^{(1)} + \alpha_n^\circ \le \eta_n$. The first-order condition reads $\widehat{S}_2 + (H + \Delta_2)(\widehat{v}_{\mathrm{p}} - v_2^\circ) = 0$, i.e. $(H + \Delta_2)\widehat{v}_{\mathrm{p}} = H\widehat{U}_2 + \Delta_2 v_2^\circ$; apply Lemma 32.

(b) Localization: $\widehat{Q}_{\mathrm{TSE}}(\widehat{\beta}_{\mathrm{TSE}}) \le \widehat{Q}_{\mathrm{TSE}}(\beta^\star)$ gives

$$\begin{aligned} \widehat{L}_{1,h_1}(\widehat{\beta}_{\mathrm{TSE}}) - \widehat{L}_{1,h_1}(\beta^\star) &\le \rho_{\mathrm{TSE}}\big[\|\beta^\star - \widehat{\beta}_{\mathrm{proxy}}\|^2 - \|\widehat{\beta}_{\mathrm{TSE}} - \widehat{\beta}_{\mathrm{proxy}}\|^2\big] \\ &\le \overline{\rho}_{\mathrm{TSE}}\,\angle(\widehat{\beta}_{\mathrm{proxy}}, \beta^\star)^2 \le \overline{\rho}_{\mathrm{TSE}}\big(\eta_n^{(1)} + \alpha_n^\circ\big)^2 \end{aligned} \tag{119}$$

(using $\|\beta - \beta'\| \le \angle(\beta, \beta')$). Therefore $c_{\mathrm{q}}\angle(\widehat{\beta}_{\mathrm{TSE}}, \beta^\star)^2 \le L_{1,0}(\widehat{\beta}_{\mathrm{TSE}}) - L_{1,0}(\beta^\star) \le \overline{\rho}_{\mathrm{TSE}}(\eta_n^{(1)} + \alpha_n^\circ)^2 + 2\epsilon_n^{\mathrm{L}} + 2\bar{f}_1 \log 2\,\bar{c}h_n$, whence $\alpha \le \eta_n$. First-order condition (with $\widehat{\beta}_{\mathrm{TSE}} = \phi(\widehat{v})$ and the penalty gradient of Lemma 14(d) with $v_1 = \widehat{v}$, $v_2 = \widehat{v}_{\mathrm{p}}$, $\lambda = \rho_{\mathrm{TSE}}$):

$$\widehat{S}_1 + (H + \Delta_1)(\widehat{v} - v_1^\circ) + 2\rho_{\mathrm{TSE}}(\widehat{v} - \widehat{v}_{\mathrm{p}}) + \epsilon = 0, \qquad \|\epsilon\| \le 4\overline{\rho}_{\mathrm{TSE}}\eta_n^2\|\widehat{v}\|. \tag{120}$$

That is, $(H + 2\rho_{\mathrm{TSE}} I + \Delta_1)\widehat{v} + \epsilon = H\widehat{U}_1 + \Delta_1 v_1^\circ + 2\rho_{\mathrm{TSE}}\widehat{U}_2 + 2\rho_{\mathrm{TSE}}\rho_{\mathrm{p}} =: \mathbf{r}$, with $\mathcal{A} = H + 2\rho_{\mathrm{TSE}} I \succeq \underline{h} I$. By Lemma 32, $\widehat{v} = \mathcal{A}^{-1}\mathbf{r} + \mathbf{e}$, where $\mathcal{A}^{-1}(H\widehat{U}_1 + 2\rho_{\mathrm{TSE}}\widehat{U}_2) = (I - W_{\mathrm{TSE}})\widehat{U}_1 + W_{\mathrm{TSE}}\widehat{U}_2$, and the remaining terms $\mathcal{A}^{-1}(\Delta_1 v_1^\circ + 2\rho_{\mathrm{TSE}}\rho_{\mathrm{p}})$ and $\mathbf{e}$ obey the stated bound. □

Proposition 5 (Convergence in distribution of the three estimators and a norm bound on the good event). *Let $(\theta_n, \zeta_n) \in \mathcal{K}_{\mathcal{M}} \times \mathcal{Z}$ be any sequence with $(\theta_n, \zeta_n) \to (\theta^\infty, \zeta^\infty)$. Then for $\mathcal{M} \in \{\mathrm{WE}, \mathrm{THE}, \mathrm{TSE}\}$ (THE meaning its first head):*

(a) $r_n^{-1}\widehat{v}_{\mathcal{M}} \Rightarrow U_{\mathcal{M}}(\theta^\infty, \zeta^\infty) = (I - W_{\mathcal{M}}(\theta^\infty))U_1 + W_{\mathcal{M}}(\theta^\infty)U_2$, and consequently $n^{2/5}\angle(\widehat{\beta}_{\mathcal{M}}, \beta^\star) \Rightarrow \|U_{\mathcal{M}}(\theta^\infty, \zeta^\infty)\|$.

(b) On $\mathcal{E}_n$ (for $n \ge n_0$): $n^{2/5}\angle(\widehat{\beta}_{\mathcal{M}}, \beta^\star) \le C_{\mathrm{UI}}\big(1 + \|r_n^{-1}\widehat{S}_1\| + \|r_n^{-1}\widehat{S}_2\|\big)$, where C_{UI} depends only on the DGP and the compact parameter sets.

Proof. (a) By Lemmas 33–35, on $\mathcal{E}_n$, $\widehat{v}_{\mathcal{M}} = (I - W_{\mathcal{M}}(\theta_n))\widehat{U}_1 + W_{\mathcal{M}}(\theta_n)\widehat{U}_2 + \rho_{\mathcal{M}}$ with $\|\rho_{\mathcal{M}}\| \le C(\epsilon_n + \eta_n^2)(\|\widehat{U}_1\| + \|\widehat{U}_2\| + \|v_1^\circ\| + \|v_2^\circ\|)$. By (103), $r_n^{-1}(\widehat{U}_1, \widehat{U}_2) \Rightarrow (U_1, U_2)$, $r_n^{-1}\|v_k^\circ\| = O(1)$ and $\epsilon_n + \eta_n^2 \to 0$, so $r_n^{-1}\rho_{\mathcal{M}} \to 0$ in probability; $W_{\mathcal{M}}(\theta_n) \to W_{\mathcal{M}}(\theta^\infty)$ (matrix inversion is continuous on the compact set; $\lambda \ge \underline{\lambda} > 0$ and $\rho_{\mathrm{TSE}} \ge 0$ keep the relevant matrices

uniformly invertible); and $\Pr(\mathcal{E}_n^c) \to 0$. Slutsky's lemma gives $r_n^{-1}\widehat{v}_{\mathcal{M}} \Rightarrow U_{\mathcal{M}}(\theta^\infty, \zeta^\infty)$. On $\mathcal{E}_n$, $\angle(\widehat{\beta}_{\mathcal{M}}, \beta^\star) = \arcsin\|\widehat{v}_{\mathcal{M}}\| = \|\widehat{v}_{\mathcal{M}}\|(1+O(\eta_n^2))$ (Lemma 14(a)); conclude by the continuous mapping theorem. (b) From the same representation, $\|W\|, \|I - W\| \le C$, $\|\widehat{U}_k\| \le \|v_k^\circ\| + \underline{h}^{-1}\|\widehat{S}_k\|$, $\|v_k^\circ\| \le Cr_n$ and $\alpha \le \frac{\pi}{2}\|\widehat{v}\|$. □

D.8 From convergence in distribution to expectations

Lemma 36 (Uniform integrability of $n^{2/5}\alpha$)**.** *For each $\mathcal{M}$, the family $\{Z_n := n^{2/5}\angle(\widehat{\beta}_{\mathcal{M}}, \beta^\star) : n \ge n_0,\ \theta \in \mathcal{K}_{\mathcal{M}},\ \zeta_n \in \mathcal{Z}\}$ is uniformly integrable:*

$$\sup_{n\ge n_0,\ \theta,\ \zeta_n} \mathbb{E}\big[Z_n \mathbf{1}\{Z_n > M\}\big] \le \frac{C}{M} + \sup_{\substack{n\ge n_0 \\ \pi n^{2/5} > M}} \pi n^{2/5}\Pr(\mathcal{E}_n^c) \le \frac{C}{M} + C' M^{-13/2} \xrightarrow[M\to\infty]{} 0. \tag{121}$$

Proof. Decompose $Z_n = Z_n\mathbf{1}_{\mathcal{E}_n} + Z_n\mathbf{1}_{\mathcal{E}_n^c}$. On $\mathcal{E}_n^c$ use $\alpha \le \pi$: $\mathbb{E}[Z_n\mathbf{1}_{\mathcal{E}_n^c}] \le \pi n^{2/5}\Pr(\mathcal{E}_n^c) \le 4\pi(1+d+d^2)n^{-13/5}$ (Lemma 27). For uniform integrability retain the tail indicator: $Z_n\mathbf{1}\{Z_n > M\}$ vanishes whenever $\pi n^{2/5} \le M$. Thus the bad-event contribution is bounded by the supremum over $n > (M/\pi)^{5/2}$, which is at most $C'M^{-13/2}$. This bound, unlike the unrestricted supremum, tends to zero. On $\mathcal{E}_n$, by Proposition 5(b), $Z_n \le Y_n := C_{\mathrm{UI}}(1 + \|r_n^{-1}\widehat{S}_1\| + \|r_n^{-1}\widehat{S}_2\|)$, and $\sup\mathbb{E}Y_n^2 \le 3C_{\mathrm{UI}}^2(1 + 2\sup_k \mathbb{E}\|r_n^{-1}\widehat{S}_k\|^2) < \infty$ (Lemma 30(c)). Hence $\mathbb{E}[Z_n\mathbf{1}_{\mathcal{E}_n}\mathbf{1}\{Z_n > M\}] \le \mathbb{E}[Y_n\mathbf{1}\{Y_n > M\}] \le \mathbb{E}Y_n^2/M$. □

D.9 Uniformity and the main theorem

Theorem 20 (Asymptotic risk)**.** *Assume the common assumptions of Section D.2 together with the additional conditions $f_X \in C_c^2(\mathbb{R}^d)$ and $\mathbb{E}\|\varepsilon\|^2 < \infty$, and the asymptotic framework: $N_k/n \to \nu_k > 0$, $h_k = c_k n^{-1/5}$, $r_n = n^{-2/5}$, $\beta^\star + \delta_n = \phi(r_n\zeta_n)$ with $\zeta_n \in \mathcal{Z}$ compact. Let $\widehat{\beta}_{\mathcal{M}}$ be WE, the first head of THE, or TSE as defined in Section D.2, with a measurable selection rule fixed in advance in case of multiple minimizers. Then for each $\mathcal{M}$,*

$$\lim_{n\to\infty} \sup_{\theta\in\mathcal{K}_{\mathcal{M}}} \sup_{\zeta\in\mathcal{Z}} \Big| n^{2/5}\,\mathbb{E}\,\alpha\big(\widehat{\beta}_{\mathcal{M}}, \beta^\star\big) - \mathbb{E}\big\|U_{\mathcal{M}}(\theta,\zeta)\big\| \Big| = 0, \tag{122}$$

where the expectation $\mathbb{E}\alpha$ is computed under the bias $\delta_n = \phi(r_n\zeta) - \beta^\star$ and the hyperparameters θ, and $U_{\mathcal{M}}(\theta,\zeta) = (I - W_{\mathcal{M}})U_1 + W_{\mathcal{M}}U_2$ with $U_1 = c_1^2 b + V^{1/2}G_1/\sqrt{\nu_1 c_1}$, $U_2 = \zeta + c_2^2 b + V^{1/2}G_2/\sqrt{\nu_2 c_2}$ and $W_{\mathcal{M}}$ as in (53)*. Equivalently,*

$$\mathbb{E}\,\angle(\widehat{\beta}_{\mathcal{M}}, \beta^\star) = n^{-2/5}\,\mathbb{E}\|U_{\mathcal{M}}\| + o(n^{-2/5}), \tag{123}$$

with the $o(\cdot)$ uniform over $\theta \in \mathcal{K}_{\mathcal{M}}$ and $\zeta \in \mathcal{Z}$ (including $\zeta = r\Theta$, $\Theta \in \mathbb{S}^{p-1}$, r fixed). The theorem does not require H to be non-scalar, $b \neq 0$, or any condition relating the curvature H to the noise covariance V, and it allows $\zeta = 0$.

Proof. Write $F_n(\theta,\zeta) := n^{2/5}\mathbb{E}\angle(\widehat{\beta}_{\mathcal{M}}, \beta^\star)$ and $F(\theta,\zeta) := \mathbb{E}\|U_{\mathcal{M}}(\theta,\zeta)\|$.

Step 1 (F is continuous). For fixed (G_1, G_2), the map $(\theta,\zeta) \mapsto U_{\mathcal{M}}(\theta,\zeta)$ is continuous on $\mathcal{K}_{\mathcal{M}} \times \mathcal{Z}$: $W_{\mathrm{WE}} = \rho I$; W_{THE} is a continuous function of (ρ,λ) (as $\lambda \ge \underline{\lambda} > 0$ and $I + \frac{\rho(1-\rho)}{2\lambda}H \succeq I$); W_{TSE} is continuous (as $H + 2\rho_{\mathrm{TSE}}I \succeq \underline{h}I$); $c_k \mapsto c_k^2 b$ and $(\nu_k c_k)^{-1/2}$ are continuous (as $c_k \ge \underline{c}$). Moreover $\|U_{\mathcal{M}}\| \le C(1 + \|G_1\| + \|G_2\|)$ is integrable, so dominated convergence gives continuity of F.

Step 2 (contradiction). If the conclusion failed, there would be $\epsilon_0 > 0$, a subsequence $n_j \to \infty$ and $(\theta_{n_j}, \zeta_{n_j})$ with $|F_{n_j}(\theta_{n_j}, \zeta_{n_j}) - F(\theta_{n_j}, \zeta_{n_j})| > \epsilon_0$. By compactness, pass to a further subsequence with $(\theta_{n_j}, \zeta_{n_j}) \to (\theta^\infty, \zeta^\infty) \in \mathcal{K}_{\mathcal{M}} \times \mathcal{Z}$. Along it, Proposition 5(a) gives $Z_{n_j} \Rightarrow \|U_{\mathcal{M}}(\theta^\infty, \zeta^\infty)\|$; Lemma 36 gives uniform integrability of $\{Z_{n_j}\}$; by Vitali's theorem (convergence in distribution together with uniform integrability implies convergence of expectations; see, e.g., the corollary form of Theorem 25.12 in Billingsley's *Probability and Measure*), $F_{n_j}(\theta_{n_j}, \zeta_{n_j}) = \mathbb{E}Z_{n_j} \to F(\theta^\infty, \zeta^\infty)$; and Step 1 gives $F(\theta_{n_j}, \zeta_{n_j}) \to F(\theta^\infty, \zeta^\infty)$. This is a contradiction. □

Corollary 4 (Direction-averaged version). *Fix $r > 0$, let $\Theta \sim \mathrm{Unif}(\mathbb{S}^{p-1})$ be independent of all samples, $\delta_n = \phi(r_n r\Theta) - \beta^\star$, and let the hyperparameters not depend on Θ. Then*

$$\mathbb{E}_\Theta \, \mathbb{E} \angle(\widehat{\beta}_{\mathcal{M}}, \beta^\star) = n^{-2/5} \, \mathbb{E}_{\Theta,G} \big\| U_{\mathcal{M}}(\theta, r\Theta) \big\| + o(n^{-2/5}), \tag{124}$$

with the $o(\cdot)$ uniform over $\theta \in \mathcal{K}_{\mathcal{M}}$.

Proof. Take $\mathcal{Z} = r\mathbb{S}^{p-1}$; the supremum form of Theorem 20 holds pointwise and uniformly in Θ; integrate over Θ (the integrand is uniformly bounded). □

Remark 5 (The statement when H is scalar). *Theorem 20 also holds when $H = \mathsf{h}I$; then $W_{\mathrm{THE}} = \frac{\rho}{1+\mathsf{h}\rho(1-\rho)/(2\lambda)} I$ and $W_{\mathrm{TSE}} = \frac{2\rho_{\mathrm{TSE}}}{\mathsf{h}+2\rho_{\mathrm{TSE}}} I$ are scalar and the limits of the three families coincide (the isotropic case). No conclusion about equality of the actual risks is drawn from this.*

D.10 Multiple minimizers

Lemma 37 (Uniqueness in the local region with probability tending to one; arbitrary measurable selection). *On $\mathcal{E}_n$ (with $n \geq n_0$) the following uniqueness statements hold:*

(a) the minimizer of $\widehat{L}_{k,h_k}$ ($k = 1, 2$) on $\mathbb{S}^{d-1}$ is unique;

(b) the minimizer of $\widehat{Q}_{\mathrm{WE}}$ is unique;

(c) the joint minimizer of $\widehat{Q}_{\mathrm{THE}}$ on $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ is unique;

(d) given the first-stage output, the minimizer of $\widehat{Q}_{\mathrm{TSE}}$ is unique.

Consequently $\Pr($the minimizer of the corresponding estimator is unique$) \geq 1 - 4(1 + d + d^2)n^{-3}$, uniformly over the compact parameter sets; any measurable selection rule fixed in advance returns the same value on $\mathcal{E}_n$, while its contribution on $\mathcal{E}_n^c$ is already controlled by $\alpha \leq \pi$ and $\Pr(\mathcal{E}_n^c)$ (Lemma 36). Hence Theorem 20 holds for every such selection rule.

Proof. By the localization statements of Lemmas 28, 33(a), 34(a) and 35, all minimizers lie in the chart ball $B_n := \{\|v\| \leq \eta_n\}$ (in $B_n \times B_n$ for THE), which is convex. On B_n the Hessians of the chart criteria are: (a) $H + \Delta$ with $\|\Delta\| \leq \epsilon_n$ (Lemma 31); (b) $(1-\rho)(H + \Delta_1) + \rho(H + \Delta_2)$; (c) $\mathrm{diag}((1-\rho)(H + \Delta_1), \rho(H + \Delta_2)) + \nabla^2 P = \mathcal{A} + \Delta''$ with $\|\Delta''\| \leq \epsilon_n + C\overline{\lambda}\eta_n^2$ (Lemma 14(d)); (d) $H + \Delta_1 + \nabla^2_{v_1 v_1} P|_{\lambda = \rho_{\mathrm{TSE}}} = H + 2\rho_{\mathrm{TSE}} I + \Delta'''$ with $\|\Delta'''\| \leq \epsilon_n + C\overline{\rho}_{\mathrm{TSE}}\eta_n^2$. By the choice of n_0,[4] each Hessian is $\succeq \frac{1}{2}\lambda_{\min}(\mathcal{A}) I \succ 0$ on the whole convex ball, so each criterion is strictly convex there and has at most one critical point; a minimizer is a critical point. □

Remark 6 (Existence of a measurable selection rule). *Each criterion is jointly continuous in (parameter, sample) with a compact feasible set, so the argmin correspondence is nonempty, compact-valued and upper hemicontinuous, and the measurable maximum theorem (e.g. Theorem 18.19 in Aliprantis and Border's* Infinite Dimensional Analysis*) yields a measurable selection. We quote this standard result without proof; the selection rule is part of the definition of the estimators.*

E Direction-averaged comparison of the smoothed estimators

This appendix works at the level of the asymptotic constants of Appendix D. Throughout, $p = d - 1 \geq 2$, $\nu_1 + \nu_2 = 1$, $\nu_2 > \nu_1 > 0$, and $H \succ 0$, $V = H^{-1}\Sigma H^{-1} \succ 0$, $b \in \mathbb{R}^p$ are the three population quantities of Appendix D. G_1, G_2 are independent standard Gaussian vectors, Θ is uniform on $\mathbb{S}^{p-1}$ and independent of them, and $r > 0$ is the scaled bias magnitude. With

$$U_1 = c_1^2 b + \widetilde{N}_1, \qquad U_2 = r\Theta + c_2^2 b + \widetilde{N}_2, \qquad \widetilde{N}_k = \frac{V^{1/2} G_k}{\sqrt{\nu_k c_k}}, \qquad \Sigma_k := \frac{V}{\nu_k c_k}, \tag{125}$$

[4] If the constant C of Lemma 14(d) exceeds the constant 4 in the last condition on n_0, then n_0 is enlarged accordingly. This is possible because $\epsilon_n + \eta_n^2 \to 0$, and it does not affect any other statement.

every symmetric W with $0 \preceq W \preceq I$ defines

$$\begin{aligned} U_W &= (I-W)U_1 + WU_2 = \mu_W + rW\Theta + N_W, \\ \mu_W &= c_1^2 b + W(c_2^2 - c_1^2)b, \qquad N_W = (I-W)\widetilde{N}_1 + W\widetilde{N}_2, \end{aligned} \tag{126}$$

with $\Sigma_W := \mathrm{Cov}(N_W) = \frac{(I-W)V(I-W)}{\nu_1 c_1} + \frac{WVW}{\nu_2 c_2}$. The direction-averaged asymptotic constant is $\bar{C}(W, c_1, c_2; r) := \mathbb{E}\|U_W\|$. The three families are

$$\begin{aligned} &\mathcal{W}_{\mathrm{WE}} = \{\rho I\}, \qquad \mathcal{W}_{\mathrm{THE}} = \{\rho(I+tH)^{-1} : \rho \in (0,1),\, t \geq 0\}, \\ &\mathcal{W}_{\mathrm{TSE}} = \{s(H+sI)^{-1} : s > 0\}, \end{aligned} \tag{127}$$

where $t = \rho(1-\rho)/(2\lambda)$ and $s = 2\rho_{\mathrm{TSE}}$; $W = 0$ is gold-only. $\bar{C}^\star_{\mathcal{M}}(r)$ denotes the infimum over the family and over $c_1, c_2 > 0$. These are infima of limiting risks. Compact-uniform convergence in Appendix D allows optimization on fixed compact interior parameter sets; it does not by itself justify the same operation over all bandwidths or boundary parameters. Our strict-improvement result exhibits a fixed admissible point below the WE limiting-risk infimum.

E.1 Gaussian norm expectations

For $\Sigma_0 \succ 0$ and $\mu \in \mathbb{R}^p$ let $N \sim \mathcal{N}(0, \Sigma_0)$, $U = \mu + N$, $\widehat{U} = U/\|U\|$, and

$$\phi_{\Sigma_0}(\mu) = \mathbb{E}\|\mu + N\|, \qquad m_{\Sigma_0}(\mu) = \mathbb{E}\widehat{U}, \qquad J_{\Sigma_0}(\mu) = \mathbb{E}\Big[\frac{I - \widehat{U}\widehat{U}^\top}{\|U\|}\Big]. \tag{128}$$

Lemma 38. *(i) ϕ_{Σ_0} is finite, even, convex and C^∞; (ii) $\nabla\phi_{\Sigma_0} = m_{\Sigma_0}$; (iii) $\nabla^2\phi_{\Sigma_0} = J_{\Sigma_0} \succ 0$, so ϕ_{Σ_0} is strictly convex; (iv) $\phi_{\Sigma_0}(\mu) > \phi_{\Sigma_0}(0)$ and $\mu^\top m_{\Sigma_0}(\mu) > 0$ for $\mu \neq 0$; (v) for every matrix M, $\mathbb{E}[\widehat{U}^\top M N] = \mathrm{tr}(M\Sigma_0 J_{\Sigma_0}(\mu))$; in particular $\phi_{\Sigma_0}(\mu) - \mu^\top m_{\Sigma_0}(\mu) = \mathrm{tr}(\Sigma_0 J_{\Sigma_0}(\mu)) > 0$; (vi) $\Sigma_0 \preceq \Sigma_0'$ implies $\phi_{\Sigma_0} \leq \phi_{\Sigma_0'}$, strictly if $\Sigma_0 \neq \Sigma_0'$.*

Proof. (i) $\phi_{\Sigma_0}(\mu) = \int \|x\|\varphi_{\Sigma_0}(x-\mu)dx$ can be differentiated under the integral arbitrarily often; convexity and evenness follow from those of the norm and the symmetry of N. (ii) The norm is Lipschitz and differentiable off the origin with gradient $\widehat{x}$; $U \neq 0$ a.s.; dominated convergence. (iii) the map $x \mapsto x/\|x\|$ has derivative $(I - \widehat{x}\widehat{x}^\top)/\|x\|$, and $\mathbb{E}\|U\|^{-1} < \infty$ for $p \geq 2$. For $u \neq 0$, $u^\top J u = \mathbb{E}[(\|u\|^2 - (u^\top\widehat{U})^2)/\|U\|] > 0$ since $\widehat{U} \parallel u$ is a null event. (iv) An even strictly convex function is minimized only at 0; convexity gives $\phi(0) \geq \phi(\mu) - \mu^\top\nabla\phi(\mu)$. (v) Stein's identity $\mathbb{E}[Ng(\mu+N)^\top] = \Sigma_0\mathbb{E}[\nabla g(\mu+N)]^\top$ for Lipschitz g gives $\mathbb{E}[\widehat{U}^\top M N] = \mathrm{tr}(M\,\mathbb{E}[N\widehat{U}^\top]) = \mathrm{tr}(M\Sigma_0 J)$; take $M = I$ and use $\mathbb{E}[\widehat{U}^\top U] = \phi$, $\mathbb{E}[\widehat{U}^\top\mu] = \mu^\top m$. (vi) Write $N' = N + N''$ with $N'' \sim \mathcal{N}(0, \Sigma_0' - \Sigma_0)$ independent; conditional Jensen gives $\phi_{\Sigma_0'}(\mu) = \mathbb{E}\phi_{\Sigma_0}(\mu + N'') \geq \phi_{\Sigma_0}(\mu)$, strictly when N'' is nondegenerate because ϕ_{Σ_0} is strictly convex. □

Lemma 39. *For every W, c_1, c_2: (i) $r \mapsto \bar{C}(W, c_1, c_2; r)$ is nondecreasing and $\bar{C}(\cdot; r) \geq \bar{C}(\cdot; 0)$; (ii) $\bar{C}(W, c_1, c_2; r) = \phi_{\Sigma_W}(\mu_W) + \frac{r^2}{2p}\,\mathrm{tr}(W^\top J_{\Sigma_W}(\mu_W)W) + O(r^4)$; (iii) $|\bar{C}(\cdot; r) - \bar{C}(\cdot; r')| \leq |r - r'|$, so every $\bar{C}^\star_{\mathcal{M}}$ is 1-Lipschitz and nondecreasing in r.*

Proof. (i) Jensen in Θ ($\mathbb{E}\Theta = 0$) and convexity along the segment $\{\mu_W + \tau W\Theta\}$. (ii) Taylor expansion of ϕ_{Σ_W} at μ_W; odd orders vanish under $\Theta \to -\Theta$ and $\mathbb{E}\Theta\Theta^\top = I/p$. (iii) $\|U_W(r) - U_W(r')\| \leq \|W\||r - r'|$. □

E.2 The weighted estimator

For $W = \rho I$: $U_\rho = Ab + \rho r\Theta + N_\rho$ with $A = (1-\rho)c_1^2 + \rho c_2^2$, $N_\rho \sim \mathcal{N}(0, \sigma_\rho^2 V)$, $\sigma_\rho^2 = \frac{(1-\rho)^2}{\nu_1 c_1} + \frac{\rho^2}{\nu_2 c_2}$.

Lemma 40 (Bandwidth elimination)**.** *Fix $\rho \in [0,1]$ and $A > 0$. Under $(1-\rho)c_1^2 + \rho c_2^2 = A$ the minimum of σ_ρ^2 is $S(\rho)^{3/2}A^{-1/2}$, $S(\rho) = (1-\rho)^{5/3}\nu_1^{-2/3} + \rho^{5/3}\nu_2^{-2/3}$, attained at $c_1 = \sqrt{A/S(\rho)}((1-\rho)/\nu_1)^{1/3}$, $c_2 = \sqrt{A/S(\rho)}(\rho/\nu_2)^{1/3}$. Since $\bar{C}_{\mathrm{WE}}$ is strictly increasing in σ_ρ^2 for fixed (ρ, A) (Lemma 38(vi)),*

$$\bar{C}^\star_{\mathrm{WE}}(r) = \min_{\rho\in[0,1],\,A>0} \bar{F}(\rho, A; r), \qquad \bar{F}(\rho, A; r) = \mathbb{E}\big\|Ab + \rho r\Theta + S(\rho)^{3/4}A^{-1/4}V^{1/2}G\big\|. \tag{129}$$

Proof. The Lagrange conditions give $c_k^3 \propto (\text{weight}_k/\nu_k)$; substituting in the constraint yields the factor $\sqrt{A/S}$; $\sigma_\rho^2 \to \infty$ as either $c_k \to 0$, and the constraint set has a single stationary point. □

Lemma 41 (Bandwidth cancellation)**.** *At the optimal bandwidths of Lemma 40, with* $\kappa' := \frac{\rho}{\nu_2 c_2} - \frac{1-\rho}{\nu_1 c_1}$ *(the coefficient in* $\mathrm{Cov}(\widetilde{N}_2 - \widetilde{N}_1, N_\rho) = \kappa' V$*),*

$$\frac{\kappa'}{\sigma_\rho^2} = \frac{c_2^2 - c_1^2}{A}, \qquad \sigma_\rho^2 - \kappa'\rho = \frac{1-\rho}{\nu_1 c_1}, \qquad \operatorname{sign}\kappa' = \operatorname{sign}(\rho - \nu_2). \tag{130}$$

Proof. The Lagrange conditions read $\frac{1-\rho}{\nu_1 c_1} = 2\mu_L c_1^2$ and $\frac{\rho}{\nu_2 c_2} = 2\mu_L c_2^2$. Subtracting gives $\kappa' = 2\mu_L(c_2^2 - c_1^2)$; the weighted sum gives $\sigma_\rho^2 = 2\mu_L A$. The second identity is direct. For the sign, $\frac{\rho}{\nu_2 c_2} = \sqrt{S/A}(\rho/\nu_2)^{2/3}$ and $\frac{1-\rho}{\nu_1 c_1} = \sqrt{S/A}((1-\rho)/\nu_1)^{2/3}$, so $\kappa' < 0 \iff \rho/\nu_2 < (1-\rho)/\nu_1 \iff \rho < \nu_2$. □

Lemma 42 (Existence and interiority)**.** *Let* $b \neq 0$ *and* $r > 0$*.* $\bar{F}(\cdot; r)$ *attains its minimum on* $[0,1] \times (0,\infty)$*; every minimizer* $(\rho^\star, A^\star)$ *satisfies* $0 < \rho^\star < \nu_2$*; and* $\bar{C}^\star_{\mathrm{WE}}(r) < \bar{C}^\star_{\mathrm{gold}} := \min_{c_1} \mathbb{E}\|c_1^2 b + \widetilde{N}_1\|$*.*

Proof. Existence: $\bar{F} \to \infty$ as $A \to 0$ (the noise term diverges) and as $A \to \infty$ ($\|Ab\| \to \infty$ while the noise is $O(A^{-1/4})$). $\rho^\star > 0$: at $(0, A)$ take $c_2 = c_1$ and differentiate at fixed bandwidths, $\partial_\rho \bar{C} = \mathbb{E}[\widehat{U}_\rho^\top D]$, $D = r\Theta + \widetilde{N}_2 - \widetilde{N}_1$. At $\rho = 0$ the Θ and $\widetilde{N}_2$ terms vanish by independence, and $\mathbb{E}[U_1^\top \widetilde{N}_1/\|U_1\|] = \mathrm{tr}(\Sigma_1 J) > 0$, so $\partial_\rho \bar{C}|_{\rho=0} < 0$. Taking c_1 optimal for gold-only gives the last claim. $\rho^\star < 1$: at $\rho = 1$ the risk does not depend on c_1; set $c_1 = c_2 = c$. Given Θ, with $\mu_\Theta = c^2 b + r\Theta$ and $\phi = \phi_{\Sigma_2}$,

$$\mathbb{E}[\|U_2\| - U_2^\top c^2 b/\|U_2\| \mid \Theta] = \phi(\mu_\Theta) - \mu_\Theta^\top \nabla\phi(\mu_\Theta) + r\Theta^\top \nabla\phi(\mu_\Theta).$$

The first difference is $\mathrm{tr}(\Sigma_2 J) > 0$; for the last term, $\tau \mapsto \phi(c^2 b + \tau\Theta)$ is convex, so $\Theta^\top \nabla\phi(c^2 b + r\Theta) \geq \Theta^\top \nabla\phi(c^2 b)$, whose Θ-average is zero. Hence $\partial_\rho \bar{C}|_{\rho=1} > 0$. $\rho^\star < \nu_2$: at an interior minimizer the ρ-derivative at fixed bandwidths vanishes, $\mathbb{E}[\widehat{U}_0^\top D] = 0$. Identity (132) below with $M = I$ gives

$$\frac{r(1-\rho)}{\nu_1 c_1}\,\mathbb{E}[\widehat{U}_0^\top \Theta] = -\kappa'\,\mathbb{E}\|U_0\|, \tag{131}$$

and $\mathbb{E}[\widehat{U}_0^\top \Theta] = \mathbb{E}_\Theta[\nabla\phi(Ab + \rho r\Theta)^\top \Theta] > \mathbb{E}_\Theta[\nabla\phi(Ab)^\top \Theta] = 0$ by strict convexity. Thus $\kappa' < 0$ and Lemma 41 gives $\rho^\star < \nu_2$. □

E.3 Keeping the weight and the bandwidths, tuning only the penalty

Fix a WE minimizer (ρ, c_1, c_2), let $U_0 = Ab + \rho r\Theta + N_\rho$ (the main text's U), $D = U_2 - U_1 = r\Theta + (c_2^2 - c_1^2)b + \widetilde{N}_2 - \widetilde{N}_1$, $W_t = \rho(I + tH)^{-1}$, $U_t = U_1 + W_t D$, $g(t) = \mathbb{E}\|U_t\|$ for $t \in [0,\infty]$. Then $g(0) = \bar{C}^\star_{\mathrm{WE}}(r)$ and $g(\infty) = \mathbb{E}\|U_1\| \geq \bar{C}^\star_{\mathrm{gold}}$.

Lemma 43. $g \in C^1[0,\infty)$ *with* $g'(t) = -\rho\,\mathbb{E}[\widehat{U}_t^\top H(I + tH)^{-2} D]$*, and for every matrix* M

$$\mathbb{E}[\widehat{U}_0^\top M D] = \frac{r(1-\rho)}{\nu_1 c_1 \sigma_\rho^2}\,\mathbb{E}[\widehat{U}_0^\top M\Theta] + \frac{\kappa'}{\sigma_\rho^2}\,\mathbb{E}[\widehat{U}_0^\top M U_0]. \tag{132}$$

Proof. Differentiation under the expectation is justified by $|\partial_t \|U_t\|| \leq \rho\|H\|\|D\|$. For (132), the pair $(\widetilde{N}_2 - \widetilde{N}_1, N_\rho)$ is jointly Gaussian and independent of Θ, so $\widetilde{N}_2 - \widetilde{N}_1 = \frac{\kappa'}{\sigma_\rho^2} N_\rho + \widetilde{D}^\perp$ with $\widetilde{D}^\perp$ centered and independent of (N_ρ, Θ). Hence $\mathbb{E}[\widehat{U}_0^\top M D] = \mathbb{E}[\widehat{U}_0^\top M\{r\Theta + (c_2^2 - c_1^2)b\}] + \frac{\kappa'}{\sigma_\rho^2}\mathbb{E}[\widehat{U}_0^\top M(U_0 - Ab - \rho r\Theta)]$. The coefficient of Θ is $r(1 - \kappa'\rho/\sigma_\rho^2) = \frac{r(1-\rho)}{\nu_1 c_1 \sigma_\rho^2}$ and the coefficient of b is $c_2^2 - c_1^2 - \kappa' A/\sigma_\rho^2 = 0$, both by Lemma 41. □

Define the two curvature averages

$$\bar{h}_{\mathrm{bias}} = \frac{\mathbb{E}[\widehat{U}_0^\top H\Theta]}{\mathbb{E}[\widehat{U}_0^\top \Theta]}, \qquad \bar{h}_{\mathrm{noise}} = \frac{\mathbb{E}[\widehat{U}_0^\top H U_0]}{\mathbb{E}\|U_0\|} \in [\lambda_{\min}(H), \lambda_{\max}(H)]. \tag{133}$$

Theorem 21 (Exact criterion). *Let $b \neq 0$, $r > 0$, $p \geq 2$. At any WE minimizer,*

$$g'(0) = -\frac{\rho(1-\rho)\, r}{\nu_1 c_1 \sigma_\rho^2}\, \mathbb{E}[\widehat{U}_0^\top \Theta]\, \big(\bar{h}_{\mathrm{bias}} - \bar{h}_{\mathrm{noise}}\big), \qquad \mathbb{E}[\widehat{U}_0^\top \Theta] > 0. \tag{134}$$

Hence $g'(0) < 0$ if and only if $\bar{h}_{\mathrm{bias}} > \bar{h}_{\mathrm{noise}}$. If H is a scalar matrix the two averages coincide and $g'(0) = 0$.

Proof. Apply (132) with $M = H$ and eliminate κ' through (131). □

Theorem 22 (Interior optimal penalty). *Under the assumptions of Theorem 21, if $\bar{h}_{\mathrm{bias}} > \bar{h}_{\mathrm{noise}}$ then g attains its minimum over $[0, \infty]$ at some $t_\star \in (0, \infty)$, $\lambda_\star = \rho(1-\rho)/(2t_\star) \in (0, \infty)$, and*

$$\bar{C}^\star_{\mathrm{THE}}(r) \leq g(t_\star) < g(0) = \bar{C}^\star_{\mathrm{WE}}(r) < \bar{C}^\star_{\mathrm{gold}}. \tag{135}$$

The point $t_\star$ solves $\mathbb{E}[\widehat{U}_t^\top H(I + tH)^{-2} D] = 0$. Moreover $g \in C^2[0, \infty)$ with $g''(t) = \mathbb{E}[\dot{U}_t^\top J_{U_t} \dot{U}_t + \widehat{U}_t^\top \ddot{U}_t]$, where $\dot{U}_t = -\rho H(I + tH)^{-2} D$, $\ddot{U}_t = 2\rho H^2 (I + tH)^{-3} D$ and $J_x = (I - \widehat{x}\widehat{x}^\top)/\|x\|$; the supremum $\sup_{t \geq 0} g''(t)$ is finite for $p \geq 2$, and $g(0) - g(t_\star) \geq g'(0)^2/(2 \sup_{t \geq 0} g''(t))$. The small-$r$ expansion is in Corollary 5.

Proof. g is continuous on $[0, \infty]$, $g(\infty) \geq \bar{C}^\star_{\mathrm{gold}} > g(0)$ by Lemma 42, and $g'(0) < 0$; so the minimum is interior and stationary. For the second derivative, write $\phi(s) = \|U_s\|$. Conditionally on $(\Theta, \widetilde{N}_2)$, $U_s = (I - W_s)\widetilde{N}_1 + R_s$ with $R_s = c_1^2 b + W_s(D + \widetilde{N}_1)$ fixed, so U_s is Gaussian with covariance $(I - W_s)V(I - W_s)/(\nu_1 c_1) \succeq \sigma^2 I$, $\sigma^2 = (1-\rho)^2 \lambda_{\min}(V)/(\nu_1 c_1)$. By Anderson's inequality, whose hypotheses hold because the superlevel sets of $x \mapsto \|x\|^{-q}$ are balls, $\mathbb{E}[\|U_s\|^{-q} \mid \Theta, \widetilde{N}_2] \leq \sigma^{-q}\, \mathbb{E}\|G_1\|^{-q}$, which is finite for $q < p$; take $q = 3/2$. Since $\|D\| \leq \|D + \widetilde{N}_1\| + \|\widetilde{N}_1\|$ with $D + \widetilde{N}_1$ measurable in $(\Theta, \widetilde{N}_2)$ and $\|\widetilde{N}_1\| \leq (1-\rho)^{-1}(\|U_s\| + \|R_s\|)$, this bounds $\sup_{s \geq 0} \mathbb{E}[(\|D\|^2/\|U_s\|)^{3/2}]$ by a continuous function of (ρ, c_1, c_2, r). For $p \geq 2$ the path $s \mapsto U_s$ avoids the origin almost surely (it hits 0 only if $\widetilde{N}_1$ lies on a smooth curve), so ϕ is smooth with $|\phi''(s)| \leq \rho^2 \|H\|^2 \|D\|^2/\|U_s\| + 2\rho \|H\|^2 \|D\|$. Hence $\{\phi''(s)\}_{s \geq 0}$ is uniformly integrable, $s \mapsto \mathbb{E}\phi''(s)$ is continuous and bounded, and integrating $\phi'(t) - \phi'(s) = \int_s^t \phi''$ (Fubini) together with Lemma 43 gives $g \in C^2$, $g'' = \mathbb{E}\phi''$ and $\sup g'' < \infty$. The gain bound is $g(t) \leq g(0) + g'(0)t + \frac{1}{2} \sup g''\, t^2$ minimized over t. □

Remark 7. *$\bar{h}_{\mathrm{noise}}$ is the curvature along the direction of the WE error U_0, averaged with weight $\|U_0\|$, which puts more weight on noisy directions; $\bar{h}_{\mathrm{bias}}$ is the curvature along the bias direction Θ, averaged through the inner product with $\widehat{U}_0$, and the $1/\|U_0\|$ weighting emphasizes small-error realizations. The criterion characterizes the sign of the derivative at infinite penalty and gives a sufficient condition for a finite improving penalty. It is not a global necessity claim or a general equivalence with reverse eigenvalue ordering. The smoothing bias b does not appear explicitly: it has been canceled by Lemma 41 and only enters through the mean Ab of U_0.*

Lemma 44 (Interval of weights). *Fix $r > 0$. For any (ρ, c_1, c_2) let $g_{\rho,c}(t) = \bar{C}(\rho(I + tH)^{-1}, c_1, c_2; r)$; then $g'_{\rho,c}(0) = -\rho\, \mathbb{E}[\widehat{U}^\top H D]$ with $U = U_1 + \rho D$, a continuous function of (ρ, c_1, c_2). Suppose that at $\rho^\star$ every bandwidth pair that is optimal for WE at weight $\rho^\star$ has $g'(0) < 0$. Then there is an open interval containing $\rho^\star$ such that for every ρ in it and every bandwidth pair $c^{\mathrm{WE}}(\rho)$ that is optimal for WE at weight ρ, $g'_{\rho, c^{\mathrm{WE}}(\rho)}(0) < 0$; hence some $\lambda \in (0, \infty)$ makes THE at $(\rho, \lambda, c^{\mathrm{WE}}(\rho))$ strictly better than WE at $(\rho, c^{\mathrm{WE}}(\rho))$.*

Proof. The derivative formula is the chain rule with $\frac{d}{dt}\rho(I + tH)^{-1}|_{t=0} = -\rho H$ and Lemma 38; continuity is Gaussian smoothing. By the coercivity in the proof of Lemma 42, uniform for ρ in a compact subinterval of $(0, 1)$, the bandwidth pairs that are optimal for WE at weights near $\rho^\star$ lie in one compact set, and the argmin correspondence is upper semicontinuous. If the claim failed there

would be $\rho_n \to \rho^\star$ and pairs c_n optimal for WE at ρ_n with $g'_{\rho_n,c_n}(0) \geq 0$; a subsequence converges to a pair c^∞ that is optimal for WE at $\rho^\star$, and continuity gives $g'_{\rho^\star,c^\infty}(0) \geq 0$, a contradiction. A negative one-sided derivative at $t = 0$ gives $g(t) < g(0)$ for some $t > 0$, and $\lambda = \rho(1-\rho)/(2t)$. □

E.4 Small bias

Lemma 45. *S is strictly convex on $[0,1]$ with unique minimizer $\rho = \nu_2$ and $S(\nu_2) = 1$, and $A \mapsto \phi_{A^{-1/2}V}(Ab)$ has a unique minimizer $A_0 > 0$. Hence at $r = 0$ the WE optimum is (ν_2, A_0), with equal bandwidths $c_1 = c_2 = A_0^{1/2}$ and covariance $\Sigma_0 = A_0^{-1/2}V$.*

Proof. $S''(\rho) = \frac{10}{9}[(1-\rho)^{-1/3}\nu_1^{-2/3} + \rho^{-1/3}\nu_2^{-2/3}] > 0$ on $(0,1)$, and $S'(\rho) = 0$ iff $(\rho/\nu_2)^{2/3} = ((1-\rho)/\nu_1)^{2/3}$, i.e. $\rho\nu_1 = (1-\rho)\nu_2$, i.e. $\rho = \nu_2$; there $S(\nu_2) = \nu_1 + \nu_2 = 1$. At $r = 0$ Lemma 40 gives $\bar{F}(\rho, A; 0) = \phi_{S(\rho)^{3/2}A^{-1/2}V}(Ab)$, which by Lemma 38(vi) is $\geq \phi_{A^{-1/2}V}(Ab)$ with equality iff $S(\rho) = 1$, i.e. iff $\rho = \nu_2$. The map $A \mapsto \phi_{A^{-1/2}V}(Ab) = F(Ab, A^{-1/4})$ with $F(\mu, s) = \mathbb{E}\|\mu + sV^{1/2}G\|$ is strictly convex on $(0,\infty)$: F is jointly convex (the norm of an affine map) and strictly increasing in s (Lemma 38(vi)), $A \mapsto Ab$ is affine and $A \mapsto A^{-1/4}$ is strictly convex, as in the proof of Theorem 26. With the coercivity used in the proof of Lemma 42 it has a unique minimizer A_0. Finally the bandwidth formulas of Lemma 40 at $\rho = \nu_2$ read $c_1 = c_2 = A_0^{1/2}$ and $\sigma_{\nu_2}^2 = A_0^{-1/2}$, i.e. $\Sigma_0 = A_0^{-1/2}V$. □

Let $Y = A_0 b + N_0$, $N_0 \sim \mathcal{N}(0, \Sigma_0)$, $m_0 = m_{\Sigma_0}(A_0 b)$, $J_0 = J_{\Sigma_0}(A_0 b)$, $\bar{h}^{(0)}_{\mathrm{noise}} = \mathbb{E}[Y^\top HY/\|Y\|]/\mathbb{E}\|Y\|$, and

$$\Delta h(0) := \frac{\operatorname{tr}(HJ_0)}{\operatorname{tr} J_0} - \frac{A_0\, m_0^\top Hb + \operatorname{tr}(H\Sigma_0 J_0)}{A_0\, m_0^\top b + \operatorname{tr}(\Sigma_0 J_0)} = \frac{\operatorname{tr}(HJ_0)}{\operatorname{tr} J_0} - \frac{\mathbb{E}[Y^\top HY/\|Y\|]}{\mathbb{E}\|Y\|}. \tag{136}$$

Theorem 23 (Small-bias criterion)**.** *Let $b \neq 0$, $p \geq 2$. As $r \to 0$ every WE minimizer converges to (ν_2, A_0) and*

$$\begin{aligned} &\bar{h}_{\mathrm{bias}} - \bar{h}_{\mathrm{noise}} = \Delta h(0) + O(r^2), \\ &g'(0) = -\frac{\nu_2^2 r^2 \operatorname{tr} J_0}{p}\, \Delta h(0) + O(r^4), \qquad c_0 = A_0^{1/2}. \end{aligned} \tag{137}$$

If $\Delta h(0) > 0$ there is $r_0 > 0$ such that Theorem 22 applies for all $0 < r < r_0$; if $\Delta h(0) < 0$ then $g'(0) > 0$ for small r.

Proof. Coercivity and upper semicontinuity give convergence of every minimizer. To justify its rate, put $F(\mu, s) = \mathbb{E}\|\mu + sV^{1/2}G\|$. It is jointly convex and $F_s > 0$ for $s > 0$, by the Gaussian heat identity. At $r = 0$, $\bar{F}(\rho, A; 0) = F(Ab, S(\rho)^{3/4}A^{-1/4})$. At $\rho = \nu_2$, $S' = 0$ and $S'' > 0$, so $\bar{F}_{\rho\rho} > 0$ and $\bar{F}_{\rho A} = 0$. The A second derivative is the sum of a nonnegative Hessian quadratic form of F and $F_s(5/16)A^{-9/4} > 0$ at $S(\nu_2) = 1$. Thus the Hessian in (ρ, A) is positive definite. Gaussian smoothing and symmetry in Θ make the objective smooth and even in r near this point. The implicit function theorem, together with localization of every global minimizer, gives $(\rho_r, A_r) = (\nu_2, A_0) + O(r^2)$ and the same order for $c_k - c_0$. For the expansion, $\mathbb{E}[\widehat{U}_0^\top \Theta] = \mathbb{E}_\Theta[\nabla\phi(A_r b + \rho_r r\Theta)^\top \Theta] = \frac{\rho_r r}{p} \operatorname{tr} J_r + O(r^3)$ (the second-order term is odd in Θ), likewise $\mathbb{E}[\widehat{U}_0^\top H\Theta] = \frac{\rho_r r}{p} \operatorname{tr}(HJ_r) + O(r^3)$, and Gaussian smoothing gives $\bar{h}_{\mathrm{noise}} = \bar{h}^{(0)}_{\mathrm{noise}} + O(r^2)$. Insert these expansions and the $O(r^2)$ parameter displacement in (134). The second form of (136) is Lemma 38(v). □

Lemma 46 (Weighted Chebyshev)**.** *If $J = \operatorname{diag}(J_i)$ with $J_i > 0$, $h_1 \geq \cdots \geq h_p$, $v_1 \leq \cdots \leq v_p$ and $\Sigma = \sigma^2 \operatorname{diag}(v_i)$, then $\operatorname{tr}(HJ)\operatorname{tr}(\Sigma J) \geq \operatorname{tr} J \operatorname{tr}(H\Sigma J)$, strictly when both sequences are nonconstant.*

Proof. $\sum_{i,j} J_i J_j (h_i - h_j)(v_i - v_j) \leq 0$ term by term. □

Theorem 24 (Sufficient conditions and the reverse case)**.** *Let H, V commute with common eigenbasis $\{e_i\}$, $He_i = h_i e_i$, $Ve_i = v_i e_i$.*

1. *For a centered Gaussian with a fixed covariance scale $A_0 > 0$, the trace difference in* (136) *with its mean terms removed is positive under reverse ordering and negative under the same ordering, strictly when H, V are nonscalar. This is a fixed-bandwidth covariance calculation; it does not assert existence of a bandwidth optimum when $b = 0$.*

2. *If $b \parallel e_j$ then J_0 is diagonal, $m_0 \parallel e_j$, and $\Delta h(0) > 0$ if and only if*

$$A_0 m_0^\top b\,[\operatorname{tr}(HJ_0) - h_j \operatorname{tr} J_0] + [\operatorname{tr}(HJ_0)\operatorname{tr}(\Sigma_0 J_0) - \operatorname{tr} J_0 \operatorname{tr}(H\Sigma_0 J_0)] > 0. \tag{138}$$

   Under reverse ordering and $h_j \le \operatorname{tr}(HJ_0)/\operatorname{tr} J_0$ (in particular $h_j = h_{\min}$) both brackets are nonnegative and the second is positive.

3. *For general b, $\Delta h(0) > 0$ is the explicit inequality* (136).

Proof. (1) Lemma 46; J_0 is diagonal by reflection symmetry of a centered Gaussian with diagonal covariance. (2) Reflections of the coordinates $i \neq j$ preserve the law of Y, so J_0 is diagonal and $m_0 \parallel e_j$, whence $m_0^\top Hb = h_j m_0^\top b$; clear denominators in (136). $m_0^\top b > 0$ by Lemma 38(iv). □

Corollary 5 (Asymptotics of the optimal penalty)**.** *Under Theorem 23 with $\Delta h(0) > 0$,*

$$\begin{aligned} g''(0) &\to \frac{\nu_2}{\nu_1 c_0} \operatorname{tr}(HJ_0HV), \qquad g(0) - g(t_\star) \asymp r^4, \\ \lambda_\star &= \frac{p \operatorname{tr}(HJ_0HV)}{2c_0 r^2 \operatorname{tr} J_0\, \Delta h(0)}\{1 + o(1)\} \asymp r^{-2}. \end{aligned} \tag{139}$$

When $V = \kappa H^{-1}$ this is $\lambda_\star \approx \frac{p\kappa}{2c_0 r^2} \cdot \frac{\operatorname{tr}(HJ_0)}{\operatorname{tr} J_0\, \Delta h(0)}$, the squared-loss Bayes value of Appendix J times a factor larger than one.

Proof. At $r \to 0$: $\rho \to \nu_2$, $\kappa' \to 0$, $c_2^2 - c_1^2 \to 0$, so D and U_0 become independent with $\mathbb{E}D \to 0$; the integrands converge almost surely and are uniformly integrable by the bound in the proof of Theorem 22, so $\mathbb{E}[\widehat{U}_0^\top \ddot{U}_0] \to 0$ and $\mathbb{E}[\dot{U}_0^\top J_{U_0} \dot{U}_0] \to \nu_2^2 \operatorname{tr}(HJ_0H \operatorname{Cov} D)$ with $\operatorname{Cov} D \to V/(\nu_1\nu_2 c_0)$, giving the stated positive limit of $g''(0)$. It remains to locate the optimal penalty along this fixed-WE-parameter path. At $r = 0$, the common-bandwidth mean is $A_0 b$, and completing the covariance square gives

$$\Sigma_W = V/c_0 + (W - \nu_2 I)V(W - \nu_2 I)/(\nu_1\nu_2 c_0).$$

Along $W = \nu_2(I + tH)^{-1}$ this has strictly greater Gaussian norm risk for every $t > 0$, including the limit $t = \infty$. As the WE parameters tend to (ν_2, c_0, c_0), U_t is affine in (G_1, G_2, Θ) with coefficients converging uniformly in $t \in [0, \infty]$ ($W_\infty = 0$), so g converges uniformly on the compactified path to its $r = 0$ value, whose infimum over $t \ge \epsilon$ exceeds its value at 0 for every $\epsilon > 0$ by continuity and compactness. Hence every global path minimizer $t_\star(r)$ tends to zero. Moreover g'' is continuous at 0 uniformly in small r: by $\|U_s - U_0\| \le s\rho\|H\|\|D\|$, outside the event that $\|U_0\| < 2\eta$ or $s\rho\|H\|\|D\| > \eta$ the path stays at distance at least η from the origin on $[0, s]$, where ϕ'' is Lipschitz in s with a constant that is a cubic polynomial in $\|D\|$; that event has small probability uniformly in r (Gaussian anticoncentration of U_0, as in the proof of Theorem 22, and the Gaussian tail of D), and its contribution is controlled by the uniform integrability there; let $s \to 0$ and then $\eta \to 0$. Hence $g'' > 0$ near 0 for small r, and the interior first-order condition $0 = g'(t_\star) = g'(0) + t_\star\{g''(0) + o(1)\}$ with $g(0) - g(t_\star) = -\int_0^{t_\star} g'$ yield

$$t_\star = -\frac{g'(0)}{g''(0)}\{1 + o(1)\}, \qquad g(0) - g(t_\star) = \frac{g'(0)^2}{2g''(0)}\{1 + o(1)\}.$$

Theorem 23 now gives the result, including $\lambda_\star = \rho(1 - \rho)/(2t_\star)$. These are path optima with the WE-optimal weight and bandwidths retained, not an assertion about the penalty at a jointly optimized THE point. □

E.5 Family structure and the full THE infimum

Proposition 6. *(i) $\mathcal{W}_{\mathrm{WE}} \subset \overline{\mathcal{W}_{\mathrm{THE}}}$ $(t = 0)$. (ii) For every $s > 0$, $\rho(I + H/s)^{-1} \to s(H + sI)^{-1}$ as $\rho \uparrow 1$ with $\lambda = \rho(1-\rho)s/2 \downarrow 0$, each point being an admissible THE parameter; so $\mathcal{W}_{\mathrm{TSE}} \subset \overline{\mathcal{W}_{\mathrm{THE}}}$. Equivalently,*

$$W_{\mathrm{THE}}(\rho, \lambda) = \rho\, W_{\mathrm{TSE}}\Big(\rho_{\mathrm{TSE}} = \frac{\lambda}{\rho(1-\rho)}\Big). \tag{140}$$

(iii) $\bar{C}$ is continuous in W, hence $\bar{C}^{\star}_{\mathrm{THE}}(r) \le \min\{\bar{C}^{\star}_{\mathrm{WE}}(r), \bar{C}^{\star}_{\mathrm{TSE}}(r), \bar{C}^{\star}_{\mathrm{gold}}\}$. (iv) If H is scalar the three families have identical sets of attainable constants and $\bar{C}^{\star}_{\mathrm{THE}} = \bar{C}^{\star}_{\mathrm{WE}} = \bar{C}^{\star}_{\mathrm{TSE}}$.

Proof. (i) Taking $t = 0$ gives $W = \rho I$ for every $\rho \in (0, 1)$, and $\rho \in \{0, 1\}$ lies in the closure. (ii) For fixed $s > 0$ take $t = 1/s$: then $\rho(I + H/s)^{-1} \to (I + H/s)^{-1} = s(H + sI)^{-1}$ as $\rho \uparrow 1$, while $\lambda = \rho(1-\rho)/(2t) = \rho(1-\rho)s/2 \downarrow 0$, and every (ρ, t) on this path is an admissible THE parameter. For (140), substituting $s = 2\rho_{\mathrm{TSE}} = 2\lambda/(\rho(1-\rho))$ into $s(H + sI)^{-1} = (I + H/s)^{-1}$ produces $(I + \frac{\rho(1-\rho)}{2\lambda} H)^{-1}$, whose ρ-multiple is $W_{\mathrm{THE}}(\rho, \lambda)$. (iii) $\|U_W - U_{W'}\| \le \|W - W'\|\,\|D\|$ with $\mathbb{E}\|D\| < \infty$, so at fixed bandwidths $\bar{C}$ is Lipschitz in W; hence the infimum over $\mathcal{W}_{\mathrm{THE}}$ equals the infimum over its closure, which by (i), (ii) and $W = 0$ (the limit $t \to \infty$) contains $\mathcal{W}_{\mathrm{WE}}$, $\mathcal{W}_{\mathrm{TSE}}$ and gold-only. (iv) If $H = \mathsf{h} I$ then $\mathcal{W}_{\mathrm{THE}} = \{\frac{\rho}{1+t\mathsf{h}} I\}$ and $\mathcal{W}_{\mathrm{TSE}} = \{\frac{s}{\mathsf{h}+s} I\}$ are contained in $\mathcal{W}_{\mathrm{WE}}$ and each realizes every scalar weight in $(0, 1)$, over the same range of bandwidths; so the three attainable sets of constants coincide. □

Lemma 47 (Model averaging)**.** *Let $\widehat{\beta}_{\mathrm{avg}}(\rho)$ be the normalized convex combination $((1-\rho)\widehat{\beta}_{\mathrm{gold}} + \rho\widehat{\beta}_{\mathrm{proxy}})/\|(1-\rho)\widehat{\beta}_{\mathrm{gold}} + \rho\widehat{\beta}_{\mathrm{proxy}}\|$ with $\rho \in (0, 1)$ fixed. Under the assumptions of Theorem 20, $n^{2/5}\,\mathbb{E}\angle(\widehat{\beta}_{\mathrm{avg}}(\rho), \beta^{\star}) \to \mathbb{E}\|(1-\rho)U_1 + \rho U_2\|$, the constant of WE with the same weight.*

Proof. Write $\widehat{\beta}_{\mathrm{gold}} = \phi(\widehat{v}_1)$ and $\widehat{\beta}_{\mathrm{proxy}} = \phi(\widehat{v}_2)$ with $\phi(v) = \sqrt{1 - \|v\|^2}\beta^{\star} + Ev$, so that the tangential coordinate of $\phi(v)$ is v and $\angle(\phi(v), \beta^{\star}) = \arcsin\|v\|$. On the good event $\mathcal{E}_n$ both $\|\widehat{v}_k\| \to 0$ (Lemma 28) and $\widehat{v}_k = O_P(r_n)$ (Lemma 33). Hence $(1-\rho)\phi(\widehat{v}_1) + \rho\phi(\widehat{v}_2) = (1 - O(\|\widehat{v}\|^2))\beta^{\star} + E\,w$ with $w = (1-\rho)\widehat{v}_1 + \rho\widehat{v}_2$, its norm is $1 + O(\|\widehat{v}\|^2)$, and the tangential coordinate of the normalized vector is $w + O(\|\widehat{v}\|^3)$. Since $r_n^{-1}(\widehat{v}_1, \widehat{v}_2) \Rightarrow (U_1, U_2)$ by (103), the continuous mapping theorem gives $r_n^{-1}\angle(\widehat{\beta}_{\mathrm{avg}}, \beta^{\star}) \Rightarrow \|(1-\rho)U_1 + \rho U_2\|$, and uniform integrability follows as in Lemma 36: on $\mathcal{E}_n$ the angle is at most $\|\widehat{v}_1\| + \|\widehat{v}_2\|$ up to a factor $1 + o(1)$, and $\alpha \le \pi$ off it. □

Theorem 25 (Strict improvement)**.** *Let $b \neq 0$, $p \ge 2$, $r > 0$, and suppose $\bar{h}_{\mathrm{bias}} > \bar{h}_{\mathrm{noise}}$ at some WE minimizer. Then $\bar{C}^{\star}_{\mathrm{THE}}(r) \le g(t_{\star}) < \bar{C}^{\star}_{\mathrm{WE}}(r) < \bar{C}^{\star}_{\mathrm{gold}}$, where $g(t_{\star})$ is attained by the admissible THE parameter that keeps the WE weight and bandwidths and uses $\lambda_{\star} \in (0, \infty)$.*

Proof. Theorem 22 and Lemma 42. □

By Theorem 23, the hypothesis of Theorem 25 holds whenever $\Delta h(0) > 0$ and $r < r_0$. Moreover $\bar{C}^{\star}_{\mathrm{THE}}(r) > 0$. Indeed $\bar{C} \ge \|\mathbb{E}U\|$ and $\bar{C} \ge \mathbb{E}\|N_W\|$, and along any sequence with $\bar{C} \to 0$, writing $u = b/\|b\|$ and $\alpha = u^{\top}(I - W)u \in [0, 1]$, we get $u^{\top}\mathbb{E}U = \|b\|[c_1^2\alpha + c_2^2(1-\alpha)] \to 0$ while $\operatorname{tr}\Sigma_W \ge \lambda_{\min}(V)[\alpha^2/(\nu_1 c_1) + (1-\alpha)^2/(\nu_2 c_2)] \to 0$, which forces both $\alpha \to 0$ and $1 - \alpha \to 0$, a contradiction.

Theorem 26 (Bandwidths for a fixed weight matrix)**.** *Fix a symmetric $0 \prec W \prec I$ and $b \neq 0$. In the variables $a_k = c_k^2$ the map $(a_1, a_2) \mapsto \bar{C}$ is jointly strictly convex on $(0, \infty)^2$ with a unique minimizer, characterized by*

$$(c_k^{\star})^3 = \frac{\operatorname{tr}(W_k V W_k J)}{4\nu_k\, m^{\top} W_k b}, \qquad W_1 = I - W,\ W_2 = W,\ k = 1, 2, \tag{141}$$

where $m = \mathbb{E}\widehat{U}$ and $J = \mathbb{E}[(I - \widehat{U}\widehat{U}^{\top})/\|U\|]$ are evaluated at the minimizer and the denominators are positive there.

Proof. Given (Θ, G_1, G_2), $U = W_1 a_1 b + W_2 a_2 b + rW\Theta + a_1^{-1/4}\xi_1 + a_2^{-1/4}\xi_2$ with $\xi_k = W_k V^{1/2} G_k/\sqrt{\nu_k}$. The function $F(\mu, s_1, s_2) = \mathbb{E}\|\mu + rW\Theta + s_1\xi_1 + s_2\xi_2\|$ is jointly convex (norm of an affine map) and strictly increasing in each $s_k \geq 0$ (Lemma 38(vi)). Since $\mu(a)$ is affine and $a_k^{-1/4}$ is strictly convex, $\bar{C}(a) = F(\mu(a), a_1^{-1/4}, a_2^{-1/4})$ is convex, and strictly convex along every direction $\delta a \neq 0$ because some $\delta a_k \neq 0$ and F is strictly increasing in s_k. Coercivity: as $a_k \to 0$ the noise term diverges; as $a_k \to \infty$, $\mu(a)^\top b = a_1 b^\top (I - W)b + a_2 b^\top W b \to \infty$. For the first-order condition, $\partial_{c_k} U = 2c_k W_k b - \frac{1}{2c_k} W_k \widetilde{N}_k$ and $\mathbb{E}[\widehat{U}^\top W_k \widetilde{N}_k] = \frac{1}{\nu_k c_k} \operatorname{tr}(W_k V W_k J)$ by Lemma 38(v). □

Theorem 27 (Common bandwidth: criterion and localization)**.** *Let $b \neq 0$, H be nonscalar, and restrict the families to $c_1 = c_2 = c$. Use $\bar{C}^{\mathrm{c}}_{\mathcal{M}}$ for these common-bandwidth infima. At any interior common-bandwidth WE optimum let $U = c^2 b + \rho r\Theta + N_\rho$, $\bar{J} = \mathbb{E}[(I - \widehat{U}\widehat{U}^\top)/\|U\|]$, and $a_M = \mathbb{E}[\widehat{U}^\top M\Theta]$. Then $a_I > 0$ and*

$$g'(0) = -\rho r a_I \left\{ \frac{a_H}{a_I} - \frac{\operatorname{tr}(HV\bar{J})}{\operatorname{tr}(V\bar{J})} \right\}, \qquad \Delta_{\mathrm{c}} := \frac{\operatorname{tr}(HJ_0)}{\operatorname{tr} J_0} - \frac{\operatorname{tr}(HVJ_0)}{\operatorname{tr}(VJ_0)}. \tag{142}$$

As $r \to 0$, all common-bandwidth THE minimizers over the closure satisfy $W \to \nu_2 I$, $c \to c_0 = A_0^{1/2}$. For sufficiently small r: (i) $\bar{C}^{\mathrm{c}}_{\mathrm{THE}}(r) < \bar{C}^{\mathrm{c}}_{\mathrm{TSE}}(r)$; (ii) if $\Delta_{\mathrm{c}} < 0$, every THE minimizer has $t = 0$ and $\bar{C}^{\mathrm{c}}_{\mathrm{THE}} = \bar{C}^{\mathrm{c}}_{\mathrm{WE}}$; (iii) if $\Delta_{\mathrm{c}} > 0$, a finite penalty strictly improves on common-bandwidth WE. No conclusion for $\Delta_{\mathrm{c}} = 0$ is asserted. If H, V commute, neither is scalar, their eigenvalues are in reverse order, and b is along a common eigenvector, then $\Delta_{\mathrm{c}} > 0$.

Proof. Put $\kappa = \rho/(\nu_2 c) - (1-\rho)/(\nu_1 c)$ and $D = r\Theta + \widetilde{N}_2 - \widetilde{N}_1$. Gaussian integration by parts gives $\mathbb{E}[\widehat{U}^\top M D] = r a_M + \kappa \operatorname{tr}(MV\bar{J})$ for $M = I, H$. Interior weight stationarity is $r a_I + \kappa \operatorname{tr}(V\bar{J}) = 0$. Substitution in $g'(0) = -\rho\mathbb{E}[\widehat{U}^\top H D]$ proves (142); strict convexity in the conditional mean and symmetry of Θ give $a_I > 0$. Independent-bandwidth cancellation has not been used.

At $r = 0$ the mean $c^2 b$ is independent of W and

$$\Sigma_W = V/c + (W - \nu_2 I)V(W - \nu_2 I)/(\nu_1 \nu_2 c).$$

The added covariance is nonzero for $W \neq \nu_2 I$, hence strict Gaussian convexity makes $W = \nu_2 I$ the unique minimizing weight. The bandwidth minimum is unique (Lemma 45) and nondegenerate (the positive second derivative in $A = c^2$ in the proof of Theorem 23). The bounds $\bar{C} \geq c^2\|b\|$ (Jensen) and $\bar{C} \geq \mathbb{E}\|N_W\| \geq \mathbb{E}\|V^{1/2}G\|/\sqrt{c}$ (symmetry of N_W and $\Sigma_W \succeq V/c$ by the identity) confine every bounded-risk sequence to a compact bandwidth interval. The closure of the weight family is compact. Uniform continuity and the unique $r = 0$ optimum therefore localize every minimizer at $(\nu_2 I, c_0)$. The closed TSE path, including 0 and I, excludes $\nu_2 I$ when H is nonscalar. Compactness gives a strict positive gap at $r = 0$, which persists for small r, proving (i).

By the localization, for small r the common-bandwidth WE optimum is interior, so (142) applies to it. For (ii)–(iii), localization alone is insufficient, so we check local curvature. At $r = 0$, the second derivative in a nonzero symmetric weight direction D_0 is

$$\frac{\operatorname{tr}(J_0 D_0 V D_0)}{\nu_1 \nu_2 c_0} > 0.$$

Since $W = \nu_2 I$ minimizes the weight risk for every c, the W-gradient vanishes at $(\nu_2 I, c)$ for every c; hence the mixed weight–bandwidth derivatives vanish, and the (ρ, t) block is the pull-back of the W-Hessian by the tangent map $(d\rho, dt) \mapsto d\rho I - \nu_2 dt H$, which is injective for nonscalar H (the second derivatives of $W(\rho, t)$ are multiplied by the vanishing gradient). Together with positive bandwidth curvature, this proves a positive definite Hessian in (ρ, t, c) at $(\nu_2, 0, c_0)$. The Gaussian objective extends smoothly to small negative t; the implicit function theorem gives a smooth branch of (ρ, c)-minimizers of the objective, which is strictly convex in (ρ, c) near this point, and the profile (the value along the branch) has strictly positive t second derivative. At $t = 0$ the branch is the common-bandwidth WE optimum and, by the envelope theorem, the derivative of the profile equals $g'(0)$; symmetry in r gives $\rho = \nu_2 + O(r^2)$, $c = c_0 + O(r^2)$, hence $\bar{J} = J_0 + O(r^2)$ and $a_M = \frac{\rho r}{p} \operatorname{tr}(MJ_0) + O(r^3)$ (the first-order term in Θ averages to zero and $\mathbb{E}\Theta\Theta^\top = I/p$), where J_0 and c_0 are those of the small-bias subsection because the $r = 0$ optimum is the same; therefore

$$g'(0) = -\nu_2^2 r^2 \operatorname{tr}(J_0)\Delta_{\mathrm{c}}/p + O(r^4).$$

If $\Delta_c < 0$, the profile is increasing for $t \geq 0$ in this neighborhood. All global minimizers are in the neighborhood, so they have $t = 0$. If $\Delta_c > 0$, its negative boundary derivative gives a strictly improving positive t. Finally, for an eigenvector offset J_0 is diagonal with positive entries, and weighted Chebyshev (Lemma 46) gives the asserted strict trace inequality. □

Proposition 7 (Independent bandwidths: the two-headed optimum is not in the two-step family)**.** *Let $r = 0$, let H, V commute with common eigenbasis, H nonscalar, and let $b \neq 0$ with $b \parallel e_j$. Let $(s^\star, c_1^\star, c_2^\star)$ be a minimizer of $\bar{C}$ over the two-step family $W_s = s(H + sI)^{-1}$ and the bandwidths, with $s^\star \in (0, \infty)$, and write $w_i = s^\star/(h_i + s^\star)$, $\Sigma_{ii} = v_i[(1 - w_i)^2/(\nu_1 c_1^\star) + w_i^2/(\nu_2 c_2^\star)]$, $J_{ii} = (J_\Sigma(\mu_W))_{ii}$, $m_j = (m_\Sigma(\mu_W))_j$ and*

$$\Delta_\rho := \sum_i w_i^2 \, \partial_i R, \qquad \partial_i R = J_{ii} v_i \Big[\frac{w_i}{\nu_2 c_2^\star} - \frac{1 - w_i}{\nu_1 c_1^\star} \Big] + \mathbf{1}\{i = j\} \, m_j b_j \big((c_2^\star)^2 - (c_1^\star)^2 \big). \quad (143)$$

If $\Delta_\rho > 0$ then $\bar{C}^\star_{\mathrm{THE}}(0) < \bar{C}^\star_{\mathrm{TSE}}(0)$ and, by Proposition 8, $\bar{C}^\star_{\mathrm{THE}}(r) < \bar{C}^\star_{\mathrm{TSE}}(r)$ for all small r. For $b = 0$, a separate fixed-bandwidth statement holds: fix positive c_1, c_2 and minimize only over s. At every interior minimizer, $\Delta_\rho > 0$, giving a strict improvement at those bandwidths. No unrestricted-bandwidth centered optimum is asserted.

Proof. In the common eigenbasis every diagonal matrix weight $W = \mathrm{diag}(w_i)$ gives $\mu_{W,i} = b_i(c_1^2 + w_i(c_2^2 - c_1^2))$ and $\Sigma_{W,ii} = v_i[(1 - w_i)^2/(\nu_1 c_1) + w_i^2/(\nu_2 c_2)]$, so the risk $R(w) = \phi_{\Sigma_W}(\mu_W)$ is a smooth function of $w \in (0, 1)^p$ with $\partial_{w_i} R = m_i \partial_{w_i} \mu_{W,i} + \frac{1}{2} J_{ii} \partial_{w_i} \Sigma_{W,ii}$, because $\partial_{\Sigma_{ii}} \phi_\Sigma = \frac{1}{2} \partial^2_{\mu_i} \phi_\Sigma$ for a diagonal covariance (heat equation); $m_i = 0$ for $i \neq j$ by the reflection symmetry of the law of U_W. This is (143). Along the two-step family $\partial_s w_i = w_i(1 - w_i)/s$, so the optimality of $s^\star$ at fixed bandwidths reads $\sum_i w_i(1 - w_i)\partial_i R = 0$. The two-headed family contains the points $\rho W_{s^\star}$, $\rho \in (0, 1]$ (Proposition 6(ii)), whose risk has ρ-derivative $\sum_i w_i \partial_i R$ at $\rho = 1$; subtracting the optimality condition gives $\sum_i w_i \partial_i R = \sum_i w_i^2 \partial_i R = \Delta_\rho$. If $\Delta_\rho > 0$, decreasing ρ below 1 strictly decreases the risk at the two-step optimum, so $\bar{C}^\star_{\mathrm{THE}}(0) < \bar{C}_{\mathrm{TSE}}(s^\star, c_1^\star, c_2^\star; 0) = \bar{C}^\star_{\mathrm{TSE}}(0)$; the small-$r$ claim follows as in Proposition 8, since $\bar{C}^\star_{\mathrm{TSE}}(r) \geq \bar{C}^\star_{\mathrm{TSE}}(0)$ and $\bar{C}^\star_{\mathrm{THE}}(r)$ is at most the risk of a fixed two-headed point, which is continuous in r. If $b = 0$, then $\partial_i R = J_{ii} v_i \gamma_i$ with $\gamma_i = w_i/(\nu_2 c_2) - (1 - w_i)/(\nu_1 c_1)$ increasing in w_i and vanishing at $\bar{w} = \nu_2 c_2/(\nu_1 c_1 + \nu_2 c_2)$; with $a_i = w_i(1 - w_i)\partial_i R$, whose sign is that of $w_i - \bar{w}$, and $\sum_i a_i = 0$, $\Delta_\rho = \sum_i a_i \frac{w_i}{1 - w_i} = \sum_i a_i \Big(\frac{w_i}{1 - w_i} - \frac{\bar{w}}{1 - \bar{w}} \Big)$, each term nonnegative and some positive, since H is nonscalar and hence the w_i are unequal. □

The condition $\Delta_\rho > 0$ says that at the two-step optimum the coordinates moved too far toward the proxy (the least curved ones, where w_i exceeds the inverse-variance weight) outweigh, with weights w_i^2, those moved too little; a global factor $\rho < 1$ then lowers the risk. With the offset present and b along the least curved axis the two-step estimator gains from its large weight along b, which is the term $m_j b_j (c_2^2 - c_1^2) w_j^2 < 0$ in Δ_ρ, and the condition can fail.

E.6 THE TWO-STEP ESTIMATOR

Proposition 8 (Small bias)**.** *By Lemma 39, $\bar{C}^\star_{\mathrm{TSE}}(r) \geq \bar{C}^\star_{\mathrm{TSE}}(0)$ and $\bar{C}^\star_{\mathrm{WE}}(r) \leq \bar{C}^\star_{\mathrm{WE}}(0) + \frac{\nu_2^2 r^2}{2p} \operatorname{tr} J_0 + O(r^4)$. Hence if $\bar{C}^\star_{\mathrm{TSE}}(0) > \bar{C}^\star_{\mathrm{WE}}(0)$ then TSE is strictly worse than WE for all small r. This implication concerns $b \neq 0$. For a centered model with fixed positive bandwidths there is instead a strict gap between the infima over weights alone when H is nonscalar, at $r = 0$ and hence for all small r. Unrestricted centered bandwidth infima are zero (Proposition 12), so they have no such gap.*

Proof. Applying Lemma 39(i) to each member of $\mathcal{W}_{\mathrm{TSE}}$ and taking infima gives $\bar{C}^\star_{\mathrm{TSE}}(r) \geq \bar{C}^\star_{\mathrm{TSE}}(0)$. For WE, evaluate Lemma 39(ii) at the $r = 0$ optimum $(\nu_2 I, c_0, c_0)$ of Lemma 45: since $\mu_{\nu_2 I} = A_0 b$ and $\Sigma_{\nu_2 I} = \Sigma_0$, the value is $\bar{C}^\star_{\mathrm{WE}}(0) + \frac{\nu_2^2 r^2}{2p} \operatorname{tr} J_0 + O(r^4)$, and $\bar{C}^\star_{\mathrm{WE}}(r)$ is at most this. Hence a strict gap $\bar{C}^\star_{\mathrm{TSE}}(0) - \bar{C}^\star_{\mathrm{WE}}(0) > 0$ dominates the $O(r^2)$ term for all small r. In the

centered case, write $\tilde{c} = \nu_1 c_1 + \nu_2 c_2$ and $\Delta_W = W - \frac{\nu_2 c_2}{\tilde{c}} I$; completing the square in W (the cross term cancels at this value of the scalar shift),

$$\Sigma_W = \frac{V}{\tilde{c}} + \Big(\frac{1}{\nu_1 c_1} + \frac{1}{\nu_2 c_2}\Big)\Delta_W V \Delta_W \succeq \frac{V}{\tilde{c}},$$

with equality iff $\Delta_W = 0$, i.e. iff W is a scalar matrix. No $W_s = s(H + sI)^{-1}$ is scalar when H is nonscalar, so $\Sigma_{W_s} \succneqq V/\tilde{c}$ and Lemma 38(vi) gives a pointwise gap, whereas WE attains $V/\tilde{c}$ exactly. With c_1, c_2 fixed, compactify $s \in [0, \infty]$; neither endpoint $0, I$ equals $\bar{w} I$, $\bar{w} = \nu_2 c_2/\tilde{c} \in (0, 1)$. The entire compact TSE path excludes $\bar{w} I$, so continuity upgrades the pointwise gap to a strict gap between these fixed-bandwidth infima. Continuity in r then preserves it for small bias. □

Lemma 48 (Small weights). *Fix a symmetric $M \succeq 0$ and let $W = \epsilon M$. With $\Sigma_1 = V/(\nu_1 c_1)$, $J_1 = J_{\Sigma_1}(c_1^2 b)$, $m_1 = m_{\Sigma_1}(c_1^2 b)$,*

$$\partial_\epsilon \bar{C}|_0 = (c_2^2 - c_1^2) m_1^\top M b - \operatorname{tr}(M \Sigma_1 J_1), \qquad \partial_\epsilon^2 \bar{C}|_0 = \frac{r^2}{p} \operatorname{tr}(M J_1 M) + O(1). \tag{144}$$

If $\partial_\epsilon \bar{C}|_0 < 0$, $\min_\epsilon \bar{C} = \bar{C}^\star_{\mathrm{gold}}(c_1) - \frac{p\,(\partial_\epsilon \bar{C}|_0)^2}{2r^2 \operatorname{tr}(M J_1 M)}\{1 + O(r^{-2})\}$.

Proof. Here $U = U_1 + \epsilon M D$ and $\partial_\epsilon U = MD$, so $\partial_\epsilon \bar{C}|_0 = \mathbb{E}[U_1^\top M D/\|U_1\|]$. In $D = r\Theta + (c_2^2 - c_1^2) b + \widetilde{N}_2 - \widetilde{N}_1$ the term $r\Theta$ contributes nothing (Θ is independent of U_1 with $\mathbb{E}\Theta = 0$) and neither does $\widetilde{N}_2$, so $\partial_\epsilon \bar{C}|_0 = (c_2^2 - c_1^2) m_1^\top M b - \mathbb{E}[U_1^\top M \widetilde{N}_1/\|U_1\|]$, and $\mathbb{E}[U_1^\top M \widetilde{N}_1/\|U_1\|] = \operatorname{tr}(M \Sigma_1 J_1)$ by Lemma 38(v). Differentiating once more, $\partial_\epsilon^2 \bar{C}|_0 = \mathbb{E}[D^\top M J_{U_1} M D]$ with $J_{U_1} = (I - U_1 U_1^\top/\|U_1\|^2)/\|U_1\|$; writing $D = r\Theta + O(1)$ and using $\mathbb{E}_\Theta \Theta \Theta^\top = I/p$ gives $\frac{r^2}{p} \operatorname{tr}(M J_1 M) + O(1)$. The last display is the minimum of the resulting quadratic model, attained at $\epsilon_\star = -\partial_\epsilon \bar{C}|_0 / \partial_\epsilon^2 \bar{C}|_0 \asymp r^{-2}$, the factor $\{1 + O(r^{-2})\}$ absorbing the $O(1)$ term of $\partial_\epsilon^2 \bar{C}|_0$. □

Theorem 28 (Large bias). *Let $b \neq 0$ and restrict every family to a common bandwidth $c_1 = c_2 = c$; write $\bar{C}^{\mathrm{c}}_{\mathcal{M}}(r)$ for the infimum of $\bar{C}$ over $\mathcal{W}_{\mathcal{M}}$ and a common $c > 0$. For $c > 0$ let $\Sigma_1(c) = V/(\nu_1 c)$, $J_c = J_{\Sigma_1(c)}(c^2 b)$ and $\bar{C}_{\mathrm{gold}}(c) = \phi_{\Sigma_1(c)}(c^2 b)$, let c_g be the unique minimizer of $\bar{C}_{\mathrm{gold}}$, and put $\Sigma_1 = \Sigma_1(c_g)$, $J_1 = J_{c_g}$ and $\langle A, B \rangle_J = \operatorname{tr}(AJB)$ on symmetric matrices. For $M \succeq 0$, $M \neq 0$, let*

$$\gamma_c(M) = \frac{p}{2} \frac{(\langle M, \Sigma_1(c) \rangle_{J_c})_+^2}{\langle M, M \rangle_{J_c}},$$
$$\gamma_{\mathrm{WE}} = \gamma_{c_g}(I), \quad \gamma_{\mathrm{TSE}} = \gamma_{c_g}(H^{-1}), \quad \gamma_{\mathrm{THE}} = \sup_{t \in [0,\infty]} \gamma_{c_g}\big((I + tH)^{-1}\big),$$

where $x_+ = \max\{x, 0\}$ and $t = \infty$ stands for H^{-1}. As $r \to \infty$, every near-minimizer (W_r, c_r) over a family, that is, one with $\bar{C}(W_r, c_r, c_r; r) \leq \bar{C}^{\mathrm{c}}_{\mathcal{M}}(r) + o(1)$, satisfies $\|W_r\| \to 0$, $r\|W_r\| \to 0$ and $c_r \to c_g$, and for $\mathcal{M} \in \{\mathrm{WE}, \mathrm{THE}, \mathrm{TSE}\}$,

$$\bar{C}^{\mathrm{c}}_{\mathcal{M}}(r) = \bar{C}^\star_{\mathrm{gold}} - \gamma_{\mathcal{M}} r^{-2} + o(r^{-2});$$

in particular $\gamma_{\mathrm{THE}} \geq \max\{\gamma_{\mathrm{WE}}, \gamma_{\mathrm{TSE}}\}$. If H and V commute with eigenvalues in reverse order, neither is a multiple of the identity, and b is a common eigenvector of H and V, then $\gamma_{\mathrm{THE}} > \gamma_{\mathrm{WE}}$. If $V = \kappa H^{-1}$ with H not a multiple of the identity, then $\gamma_{\mathrm{THE}} = \gamma_{\mathrm{TSE}} = \frac{p}{2}\langle \Sigma_1, \Sigma_1 \rangle_{J_1} > \gamma_{\mathrm{WE}}$.

Proof. Small matrix weights. Let (W_r, c_r) be near-minimizers over a family, and write $v_{\min}$ for the smallest eigenvalue of V. Since N_W is independent of Θ and $\mathbb{E}\Theta = 0$, Jensen's inequality and the triangle inequality give $\bar{C} \geq \max\{\mathbb{E}\|c^2 b + N_W\|, r\mathbb{E}\|W\Theta\| - \mathbb{E}\|c^2 b + N_W\|\} \geq \frac{r}{2}\mathbb{E}\|W\Theta\|$, and $\mathbb{E}\|W\Theta\| \geq \mathbb{E}\|W\Theta\|^2/\|W\|_{\mathrm{op}} \geq \|W\|_{\mathrm{op}}/p$. Since $W = 0$ lies in the closure of every family, $\bar{C}^{\mathrm{c}}_{\mathcal{M}}(r) \leq \bar{C}^\star_{\mathrm{gold}}$, and hence $\|W_r\|_{\mathrm{op}} \to 0$. For $\|W\|_{\mathrm{op}} \leq \frac{1}{2}$, $\bar{C} \geq c^2\|b\|$ and $\bar{C} \geq \mathbb{E}\|N_W\| \geq \frac{1}{2}\sqrt{v_{\min}/(\nu_1 c)}\,\mathbb{E}\chi_p$, so c_r stays in a compact subset of $(0, \infty)$. Write $W = \epsilon M$ with $\|M\|_{\mathrm{op}} = 1$ and $\tau = \epsilon r$, so that $\bar{C} = G(\epsilon, \tau) := \mathbb{E}_\Theta \phi_{\Sigma(\epsilon)}(c^2 b + \tau M \Theta)$ with $\Sigma(\epsilon) = (I - \epsilon M)\Sigma_1(c)(I - \epsilon M) + \epsilon^2 M \Sigma_2(c) M$. Since $N_W - \widetilde{N}_1 = \epsilon M(\widetilde{N}_2 - \widetilde{N}_1)$, $|G(\epsilon, \tau) - G(0, \tau)| \leq \epsilon\,\mathbb{E}\|\widetilde{N}_2 - \widetilde{N}_1\|$; and $G(0, \tau) - G(0, 0)$ is nondecreasing in τ, as in Lemma 39(i), and positive for $\tau > 0$ by strict convexity, with a positive infimum over $\|M\|_{\mathrm{op}} = 1$ and c in a compact set. Hence $\epsilon_r r \to 0$ as well.

For TSE, $\|W\| \to 0$ forces $s \to 0$, and $s(H + sI)^{-1} = (I + tH)^{-1}$ with $t = 1/s \to \infty$, whose normalized shape tends to H^{-1}.

Expansion. The derivatives of ϕ_Σ are bounded uniformly over compact sets of positive definite Σ (Gaussian integration by parts), so G is smooth near $(0, 0)$ with uniformly bounded derivatives. Odd orders in τ vanish by the symmetry of Θ, $\partial_\tau^2 G(0, 0) = \operatorname{tr}(MJ_cM)/p$, and $\partial_\epsilon G(0, 0) = -\operatorname{tr}(M\Sigma_1(c)J_c)$ as in Lemma 48. Thus

$$\bar{C}(\epsilon M, c, c; r) = \bar{C}_{\text{gold}}(c) - \epsilon\langle M, \Sigma_1(c)\rangle_{J_c} + \frac{\epsilon^2 r^2}{2p}\langle M, M\rangle_{J_c} + O(\epsilon^2 + \epsilon^3 r^2 + \epsilon^4 r^4),$$

uniformly in M and in c on compact sets. For $\epsilon \to 0$ and $\epsilon r \to 0$ the remainder is $o(\epsilon^2 r^2)$, and minimizing over $\epsilon \geq 0$ gives $\bar{C}_{\text{gold}}(c) - r^{-2}\gamma_c(M) + o(r^{-2})$. The gold-only constant has the unique minimizer c_g (the argument of Theorem 26 with $W = 0$), the correction is uniformly $O(r^{-2})$, and γ_c is continuous in c, so $c_r \to c_g$; the supremum over $t \in [0, \infty]$ is attained because γ_{c_g} is scale invariant and continuous on the compactified path. Under either of the last two conditions of the theorem, $\langle (I + tH)^{-1}, \Sigma_1\rangle_{J_1} > 0$ for every t, so the positive part is inactive.

Reverse order. If $b \parallel e_j$, the law of U_1 is invariant under the reflections of the coordinates $i \neq j$, so $J_1 = \operatorname{diag}(j_i)$ with $j_i > 0$ in the common eigenbasis, and $\Sigma_1 = \operatorname{diag}(\sigma_i)$ with $\sigma_i \propto v_i$. With $f(t) = \gamma_{c_g}((I + tH)^{-1})$,

$$f'(0) = \frac{p\,\langle I, \Sigma_1\rangle_{J_1}}{\langle I, I\rangle_{J_1}^2}\Big[\langle I, \Sigma_1\rangle_{J_1}\langle I, H\rangle_{J_1} - \langle H, \Sigma_1\rangle_{J_1}\langle I, I\rangle_{J_1}\Big] > 0$$

by Lemma 46, strictly because both sequences are nonconstant; hence $\gamma_{\text{THE}} \geq f(t) > f(0) = \gamma_{\text{WE}}$ for small $t > 0$. If $V = \kappa H^{-1}$, then $\Sigma_1 M$ is symmetric positive definite for $M = (I + tH)^{-1}$, and Cauchy–Schwarz gives $\gamma_{c_g}(M) \leq \frac{p}{2}\langle \Sigma_1, \Sigma_1\rangle_{J_1}$ with equality iff $M \propto \Sigma_1$, attained by H^{-1} and not by I. □

Proposition 9 (The two-step corner for large bias)**.** *Under the conditions of Theorem 28, reverse order does not imply $\gamma_{\text{TSE}} \geq \gamma_{\text{WE}}$: for $p = 2$, common eigenvectors e_1, e_2 of H and V, b along one of them, eigenvalues $v_1 < v_2 \leq 2v_1$ of V and $h_1 > h_2$ of H, one has $\gamma_{\text{TSE}} < \gamma_{\text{WE}}$ once h_1/h_2 is large enough.*

Proof. $J_1 = \operatorname{diag}(j_1, j_2)$ and $\Sigma_1 = \operatorname{diag}(\sigma_1, \sigma_2)$ with $\sigma_1 < \sigma_2 \leq 2\sigma_1$ do not depend on H. As $h_1/h_2 \to \infty$, $\gamma_{c_g}(H^{-1}) \to \frac{p}{2} j_2\sigma_2^2$, while $(j_1\sigma_1 + j_2\sigma_2)^2 - j_2\sigma_2^2(j_1 + j_2) = j_1[j_1\sigma_1^2 + j_2\sigma_2(2\sigma_1 - \sigma_2)] > 0$, that is, $\gamma_{c_g}(I) > \frac{p}{2} j_2\sigma_2^2$. □

E.7 Fixed directions and the worst direction

Proposition 10 (Fixed direction: a local criterion)**.** *Let $b \neq 0$, fix $\zeta \in \mathbb{R}^p$, and suppose (ρ, c_1, c_2) is a finite WE minimizer with $0 < \rho < 1$, allowing separate bandwidth optimization. Put $m = \nabla\phi_{\sigma_\rho^2 V}(Ab + \rho\zeta)$ and $\bar{h}_{\text{noise}} = \mathbb{E}[U_0^\top H U_0/\|U_0\|]/\mathbb{E}\|U_0\|$. Then*

$$g'(0) = -\frac{\rho(1 - \rho)}{\nu_1 c_1 \sigma_\rho^2} m^\top (H - \bar{h}_{\text{noise}} I)\zeta.$$

Thus this numerator is positive if and only if $g'(0) < 0$, a sufficient condition for improvement by all sufficiently large finite penalties. It is not a necessary condition for improvement elsewhere on the path. Near $\zeta = 0$ the local criterion has expansion $g'(0) = -\nu_2 m_0^\top (H - \bar{h}_{\text{noise}}^{(0)} I)\zeta + O(\|\zeta\|^2)$. When its linear coefficient is nonzero, it describes a favorable half-space to first order; otherwise higher orders decide.

Proof. Bandwidth elimination remains valid for a fixed shift. Apply (132) with $r\Theta$ replaced by ζ and use the interior weight stationarity equation $(1 - \rho)m^\top\zeta/(\nu_1 c_1) = -\kappa'\mathbb{E}\|U_0\|$. The expansion follows from the nondegenerate $r = 0$ optimum proved in Theorem 23: every global WE minimizer converges to (ν_2, A_0) as $\zeta \to 0$ (coercivity and upper semicontinuity, as there), and its fixed-shift perturbation is $O(\|\zeta\|)$ by the implicit function theorem. A derivative at zero does not control every finite t, which explains the local scope. □

Proposition 11 (Worst direction)**.** *Fix finite WE minimax parameters (ρ, c_1, c_2) with $0 < \rho < 1$. Let $R(t, \theta)$ be the THE risk at these parameters for bias $r\theta$, and $\Theta^\star = \arg\max_{\|\theta\|=1} R(0, \theta)$. For $D_\theta = r\theta + (c_2^2 - c_1^2)b + \widetilde{N}_2 - \widetilde{N}_1$, $U_\theta = U_1 + \rho D_\theta$, define*

$$d(\theta) = -\rho \mathbb{E}[\widehat{U}_\theta^\top H D_\theta].$$

Then

$$\frac{d}{dt^+} \max_{\|\theta\|=1} R(t, \theta)\Big|_{t=0} = \max_{\theta \in \Theta^\star} d(\theta).$$

If this maximum is negative, a finite penalty strictly improves the worst-direction risk. A positive maximum gives local increase; a zero maximum is inconclusive.

Proof. The derivative formula is differentiation under the expectation, dominated by $\rho\|H\|\|D_\theta\|$, uniformly on the unit sphere. Gaussian smoothing makes R and its t derivative continuous in θ. Compactness of the sphere permits Danskin's theorem and gives the displayed maximum over active directions. Its negative sign implies a strict decrease for sufficiently small positive t. Minimax weight optimality need not make the weight derivative vanish separately at each active direction. Hence the fixed-direction stationarity simplification of Proposition 10 cannot be substituted here without additional assumptions. □

E.8 Degenerate smoothing bias

Proposition 12. *If $b = 0$ then for every W, $\bar{C}(W, c_1, c_2; r) = \mathbb{E}\|rW\Theta + N_W\|$ is strictly decreasing in each bandwidth whose noise coefficient is nonzero (independent of an inactive bandwidth at $W = 0$ or $W = I$); letting $c \to \infty$ and then $W \to 0$ gives infimum 0, not attained. The $n^{-2/5}$ constants of all three estimators are zero and the comparison at this scale is void.*

Proof. With $b = 0$ we have $\mu_W = 0$, so $\bar{C}(W, c_1, c_2; r) = \mathbb{E}\|rW\Theta + N_W\|$. The covariance $\Sigma_W = \frac{(I-W)V(I-W)}{\nu_1 c_1} + \frac{WVW}{\nu_2 c_2}$ is strictly decreasing in the Loewner order in c_1 (unless $W = I$) and in c_2 (unless $W = 0$), so $\bar{C}$ is strictly decreasing in each bandwidth by Lemma 38(vi), applied conditionally on Θ with mean $rW\Theta$. Letting $c_1, c_2 \to \infty$ gives $\Sigma_W \to 0$ and, by dominated convergence, $\bar{C} \to \mathbb{E}\|rW\Theta\|$, which tends to 0 as $W \to 0$. The infimum is not attained, because $\Sigma_W \succ 0$ for every admissible W (if $u^\top \Sigma_W u = 0$ then $(I - W)u = Wu = 0$, hence $u = 0$) and therefore $\bar{C} > 0$ at every finite (W, c_1, c_2). Finally gold-only, $W = 0$ with $c_1 \to \infty$, lies in the closure of all three families, so all three asymptotic constants are 0. □

F Sparse correction with the exact sign loss: the ℓ_1 two-step estimator

This appendix proves Theorem 7(a). Notation is that of Appendix B: the loss is $\ell(\beta; x, y) = -1 + 2\mathbf{1}\{(2y - 1)x^\top \beta < 0\}$, as in (2), defined for every nonzero β and invariant under positive scaling. Put $\widehat{L}_k = \mathbb{P}_{N_k}\ell$ and $L_k = P_k\ell$. Its increments lie in $\{0, \pm 2\}$; a nonzero increment requires a sign change or a zero margin. Hence the same slab envelopes and disagreement-probability bounds apply. For fixed nonzero directions, ties have probability zero under Assumption 2. Assumption 2 gives a bound $\bar{f}_1$ on the density of $X^\top u$ for every unit u, hence a bound $\bar{f}_1/\|u\|$ for every nonzero u, and the disagreement bound $\Pr\{\operatorname{sign}(X^\top \beta) \neq \operatorname{sign}(X^\top \beta')\} \leq \Gamma \angle(\beta, \beta')$ of Lemma 1(b). Assumption 3 brackets the gold population loss in the chart coordinates v of radius $\eta_0 \leq 1/2$; since for unit β the chord $\|\beta - \beta^\star\| = 2\sin(\alpha/2)$ and the chart coordinate $\|v\| = \sin\alpha$ satisfy $\frac{2}{\pi}\alpha \leq \|\beta - \beta^\star\| \leq \alpha$ and $\frac{2}{\pi}\alpha \leq \|v\| \leq \alpha$ for $\alpha \leq \pi/2$, the bracket reads, for chords $\|\beta - \beta^\star\| \leq 2\eta_0/\pi$,

$$\underline{c}\,\|\beta - \beta^\star\|^2 \leq L_1(\beta) - L_1(\beta^\star) \leq \bar{c}\,\|\beta - \beta^\star\|^2, \qquad \underline{c} = \frac{4c_{\min}}{\pi^2}, \quad \bar{c} = \frac{\pi^2 c_{\max}}{4}, \quad c_{\max} = \max_j c'_j. \tag{145}$$

The bias $\delta = \beta_2 - \beta^\star$ has support S with $|S| \leq s$; T is the set of the s largest coordinates of $|\beta^\star|$, and $S' = S \cup T$, $|S'| \leq 2s$.

The estimator. Step 1 is the proxy-only estimator with the exact sign loss $\widehat{\beta}_{\mathrm{proxy}} \in \arg\min_{\|\beta\|=1} \widehat{L}_2(\beta)$. Step 2 is

$$\widehat{\beta} \in \arg\min \Big\{ \widehat{L}_1(\beta) + \lambda \|\beta - \widehat{\beta}_{\mathrm{proxy}}\|_1 : \ \|\beta\| = 1, \ \|\beta - \widehat{\beta}_{\mathrm{proxy}}\|_2 \le r_{\mathrm{loc}} \Big\}, \tag{146}$$

with a trust region of constant radius $r_{\mathrm{loc}} \le \eta_0/\pi$. The minimum exists because the indicator of a strictly negative margin is lower semicontinuous, its finite sum plus the continuous penalty is lower semicontinuous, and the nonempty feasible set is compact. Write $e = \widehat{\beta}_{\mathrm{proxy}} - \beta_2$.

Theorem 29 (Theorem 7(a) with explicit constants)**.** *Let Assumptions 1–3 hold, $d \ge 3$, and let δ be s-sparse. Put $D = d^{1/4}(\log 2d)^{3/4}$, $C_1 = 4C_{\mathrm{ep}}\sqrt{\bar{f}_1 R}$ with C_{ep} the universal constant of Lemma 51,*

$$r_\star = \Big(\frac{4C_1 D (2s)^{1/4}}{\underline{c}\sqrt{N_1}}\Big)^{2/3} = \Big(\frac{16\sqrt{2}\, C_1^2 \sqrt{sd}\,(\log 2d)^{3/2}}{\underline{c}^{\,2} N_1}\Big)^{1/3}, \qquad \lambda = \frac{C_1 D}{2(2s)^{1/4}\sqrt{N_1 r_\star}}. \tag{147}$$

Suppose (C1) $\|\delta\| + 2\|e\| \le r_{\mathrm{loc}}/2$. Then, conditionally on the proxy sample, with probability at least $1-\eta$ over the gold sample, with $J = (\lceil \log_2 N_1 \rceil + 1)(\lceil \log_2 \sqrt{d} \rceil + 2)$ and $\Lambda = \log(2dJ/\eta)$,

$$\begin{aligned} \|\widehat{\beta} - \beta^\star\|_2 \le{}& \frac{5}{3} r_\star + \frac{\underline{c}\, r_\star^2}{4C_{\mathrm{ep}}\bar{f}_1 R} + \frac{1}{\sqrt{3}}\Big(\frac{64 C_1^2 \Lambda}{\underline{c}^{\,2} N_1}\Big)^{1/3} + \Big(1 + \sqrt{\tfrac{8(\underline{c}+\bar{c})}{3\underline{c}}}\Big)\|e\| \\ &+ \sqrt{\frac{32 C_{\mathrm{ep}}\sqrt{2s}\, D^2}{3\underline{c}\, N_1}} + \sqrt{\frac{8 C_2 \Lambda}{3\underline{c}\, N_1}}, \end{aligned} \tag{148}$$

with C_2 depending only on $\bar{f}_1 R$. On the event of Theorem 8 for the proxy sample, $\|e\| \le \frac{\pi}{\sqrt{2}}(\kappa_2 + \tau_2)$, and $\angle(\widehat{\beta}, \beta^\star) \le \frac{\pi}{2}\|\widehat{\beta} - \beta^\star\|$.

When $N_1 \ge \sqrt{sd}\,(\log 2d)^{3/2}$, the second and fifth terms are at most of the order of $r_\star$, and the bound is $\widetilde{O}((\sqrt{sd}/N_1)^{1/3} + \|e\|)$. The tuning (147) is $\lambda \asymp (d/s^2)^{1/6}\sqrt{\log 2d}\, N_1^{-1/3}$; it uses s and d through small powers, as the theoretical tuning of Bastani (2021) uses the noise levels.

F.1 Comparison point and population term

Lemma 49 (Comparison point on the sphere)**.** *There is a unit vector $\tilde{\beta}$ such that $\theta := \widehat{\beta}_{\mathrm{proxy}} - \tilde{\beta}$ is supported on S' and $\epsilon_0 := \|\tilde{\beta} - \beta^\star\| \le \|e\|$.*

Proof. A unit vector $\tilde{\beta}$ that agrees with $\widehat{\beta}_{\mathrm{proxy}}$ off S' has $\|\tilde{\beta}_{S'}\| = \|\widehat{\beta}_{\mathrm{proxy},S'}\| =: r_{S'}$; take $\tilde{\beta}_{S'} = r_{S'}\beta^\star_{S'}/\|\beta^\star_{S'}\|$, where $\|\beta^\star_{S'}\| \ge \|\beta^\star_T\| > 0$ because T contains the largest coordinates of the nonzero vector $\beta^\star$. Since δ is supported in S', β_2 and $\beta^\star$ agree off S'; hence $\widehat{\beta}_{\mathrm{proxy},S'^c} - \beta^\star_{S'^c} = e_{S'^c}$ and, both centers being unit, $\|\beta_{2,S'}\| = \|\beta^\star_{S'}\|$. So $\|\tilde{\beta} - \beta^\star\|^2 = \|e_{S'^c}\|^2 + (r_{S'} - \|\beta_{2,S'}\|)^2$, and since $r_{S'} = \|\beta_{2,S'} + e_{S'}\|$, the reverse triangle inequality gives $|r_{S'} - \|\beta_{2,S'}\|| \le \|e_{S'}\|$, whence $\epsilon_0^2 \le \|e_{S'^c}\|^2 + \|e_{S'}\|^2 = \|e\|^2$. □

The unit-norm constraint therefore adds at most the proxy error to the bound; the comparison point is a device of the proof, and the estimator does not need S or T.

By (C1), $\|\tilde{\beta} - \widehat{\beta}_{\mathrm{proxy}}\| \le \epsilon_0 + \|\delta\| + \|e\| \le r_{\mathrm{loc}}/2$, so $\tilde{\beta}$ is feasible in (146), and every feasible β has $\|\beta - \beta^\star\| \le r_{\mathrm{loc}} + \|\delta\| + \|e\| \le \frac{3}{2} r_{\mathrm{loc}} \le 2\eta_0/\pi$, so (145) applies to $\widehat{\beta}$ and to $\tilde{\beta}$.

Lemma 50 (Population term)**.** *For feasible $\widehat{\beta}$ and $\tilde{\beta}$, with $\Delta = \widehat{\beta} - \tilde{\beta}$ and $r = \|\Delta\|$, $L_1(\widehat{\beta}) - L_1(\tilde{\beta}) \ge \frac{\underline{c}}{2} r^2 - (\underline{c} + \bar{c})\epsilon_0^2$.*

Proof. By (145), $L_1(\widehat{\beta}) - L_1(\tilde{\beta}) \ge \underline{c}\|\widehat{\beta} - \beta^\star\|^2 - \bar{c}\epsilon_0^2$, and $\|\widehat{\beta} - \beta^\star\| \ge (r - \epsilon_0)_+$ with $(r-\epsilon_0)_+^2 \ge r^2/2 - \epsilon_0^2$. □

F.2 THE INCREMENT PROCESS OVER AN ℓ_1 BALL

Fix a unit vector $\tilde\beta$ independent of the gold sample. For $\Delta \in \mathbb{R}^d$ let $g_\Delta = \ell(\tilde\beta + \Delta; \cdot) - \ell(\tilde\beta; \cdot)$ and $\mathbb{V}_N(\Delta) = (\mathbb{P}_{N_1} - P_1)g_\Delta$; for $r \le 1/2$ and $\rho_1 \ge r$ let $\mathcal{D}(r, \rho_1) = \{\Delta : \|\Delta\|_2 \le r,\ \|\Delta\|_1 \le \rho_1\}$.

Lemma 51 (Expected supremum)**.** *There is a universal constant $C_{\mathrm{ep}} \ge 5$ such that, with $D = d^{1/4}(\log 2d)^{3/4}$,*

$$\mathbb{E} \sup_{\Delta \in \mathcal{D}(r,\rho_1)} |\mathbb{V}_N(\Delta)| \le C_{\mathrm{ep}} \sqrt{\bar f_1 R}\Big[\sqrt{\frac{\rho_1}{N_1}}\, D + \sqrt{\frac{r \log(2d)}{N_1}}\Big] + C_{\mathrm{ep}}\, \frac{(\rho_1/r)D^2 + \log(2d)}{N_1}.$$

Proof. Three facts. (F1) A sign change between $\tilde\beta + \Delta$ and $\tilde\beta$ requires $|x^\top\tilde\beta| \le |x^\top\Delta| \le Rr$, so $|g_\Delta| \le G_r := 2 \cdot \mathbf{1}\{|x^\top\tilde\beta| \le Rr\}$ and $\|G_r\|_{P,2}^2 \le 8\bar f_1 Rr =: \varsigma^2$. (F2) If $\|\Delta - \Delta'\| \le \epsilon \le 1/2$, a sign change between $\tilde\beta + \Delta$ and $\tilde\beta + \Delta'$ requires $|x^\top(\tilde\beta + \Delta')| \le R\epsilon$, so with $U_{\Delta',\epsilon} := 2 \cdot \mathbf{1}\{|x^\top(\tilde\beta + \Delta')| \le R\epsilon\}$ we have $g_{\Delta'} - U_{\Delta',\epsilon} \le g_\Delta \le g_{\Delta'} + U_{\Delta',\epsilon}$, a bracket of $L_2(P)$-size $\|2U_{\Delta',\epsilon}\|_{P,2} \le 4\sqrt{2\bar f_1 R\epsilon/\|\tilde\beta + \Delta'\|} \le 8\sqrt{\bar f_1 R\epsilon}$ because $\|\tilde\beta + \Delta'\| \ge 1 - \epsilon \ge \frac12$. (F3) The ℓ_2 covering number of $\mathcal{D}(r, \rho_1)$ at scale ϵ is at most $\min\{(2d)^{\lceil \rho_1^2/\epsilon^2\rceil}, (3r/\epsilon)^d\}$: the first bound is Maurey's lemma (every point of $\rho_1 B_1^d$ is within ϵ of an average of $\lceil\rho_1^2/\epsilon^2\rceil$ signed vertices), the second is volumetric; the net points can be taken within ϵ of $\mathcal{D}$.

By (F2) and (F3), the bracketing numbers of $\mathcal{G} = \{g_\Delta : \Delta \in \mathcal{D}(r, \rho_1)\}$ satisfy, with $a = 64\bar f_1 R\rho_1$ and $b = 192\bar f_1 Rr$,

$$\log N_{[]}(\epsilon', \mathcal{G}, L_2(P)) \le \min\Big\{\Big(1 + \frac{a^2}{\epsilon'^4}\Big)\log(2d),\ d\log\frac{b}{\epsilon'^2}\Big\}.$$

Lemma 3.4.2 of van der Vaart & Wellner (1996) gives, for a class bounded by 2 with $L_2(P)$-radius ς,

$$\mathbb{E}\|\mathbb{G}_{N_1}\|_{\mathcal{G}} \le C_0 J_{[]}(\varsigma)\Big(1 + \frac{2J_{[]}(\varsigma)}{\varsigma^2\sqrt{N_1}}\Big), \qquad J_{[]}(\varsigma) = \int_0^\varsigma \sqrt{1 + \log N_{[]}(\epsilon')}\, d\epsilon'.$$

Split the integral at $\epsilon_\times^2 := a\sqrt{\log(2d)/d}$:

$$\int_{\epsilon_\times}^\varsigma \sqrt{1 + \Big(1 + \frac{a^2}{\epsilon'^4}\Big)\log 2d}\, d\epsilon' \le \sqrt{\log 2d}\Big(\frac{a}{\epsilon_\times} + 2\varsigma\Big),$$

$$\int_0^{\epsilon_\times} \sqrt{1 + d\log\frac{b}{\epsilon'^2}}\, d\epsilon' \le \epsilon_\times\Big(1 + \sqrt{d\log\frac{b}{\epsilon_\times^2}} + \sqrt{\tfrac{\pi d}{2}}\Big),$$

using $\int_0^t \sqrt{\log(b/\epsilon^2)}\, d\epsilon \le t[\sqrt{\log(b/t^2)} + \sqrt{\pi/2}]$. Since $\rho_1 \ge r$, $b/\epsilon_\times^2 = \frac{3r}{\rho_1}\sqrt{d/\log 2d} \le 3\sqrt d \le 2d$ for $d \ge 3$, so $\log(b/\epsilon_\times^2) \le \log 2d$. Both pieces are $\sqrt a\,(d\log 2d)^{1/4}$ times $O(\sqrt{\log 2d})$:

$$J_{[]}(\varsigma) \le 4\sqrt a\, d^{1/4}(\log 2d)^{3/4} + 2\varsigma\sqrt{\log 2d} = 32\sqrt{\bar f_1 R\rho_1}\, d^{1/4}(\log 2d)^{3/4} + 2\varsigma\sqrt{\log 2d}.$$

If $\epsilon_\times > \varsigma$, the second integral taken up to ς alone bounds $J_{[]}(\varsigma)$ by $4\varsigma\sqrt d \le 4\sqrt a\, d^{1/4}(\log 2d)^{3/4}$, so the bound holds in both cases. Squaring, $J_{[]}(\varsigma)^2/\varsigma^2 \le 256(\rho_1/r)D^2 + 8\log(2d)$, and dividing $C_0 J_{[]}(\varsigma)\big(1 + 2J_{[]}(\varsigma)/(\varsigma^2\sqrt{N_1})\big)$ by $\sqrt{N_1}$ gives the lemma with $C_{\mathrm{ep}} = \max\{512C_0, 5\}$. □

The factor $\sqrt{sd}$ comes from the two pieces of the entropy integral. The $L_2(P)$ distance between two loss increments is the square root of the Euclidean distance between the parameters (F2), so the ℓ_1 geometry (Maurey) controls the class only at coarse scales and the class is as rich as all halfspaces (dimension d) at fine scales; the two regimes meet at $\epsilon_\times$ and contribute equally.

Lemma 52 (Uniform version)**.** *Let $r_{\max} \le 1/2$, $J = (\lceil\log_2 N_1\rceil + 1)(\lceil\log_2\sqrt d\rceil + 2)$ and $\Lambda = \log(2dJ/\eta)$. With probability at least $1 - \eta$, simultaneously for all $0 < \|\Delta\| \le r_{\max}$,*

$$|\mathbb{V}_N(\Delta)| \le C_1\Big[\sqrt{\frac{\|\Delta\|_1}{N_1}}\, D + \sqrt{\frac{\|\Delta\|\,\Lambda}{N_1}}\Big] + C_3\,\frac{\|\Delta\|_1}{\|\Delta\|}\,\frac{D^2}{N_1} + \frac{C_2\Lambda}{N_1},$$

with $C_1 = 4C_{\mathrm{ep}}\sqrt{\bar f_1 R}$, $C_3 = 4C_{\mathrm{ep}}$ and C_2 depending only on $\bar f_1 R$.

Proof. Let $r_j = 2^{-j} r_{\max}$ for $0 \le j \le \lceil \log_2 N_1 \rceil$ and $\bar\rho_{jk} = \max\{2^{-k}\sqrt{d}\, r_{\max}, r_j\}$ for $j \le k \le j + \lceil \log_2 \sqrt{d} \rceil + 1$. Apply Bousquet's inequality (Bousquet, 2002) to each of the J classes $\mathcal{G}(r_j, \bar\rho_{jk})$ with confidence η/J: the class is bounded by 2 with variance at most $8\bar f_1 R r_j$, so

$$\sup_{\mathcal{G}(r_j,\bar\rho_{jk})} |\mathbb{V}_N| \le 2\,\mathbb{E}\sup_{\mathcal{G}(r_j,\bar\rho_{jk})} |\mathbb{V}_N| + \sqrt{\frac{16\bar f_1 R r_j \log(J/\eta)}{N_1}} + \frac{4\log(J/\eta)}{N_1}.$$

Since $\|\Delta\| \le \|\Delta\|_1 \le \sqrt{d}\|\Delta\|$, every Δ with $r_{\max}/N_1 < \|\Delta\| \le r_{\max}$ lies in one of these classes with $\|\Delta\| \le r_j \le 2\|\Delta\|$ and $\bar\rho_{jk} \le 2\|\Delta\|_1$, so that $\bar\rho_{jk}/r_j \le 2\|\Delta\|_1/\|\Delta\|$. Lemma 51 then gives the bound, because $2\sqrt{2}C_{\mathrm{ep}} + 4\sqrt{2} \le 4C_{\mathrm{ep}}$ for $C_{\mathrm{ep}} \ge 5$. For $\|\Delta\| \le r_{\max}/N_1$ the increment is bounded by $2/N_1$ times the number of observations in the slab of half-width $Rr_{\max}/N_1$ around the hyperplane of $\tilde\beta$, plus twice the probability of that slab; the count is binomial with mean at most $2\bar f_1 R r_{\max} \le \bar f_1 R$, and Bernstein's inequality bounds the sum by $C(1 + \bar f_1 R)\log(J/\eta)/N_1$. □

F.3 Proof of Theorem 29

Condition on the proxy sample and take the comparison point $\tilde\beta$ of Lemma 49; take the event of Lemma 52 with $r_{\max} = \frac{3}{2} r_{\mathrm{loc}}$. Let $\Delta = \widehat\beta - \tilde\beta$, $r = \|\Delta\| \le r_{\max}$, $q = \|\Delta_{S'^c}\|_1$, so that $\|\Delta\|_1 \le \|\Delta_{S'}\|_1 + q \le \sqrt{2s}\, r + q$.

Basic inequality. $\widehat\beta$ minimizes (146) and $\tilde\beta$ is feasible; since θ is supported on S', $\|\widehat\beta - \widehat\beta_{\mathrm{proxy}}\|_1 = \|\Delta - \theta\|_1 \ge \|\theta\|_1 - \|\Delta_{S'}\|_1 + q$, hence

$$L_1(\widehat\beta) - L_1(\tilde\beta) \le |\mathbb{V}_N(\Delta)| + \lambda(\|\Delta_{S'}\|_1 - q) \le |\mathbb{V}_N(\Delta)| + \lambda\sqrt{2s}\, r - \lambda q.$$

The penalty absorbs the part of Δ outside S'. In Lemma 52, $\sqrt{\|\Delta\|_1} \le \sqrt{\sqrt{2s}\, r} + \sqrt{q}$ and $\|\Delta\|_1/r \le \sqrt{2s} + q/r$. Now $C_1 D\sqrt{q/N_1} \le \frac{\lambda}{2} q + C_1^2 D^2/(2\lambda N_1)$, and $C_3 q D^2/(rN_1) \le \frac{\lambda}{2} q$ whenever $r \ge r_D := 2C_3 D^2/(\lambda N_1)$; for such r the terms in q cancel against $-\lambda q$. Thus, under the sign loss, the ℓ_1 penalty does not produce a cone condition; instead it dominates a process that grows like the square root of the ℓ_1 radius. With Lemma 50 and $E_N = (C_3\sqrt{2s}\, D^2 + C_2\Lambda)/N_1$, for $r \ge r_D$,

$$\frac{\underline c}{2} r^2 \le (\underline c + \bar c)\epsilon_0^2 + C_1 D (2s)^{1/4}\sqrt{\frac{r}{N_1}} + C_1\sqrt{\frac{r\Lambda}{N_1}} + E_N + \lambda\sqrt{2s}\, r + \frac{C_1^2 D^2}{2\lambda N_1}.$$

Choice of λ. By (147), $C_1 D(2s)^{1/4}/\sqrt{N_1} = \frac{\underline c}{4} r_\star^{3/2}$, $\lambda\sqrt{2s}\, r = \frac{\underline c}{8} r\, r_\star$, $C_1^2 D^2/(2\lambda N_1) = \frac{\underline c}{4} r_\star^2$ and $r_D = (C_3\underline c/C_1^2)\, r_\star^2 = \underline c\, r_\star^2/(4C_{\mathrm{ep}}\bar f_1 R)$. With $r_\Lambda = (\frac{8C_1}{\underline c}\sqrt{\Lambda/N_1})^{2/3}$ we have $C_1\sqrt{r\Lambda/N_1} \le \frac{\underline c}{8}(r^2 + r_\Lambda^2)$. Hence, for $r \ge r_D$,

$$\frac{\underline c}{2} r^2 \le (\underline c + \bar c)\epsilon_0^2 + \frac{\underline c}{4} r_\star^{3/2}\sqrt{r} + \frac{\underline c}{8} r\, r_\star + \frac{\underline c}{4} r_\star^2 + \frac{\underline c}{8}(r^2 + r_\Lambda^2) + E_N.$$

Solving. If $r \le \max\{r_\star, r_D\}$ there is nothing to prove. Otherwise $r_\star^{3/2}\sqrt{r} \le r_\star r$ and $r_\star^2 \le r_\star r$, so $\frac{3\underline c}{8} r^2 - \frac{5\underline c}{8} r_\star r \le (\underline c + \bar c)\epsilon_0^2 + \frac{\underline c}{8} r_\Lambda^2 + E_N$, and $ax^2 - bx \le c$ implies $x \le b/a + \sqrt{c/a}$:

$$r \le \frac{5}{3} r_\star + \sqrt{\frac{8(\underline c + \bar c)}{3\underline c}}\,\epsilon_0 + \frac{r_\Lambda}{\sqrt{3}} + \sqrt{\frac{8E_N}{3\underline c}}.$$

In either case r is at most r_D plus the right-hand side. Finally $\|\widehat\beta - \beta^\star\| \le r + \epsilon_0$ with $\epsilon_0 \le \|e\|$ (Lemma 49), $r_\Lambda = (64 C_1^2 \Lambda/(\underline c^{\,2} N_1))^{1/3}$, $\sqrt{E_N} \le \sqrt{C_3\sqrt{2s}\, D^2/N_1} + \sqrt{C_2\Lambda/N_1}$ and $C_3 = 4C_{\mathrm{ep}}$. □

Remark 8 (Recovering $s \log d$ changes the problem)**.** *Two modifications restore the complexity $s \log d$, and each changes the problem. Replacing the ℓ_1 penalty by the constraint $\|\beta - \widehat\beta_{\mathrm{proxy}}\|_0 \le 2s$ makes the class a union of $\binom{d}{2s}$ classes of halfspaces in $2s + 1$ variables; its bracketing entropy*

is $O(s\log(d/s)+s\log(1/\epsilon))$ at every scale, the two regimes of Lemma 51 merge, and the same proof gives the gold term $(s\log d/N_1)^{1/3}$; but the constrained problem is a best-subset problem, not a relaxation of one. Replacing the sign by the smooth surrogate of Section 4 makes the increment process linear in Δ up to higher-order terms, so that the ℓ_1 geometry enters only through the ℓ_∞ norm of a gradient and contributes a factor $\sqrt{\log d}$; with Assumptions 4–5 the corresponding two-step estimator has, for a fixed design, the gold term $(s\log d/N_1)^{2/5}$ (Appendix G), the counterpart of Feng et al. (2022).

G SPARSE CORRECTION WITH THE SMOOTHED LOSS

This appendix proves Theorem 7(b). It uses the smoothed loss of Section 4 and the population facts of Appendix D, and it reuses the sphere geometry of Appendix F. Compared with the exact sign loss, the smoothed loss has a gradient, and the ℓ_1 geometry then enters only through the ℓ_∞ norm of that gradient; but the analysis then holds only for perturbations of the order of the bandwidth, which is why the estimator has a trust region.

Setting. Assumptions 1, 2, 4 and 5 hold; X has density $f_X \le \bar f$ with support in $\bar B(0,R)$, $\bar f_1$ bounds the density of $u^\top X$ for every unit u, $R_\infty = \sup\|x\|_\infty \le R$ over the support; δ is s-sparse, and $S' = S\cup T$ is as in Appendix F. The smoothed loss is $\ell_h(\beta;x,y) = \frac12[1-(2y-1)\tanh(x^\top\beta/h)]$, defined on all of $\mathbb{R}^d$; $\widehat L_{k,h} = \mathbb{P}_{N_k}\ell_h$, $L_{k,h} = P_k\ell_h$; $\kappa_0 = \tanh'$ as in Appendix D, with $|\kappa_0'(u)|, |\kappa_0''(u)| \le c_e e^{-2|u|}$ and $\int\kappa_0^2 = \frac43$. Then $\nabla_\beta\ell_h = -\frac{2y-1}{2h}\kappa_0(x^\top\beta/h)x$ and $\nabla^2_\beta\ell_h = -\frac{2y-1}{2h^2}\kappa_0'(x^\top\beta/h)xx^\top$, and by scale invariance of the sign, $\nabla^2 L_{k,h}(\xi) = \|\xi\|^{-2}\nabla^2 L_{k,h/\|\xi\|}(\xi/\|\xi\|)$ for $\xi\neq0$. We write $m_k(x) = \mathbb{E}_k[2Y-1 \mid X=x]$, for which $|m_k(x)| \le \frac12|x^\top\beta_k|$, H for the tangential curvature of the gold sign risk, so that $\mathcal{H}_1(\beta^\star) = EHE^\top$, $\underline h > 0$ for its smallest eigenvalue on the tangent space $\beta^{\star\perp}$ (Lemma 18), ω_H for the modulus of Lemma 23 and $\omega_{\mathcal H}$ for the modulus of continuity of $\beta\mapsto\mathcal{H}_1(\beta)$.

Estimator. Step 1: $\widehat\beta_{\mathrm{proxy}} \in \arg\min_{\|\beta\|=1}\widehat L_{2,h_2}(\beta)$. Step 2, with $r_{\mathrm{loc}} = h_1/(8R)$:

$$\widehat\beta \in \arg\min\Big\{\widehat L_{1,h_1}(\beta) + \lambda\|\beta-\widehat\beta_{\mathrm{proxy}}\|_1 :\ \|\beta\|=1,\ \|\beta-\widehat\beta_{\mathrm{proxy}}\|_2 \le r_{\mathrm{loc}}\Big\}. \tag{149}$$

Write $e = \widehat\beta_{\mathrm{proxy}} - \beta_2$.

Theorem 30 (Theorem 7(b) with explicit constants)**.** *There are $h_0, r_1 > 0$ with $h_0 \le 2Rr_1$, $2h_0\bar f_1 \le 1$ and $h_0\underline h \le 64C_{\mathrm{rsc}}$, and constants $C_{\mathrm{off}}, C_H, C_{\mathrm{rsc}}$ depending only on the design and the law of ε (Lemmas 53 and 55), such that the following holds for $h_1 \le h_0$. Put $C_\lambda = \max\{4(\sqrt{2\bar f_1/3}+\frac23)R_\infty,\ 4C_{\mathrm{rsc}}/R\}$ and $\lambda = C_\lambda\sqrt{\log(8d/\eta)/(N_1h_1)}$. Suppose (C1) $\|\delta\| + 2\|e\| \le r_{\mathrm{loc}}/2$; (C2) $N_1h_1 \ge \log(8d/\eta)$ and $N_1h_1^3 \ge (64C_{\mathrm{rsc}})^2\log(2J/\eta)/\underline h^{\,2}$ with $J = (\lceil\log_2 N_1\rceil+1)(\lceil\log_2\sqrt d\rceil+2)$; (C3) $N_1h_1^2\sqrt{\underline h} \ge 1$. Then, conditionally on the proxy sample, with probability at least $1-\eta$ over the gold sample, every minimizer $\widehat\beta$ of* (149) *satisfies*

$$\|\widehat\beta-\beta^\star\|_2 \le \frac{34\sqrt s\,\lambda}{\underline h} + \Big(\frac{8C_{\mathrm{off}}}{\underline h}+1\Big)h_1^2 + \Big(1+\frac{8C_H}{\underline h}\Big)\|e\|. \tag{150}$$

With $h_1 = (s\log(8d/\eta)/N_1)^{1/5}$ the first two terms are $O((s\log(8d/\eta)/N_1)^{2/5})$; by Theorem 20 applied to the proxy population alone, $\|e\| = O_p(N_2^{-2/5})$ for fixed d when $h_2 \asymp N_2^{-1/5}$.

Condition (C1) requires the bias and the proxy error to be below a constant times the bandwidth h_1, which is above the gold error scale h_1^2 by the factor $1/h_1$; (C2) and (C3) are sample conditions. The tuning uses s and d only through h_1 and λ. The constants $\underline h$, C_{off}, C_H, C_{rsc}, h_0 and r_1 depend on the design and the law of ε, hence on d, and are not tracked here: the statement is one for a fixed design, in which the dependence on s and $\log d$ is explicit.

G.1 POPULATION FACTS NEAR THE TRUTH

Lemma 53. *There are $h_0, r_1 > 0$ and constants C_{off}, C_H such that for $h \le h_0$: (a) $\|\nabla L_{1,h}(\beta^\star)\|_2 \le C_{\mathrm{off}}h^2$; (b) for unit $\beta,\beta' \in B(\beta^\star,r_1)$ and $\Delta = \beta'-\beta$, $L_{1,h}(\beta') - L_{1,h}(\beta) - \nabla L_{1,h}(\beta)^\top\Delta \ge \frac{\underline h}{4}\|\Delta\|^2$; (c) $\|\nabla^2 L_{1,h}(\xi)\|_{\mathrm{op}} \le C_H$ whenever $\|\xi\| \in [\frac12,1]$ and $\xi/\|\xi\| \in B(\beta^\star,r_1)$.*

Proof. (a) Lemma 22 gives $\Pi_{\beta^\star}\nabla L_{1,h}(\beta^\star) = -h^2 EHb + o(h^2)$, hence $\le C_{\tan}h^2$ for $h \le h_0$. The radial component is $\beta^{\star\top}\nabla L_{1,h}(\beta^\star) = -\frac{1}{2h}\mathbb{E}[m_1(X)\, t\, \kappa_0(t/h)]$ with $t = X^\top\beta^\star$; by $|m_1| \le |t|/2$ and the marginal density bound, $|\beta^{\star\top}\nabla L_{1,h}(\beta^\star)| \le \frac{\bar f_1}{4h}\int t^2\kappa_0(t/h)dt = \frac{\pi^2\bar f_1}{24}h^2$. (b)–(c) For a unit vector $\hat\xi$ and bandwidth h', the radial row of the Hessian is bounded by

$$\begin{aligned}\|\hat\xi^\top\nabla^2 L_{1,h'}(\hat\xi)\| &\le \frac{R}{4h'^2}\mathbb{E}\big[(|t'| + R\|\hat\xi - \beta^\star\|)\,|t'\kappa_0'(t'/h')|\big]\\ &\le \frac{R\bar f_1}{4}\big[c_2 h' + c_1 R\|\hat\xi - \beta^\star\|\big] =: C_{\rm rad}(h', \hat\xi),\end{aligned}$$

$t' = X^\top\hat\xi$, $c_1 = \int|u\kappa_0'|$, $c_2 = \int u^2|\kappa_0'|$, using $|m_1(X)| \le \frac12(|t'| + R\|\hat\xi - \beta^\star\|)$. The tangential block satisfies $\|\Pi_{\hat\xi}\nabla^2 L_{1,h'}(\hat\xi)\Pi_{\hat\xi} - \mathcal{H}_1(\hat\xi)\| \le \omega_H(h')$ (Lemma 23) and $\|\mathcal{H}_1(\hat\xi) - EHE^\top\| \le \omega_{\mathcal H}(\|\hat\xi - \beta^\star\|)$. Choose h_0, r_1 with $\omega_H(2h_0) + \omega_{\mathcal H}(2r_1) + 4C_{\rm rad}(2h_0, \cdot) \le \underline h/8$; scale invariance gives (c) with $C_H = 4(\|H\| + \underline h/8 + 2C_{\rm rad})$. For (b), the points $\xi_t = \beta + t\Delta$ of the chord satisfy $\|\xi_t\|^2 = 1 - t(1-t)\|\Delta\|^2$ and $|\hat\xi_t^\top\Delta| \le \|\Delta\|^2$; decomposing Δ into its components along and orthogonal to $\hat\xi_t$, $\Delta^\top\nabla^2 L_{1,h}(\xi_t)\Delta \ge \|\xi_t\|^{-2}[(\underline h - \omega_{\mathcal H} - \omega_H)(\|\Delta\|^2 - \|\Delta\|^4) - 3C_{\rm rad}\|\Delta\|^3] \ge \frac{\underline h}{2}\|\Delta\|^2$ for r_1 small, where the rotation of the tangent space from $\beta^\star$ to $\hat\xi_t$ reduces the lower bound by a further $\underline h\|\hat\xi_t - \beta^\star\|^2 \le 4r_1^2\underline h$, absorbed for r_1 small; the integral remainder $\int_0^1(1-t)\Delta^\top\nabla^2 L_{1,h}(\xi_t)\Delta\, dt$ is at least $\frac{\underline h}{4}\|\Delta\|^2$. □

Lemma 54 (Score at a fixed point)**.** *For a unit vector β independent of the gold sample and $h \le h_0$, with probability at least $1-\eta$, $\|\nabla\widehat L_{1,h}(\beta) - \nabla L_{1,h}(\beta)\|_\infty \le \sqrt{\frac{2R_\infty^2\bar f_1\log(2d/\eta)}{3N_1 h}} + \frac{2R_\infty\log(2d/\eta)}{3N_1 h}$.*

Proof. Coordinate j averages $Z_i = -\frac{2Y_i-1}{2h}\kappa_0(X_i^\top\beta/h)X_{ij}$, with $|Z_i| \le R_\infty/(2h)$ and $\operatorname{Var} Z_i \le \frac{R_\infty^2}{4h^2}\mathbb{E}\kappa_0^2(X^\top\beta/h) \le \frac{R_\infty^2}{4h^2}\bar f_1 h\int\kappa_0^2 = \frac{R_\infty^2\bar f_1}{3h}$; the centered variables are bounded by $\frac{R_\infty}{2h}(1+2h\bar f_1) \le \frac{R_\infty}{h}$ since $2h\bar f_1 \le 1$; Bernstein's inequality and a union bound over the d coordinates. □

The variance is of order $1/h$, not $1/h^2$ as a Lipschitz bound would give, because the score is concentrated on a band of width $O(h)$ around the hyperplane; this step raises the exponent from $1/3$ to $2/5$.

G.2 Restricted strong convexity on the ℓ_1 cone

Fix $h \le h_0$ and a unit vector $\tilde\beta$ independent of the gold sample with $\|\tilde\beta - \beta^\star\| \le r_1/2$, and for $r \le h/(4R)$ and $\rho_1 > 0$ let $\mathcal{D}(r,\rho_1) = \{\Delta = \beta' - \tilde\beta :\ \|\beta'\| = 1,\ \|\Delta\|_2 \le r,\ \|\Delta\|_1 \le \rho_1\}$.

Lemma 55 (RSC)**.** *There is $C_{\rm rsc}$, depending only on $R, R_\infty, \bar f_1$, such that with probability at least $1-\eta$, simultaneously for all $\Delta \in \mathcal{D}(r,\rho_1)$,*

$$\begin{aligned}&\widehat L_{1,h}(\tilde\beta + \Delta) - \widehat L_{1,h}(\tilde\beta) - \nabla\widehat L_{1,h}(\tilde\beta)^\top\Delta\\ &\quad\ge \frac{\underline h}{4}\|\Delta\|^2 - C_{\rm rsc}\Big[\rho_1 r\sqrt{\frac{\log 2d}{N_1 h^3}} + r^2\sqrt{\frac{\log(1/\eta)}{N_1 h^3}} + \frac{r^2\log(1/\eta)}{N_1 h^2}\Big].\end{aligned}$$

Proof. For each observation $\ell_h(\cdot; x, y)$ is smooth, and Taylor's formula with integral remainder writes the left side as $\mathbb{P}_{N_1}\bar g_\Delta$ with

$$\bar g_\Delta(x,y) = -\frac{2y-1}{2h^2}\,\varphi_u(a), \qquad \varphi_u(a) = a^2\int_0^1(1-t)\,\kappa_0'\Big(\frac{u+ta}{h}\Big)dt, \qquad u = x^\top\tilde\beta,\ a = x^\top\Delta,$$

and $P\bar g_\Delta$ is the population remainder, at least $\frac{\underline h}{4}\|\Delta\|^2$ by Lemma 53(b), which applies because $h \le h_0 \le 2Rr_1$ puts $\tilde\beta + \Delta$ in $B(\beta^\star, r_1)$. It remains to bound $\sup_{\mathcal D}|(\mathbb{P}_{N_1} - P)\bar g_\Delta|$.

Expectation. By symmetrization, $\mathbb{E}\sup_{\mathcal D}|(\mathbb{P}_{N_1} - P)\bar g_\Delta| \le 2\,\mathbb{E}\sup_{\mathcal D}|\frac{1}{N_1}\sum_i \epsilon_i\bar g_\Delta(X_i, Y_i)|$, and given the data $\epsilon_i(2Y_i - 1)$ are again independent signs, so the factor $2Y_i - 1$ can be dropped. For $|a| \le Rr \le h/4$, $\varphi_{u_i}(0) = 0$ and $|\varphi_{u_i}'(a)| \le c_e(\frac{R^2r^2}{h} + 2Rr)e^{-2|u_i|/h}e^{2Rr/h} \le 4c_e Rr\, e^{-2|u_i|/h} =:$

M_i, using $|\kappa_0'|, |\kappa_0''| \le c_e e^{-2|\cdot|}$; extend each φ_{u_i} by constants outside $[-Rr, Rr]$. The contraction principle (Theorem 4.12 of Ledoux & Talagrand, 1991), applied to the maps $a \mapsto \varphi_{u_i}(a)$ with the data-dependent Lipschitz constants M_i, and the bound $\|\Delta\|_1 \le \rho_1$ give

$$\mathbb{E}_\epsilon \sup_{\mathcal{D}} \Big| \sum_i \epsilon_i \varphi_{u_i}(X_i^\top \Delta) \Big| \le 2\rho_1 \, \mathbb{E}_\epsilon \Big\| \sum_i \epsilon_i M_i X_i \Big\|_\infty \le 2\rho_1 \sqrt{2\log(2d)} \, \max_j \Big(\sum_i M_i^2 X_{ij}^2 \Big)^{1/2},$$

the last step being the maximum of $2d$ subgaussian variables. Here $\sum_i M_i^2 X_{ij}^2 \le 16 c_e^2 R^2 r^2 R_\infty^2 \sum_i e^{-4|u_i|/h}$ for every j, and since $\mathbb{E} e^{-4|X^\top \tilde\beta|/h} \le \bar f_1 h/2$, Jensen's inequality gives $\mathbb{E}(\sum_i e^{-4|u_i|/h})^{1/2} \le (N_1 \bar f_1 h/2)^{1/2}$. Collecting the factors $\frac{1}{2h^2}$ and 2,

$$\mathbb{E} \sup_{\mathcal{D}} |(\mathbb{P}_{N_1} - P)\bar g_\Delta| \le 8 \, c_e R R_\infty \bar f_1^{1/2} \, \frac{\rho_1 r}{h^{3/2}} \sqrt{\frac{\log 2d}{N_1}}.$$

The Lipschitz constants are concentrated on the band of width $O(h)$ around the hyperplane of $\tilde\beta$; the contraction principle absorbs them observation by observation, which gives the factor $h^{-3/2}$ without any net over the d-dimensional set of directions.

Concentration. The class is bounded by $U = \|\kappa_0'\|_\infty R^2 r^2/(4h^2)$, so its centered version by $2U$, and has variances at most $\sigma^2 = \frac{R^4 r^4}{4h^4} c_e^2 e \, \bar f_1 \frac{h}{2}$; since $\bar g_\Delta$ is continuous in Δ the supremum may be taken over a countable dense subset. Bousquet's inequality (Bousquet, 2002) gives, with probability at least $1-\eta$, $\sup \le \mathbb{E}\sup + \sqrt{2(\sigma^2 + 4U\,\mathbb{E}\sup)\log(1/\eta)/N_1} + 2U\log(1/\eta)/(3N_1)$, whose terms are the three terms of the lemma up to constants. □

Lemma 56 (RSC, uniform version)**.** *Let $r_{\max} \le h/(4R)$ and $J = (\lceil \log_2 N_1 \rceil + 1)(\lceil \log_2 \sqrt{d} \rceil + 2)$. With probability at least $1-\eta$, for all $\Delta = \beta' - \tilde\beta$ with β' unit and $\|\Delta\| \le r_{\max}$,*

$$\widehat L_{1,h}(\tilde\beta + \Delta) - \widehat L_{1,h}(\tilde\beta) - \nabla \widehat L_{1,h}(\tilde\beta)^\top \Delta \ge \frac{h}{4}\|\Delta\|^2 - \epsilon_{\mathrm{rsc}} \|\Delta\|_1 \|\Delta\|_2 - \epsilon'_{\mathrm{rsc}} \|\Delta\|_2^2 - \varsigma,$$

with $\epsilon_{\mathrm{rsc}} = 4C_{\mathrm{rsc}} \sqrt{\log(2d)/(N_1 h^3)}$, $\epsilon'_{\mathrm{rsc}} = 4C_{\mathrm{rsc}}[\sqrt{\log(J/\eta)/(N_1 h^3)} + \log(J/\eta)/(N_1 h^2)]$ and $\varsigma = 1/(16N_1^2)$.

Proof. Apply Lemma 55 at the radial scales $r_j = 2^{-j} r_{\max}$, $j = 0, \dots, \lceil \log_2 N_1 \rceil$, and, for each j, at the scales $\rho_{1,j,k} = 2^{-k}\sqrt{d}\, r_j$, $k = 0, \dots, \lceil \log_2 \sqrt{d} \rceil + 1$, each with confidence η/J. A Δ with $\|\Delta\| \in (r_{j+1}, r_j]$ has $\|\Delta\|_1 \in (r_{j+1}, \sqrt{d}\, r_j]$, so it lies in a cell with $r_j \le 2\|\Delta\|$ and $\rho_{1,j,k} \le 2\|\Delta\|_1$, and the bound of that cell gives the claim. The cells cover $\|\Delta\| > r_{\max}/N_1$. For $\|\Delta\| \le r_{\max}/N_1$, $|\bar g_\Delta| \le \|\kappa_0'\|_\infty R^2 \|\Delta\|^2/(4h^2) \le 0.77/(64N_1^2)$ because $R r_{\max} \le h/4$, so $|(\mathbb{P}_{N_1} - P)\bar g_\Delta| \le 1/(16N_1^2) = \varsigma$ and the claim follows from $P\bar g_\Delta \ge \frac{h}{4}\|\Delta\|^2$. □

G.3 Proof of Theorem 30

Condition on the proxy sample. Lemma 49 provides a unit vector $\tilde\beta$ with $\theta = \widehat\beta_{\mathrm{proxy}} - \tilde\beta$ supported on S', $|S'| \le 2s$, and $\epsilon_0 := \|\tilde\beta - \beta^\star\| \le \|e\|$. By (C1), $\|\tilde\beta - \widehat\beta_{\mathrm{proxy}}\| \le \epsilon_0 + \|\delta\| + \|e\| \le \|\delta\| + 2\|e\| \le r_{\mathrm{loc}}/2$, so $\tilde\beta$ is feasible, $\epsilon_0 \le r_1/2$, and $\Delta := \widehat\beta - \tilde\beta$ satisfies $\|\Delta\| \le \frac{3}{2} r_{\mathrm{loc}} =: r_{\max}$ with $r_{\max} \le h_1/(4R)$, and $\widehat\beta \in B(\beta^\star, r_1)$ because $h_1 \le h_0 \le 2Rr_1$. Take the events of Lemma 54 at $\tilde\beta$ and of Lemma 56, each with confidence $\eta/2$; by (C2) and the choice of C_λ, $\|\nabla(\widehat L_{1,h_1} - L_{1,h_1})(\tilde\beta)\|_\infty \le \lambda/4$. Write $\widehat L = \widehat L_{1,h_1}$, $L = L_{1,h_1}$, $r = \|\Delta\|$, $q = \|\Delta_{S'^c}\|_1$, so that $\|\Delta\|_1 \le \sqrt{2s}\, r + q$.

Basic inequality. $\widehat\beta$ minimizes (149) and $\tilde\beta$ is feasible; since θ is supported on S', $\|\widehat\beta - \widehat\beta_{\mathrm{proxy}}\|_1 = \|\Delta - \theta\|_1 \ge \|\theta\|_1 - \|\Delta_{S'}\|_1 + q$, hence $\widehat L(\tilde\beta + \Delta) - \widehat L(\tilde\beta) \le \lambda(\|\Delta_{S'}\|_1 - q)$.

Lower bound of the left side. $\widehat L(\tilde\beta + \Delta) - \widehat L(\tilde\beta) = \nabla \widehat L(\tilde\beta)^\top \Delta + \mathbb{P}_{N_1} \bar g_\Delta$. The gradient term is at least $-\frac{\lambda}{4}\|\Delta\|_1 - \epsilon_{\mathrm{lin}}\|\Delta\|$ with $\epsilon_{\mathrm{lin}} := \|\nabla L(\tilde\beta)\|_2 \le C_{\mathrm{off}} h_1^2 + C_H \|e\|$ by Lemma 53(a)(c) along the segment from $\beta^\star$ to $\tilde\beta$. The second-order term is bounded below by Lemma 56; since $\|\Delta\| \le r_{\max}$,

$\epsilon_{\rm rsc}\|\Delta\| \le \frac{C_{\rm rsc}}{R}\sqrt{\log(2d)/(N_1 h_1)} \le \lambda/4$, and $\epsilon'_{\rm rsc} \le \underline{h}/8$ by (C2) and $h_0\underline{h} \le 64C_{\rm rsc}$, the lemma being applied with confidence $\eta/2$. Altogether

$$\frac{\underline{h}}{8}\|\Delta\|^2 \le \frac{3\lambda}{2}\|\Delta_{S'}\|_1 - \frac{\lambda}{2}\|\Delta_{S'^c}\|_1 + \epsilon_{\rm lin}\|\Delta\| + \varsigma,$$

and Young's inequality $\epsilon_{\rm lin}\|\Delta\| \le \frac{\underline{h}}{16}\|\Delta\|^2 + \frac{4\epsilon_{\rm lin}^2}{\underline{h}}$ turns the left side into $\frac{\underline{h}}{16}\|\Delta\|^2$ at the cost of $\frac{4\epsilon_{\rm lin}^2}{\underline{h}}$.

Two cases. If $\|\Delta_{S'^c}\|_1 \ge 3\|\Delta_{S'}\|_1$ the penalty terms are nonpositive and $\|\Delta\| \le \frac{8\epsilon_{\rm lin}}{\underline{h}} + \sqrt{16\varsigma/\underline{h}}$. Otherwise $\|\Delta\|_1 \le 4\|\Delta_{S'}\|_1 \le 4\sqrt{2s}\|\Delta\|$, so $\frac{\underline{h}}{16}\|\Delta\|^2 \le \frac{3\lambda}{2}\sqrt{2s}\|\Delta\| + \frac{4\epsilon_{\rm lin}^2}{\underline{h}} + \varsigma$ and $\|\Delta\| \le \frac{24\sqrt{2s}\lambda}{\underline{h}} + \frac{8\epsilon_{\rm lin}}{\underline{h}} + \sqrt{16\varsigma/\underline{h}}$. In both cases $\|\Delta\| \le \frac{34\sqrt{s}\lambda}{\underline{h}} + \frac{8\epsilon_{\rm lin}}{\underline{h}} + \sqrt{16\varsigma/\underline{h}}$, and (C3) makes $\sqrt{16\varsigma/\underline{h}} = 1/(N_1\sqrt{\underline{h}}) \le h_1^2$.

Conclusion. The argument applies to every minimizer on the intersection of the two events, and $\|\widehat{\beta} - \beta^\star\| \le \|\Delta\| + \epsilon_0$; inserting the bounds on $\epsilon_{\rm lin}$ and ϵ_0 gives (150). With $h_1 = (s\log(8d/\eta)/N_1)^{1/5}$ the terms $\lambda\sqrt{s} \asymp \sqrt{s\log(8d/\eta)/(N_1 h_1)}$ and h_1^2 balance at $(s\log(8d/\eta)/N_1)^{2/5}$. □

Remark 9. *The analysis requires $\|\Delta\|$ to be at most of the order of the bandwidth, for two reasons: the contraction step of Lemma 55 needs $|x^\top\Delta| \le h/4$, and the tolerance $\epsilon_{\rm rsc}\|\Delta\|_1\|\Delta\|$ is absorbed by the penalty only when $\|\Delta\| \lesssim h$. The trust region enforces this bound, and it is why Theorem 7(b) requires $\|\delta\| \lesssim h_1$ rather than a constant bound on $\|\delta\|$. The proxy error $\|e\|$ enters as a number; a finite-sample bound for it with explicit dependence on d would follow from the uniform deviation lemmas of Appendix D by tracking their constants, which we do not do, and we use the fixed-d asymptotic form of Theorem 20.*

H Sparse correction with a known Gaussian heterogeneity law: proof of Theorem 7(c)

The assumption of Theorem 7(c) is the following.

Assumption 6 (Known Gaussian heterogeneity)**.** *X is uniform on the ball of radius $\sqrt{d+2}$ (so that $\mathbb{E}XX^\top = I$), $\varepsilon \sim \mathcal{N}(0, \tau^2 I/(d+2))$ with τ known, both mean utility vectors lie in the ball $\mathcal{M}_R$ of radius R at a fixed distance from its boundary, and the shift $\delta_2 = \beta_2 - \beta_1$ between the centers has at most s nonzero coordinates.*

This appendix proves Theorem 7(c). The argument has three parts: the marginal likelihood is a correctly specified convex model whose Fisher information is bounded below by a constant free of the dimension (§H.1); its maximum likelihood estimator on a ball has an explicit finite-sample rate (§H.2); and the ℓ_1 correction of a proxy fit inherits the sparsity of the bias through a comparison with the same problem centered at the true proxy center β_2 and a stability bound in the penalty center (§H.3). Throughout, σ is the logistic function, $\sigma' = \sigma(1-\sigma)$, and $\|\cdot\|$ the Euclidean norm.

H.1 Model, likelihood and information

Under Assumption 6, $X \sim \mathrm{Unif}(B(0, \sqrt{d+2}))$, so that $\mathbb{E}[XX^\top] = I_d$ and $\|X\|^2 \le d+2$; $\varepsilon \sim \mathcal{N}(0, \frac{\tau^2}{d+2} I_d)$ with τ known; the two centers $\beta_1 = \bar\beta$ and $\beta_2 = \bar\beta + \delta_2$ lie in $\mathcal{M}_R = \{\|\beta\| \le R\}$ with $\|\beta_k\| \le R - \Delta$; and δ_2 has support S with $|S| \le s$. Given $X = x$, the conditional law of $x^\top\varepsilon$ is $\mathcal{N}(0, a(x)^2)$ with $a(x) = \tau\|x\|/\sqrt{d+2} \le \tau$, so

$$\Pr(Y = 1 \mid X = x) = G_x(x^\top\beta_k), \qquad G_x(t) = \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\,\sigma\big(t + a(x)Z\big). \tag{151}$$

With $Q_1(t;x) = G_x(t)$ and $Q_0(t;x) = 1 - G_x(t)$, the marginal negative log-likelihood of one observation is $\ell(y, x; \beta) = -\log Q_y(x^\top\beta; x)$, and $\widehat{L}_k(\beta) = \frac{1}{N_k}\sum_j \ell(Y_{k,j}, X_{k,j}; \beta)$ (in this appendix $\widehat{L}_k$ is the marginal likelihood, not the sign loss), $L_k = \mathbb{E}\widehat{L}_k$, and

$$c = \Pr_{Z\sim\mathcal{N}(0,1)}(|Z| \le 1)^2\,\sigma'(2R+\tau)^2, \qquad r_0 = \min\Big\{\frac{\Delta}{2},\ \frac{c}{24(d+2)^{3/2}}\Big\}. \tag{152}$$

Lemma 57 (Scalar facts). *Fix x and $y \in \{0,1\}$ and write $\ell_y(t) = -\log Q_y(t;x)$.*

(a) ℓ_y is convex, $0 \le \ell_y'' \le \frac{1}{4}$ and $|\ell_y'''| \le 6$. Hence $\beta \mapsto \ell(y,x;\beta)$ is convex on $\mathbb{R}^d$, $0 \preceq \nabla^2_\beta \ell \preceq \frac{1}{4}xx^\top$, and $\|\nabla^2 \widehat{L}_k(\beta) - \nabla^2 \widehat{L}_k(\beta')\|_{\mathrm{op}} \le 6(d+2)^{3/2}\|\beta - \beta'\|$.

(b) At the center β_k of population k, $\nabla_\beta \ell(Y,X;\beta_k) = \psi X$ with $\mathbb{E}[\psi \mid X] = 0$ and $|\psi - \psi'| \le 2$ for the two values ψ, ψ' that ψ takes given X; consequently $\mathbb{E}[e^{\theta h^\top \psi X} \mid X] \le e^{\theta^2 (h^\top X)^2/2}$ for every unit h and real θ.

(c) For every unit h, $\mathbb{E}[(h^\top X)^{2k}] \le (2k-1)!!$, hence $\mathbb{E}\, e^{\theta^2 (h^\top X)^2/2} \le (1-\theta^2)^{-1/2} \le e^{\theta^2}$ for $\theta^2 \le \frac{1}{2}$.

(d) For every $\beta \in \mathcal{M}_R$ the Fisher information $\mathcal{I}(\beta) = \mathbb{E}[w(X,\beta)XX^\top]$, $w(x,\beta) = \frac{(\partial_t G_x(x^\top\beta))^2}{G_x(x^\top\beta)(1-G_x(x^\top\beta))}$, satisfies $\mathcal{I}(\beta) \succeq cI_d$, and $\nabla^2 L_k(\beta_k) = \mathcal{I}(\beta_k)$.

Proof. (a) With $a = a(x)$ and $q_y(z;t) = \sigma((2y-1)(t+az))$, $Q_y(t;x) = \mathbb{E}q_y(Z;t)$. Since $\log\sigma$ is concave and the normal density is log-concave, $(t,z) \mapsto q_y(z;t)\varphi(z)$ is jointly log-concave, and Prékopa's theorem makes its marginal $Q_y(\cdot;x)$ log-concave: ℓ_y is convex. Under the tilted law with density proportional to $q_y(z;t)\varphi(z)$ let $U = 1 - q_y(Z;t)$ and $\mu = \mathbb{E}_{\mathrm{tilt}}U$; since $\partial_t q_y = (2y-1)q_y(1-q_y)$ and $\partial_t \log q_y = (2y-1)U$, one finds $\partial_t \log Q_y = (2y-1)\mu$ and $\ell_y'' = \mu(1-\mu) - 2\,\mathrm{Var}_{\mathrm{tilt}}(U) \le \mu(1-\mu) \le \frac{1}{4}$, while $\ell_y'' \ge 0$ by convexity. For the third derivative, $|q_y^{(j)}| \le q_y$ for $j \le 3$ (the derivatives are $q_y(1-q_y)$ times polynomials in q_y bounded by one on $[0,1]$), so $|Q_y^{(j)}| \le Q_y$ and $-(\log Q_y)''' = -[Q_y'''/Q_y - 3Q_y'Q_y''/Q_y^2 + 2Q_y'^3/Q_y^3]$ is bounded by $1+3+2=6$. The statements for $\ell(y,x;\beta) = \ell_y(x^\top\beta)$ follow from $\nabla^2_\beta \ell = \ell_y''(x^\top\beta)xx^\top$, $|x^\top(\beta-\beta')| \le \sqrt{d+2}\|\beta-\beta'\|$ and $\|xx^\top\|_{\mathrm{op}} \le d+2$. (b) With $p = G_x(x^\top\beta_k)$ and $v = \partial_t G_x(x^\top\beta_k) = \mathbb{E}\sigma'(x^\top\beta_k + aZ)$, the score is $-\frac{v}{p}x$ if $Y=1$ and $\frac{v}{1-p}x$ if $Y=0$, with conditional mean zero under (151). Since $\sigma' \le \min\{\sigma, 1-\sigma\}$, $v \le \min\{p, 1-p\}$, so the two values differ by $\frac{v}{p(1-p)} \le \frac{1}{\max\{p,1-p\}} \le 2$; Hoeffding's lemma for a centered variable with range $2|h^\top x|$ gives the exponential bound. (c) For X uniform on the ball of radius $\sqrt{d+2}$ and unit h, $\mathbb{E}(h^\top X)^{2k} = (d+2)^k(2k-1)!!/\prod_{j=1}^k(d+2j) \le (2k-1)!!$; then $\mathbb{E}e^{\theta^2(h^\top X)^2/2} \le \sum_k \frac{\theta^{2k}(2k-1)!!}{2^k k!} = \sum_k \binom{2k}{k}\frac{\theta^{2k}}{4^k} = (1-\theta^2)^{-1/2}$, and $(1-u)^{-1/2} \le e^u$ for $u \le \frac{1}{2}$. (d) The formula for $\mathcal{I}$ is the Bernoulli information with the chain rule, and $\nabla^2 L_k(\beta_k) = \mathcal{I}(\beta_k)$ is the information identity of a correctly specified model. For the lower bound fix $\beta \in \mathcal{M}_R$, let u be a unit vector along β (any unit vector if $\beta = 0$), $X_u = u^\top X$, $\mathcal{E} = \{|X_u| \le 2\}$. On $\mathcal{E}$, $|X^\top\beta| \le 2R$ and $a(X) \le \tau$, so on the event $|Z| \le 1$ the argument of σ' in $v = \mathbb{E}\sigma'(X^\top\beta + a(X)Z)$ has absolute value at most $2R+\tau$; as σ' is even and decreasing in $|\cdot|$, $v \ge \Pr(|Z| \le 1)\sigma'(2R+\tau)$ on $\mathcal{E}$, and with $G(1-G) \le \frac{1}{4}$, $w(X,\beta) \ge 4\Pr(|Z| \le 1)^2\sigma'(2R+\tau)^2$ on $\mathcal{E}$. For the uniform ball, $\mathbb{E}X_u^2 = 1$, $\mathbb{E}X_u^4 = \frac{3(d+2)}{d+4} \le 3$, and for a unit $u' \perp u$ with $X_{u'} = u'^\top X$, $\mathbb{E}X_{u'}^2 = 1$, $\mathbb{E}X_u^2X_{u'}^2 = \frac{d+2}{d+4} \le 1$, the cross moments vanishing by symmetry. Hence $\mathbb{E}[X_u^2\mathbf{1}_\mathcal{E}] \ge 1 - \frac{1}{4}\mathbb{E}X_u^4 \ge \frac{1}{4}$ and $\mathbb{E}[X_{u'}^2\mathbf{1}_\mathcal{E}] \ge 1 - \frac{1}{4}\mathbb{E}X_u^2X_{u'}^2 \ge \frac{3}{4}$, so $\mathbb{E}[XX^\top\mathbf{1}_\mathcal{E}] \succeq \frac{1}{4}I_d$ and $\mathcal{I}(\beta) \succeq cI_d$. □

H.2 The marginal maximum likelihood estimator

For a sample of size N from population k and $\eta' \in (0,1)$ let $A(\eta') = d\log 5 + \log(4/\eta')$.

Lemma 58 (Events). *Let $\widehat{L}$ be the marginal negative log-likelihood of a sample of size N from a population with center $\beta_0 \in \mathcal{M}_R$, $\|\beta_0\| \le R - \Delta$.*

(E1) If $N \ge 2A(\eta')$, then $\|\nabla\widehat{L}(\beta_0)\| \le 4\sqrt{A(\eta')/N}$ with probability at least $1-\eta'$.

(E2) If $N \ge 2\log(2d/\eta')$, then $\|\nabla\widehat{L}(\beta_0)\|_\infty \le 2\sqrt{\log(2d/\eta')/N}$ with probability at least $1-\eta'$.

(E3) If $N \ge \frac{2(d+2)}{c}\log\frac{d}{\eta'}$, then with probability at least $1-\eta'$, $\lambda_{\min}(\nabla^2\widehat{L}(\beta)) \ge \frac{c}{4}$ for all β with $\|\beta - \beta_0\| \le r_0$.

(E4) If $N \geq \frac{d+2}{2\log 2-1}\log\frac{d}{\eta'}$, then with probability at least $1-\eta'$, $\lambda_{\max}(\frac{1}{N}\sum_j X_j X_j^\top) \leq 2$, hence $\nabla^2\widehat{L}(\beta) \preceq \frac{1}{2}I_d$ for all β.

Proof. (E1)–(E2) By Lemma 57(b)(c), for a unit h and $W_j = h^\top\nabla\ell(Y_j, X_j;\beta_0)$, $\mathbb{E}e^{\theta W_j} \leq e^{\theta^2}$ for $|\theta| \leq 1/\sqrt{2}$, so by the Chernoff bound with $\theta = t/2$, $\Pr(|h^\top\nabla\widehat{L}(\beta_0)| \geq t) \leq 2e^{-Nt^2/4}$ for $0 < t \leq \sqrt{2}$. For (E2) take $h = e_j$, $t = 2\sqrt{\log(2d/\eta')/N}$ and a union bound over the d coordinates. For (E1) take a $\frac{1}{2}$-net of the unit sphere with at most 5^d points, for which $\|z\| \leq 2\max_h |h^\top z|$, and $t = 2\sqrt{A(\eta')/N}$. (E3) Let $B_j = \ell''_{Y_j}(X_j^\top\beta_0)X_j X_j^\top$; then $0 \preceq B_j \preceq \frac{d+2}{4}I$ and $\mathbb{E}B_j = \mathcal{I}(\beta_0) \succeq cI$ by Lemma 57(a)(d). The matrix Chernoff lower tail gives $\Pr(\lambda_{\min}(\frac{1}{N}\sum B_j) < \frac{c}{2}) \leq d\exp(-\frac{Nc}{2(d+2)}) \leq \eta'$. On this event, for $\|\beta-\beta_0\| \leq r_0$ by the Lipschitz bound of Lemma 57(a) the Hessian changes by at most $6(d+2)^{3/2}r_0 \leq \frac{c}{4}$. (E4) $0 \preceq X_j X_j^\top \preceq (d+2)I$ with mean I; the matrix Chernoff upper tail gives $\Pr(\lambda_{\max} \geq 2) \leq d\,(e/4)^{N/(d+2)} \leq \eta'$, and $\nabla^2\widehat{L} \preceq \frac{1}{4}\cdot\frac{1}{N}\sum X_j X_j^\top$ by Lemma 57(a). □

Lemma 59 (Localization and rate of the constrained MLE)**.** *On the events (E1) and (E3), if moreover $N \geq \frac{4096A(\eta')}{c^2 r_0^2}$, every minimizer of $\widehat{L}$ over $\mathcal{M}_R$ lies in the open ball of radius r_0 around β_0, is unique, and satisfies $\|\widehat{\beta}-\beta_0\| \leq \frac{16}{c}\sqrt{A(\eta')/N}$.*

Proof. The closed ball of radius $r_0 \leq \Delta/2$ around β_0 lies in $\mathcal{M}_R$. For $\|h\| = r_0$, the integral form of Taylor's theorem and (E3) give $\widehat{L}(\beta_0+h) - \widehat{L}(\beta_0) \geq \langle\nabla\widehat{L}(\beta_0), h\rangle + \frac{c}{8}\|h\|^2$, and the sample condition makes $\|\nabla\widehat{L}(\beta_0)\| \leq \frac{cr_0}{16}$, so the right side is at least $\frac{cr_0^2}{16} > 0$: on the sphere of radius r_0 the objective exceeds its value at β_0. Since $\widehat{L}$ is convex on $\mathbb{R}^d$, the segment from β_0 to any point of $\mathcal{M}_R$ outside the ball crosses the sphere, and the secant inequality shows that such a point is not a minimizer. A minimizer exists by compactness and is unique by (E3). It is interior, so $\nabla\widehat{L}(\widehat{\beta}) = 0$, and strong monotonicity of the gradient on the ball gives $\frac{c}{4}\|\widehat{\beta}-\beta_0\|^2 \leq \langle\nabla\widehat{L}(\widehat{\beta}) - \nabla\widehat{L}(\beta_0), \widehat{\beta}-\beta_0\rangle = -\langle\nabla\widehat{L}(\beta_0), \widehat{\beta}-\beta_0\rangle \leq 4\sqrt{A(\eta')/N}\,\|\widehat{\beta}-\beta_0\|$. □

H.3 The ℓ_1 correction of the proxy fit

Let $\widehat{\beta}_{\text{proxy}}$ be the marginal MLE of the proxy sample and, for $\lambda \geq 0$,

$$\widehat{\beta}_T(\lambda) \in \arg\min_{\beta\in\mathcal{M}_R}\Big\{\widehat{L}_1(\beta) + \lambda\|\beta - \widehat{\beta}_{\text{proxy}}\|_1\Big\}, \tag{153}$$

so that $\widehat{\beta}_T(0) = \widehat{\beta}_{\text{gold}}$ is the gold-only MLE. Let z_1 be the bound of (E2) for the gold sample.

Lemma 60 (Comparison with the penalty centered at β_2)**.** *Let (E2) and (E3) hold for the gold sample, and let $(z_1+\lambda)\sqrt{d} \leq \frac{cr_0}{16}$. Then every minimizer $\tilde{\beta}$ of $\widehat{L}_1(\beta) + \lambda\|\beta-\beta_2\|_1$ over $\mathcal{M}_R$ (the problem centered at the unknown proxy center) lies in the open ball of radius r_0 around $\beta^\star = \bar{\beta}$ and satisfies*

$$\|\tilde{\beta}-\bar{\beta}\| \leq \frac{8}{c}\Big[(\lambda+z_1)\sqrt{s} + (z_1-\lambda)_+\sqrt{d-s}\Big]. \tag{154}$$

The same localization holds for the minimizers of (153)*, whatever the center.*

Proof. For a center γ and $F(\beta) = \widehat{L}_1(\beta) + \lambda\|\beta-\gamma\|_1$, on the sphere $\|h\| = r_0$, $F(\bar{\beta}+h) - F(\bar{\beta}) \geq \langle\nabla\widehat{L}_1(\bar{\beta}), h\rangle + \frac{c}{8}r_0^2 - \lambda\|h\|_1 \geq -(z_1+\lambda)\sqrt{d}\,r_0 + \frac{c}{8}r_0^2 > 0$ by (E2), (E3) and the hypothesis; F is convex, and the secant argument of Lemma 59 confines all minimizers to the ball. For the oracle center $\gamma = \beta_2$, compare $\tilde{\beta}$ with $\bar{\beta}$, whose penalty is $\lambda\|\delta_2\|_1$: with $h = \tilde{\beta}-\bar{\beta}$ and $\tilde{\beta}-\beta_2 = h-\delta_2$, $\|h-\delta_2\|_1 \geq \|\delta_2\|_1 - \|h_S\|_1 + \|h_{S^c}\|_1$ because δ_2 is supported on S, so $\frac{c}{8}\|h\|^2 \leq -\langle\nabla\widehat{L}_1(\bar{\beta}), h\rangle + \lambda(\|h_S\|_1 - \|h_{S^c}\|_1) \leq (z_1+\lambda)\|h_S\|_1 + (z_1-\lambda)\|h_{S^c}\|_1$, and $\|h_S\|_1 \leq \sqrt{s}\|h\|$, $\|h_{S^c}\|_1 \leq \sqrt{d-s}\|h\|$ give (154). □

Lemma 61 (Stability in the penalty center)**.** *Let (E3) and (E4) hold for the gold sample, and let both $\widehat{\beta}_T(\lambda)$ and $\tilde{\beta}$ lie in the open ball of radius r_0 around $\bar{\beta}$. Then $\|\widehat{\beta}_T(\lambda) - \tilde{\beta}\| \leq \frac{2}{c}\|\widehat{\beta}_{\text{proxy}} - \beta_2\|$.*

Proof. Write $b = \widehat{\beta}_T(\lambda)$, $b_0 = \tilde{\beta}$, $z = \widehat{\beta}_{\mathrm{proxy}}$, $z_0 = \beta_2$, $q = b - b_0$, $e = z - z_0$. Both minimizers are interior points of $\mathcal{M}_R$, so there are $u \in \partial\|b - z\|_1$ and $u_0 \in \partial\|b_0 - z_0\|_1$ with $\nabla\widehat{L}_1(b) + \lambda u = 0$ and $\nabla\widehat{L}_1(b_0) + \lambda u_0 = 0$. The subdifferential of a convex function is monotone: $\langle u - u_0, (b - z) - (b_0 - z_0)\rangle \geq 0$, that is, $\langle u - u_0, q\rangle \geq \langle u - u_0, e\rangle$. By (E3), the gradient is $\frac{c}{4}$-strongly monotone on the ball; by (E4), it is $\frac{1}{2}$-Lipschitz on $\mathcal{M}_R$. Hence

$$\begin{aligned}\frac{c}{4}\|q\|^2 &\leq \langle \nabla\widehat{L}_1(b) - \nabla\widehat{L}_1(b_0), q\rangle = -\lambda\langle u - u_0, q\rangle \leq -\lambda\langle u - u_0, e\rangle \\ &= \langle \nabla\widehat{L}_1(b) - \nabla\widehat{L}_1(b_0), e\rangle \leq \tfrac{1}{2}\|q\|\,\|e\|.\end{aligned}$$

□

Theorem 31 (Theorem 7(c) with explicit constants)**.** *Let Assumptions 1, 2 and 6 hold and put* $A_\eta = d\log 5 + \log(32/\eta)$, $\ell_\eta = \log(32d/\eta)$, $z_1 = 2\sqrt{\ell_\eta/N_1}$,

$$G = \frac{16}{c}\sqrt{\frac{A_\eta}{N_1}}, \qquad T = \frac{32}{c}\sqrt{\frac{s\,\ell_\eta}{N_1}} + \frac{32}{c^2}\sqrt{\frac{A_\eta}{N_2}}. \tag{155}$$

Suppose the gold sample satisfies $N_1 \geq \max\{2A_\eta,\ \frac{2(d+2)}{c}\log\frac{8d}{\eta},\ \frac{4096A_\eta}{c^2r_0^2},\ \frac{d+2}{2\log 2-1}\log\frac{8d}{\eta}\}$ *and the proxy sample* $N_2 \geq \max\{2A_\eta,\ \frac{2(d+2)}{c}\log\frac{8d}{\eta},\ \frac{4096A_\eta}{c^2r_0^2}\}$. *Then, with probability at least* $1-\eta$*, the gold-only MLE satisfies* $\|\widehat{\beta}_{\mathrm{gold}} - \bar{\beta}\| \leq G$*, and if moreover* $N_1 \geq \frac{4096\,d\,\ell_\eta}{c^2r_0^2}$*, the two-step estimator with* $\lambda = z_1$ *satisfies* $\|\widehat{\beta}_T(z_1) - \bar{\beta}\| \leq T$*. Consequently the estimator that takes* $\widehat{\beta}_T(z_1)$ *when* $T < G$ *and the sample conditions hold, and* $\widehat{\beta}_{\mathrm{gold}}$ *otherwise, a choice made from* $(N_1, N_2, d, s, \eta, c, r_0)$ *alone, has error at most* $\min\{G, T\}$ *when the additional correction sample condition holds, and at most* G *otherwise, with probability at least* $1-\eta$*.*

Proof. Allocate the confidence as follows: (E1) and (E3) for the gold sample and (E1) and (E3) for the proxy sample with $\eta' = \eta/8$ each, so that $A(\eta/8) = A_\eta$; (E2) for the gold sample with $\eta' = \eta/16$, so that its bound is z_1; and (E4) for the gold sample with $\eta' = \eta/8$. The sample conditions of the theorem are those of Lemma 58 at these levels, and the total failure probability is at most $\frac{7\eta}{8}$. On the intersection, Lemma 59 gives $\|\widehat{\beta}_{\mathrm{gold}} - \bar{\beta}\| \leq G$ and $\|\widehat{\beta}_{\mathrm{proxy}} - \beta_2\| \leq \frac{16}{c}\sqrt{A_\eta/N_2}$. If $N_1 \geq 4096d\ell_\eta/(c^2r_0^2)$ then $2z_1\sqrt{d} \leq \frac{cr_0}{16}$, so Lemma 60 with $\lambda = z_1$ gives $\|\tilde{\beta} - \bar{\beta}\| \leq \frac{16z_1\sqrt{s}}{c} = \frac{32}{c}\sqrt{s\ell_\eta/N_1}$ and places both $\tilde{\beta}$ and $\widehat{\beta}_T(z_1)$ in the ball; Lemma 61 adds $\frac{32}{c^2}\sqrt{A_\eta/N_2}$. The rule compares deterministic quantities and uses, on the event, the bound of the branch it selects. □

The gold term depends on s instead of d because the ℓ_∞ bound on the score of the marginal likelihood depends on d only through $\log d$ and the ℓ_1 penalty is decomposable across the support of δ_2; the proxy term depends on d because the proxy stage estimates a dense center, and its error enters through the Euclidean stability of Lemma 61 rather than through the ℓ_1 norm of the center perturbation, which is what keeps the proxy term at $\sqrt{d/N_2}$ instead of $(d^2/(N_1N_2))^{1/4}$. No upper bound on $\|\delta_2\|$ is needed: the comparison in Lemma 60 is with $\bar{\beta}$ itself, whose penalty $\lambda\|\delta_2\|_1$ cancels. The sample thresholds are conservative (of order d^4 for fixed R, τ, Δ) and are not claimed to be sharp.

H.4 Comparison with the unpenalized procedures

The procedures below use no penalty on the difference between the two fits: gold-only and proxy-only maximum likelihood, the average $\widehat{\beta}_{\mathrm{avg}}(\varpi) = (1-\varpi)\widehat{\beta}_{\mathrm{gold}} + \varpi\widehat{\beta}_{\mathrm{proxy}}$, $\varpi \in [0,1]$, and re-weighting $\widehat{\beta}_{\mathrm{WE}}(\rho) \in \arg\min_{\mathcal{M}_R}\{(1-\rho)\widehat{L}_1 + \rho\widehat{L}_2\}$, $\rho \in [0,1]$, with pooling at $\rho = N_2/N$. As elsewhere in the paper, an upper bound for the proposed estimator is compared with a lower bound for each of these procedures at its best tuning, here both for the expected Euclidean error. The lower bounds are limits, so d, s, R, τ and Δ are fixed, $\|\bar{\beta}\| \leq R - 2\Delta$, $N \to \infty$ with $N_k/N \to \nu_k \in (0,1)$, and the shift is local: $\delta_2 = \delta_{2,N}$ is s-sparse with $\sqrt{N}\delta_{2,N} \to \xi$. Write $\mathbb{P}_{N_k}$ for the empirical measure of sample k, $\mathcal{I} = \mathcal{I}(\bar{\beta})$, $Z \sim \mathcal{N}(0, \mathcal{I}^{-1})$, $\sigma_\rho^2 = (1-\rho)^2/\nu_1 + \rho^2/\nu_2$, and χ_d for the norm of a standard Gaussian vector in $\mathbb{R}^d$. Since $\sigma' = \sigma(1-\sigma)$ and $u \mapsto u(1-u)$ is concave, $\partial_t G_x \leq G_x(1-G_x)$, so the weight in Lemma 57(d) is at most $\partial_t G_x \leq \frac{1}{4}$ and $\mathcal{I}(\beta) \preceq \frac{1}{4}I_d$. Hence

$\mathcal{I}^{-1} \succeq 4I_d$, so that $Z \overset{d}{=} 2Z_0 + Z'$ with $Z_0 \sim \mathcal{N}(0, I_d)$ and $Z' \sim \mathcal{N}(0, \mathcal{I}^{-1} - 4I_d)$ independent, and Jensen's inequality given Z_0 yields $\mathbb{E}\|Z\| \geq 2\,\mathbb{E}\chi_d$.

Lemma 62 (Local limits). *For weights $\rho_N \to \rho \in [0,1]$ and $\varpi_N \to \varpi \in [0,1]$, $\sqrt{N}(\widehat{\beta}_{\mathrm{WE}}(\rho_N) - \bar{\beta}) \Rightarrow \rho\xi + \sigma_\rho Z$ and $\sqrt{N}(\widehat{\beta}_{\mathrm{avg}}(\varpi_N) - \bar{\beta}) \Rightarrow \varpi\xi + \sigma_\varpi Z$.*

Proof. First let $\delta_2 = 0$, so that both samples have center $\bar{\beta}$. The weighted criterion is convex with a C^3 population version whose Hessian at the interior minimizer $\bar{\beta}$ is $\mathcal{I}$ (Lemma 57), the estimator is consistent by the uniform law of large numbers over $\mathcal{M}_R$, and the proof of Theorem 5.23 of van der Vaart (1998), which applies to each sample separately with the bounded factors $\sqrt{N/N_k}$ and to weights $\rho_N \to \rho$, gives $\sqrt{N}(\widehat{\beta}_{\mathrm{WE}}(\rho_N) - \bar{\beta}) = \mathcal{I}^{-1}\sqrt{N}[(1-\rho_N)\mathbb{P}_{N_1}u + \rho_N\mathbb{P}_{N_2}u] + o_p(1)$ with $u = -\nabla_\beta \ell(\cdot;\bar{\beta})$. The cases $\rho_N \equiv 0, 1$ are the gold-only and proxy-only estimators, so the average has the same expansion with ϖ_N in place of ρ_N. The marginal model is differentiable in quadratic mean at $\bar{\beta}$, since $\sqrt{Q_y}$ is continuously differentiable in β and $\mathcal{I}(\beta)$ is continuous (van der Vaart, 1998, Lemma 7.6); so under $\delta_{2,N}$ the log-likelihood ratio of the proxy sample against $\delta_2 = 0$ is $h_N^\top\sqrt{N_2}\mathbb{P}_{N_2}u - \frac{1}{2}h_N^\top\mathcal{I}h_N + o_p(1)$ with $h_N = \sqrt{N_2}\delta_{2,N}$ (van der Vaart, 1998, Theorem 7.2). The covariance of $\rho_N\sqrt{N/N_2}\,\mathcal{I}^{-1}\sqrt{N_2}\mathbb{P}_{N_2}u$ with this linear term is $\rho_N\sqrt{N}\delta_{2,N} \to \rho\xi$, and Le Cam's third lemma (van der Vaart, 1998, Example 6.7) shifts the Gaussian limit by $\rho\xi$; the remainders stay negligible by contiguity. The covariance of the limit is $\mathcal{I}^{-1}[(1-\rho)^2\nu_1^{-1}\mathcal{I} + \rho^2\nu_2^{-1}\mathcal{I}]\mathcal{I}^{-1} = \sigma_\rho^2\mathcal{I}^{-1}$. □

Theorem 32 (Unpenalized procedures at their best tuning). *Let $\Upsilon(\xi) = \min_{\rho\in[0,1]} \max\{\rho\|\xi\|,\, 2\sigma_\rho\mathbb{E}\chi_d\}$. Then*

$$\liminf_N \sqrt{N} \inf_{\rho\in[0,1]} \mathbb{E}\|\widehat{\beta}_{\mathrm{WE}}(\rho) - \bar{\beta}\| \geq \Upsilon(\xi), \qquad \liminf_N \sqrt{N} \inf_{\varpi\in[0,1]} \mathbb{E}\|\widehat{\beta}_{\mathrm{avg}}(\varpi) - \bar{\beta}\| \geq \Upsilon(\xi),$$

and $\Upsilon(\xi) \geq \min\{\mathbb{E}\chi_d/\sqrt{\nu_1},\ \|\xi\|/4 + \mathbb{E}\chi_d/(2\sqrt{\nu_2})\}$; in the original units both errors are at least

$$(1+o(1))\min\Big\{\frac{\mathbb{E}\chi_d}{\sqrt{N_1}},\ \frac{\|\delta_2\|}{4} + \frac{\mathbb{E}\chi_d}{2\sqrt{N_2}}\Big\}.$$

Moreover, $\liminf \sqrt{N_1}\,\mathbb{E}\|\widehat{\beta}_{\mathrm{gold}} - \bar{\beta}\| \geq 2\mathbb{E}\chi_d$ and $\liminf \sqrt{N}\,\mathbb{E}\|\widehat{\beta}_{\mathrm{proxy}} - \bar{\beta}\| \geq \max\{\|\xi\|, 2\mathbb{E}\chi_d/\sqrt{\nu_2}\}$. If $\|\xi\| \geq 2m\,\mathbb{E}\chi_d/\sqrt{\nu_1}$ with $m > 0$, the lower bounds of re-weighting and averaging at their best weights are at least $\frac{m}{m+1}$ times the gold-only lower bound.

Proof. For $\rho_N \to \rho$, the portmanteau theorem for the continuous function $\|\cdot\|$ and Lemma 62 give $\liminf \sqrt{N}\,\mathbb{E}\|\widehat{\beta}_{\mathrm{WE}}(\rho_N) - \bar{\beta}\| \geq \mathbb{E}\|\rho\xi + \sigma_\rho Z\|$, which is at least $\rho\|\xi\|$ by Jensen's inequality and at least $\sigma_\rho\mathbb{E}\|Z\| \geq 2\sigma_\rho\mathbb{E}\chi_d$ by the symmetry of Z. Taking $1/N$-approximate minimizers and a convergent subsequence of the weights gives the infimum statements; the same argument applies to the average. For $\rho \leq \frac{1}{2}$, $2\sigma_\rho\mathbb{E}\chi_d \geq \mathbb{E}\chi_d/\sqrt{\nu_1}$; for $\rho \geq \frac{1}{2}$ a maximum is at least the average of $\|\xi\|/2$ and $\mathbb{E}\chi_d/\sqrt{\nu_2}$. For the last claim, $\sigma_\rho \geq (1-\rho)/\sqrt{\nu_1}$, and with $a = 2\mathbb{E}\chi_d/\sqrt{\nu_1}$, $\min_\rho \max\{\rho\|\xi\|, (1-\rho)a\} = \|\xi\|a/(\|\xi\| + a) \geq \frac{m}{m+1}a$. □

Proposition 13 (Expected error of the ℓ_1 correction). *Write $z_1(\eta)$, ℓ_η, A_η and $T(\eta)$ for the quantities of Theorem 31 at level η. Let $\eta_0 = \sqrt{s/d}$ and $\lambda = z_1(\eta_0)$, suppose the sample conditions of Theorem 31, including $N_1 \geq 4096\,d\,\ell_{\eta_0}/(c^2 r_0^2)$, hold at the level η_0, and let $\eta_{\min}$ be the smallest level at which they hold. Then*

$$\mathbb{E}\|\widehat{\beta}_T(\lambda) - \bar{\beta}\| \leq \Big(1 + \frac{1}{2\ell_{\eta_0}} + \frac{1}{2d\log 5}\Big)T(\eta_0) + 2R\,\eta_{\min}.$$

Proof. Fix λ. For $\eta \in [\eta_{\min}, \eta_0]$, on the events of the proof of Theorem 31 at the level η, Lemma 60 applies, since $\lambda \leq z_1(\eta)$ and $2z_1(\eta)\sqrt{d} \leq cr_0/16$, and gives $\|\tilde{\beta} - \bar{\beta}\| \leq \frac{8}{c}[(\lambda + z_1(\eta))\sqrt{s} + (z_1(\eta) - \lambda)\sqrt{d-s}]$; Lemmas 61 and 59 add $\frac{32}{c^2}\sqrt{A_\eta/N_2}$. Call the sum $T_\lambda(\eta)$: it is continuous and decreasing, $T_\lambda(\eta_0) = T(\eta_0)$, and the error exceeds $T_\lambda(\eta)$ with probability at most η. The error is at most $2R$, so

$$\mathbb{E}\|\widehat{\beta}_T - \bar{\beta}\| \leq T_\lambda(\eta_0) + \int_{\eta_{\min}}^{\eta_0} \eta\,|T_\lambda'(\eta)|\,d\eta + 2R\,\eta_{\min},$$

and $\eta|T'_\lambda(\eta)| = \frac{8}{c}(\sqrt{s}+\sqrt{d-s})(N_1\ell_\eta)^{-1/2} + \frac{16}{c^2}(N_2 A_\eta)^{-1/2}$. With $\ell_\eta \geq \ell_{\eta_0}$, $A_\eta \geq d\log 5$ and $\eta_0(\sqrt{s}+\sqrt{d-s}) \leq 2\sqrt{s}$, the integral is at most $T(\eta_0)[1/(2\ell_{\eta_0}) + 1/(2d\log 5)]$. □

Theorem 33 (The ℓ_1 correction against the unpenalized procedures)**.** *Let $\|\xi\| \geq 4\,\mathbb{E}\chi_d/\sqrt{\nu_1}$, so that the shift is at least twice the gold-only lower bound. Then gold-only, proxy-only, averaging and re-weighting at their best tuning satisfy $\liminf_N \sqrt{N_1}\,\mathbb{E}\|\cdot - \bar\beta\| \geq \mathbb{E}\chi_d$, while the estimator of Proposition 13 satisfies*

$$\limsup_N \sqrt{N_1}\,\mathbb{E}\|\widehat\beta_T - \bar\beta\| \leq \Big(1 + \frac{1}{2\ell_{\eta_0}} + \frac{1}{2d\log 5}\Big)\Big[\frac{32}{c}\sqrt{s\,\ell_{\eta_0}} + \frac{32}{c^2}\sqrt{\frac{A_{\eta_0}\nu_1}{\nu_2}}\Big].$$

The ratio of the upper bound to the lower bounds is at most $C[c^{-1}\sqrt{s\log(32d)/d} + c^{-2}\sqrt{\nu_1/\nu_2}]$ with C absolute; when $\nu_2/\nu_1 \geq d/(c^2 s\log(32d))$ it is at most $2Cc^{-1}\sqrt{s\log(32d)/d}$.

Proof. The lower bounds are Theorem 32 in units of N_1: the minimum there equals its first branch when $\|\xi\| \geq 4\mathbb{E}\chi_d/\sqrt{\nu_1}$. The upper bound is Proposition 13 multiplied by $\sqrt{N_1}$, where $\sqrt{N_1}\eta_{\min} \to 0$ because the sample conditions have the form $N_k \geq a(b+\log(1/\eta))$ and d is fixed. The ratio uses $\mathbb{E}\chi_d \geq \sqrt{d-1/2}$, $\ell_{\eta_0} \leq \frac{3}{2}\log(32d)$ and $A_{\eta_0} \leq d\log 160$. □

For shifts of order $N^{-1/2}$ that are at least twice the gold-only lower bound, averaging and re-weighting at their best weights keep an error of order at least $\sqrt{d/N_1}$, whereas the ℓ_1 penalty on the difference between the fits lowers the gold term to $\sqrt{s\log d/N_1}$, and the factor d remains only in the proxy term.

I Several proxy populations

With $M-1$ proxy populations, each with its own bias δ_k and sample size N_k, the two-headed estimator becomes the multi-headed estimator, with one head per population and one penalty per proxy,

$$\widehat\beta_{\mathrm{MHE}} \in \arg\min_u \min_{v_2,\dots,v_M} \Big\{\rho_1\widehat L_1(u) + \sum_{k=2}^{M}\big[\rho_k\widehat L_k(v_k) + \lambda_k\|u-v_k\|^2\big]\Big\} \tag{156}$$

over unit vectors, with weights ρ_k summing to one.

Proposition 14 (Several proxies; proved in Appendix J)**.** *Linearizing the first-order conditions of* (156) *as in Proposition 2 gives the asymptotic error $\sum_k W_k U_k$ of the gold head, a combination of the single-population asymptotic errors with matrix weights determined by H, the ρ_k and the λ_k. If the biases are independent with isotropic covariances $\tau_k^2 I$ and the gold noise covariance is proportional to H^{-1}, the squared-loss Bayes-optimal linear combination of the single-population estimates belongs to this family exactly when the weights are proportional to the effective sample sizes and $\lambda_k \propto \tau_k^{-2}$: each proxy is penalized in inverse proportion to its squared bias.*

This proposition concerns only the asymptotic error; the theorems above are not restated for M populations. The optimization decouples: for a fixed gold head each proxy head solves a single-population problem, the correction (6) with the roles of the populations exchanged, and for fixed proxy heads so does the gold head, so that adding a proxy population adds one single-population fit.

J Several proxy populations: asymptotic matrix weights and the Bayes-optimal combination

This appendix records the linear algebra behind the multi-proxy extension of the two-headed estimator. It concerns the asymptotic matrix weights; the corresponding statistical theorem for M populations is not developed in this paper.

J.1 Multi-headed estimator and its linearized optimality system

Let $k = 1$ index gold and $k = 2, \dots, M$ index proxies with tangential biases ζ_k (in the $n^{-2/5}$ scaling of Appendix D). The multi-headed estimator keeps one gold head u and $M-1$ proxy heads v_k,

$$\min_{u, v_2, \dots, v_M} \Big\{ \rho_1 \widehat{L}_{1,h_1}(u) + \sum_{k \geq 2} \rho_k \widehat{L}_{k,h_k}(v_k) + \sum_{k \geq 2} \lambda_k \|u - v_k\|_2^2 \Big\}, \qquad \rho_k \geq 0,\ \sum_k \rho_k = 1, \tag{157}$$

and outputs u. Its limiting first-order conditions in tangent coordinates, with the common curvature H and the limits U_k of the estimators that use population k alone, are

$$\rho_1 H(x - U_1) + \sum_{k \geq 2} 2\lambda_k (x - y_k) = 0, \qquad \rho_k H(y_k - U_k) + 2\lambda_k (y_k - x) = 0 \quad (k \geq 2). \tag{158}$$

Proposition 15 (Asymptotic matrix weights for M populations)**.** *The system* (158) *has the unique solution*

$$\begin{aligned} x &= \sum_{k=1}^{M} W_k U_k, \qquad W_1 = \Big(\rho_1 I + \sum_{j \geq 2} \widetilde{W}_j \Big)^{-1} \rho_1, \\ W_k &= \Big(\rho_1 I + \sum_{j \geq 2} \widetilde{W}_j \Big)^{-1} \widetilde{W}_k, \qquad \widetilde{W}_k = \rho_k \Big(I + \frac{\rho_k}{2\lambda_k} H \Big)^{-1}, \end{aligned} \tag{159}$$

with $\sum_k W_k = I$; all W_k are functions of H, hence symmetric, positive definite and mutually commuting. For $M = 2$, $\rho_1 = 1 - \rho$, $\rho_2 = \rho$, $\lambda_2 = \lambda$ this is $W_2 = \rho[I + \frac{\rho(1-\rho)}{2\lambda} H]^{-1}$.

Proof. From the kth equation, $y_k = (\rho_k H + 2\lambda_k I)^{-1}(\rho_k H U_k + 2\lambda_k x)$, so $x - y_k = (\rho_k H + 2\lambda_k I)^{-1} \rho_k H (x - U_k)$. Substituting in the first equation and multiplying by H^{-1} (all matrices are functions of H) gives $\rho_1 (x - U_1) + \sum_k \widetilde{W}_k (x - U_k) = 0$ with $\widetilde{W}_k = 2\lambda_k \rho_k (\rho_k H + 2\lambda_k I)^{-1} = \rho_k (I + \frac{\rho_k}{2\lambda_k} H)^{-1}$. The matrix $\rho_1 I + \sum_j \widetilde{W}_j$ is positive definite. □

J.2 Precision weighting and the matching conditions

Proposition 16 (Squared-loss Bayes-optimal combination)**.** *Let $U_k = \zeta_k + \widetilde{N}_k$ with $\zeta_1 = 0$, the ζ_k ($k \geq 2$) and $\widetilde{N}_k$ mutually independent and centered,* $\operatorname{Cov} \zeta_k = \tau_k^2 I$, $\operatorname{Cov} \widetilde{N}_k = \Sigma_k \succ 0$. *The unique minimizer of $\mathbb{E} \| \sum_k W_k U_k \|^2$ subject to $\sum_k W_k = I$ is the precision weighting*

$$W_k^{\mathrm{B}} = \Big(\sum_j P_j \Big)^{-1} P_k, \qquad P_k = (\Sigma_k + \tau_k^2 I)^{-1}, \quad \tau_1 = 0. \tag{160}$$

Proof. The objective is $\sum_k \operatorname{tr}(W_k P_k^{-1} W_k^\top)$. With a matrix multiplier Λ for the constraint, stationarity gives $2 W_k P_k^{-1} = \Lambda$, so $W_k = \frac{1}{2} \Lambda P_k$, and the constraint fixes Λ. The objective is strictly convex, so the stationary point is the unique minimizer. □

Theorem 34 (Matching of the multi-headed family and the Bayes weights)**.** *Let $\Sigma_k = V/(\nu_k c_k)$ and $V = \kappa H^{-1}$. The precision weights of Proposition 16 belong to the family of Proposition 15 if and only if*

$$\frac{\rho_k}{\rho_1} = \frac{\nu_k c_k}{\nu_1 c_1}, \qquad \lambda_k = \frac{\rho_k \, \kappa}{2 \, \nu_k c_k \, \tau_k^2}, \qquad k = 2, \dots, M. \tag{161}$$

Weights are proportional to effective sample sizes and each proxy's penalty is inversely proportional to its bias variance. For two populations with equal bandwidths c and $\tau^2 = r^2/p$ this gives $\rho = \nu_2$ and $\lambda_{\mathrm{B}} = p\kappa/(2cr^2)$.

Proof. In both families each W_k is a common matrix factor on the left times a relative weight, so only the relative weights $W_k W_1^{-1}$ must agree. For the Bayes weights, $W_k^{\mathrm{B}} (W_1^{\mathrm{B}})^{-1} = P_k P_1^{-1} = \frac{\nu_k c_k}{\nu_1 c_1} (I + \frac{\nu_k c_k \tau_k^2}{\kappa} H)^{-1}$ using $V^{-1} = H/\kappa$; for the multi-headed family, $W_k W_1^{-1} = \frac{\rho_k}{\rho_1} (I + \frac{\rho_k}{2\lambda_k} H)^{-1}$. Two such analytic functions of H coincide if and only if their scalar coefficients coincide. □

Remark 10. *The theorem concerns the squared loss in the Gaussian limit. For the angular loss it is only a heuristic; for two populations, Appendix E shows that the optimal penalty under the angular loss has the same r^{-2} scaling with an explicit correction factor. The statistical theorem for M populations (localization, uniform integrability and tuning) is not developed in this paper.*

K EXPERIMENTAL DETAILS

This appendix describes the designs, estimators and tuning of Section 6 and reports all tables. Errors are angular errors in degrees unless stated otherwise; paired gains are relative differences of mean errors over common replications, with 95% intervals.

K.1 DESIGNS, ESTIMATORS AND TUNING

Synthetic designs. The design density is a product of triweight marginals $\frac{35}{32a_j}(1-(x_j/a_j)^2)_+^3$ with half-widths a_j, times a tilt factor $1+\vartheta(x^\top\beta^\star)(x^\top e)$, with e the least curved tangent axis, which makes the smoothing offset nonzero; the density is twice continuously differentiable with compact support, as Assumption 4 requires. The target $\beta^\star$ is the first coordinate axis, with half-width 0.5. In the smoothed designs, $d=8$, five tangent axes have half-width 4 and two have half-width 0.25, with a tilt. In the design of Figure 2(b), $d=3$, the two tangent half-widths are 4 and 0.5, without tilt. Isotropic controls replace the tangent half-widths by their geometric mean. A comparison from population k is answered with utility vector $4\beta_k+\varepsilon$, $\varepsilon\sim\mathcal{N}(0,0.5^2I_d)$, through the logistic link, where β_k is the unit center; in Figure 2(a) the utility vector is $8\beta_k+\varepsilon$ with $\varepsilon\sim\mathcal{N}(0,0.25^2I_d)$, $N_1=8000$ and $N_2=10N_1$, and in the design of Table 4 $N_1=4000$ and $N_2=3N_1$.

Designs with real features. Answer pairs are taken from the helpful-base part of HH-RLHF (Bai et al., 2022a). Each answer, preceded by the last 1,500 characters of its prompt, is embedded by the sentence encoder all-MiniLM-L6-v2 (Reimers & Gurevych, 2019; Wang et al., 2020) with normalized output, the feature of a pair is the difference of the two embeddings, and the features are reduced to d principal components and rescaled to unit average squared coordinate; comparisons draw features with replacement from this pool. The target direction is a fixed random unit vector, and the preferences are simulated from the model above. The bias direction is uniform on the tangent sphere, or the tangent direction most aligned with the difference in answer length, or uniform within the span of the top four principal components, projected on the tangent space.

Bias. The proxy center lies at angle $r\alpha_0$ from $\beta^\star$ along a tangent direction drawn afresh in every replication, where α_0 is the gold-only error calibrated in pilot replications, 24 for Figure 2(a) and 40 for the other two designs: $\alpha_0=1.63^\circ$ in Figure 2(a), realized 1.56°, $\alpha_0=4.57^\circ$ in the design of Table 4, where the realized gold-only error is 4.77°, and $\alpha_0=6.01^\circ$ in Figure 2(b), realized 6.39°.

Estimators. The smoothed losses replace the sign by $\tanh(u/h)$ with one bandwidth $h=cN_1^{-1/5}$ for both samples; $c=1$ in Figure 2(a), and $c\in\{1,2,4\}$ is chosen by five-fold cross-validation of the gold-only fit for the real features. THE is fitted from two starting points, the weighted fit and the pair of separate fits, and the smaller objective is kept, because a start far from the gold optimum can stall on a plateau of the bounded loss. For the smoothed losses, the oracle grids are $\rho\in\{0,0.1,\dots,0.9,0.95,0.97,0.99,1\}$ with eleven values of λ from 10^{-4} to 10 on a logarithmic scale in the design of Table 4, and $\rho\in\{0,0.3,0.6,0.8,0.9,0.95,0.97,0.99,1\}$, with nine further weights from 0.001 to 0.4 for WE, and eleven values of λ from 10^{-6} to 10 in Figure 2(a). The implementable rule searches 41 equally spaced weights in $[0,1]$ and 25 logarithmically spaced values of λ from 10^{-4} (10^{-3} with real features) to 10. With the exact sign loss ($d=3$) every estimator minimizes its objective globally over the points of a grid of the chart $\beta(v)=(\beta^\star+Ev)/\sqrt{1+\|v\|^2}$, on which the sign of each observation is a half-plane. The grid is centered at the truth, its spacing along each axis is one fifth of the smaller root-mean-square error of the gold-only and proxy-only fits along that axis, and minimizers reach its edge only in 48 of the 2.5×10^5 anisotropic fits at $r=4$; tied grid points are replaced by their centroid unless they spread over more than ten steps of the coarser axis, in which case the first tied point is used. The hyperparameter grids are 24 weights in $[0,1]$ for WE; 13 weights from 0.1 to 0.99 and 16 values of λ/ρ from 10^{-2} to 10^3 for THE; and 19 penalties from 10^{-3} to 10^3 for the two-step correction. The inner minimization of THE over the

proxy head is the Moreau envelope (Rockafellar & Wets, 1998) of the proxy loss, computed exactly by two separable one-dimensional passes. It uses the chart penalty $\lambda\|v_1 - v_2\|^2$, which has the Hessian of $\lambda\|\beta_1 - \beta_2\|^2$ at the truth but differs from it away from the truth. Loss grids and envelopes agree with brute-force evaluation.

Tuning. The cross-fitted oracle selects the hyperparameters minimizing the mean error over the even replications and evaluates them on the odd ones, and conversely, so that selection noise does not lower the reported error of a larger grid. The implementable rule minimizes, over the matrix weights W of each family (Proposition 2), $\widehat{R}^2(W) + \mathrm{sd}(\widehat{R}^2(W))$, where $\widehat{R}^2(W) = \|W\widehat{\delta}\|^2 + \mathrm{tr}\, V_1 - 2\,\mathrm{tr}(WV_1)$ estimates the squared asymptotic risk from the fitted difference $\widehat{\delta}$ of the two directions and the gold sandwich covariance V_1, in the manner of Stein (1981). The proxy noise needs no estimate in $\widehat{R}^2(W)$, because $\widehat{\delta}$ carries it into the score (the standard deviation of the score uses the sum of the gold and proxy sandwich covariances on the tangent space), but the smoothing offset shared by the two fits cancels in $\widehat{\delta}$, so that even with V_1 known the expected score falls short of the risk by the squared offset plus twice its inner product with $W\delta$. Only the second term depends on W; in our designs, adding it at its noise-free value changes the mean error of THE by at most 1% and its gain over WE by at most 0.6 points. The rule also targets the squared risk rather than the reported mean angle and uses the estimated V_1; the standard-deviation term protects against selecting on noise.

K.2 Effect of smoothing on the population target

In the design of Table 4 with $c = 1$, the population target of the smoothed proxy loss, computed without label noise from expected labels on 10^6 covariates, moves by 1.82 times the nominal shift along the most curved axis, 0.39 times along the flattest, and 1.54 times on average over 40 random directions; the gold offset is $0.27°$. With half the bandwidth the ratios are 1.20, 0.60 and 1.07. At a finite bandwidth the smoothed loss favors directions with large projected variance, which moves the population minimizer by a term of order h^2 times the difference in variance; the analysis takes $h \to 0$, where this term is of higher order. In the designs with real features the ratio is 1.04 to 1.09, and 1.18 along answer length. In that design the effective shift therefore concentrates on curved directions, where the penalty acts; Table 4 reports predictions with the measured and with a uniform effective shift. Figure 2(a) removes this distortion: in every replication the proxy center is placed so that the effective shift of the smoothed population fits, computed without label noise, equals the target to within 2%. A single linear correction measured at one shift size leaves the effective shift at 0.88 to 1.02 times the target for $r \le 4$ and 0.98 to 1.16 at $r = 8$, because the distortion depends on the size of the shift; with it the gain of THE over WE is 0.8 to 4.1 points lower. Without compensation the ratio of effective to target shift is 1.49 to 1.73 along curved directions and 0.42 along flat ones, with direction errors of up to $34°$, and the gain of THE over WE rises to 27–38%. The gold offset of this design is $0.25°$. The predictions in Figure 2(a) and Table 3 are evaluated at the effective shifts of the simulation, with the curvature, noise covariance and smoothing offset computed without label noise from expected labels on 10^6 covariates.

K.3 Tables and figures

Table 2 summarizes the settings of Section 6; Figures 2, 3 and 4 show the error against the bias, the scan over the proxy share, the designs with real features and the sparse designs. Table 3 reports the design of Figure 2(a). The 95% bootstrap intervals of the improvement ratio are $[1.42, 1.61]$, $[1.86, 2.39]$ and $[2.88, 8.99]$ at $r = 1$, 2 and 4; at $r = 8$ re-weighting cannot be told from gold-only, and the ratio is undefined.

Table 4 reports an uncompensated design with $N_2 = 3N_1$. In the isotropic control (200 replications) all paired differences between THE and WE are within 1%. With the implementable rule, THE improves on WE by $1.1 \pm 0.6\%$, $5.8 \pm 1.0\%$ and $7.6 \pm 1.0\%$ at $r = 0.5$, 1 and 2, and on the two-step correction by 8–11% for $r \le 0.25$, while the two-step correction is better by $0.9 \pm 0.6\%$ and $2.5 \pm 0.5\%$ at $r = 1$ and 2; at $r = 0$ the rule gives $3.43°$ for WE and $3.45°$ for THE against $2.29°$ for pooling, because the rule estimates the bias only from one noisy fitted difference and its standard-deviation term shrinks the weight, to 0.40 on average against the optimal share 0.75.

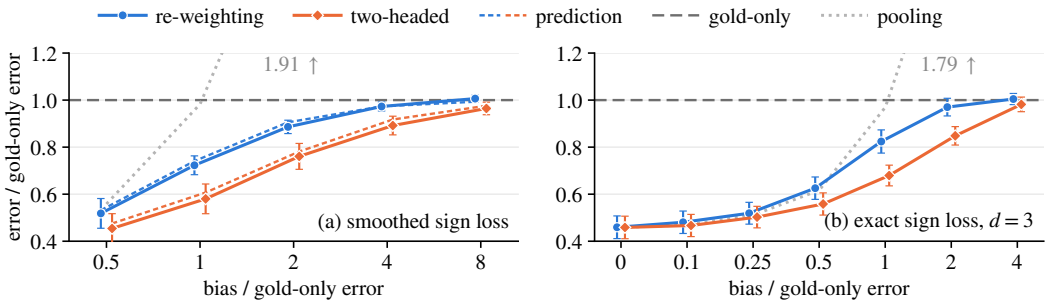


Figure 2: Oracle error relative to gold-only against the bias r, in units of the calibrated gold-only error; bars: 95% paired intervals. (a) Smoothed loss, $d = 8$, $N_2 = 10N_1$, effective bias made uniform (Appendix K.2), 198 replications; dashed: asymptotic prediction at the simulated effective bias. (b) Exact sign, $d = 3$, $N_2 = 10N_1$, 1000 replications.

Table 2: Gain of THE (% lower mean angular error, paired) over gold-only and over the best re-weighting at $r = 0.5, 1, 2$. Cross-fitted oracle, except the rows with real features (implementable rule); the last row uses THE's ℓ_1 special case (Section 5), fits over the parameter ball $\|\beta\| \le 3$ and the Euclidean error of the gold center.

setting	loss	d	N_2/N_1	vs gold-only	vs best re-weighting
synthetic, anisotropic	exact sign	3	10	44 / 32 / 15	11 / 18 / 13
synthetic, isotropic	exact sign	3	10	–	−0.5 / 1.1 / 1.9
synthetic, anisotropic	smoothed	8	10	55 / 42 / 24	12 / 20 / 14
real features, uniform bias	smoothed	128	10	36 / 20 / 6	1 / 2 / 1
real features, answer-length bias	smoothed	128	10	39 / 25 / 10	7 / 9 / 5
real features, bias in top 4 components	smoothed	128	30	45 / 32 / 15	18 / 20 / 12
sparse shift ($s = 2$), ℓ_1 correction	likelihood	20–1280	32–128	32–64 over the best unpenalized baseline	

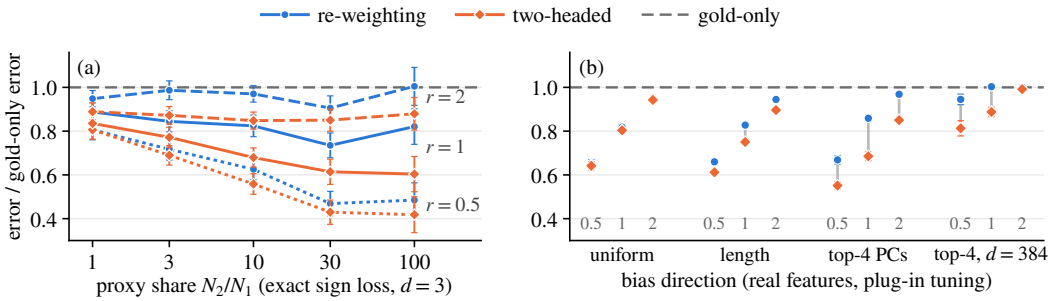


Figure 3: Error relative to gold-only. (a) Exact sign loss, design of Figure 2(b), against the proxy share at $r = 0.5, 1, 2$ (dotted, solid, dashed). (b) Real features, implementable rule, by bias direction ($d = 128$; last group $d = 384$, gold-only error 47°). Bars: 95% paired intervals.

Table 5 reports the design of Figure 2(b). The error of the best weight stays above the lower bound (7) at every bias, with K estimated as $N_1^{1/3}$ times the gold-only error; for instance 5.26° against 2.44° at $r = 1$. Figure 3(a) varies N_2/N_1 over 1, 3, 10, 30 and 100 in this design, with 1000 replications for $N_2/N_1 \le 10$, 600 for 30 and 300 for 100; at $N_2/N_1 = 100$ the gain of THE over WE is $26.4 \pm 7.9\%$ at $r = 1$ and $12.5 \pm 7.6\%$ at $r = 2$. At $r = 2$, WE's reduction of the gold-only error is at most 9% and vanishes at $N_2/N_1 = 100$, while THE's stays at 11–15%. With $N_1 = 32{,}000$ (400 replications), where the bias unit is 0.78 times the realized gold-only error instead of 0.94, the ratios of the errors of WE and THE to gold-only at $r = 0.5, 1, 2, 4$ are 0.56, 0.76, 0.95, 1.02 and 0.54, 0.68, 0.86, 0.98, and the gain of THE over WE is $9.6 \pm 6.0\%$ and $9.6 \pm 6.6\%$ at $r = 1$ and 2. In a second design with tangent half-widths 4 and 1 and $N_2 = 3N_1$, THE gains $7.0 \pm 3.4\%$ and $7.4 \pm 3.4\%$ over the two-step correction at $r = 0$ and 0.1, and $10.4 \pm 3.4\%$ over WE at $r = 1$. In

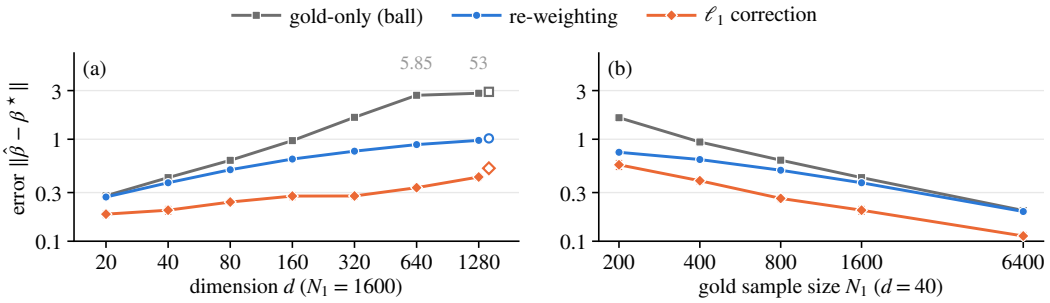


Figure 4: Sparse shift with a known Gaussian heterogeneity law ($s = 2$, $\|\delta\| = 1$): Euclidean error of gold-only maximum likelihood and re-weighting at its oracle weight, both over the parameter ball $\|\beta\| \le 3$, and of the ℓ_1 correction with a cross-validated penalty. (a) Against d at $N_1 = 1600$ ($N_2 = 51{,}200$ for $d \le 160$, $102{,}400$ above; open markers: $d = 1280$, $N_1 = 800$); numbers above the panel: gold-only without the constraint, where it binds. (b) Against N_1 at $d = 40$, $N_2 = 51{,}200$; slopes -0.50 (ℓ_1) and -0.33 (re-weighting) for $N_1 \le 1600$. The comparison of Theorem 7(c) covers $d \le 80$ in (a) and $N_1 \ge 800$ in (b). Bars: 95% intervals over 20–40 replications.

Table 3: Smoothed sign loss, design of Figure 2(a): $d = 8$, $N_1 = 8000$, $N_2 = 10N_1$, effective bias made uniform, 198 replications; cross-fitted oracle. Error ratios are relative to gold-only, with asymptotic predictions at the simulated effective bias in parentheses; the improvement ratio is (gold − THE)/(gold − WE).

r	0.5	1	2	4	8
THE / gold-only	0.454 (0.475)	0.580 (0.607)	0.761 (0.781)	0.892 (0.917)	0.964 (0.976)
WE / gold-only	0.518 (0.532)	0.723 (0.741)	0.886 (0.906)	0.973 (0.973)	1.006 (0.993)
THE vs gold-only (%)	54.6 ± 6.3	42.0 ± 6.3	23.9 ± 5.5	10.8 ± 4.0	3.6 ± 2.7
WE vs gold-only (%)	48.2 ± 6.3	27.7 ± 4.0	11.4 ± 2.8	2.7 ± 1.9	-0.6 ± 1.4
THE vs WE (%)	12.4 ± 1.6	19.7 ± 4.3	14.1 ± 3.8	8.3 ± 2.6	4.2 ± 1.7
improvement ratio (predicted)	1.13 (1.12)	1.52 (1.52)	2.10 (2.32)	3.99 (3.09)	–
THE vs two-step (%)	4.4 ± 3.2	3.1 ± 1.6	4.5 ± 2.7	0.4 ± 1.4	-0.8 ± 0.7

Table 4: Smoothed sign loss, uncompensated design: $d = 8$, $N_1 = 4000$, $N_2 = 3N_1$, 400 replications; cross-fitted oracle errors. Predictions are asymptotic gains of THE over WE with the measured and with a uniform effective shift.

r	0	0.1	0.25	0.5	1	2
gold-only	4.77	4.77	4.77	4.77	4.77	4.77
pooling	2.29	2.36	2.67	3.47	5.69	11.07
WE	2.30	2.38	2.65	3.24	4.03	4.53
THE	2.30	2.36	2.58	2.85	3.14	3.49
THE vs gold-only (%)	51.8 ± 4.0	50.4 ± 4.0	45.8 ± 3.9	40.3 ± 3.8	34.1 ± 3.8	26.8 ± 3.5
WE vs gold-only (%)	51.7 ± 4.0	50.1 ± 3.8	44.5 ± 3.7	32.1 ± 3.2	15.5 ± 2.3	5.0 ± 1.2
THE vs WE (%)	0.1 ± 0.1	0.6 ± 0.8	2.5 ± 1.1	12.0 ± 1.9	22.0 ± 2.7	22.9 ± 3.2
predicted, measured shift	0.0	0.1	3.0	12.9	24.8	27.9
predicted, uniform shift	0.0	0.0	0.4	2.9	8.6	9.0
THE vs two-step (%)	9.2 ± 2.4	8.7 ± 2.3	7.2 ± 2.0	4.3 ± 1.0	1.6 ± 1.1	0.0 ± 0.1

the isotropic control of Table 5 THE gains 4–6% over the two-step correction for $r \le 1$, because the two-step correction penalizes the distance to the noisy proxy minimizer instead of using the proxy loss itself. In an isotropic design, 200 replications per sample size, the gold-only error is $5.73°$, $4.66°$, $3.51°$, $2.80°$, $2.27°$, $1.79°$ and $1.32°$ for N from 1000 to 64,000 in factors of two; the slope in $\log N$ is -0.35, and $N^{1/3}$ times the error, the constant K of Theorem 4, is about 0.98 radians.

Table 6 repeats the first 200 replications with the spherical penalty, whose envelope is computed exactly by branch and bound, with half the grid spacing, with 1.5 times the half-widths, and with the grid centered at the gold-only logistic fit together with the spherical penalty, retuning every family.

Table 5: Exact sign loss, design of Figure 2(b): $d = 3$, $N_1 = 4000$, $N_2 = 10N_1$, 1000 replications and 1000 isotropic ones; global minimizers over a grid, cross-fitted oracle.

r	0	0.1	0.25	0.5	1	2	4
gold-only	6.39	6.39	6.39	6.39	6.39	6.39	6.39
pooling	2.93	3.00	3.27	3.94	6.19	11.45	22.30
WE	2.93	3.07	3.32	4.00	5.26	6.20	6.42
THE	2.93	2.98	3.21	3.57	4.34	5.42	6.27
THE vs gold-only (%)	54.1 ± 4.8	53.3 ± 4.7	49.7 ± 4.6	44.2 ± 4.7	32.1 ± 4.4	15.2 ± 3.9	1.8 ± 3.1
WE vs gold-only (%)	54.1 ± 4.8	51.9 ± 4.8	48.1 ± 4.6	37.4 ± 4.8	17.6 ± 4.9	3.0 ± 3.8	-0.5 ± 2.4
THE vs WE (%)	0.2 ± 1.6	2.9 ± 1.8	3.2 ± 3.2	10.7 ± 3.3	17.6 ± 4.1	12.6 ± 3.9	2.3 ± 2.6
isotropic: THE vs WE (%)	0.2	-0.2	1.6	-0.5	1.1	1.9	0.5
THE vs two-step (%)	2.3 ± 2.7	2.5 ± 3.0	3.7 ± 3.6	1.2 ± 3.3	8.3 ± 3.6	2.5 ± 2.7	0.3 ± 2.5

Table 6: Checks of the design of Table 5 on its first 200 replications: gain of THE over WE (%, ± 1.96 paired standard errors), every family retuned by the cross-fitted oracle.

	anisotropic, $r =$				isotropic, $r =$			
	0.5	1	2	4	0.5	1	2	4
grid and penalty of Table 5	6.1 ± 8.0	17.3 ± 7.6	8.4 ± 8.0	2.5 ± 5.1	-0.3 ± 2.9	-0.1 ± 2.2	1.3 ± 4.4	1.8 ± 3.3
spherical penalty	5.9 ± 8.0	16.6 ± 7.6	9.1 ± 9.4	1.0 ± 7.9	-0.4 ± 2.9	-0.1 ± 2.2	1.6 ± 4.3	0.3 ± 1.5
half the spacing	9.6 ± 6.2	17.0 ± 8.8	11.1 ± 9.7	5.3 ± 7.1	-7.4 ± 4.9	-3.2 ± 3.4	2.8 ± 4.1	-1.2 ± 4.0
1.5 times the half-widths	10.9 ± 7.6	11.1 ± 8.6	11.1 ± 8.0	9.7 ± 7.4	2.0 ± 2.5	-0.8 ± 3.8	-1.4 ± 3.6	-0.5 ± 2.3
centered at the gold-only fit, spherical penalty	11.5 ± 7.4	10.3 ± 8.8	11.6 ± 8.4	4.4 ± 6.8	-6.7 ± 4.8	-1.1 ± 3.4	-0.4 ± 2.2	0.1 ± 2.4

Single minimizers move with the grid because the empirical sign loss is minimized on cells narrower than the grid spacing and nearly tied cells can lie degrees apart: a grid with a quarter of the spacing finds a lower gold-only loss in 85–89% of replications, by at most 10 of 4000 comparisons, and halving the spacing moves the gold-only estimate by a median of 0.1–0.4°. The gains of THE over WE under anisotropic curvature change within Monte Carlo error, and at the selected hyperparameters the chart penalty exceeds the spherical one by a median of 0.5–2.7%. Under isotropic curvature THE trails WE by 7% at $r = 0.5$ on two of the variants. This is the cost of selecting from THE's larger grid with 100 replications per half: at the best hyperparameters over all 200 replications, THE and WE differ by at most 3.1% in every isotropic cell.

Table 7: Real features, $d = 128$, implementable tuning. Uniform and answer-length bias: $N_1 = 32{,}000$, $N_2 = 10N_1$, 60 replications, gold-only error 13.2°. Top four principal components: $N_1 = 16{,}000$, $N_2 = 30N_1$, 54 replications, gold-only error 17.9°. Brackets: 95% lower confidence bounds.

bias direction	r	THE vs gold-only (%)	THE vs WE (%)	predicted THE vs WE (%)
uniform	0.5 / 1 / 2	35.8 / 19.6 / 5.7	1.4 (0.8) / 1.7 (1.2) / 0.7 (0.2)	0.4 / 0.8 / 0.5
answer length	0.5 / 1 / 2	38.8 / 25.0 / 10.4	7.3 (6.5) / 9.3 (8.6) / 5.1 (4.3)	4.0 / 7.2 / 5.2
top four components	0.5 / 1 / 2	44.9 / 31.5 / 15.0	17.6 (16.5) / 20.2 (18.9) / 12.2 (10.9)	12.5 / 20.8 / 20.2

Table 7 reports the designs with real features. Along answer length, THE and WE reduce the gold-only error at $r = 1$ by 25% and 17% (Figure 3(b)). The measured curvature ratio is 15 in 128 dimensions. Across evaluations of the asymptotic constants from 7 to 383 dimensions with the same spectral shape, the gain over WE hardly changes; it is determined by the shape of the spectrum and by how much of the bias lies in the curved directions.

Table 8 reports the sparse designs. Without the constraint, gold-only maximum likelihood fails at $d = 320$ and 640, where its fit exists but lies far from $\beta^\star$, and diverges at $d = 1280$, where the gold data are linearly separable in every replication. Over the ball it stops on the boundary, with error 2.7–2.9, still larger than that of the zero vector. Re-weighting at its best weight and the ℓ_1 correction stay inside the ball, and no selected weight lies at an end of the grid.

Misspecified heterogeneity law. Theorem 7(c) assumes the heterogeneity law known. In an earlier version of the design of Figure 4, without the ball constraint, with six weights for re-weighting and nine penalty levels for the correction, $N_1 = 1600$, $N_2 = 51{,}200$, $s = 2$ and $\|\delta\| = 1$, for $d = 40$ (40 replications) and $d = 160$ (30), every estimator uses the marginal likelihood of an assumed law $\mathcal{N}(0, \tau_a^2 I/(d+2))$ while the data come from another: a Laplace, t_3, uniform or anisotropic law of

Table 8: Sparse shift with a known Gaussian heterogeneity law: $s = 2$, $\|\delta_2\| = 1$, $N_2 = 102{,}400$, 20 replications; Euclidean error of the gold center. Gold-only and re-weighting are fitted over the ball $\|\beta\| \le 3$ of Theorem 7(c), and gold-only also without it. The baselines are tuned by an oracle, re-weighting over $\rho \in \{0.01, 0.02, 0.05, 0.1, 0.15, 0.2, 0.3, \ldots, 0.9, 0.95, 0.99, 0.995\}$, and the ℓ_1 correction by cross-validation.

d / N_1	gold-only	gold-only, ball	best weight	ℓ_1 correction	gain over best baseline
320 / 1600	1.64	1.64	0.76	0.28	64%
640 / 1600	5.85	2.70	0.89	0.33	62%
1280 / 1600	53	2.82	0.97	0.43	56%
1280 / 800	19.9	2.92	1.02	0.52	49%

the same scale ($\tau_a = 1$), or a Gaussian law with $\tau = 1$ and $\tau_a = 0.5$, 2 or 0, the last being the logistic likelihood. The ℓ_1 correction is tuned by cross-validating the assumed likelihood and re-weighting by the oracle. The gain of the correction over re-weighting in Euclidean error is 41.8–46.7% under a misspecified shape and 29.8–38.2% under a misspecified scale at $d = 40$, and 56.0–57.0% and 44.5–54.0% at $d = 160$, against 45.4% and 55.7% under the correct law; in angular error it is 40.8–54.8% in every case.

K.4 Sparse shift with the smoothed and the exact sign loss

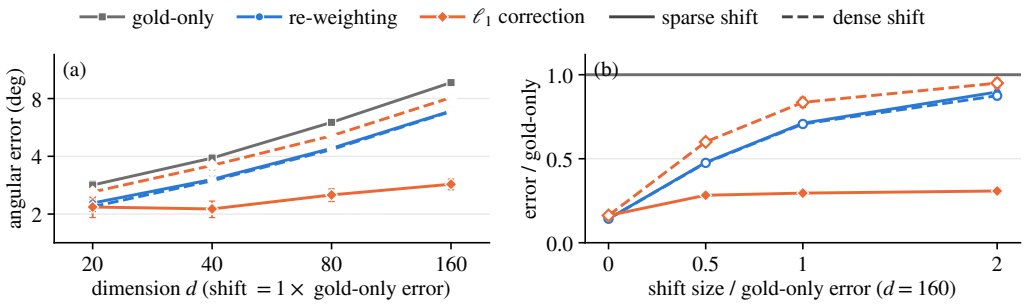


Figure 5: Sparse shift with the smoothed sign loss: $N_1 = 4000$ gold and $N_2 = 120{,}000$ proxy comparisons, 30 replications. (a) Angular error against d when the directions differ by one gold-only error; solid: difference on $s = 2$ coordinates, dashed: dense difference of the same size. (b) Error relative to gold-only at $d = 160$ against the size r of the difference. Re-weighting at its oracle weight and bandwidth; gold-only and the ℓ_1 correction tuned by 5-fold cross-validation on the gold sample. Bars: 95% intervals.

Figure 5 uses $d \in \{20, 40, 80, 160\}$, $\beta^\star$ a coordinate axis and a compactly supported smooth design density with a tilt, so that the smoothing offset is nonzero. The directions differ on $s = 2$ coordinates, or on all of them, by r times the gold-only error calibrated in pilot replications, $2.91°$, $4.10°$, $5.85°$ and $9.61°$ for the four dimensions (realized $2.83°$, $3.92°$, $6.01°$ and $9.69°$); smoothing distorts the difference by a factor of about 1.1 along the coordinate directions, so a sparse difference stays sparse. The ℓ_1 correction (5) is computed by proximal gradient steps on the sphere followed by renormalization, from two starting points, keeping the lower objective; the two fits differ in 3–29% of the cases. Its bandwidth is $h_1 = c(s \log d/N_1)^{1/5}$, and c and λ are chosen by 5-fold cross-validation on the gold sample; gold-only and pooling are cross-validated on the same folds, and re-weighting is also given its oracle weight and bandwidth, the comparison most favorable to it. The bandwidth grid was extended downward because cross-validation often selected its smallest value; the oracle choices, which lie inside the grid, did not change, and the gains changed by at most 2.6 points except at $r = 0.5$ and with no difference.

With a difference on two coordinates, the ℓ_1 correction improves on re-weighting at its oracle tuning by 5, 30, 43 and 58% at $r = 1$ for $d = 20$, 40, 80 and 160, the first not significant, by 25, 41, 53 and 66% at $r = 2$, and by -12, -2, 11 and 41% at $r = 0.5$; at $r = 1$ it improves on gold-only by 23, 46,

58 and 70%. The slopes of the logarithm of the error against $\log d$ at $r = 1$ are 0.59 for gold-only, 0.53 for re-weighting and 0.14 for the ℓ_1 correction, whose slope is 0.54 under a dense difference. At $d = 80$ and $r = 1$ the gain over re-weighting is 51, 43 and 31% for $s = 1$, 2 and 4. Under a dense difference the ℓ_1 correction is 6–26% worse than re-weighting at its oracle tuning, and with no difference it is close to pooling, 14–68% worse than re-weighting at its oracle tuning and 7–21% better than cross-validated re-weighting.

The trust region and penalty level of Theorem 7(b) are not used. At these sample sizes the radius $h_1/(8R)$ is 0.02° to 0.13°, which excludes the truth at every nonzero difference, and the stated penalty level returns the proxy fit. The experiment uses a trust region of radius at least 0.47 radians, which is never active, and a cross-validated penalty, so it tests the estimator and its dependence on d, not the conditions of the theorem.

With the exact sign loss (Theorem 7(a)), $d = 3$ and a difference along one coordinate axis, 300 replications with isotropic tangent directions show no gain of the ℓ_1 correction over re-weighting (-8 to 0%), consistent with the factor $(\sqrt{sd}/d)^{1/3} \approx 0.83$ between the bounds of Theorem 7(a) and of gold-only at $s = 1$. Exact minimization is not feasible for larger d, since the number of cells of the sign loss grows as $N^{d-1}/(d-1)!$, about 8.5×10^{15} for $N = 4000$ and $d = 6$.

K.5 Real preference labels

This subsection repeats the comparisons of Section 6 with human preference labels as the gold sample and as the held-out reference. Table 9 reports the main cells, and Figure 1(b) shows five of them.

Data and features. PRISM (Kirk et al., 2024): the 39,843 opening-turn comparisons of the 1,394 participants who rated at least one pair, obtained from ratings of up to four responses of different models to an opening prompt, ties dropped. Community Alignment (Zhang et al., 2026): the English first-turn choices of 1,272 annotators in the US and in India among four pre-generated responses, in the 602 response groups seen in both countries; a choice gives three pairs. HH-RLHF (Bai et al., 2022a): the pairs of its helpful-base part, which record no annotators. The reward-model judge, `reward-model-deberta-v3-base`, released by the OpenAssistant project (Köpf et al., 2023), was trained on the WebGPT and summarize-from-feedback comparisons and on synthetic instruction pairs, not on PRISM or HH-RLHF; it prefers the response with the larger reward and agrees with 59–62% of the human labels. A pair is represented by the leading principal components of the difference of its responses' embeddings by the sentence encoder of Appendix K.1, computed without labels (8 on PRISM, or 16 and 32 for two injected judges; 16 on Community Alignment); by model identity (PRISM, $d = 20$); by the leading 16 principal components of the difference of the reward model's own representations, on which its reward is linear.

Designs. (i) Natural populations: the gold sample comes from one of 18 PRISM cohorts or two random placebo groups ($N_1 = 20$, 40 or 80 participants), with the other participants as the proxy, or from the US annotators of Community Alignment ($N_1 = 10$, 20, 40 or 80), with about 15,000 comparisons of annotators in India as the proxy. (ii) Reward-model judge on another population's proxy: the populations are as in (i), with three PRISM gold cohorts chosen by a rule fixed before the comparisons (female, non-native English speakers and religious participants), and the reward model relabels half, three quarters or all of the proxy with its own representation as features, or a quarter, half or all of it with the sentence features; the three-quarters share was added after the training-data diagnostics placed the largest predicted gain there, before any comparison run. (iii) Rule-based judges prefer the response with the larger coordinate along the most curved, or the flattest, principal direction. They relabel 10% or 20% (PRISM) and 20% or 35% (Community Alignment) of another population's proxy as in (ii), or, as injected judges, 10% to 50% of a proxy from the gold population itself: the participants or annotators that are not held out are split at random into a gold side ($N_1 = 40$, 80 or 160 on PRISM, 10, 20 or 40 on Community Alignment) and a proxy side of about 15,000 comparisons, and an injected rule may also follow a random direction or prefer the longer answer. (iv) Reward model only: 80 random PRISM participants form the gold sample, and the reward model labels all of 3,000 to about 24,000 proxy comparisons. (v) HH-RLHF, same pool: the training pairs are split by conversation into a gold and a proxy side, the gold sample is $N_1 = 1000$ pairs, and the reward model relabels a fixed random half of 15,000 proxy pairs, with

its representation as features. The other half keeps its human labels, which come from the gold population, so the proxy contains a second gold sample, and the design is evidence only for the comparison of THE with WE, which see the same data; without annotator records, a version with another population is not possible. Cells have 30 to 50 replications. Except for the exploratory ℓ_1 reruns of the reward-model designs, every design reported here, its hypotheses and its predictions were written into a plan committed to version control before its first comparison run; several were chosen after earlier results (the three-quarters share, the HH-RLHF design, and the ℓ_1 and smoothed-sign designs below). Before the comparisons, the training data of each cell gave the shift between the gold and the proxy fits, in units of the gold-only error, the curvature of the gold loss along this shift relative to its mean over tangent directions, and an asymptotic prediction of THE's gain over WE.

Estimators and tuning. Every estimator uses the logistic likelihood of the Bradley–Terry model (Bradley & Terry, 1952) with a ridge penalty in place of the sign loss. Gold-only, pooling, WE, THE, which reports its gold head, and the two-step correction share one grid of ridge penalties, weights and penalty levels, which for THE reaches the limits where it coincides with WE and with the two-step correction; where a family's choices sat at another edge, the grid was extended and the affected cells were rerun with the same seeds. The oracle rule is cross-fitted over seeds: it chooses the hyperparameters on half of the seeds by the angle to a tuning reference, fitted to held-out labels that are not used for evaluation, evaluates them on the other half, and averages over 200 random halvings. The implementable rule uses the training data only: it is the rule of Appendix K.1, with the covariance of the gold fit clustered by participant or annotator where these are recorded, and every method, gold-only included, uses the ridge penalty chosen by cross-validating gold-only. Gold-only is then tuned less well, so improvements over gold-only are reported under the oracle rule; THE's gain over WE, a comparison under one rule, is reported under both.

Evaluation and intervals. The error of an estimate is its angle to the evaluation reference, the maximum-likelihood direction of held-out human labels that never enter training or tuning: those of held-out participants, annotators or expert votes, and for HH-RLHF the 9,912 validation and test pairs. We report the improvement over gold-only, $1 - E/E_{\mathrm{gold}}$ for the mean angle E, and THE's gain over WE, $1 - E_{\mathrm{THE}}/E_{\mathrm{WE}}$, both averaged over the cells of a row, and the ratio of THE's improvement to WE's where the lower end of WE's interval exceeds 2 percentage points (for HH-RLHF, 0). The error of the reference adds to that of every method and compresses relative gains. Its noise floor, the angle between references fitted to two random halves of the evaluation units, is 15° to 35° on PRISM and Community Alignment, and the full reference errs by about half of that. As one reference serves all seeds, intervals over seeds alone are too narrow, and we use a joint bootstrap: each draw resamples the evaluation units, and the units behind the tuning references, within each fold or test split, refits the references, resamples the seeds and repeats the cross-fitting of the oracle rule (400 draws; 200 for HH-RLHF). A resampled reference is noisier than the original one, which shifts the draws toward no effect, so the 95% intervals are centered at the point estimate: their ends are the point estimate plus the differences between the 2.5% and 97.5% quantiles of the draws and their median. The joint intervals are wider than those over seeds, for instance $[4.5, 12.1]$ against $[7.0, 9.3]$ for THE's gain over WE with the reward-model judge on Community Alignment, but they change no conclusion on PRISM and Community Alignment.

Proxy from another population. Without a judge, a proxy from another population, in design (i) or in the unrelabelled controls of design (ii), reduces the gold-only error by 27–54% under the oracle rule, equally for WE and THE (Table 9; 39.1% $[37.3, 40.7]$ between PRISM cohorts with model-identity features), and THE's gain over WE lies between -0.1% and 0.0%. The training-data diagnostics predicted this tie, as the theory does: the estimated shift is below the gold-only error in every cell and is not concentrated in curved directions (curvature 0.86 to 0.94 times the mean in the sentence features, 0.03 to 0.09 in the reward model's representation), so the penalty has no bias to act on. Under the implementable rule the two methods remain equal, and between natural populations both are 9–22% worse than pooling, the price of the rule's protection against a large bias.

Reward-model judges. When the reward model relabels half of another population's proxy, its bias lies in curved directions of its own representation (curvature 2.27 times the mean on Community Alignment and 1.09 to 1.59 on PRISM), and THE reduces the gold-only error 1.42 $[1.24, 2.13]$ and

Table 9: Real preference labels. Improvement over gold-only ($1 - E/E_{\text{gold}}$, %) of pooling, WE and THE, and the ratio of THE's improvement to WE's, under the oracle rule; THE's gain over WE ($1 - E_{\text{THE}}/E_{\text{WE}}$, %) under the oracle and the implementable rule. Second lines: 95% intervals of the joint bootstrap over the held-out reference and the seeds, centered at the point estimate (*percentile intervals); pooling without an interval was not part of the joint bootstrap. Ratios are shown where the lower end of WE's interval exceeds 2 points (for HH-RLHF, 0). Curvature: of the gold loss along the shift estimated from training data, relative to its mean over tangent directions. Rows with several cells average over gold sizes and cohorts; injected judges: cells whose estimated shift is 0.5 to 2 gold-only errors. †Added after the training-data diagnostics, before any comparison run. ‡The unrelabelled proxy comparisons come from the gold population, which inflates the improvements over gold-only.

		oracle rule				THE vs WE (%)	
design, proxy labels	curvature	pooling	WE	THE	ratio	oracle	implementable
No judge: proxy from another population							
PRISM, sentence features		35.8	36.1	36.1	1.00	−0.0	−0.4
			[33.3, 38.8]	[33.3, 38.8]	[1.00, 1.00]	[−0.1, 0.1]	[−0.5, −0.2]
Community Alignment, US gold, India proxy	0.94	27.0	27.2	27.1	1.00	−0.1	−0.1
		[17.0, 38.6]	[18.6, 37.5]	[18.4, 37.3]	[0.98, 1.00]	[−0.5, 0.1]	[−0.5, 0.2]
Reward model relabels part of another population's proxy; its own representation							
Community Alignment, half	2.27	−0.1	16.0	22.6	1.42	8.1	4.6
		[−20.5, 18.3]	[5.2, 25.1]	[10.4, 34.3]	[1.24, 2.13]	[4.5, 12.1]	[2.4, 6.8]
PRISM, half	1.09–1.59	17.0	21.7	24.4	1.13	3.5	3.1
		[10.0, 23.4]	[17.4, 26.1]	[19.9, 29.0]	[1.09, 1.18]	[2.6, 4.4]	[2.4, 3.9]
Community Alignment, three quarters†	2.97	−65.6	−1.5	4.8	–	6.2	2.5
		[−86.8, −42.2]	[−12.3, 1.3]	[−6.8, 14.3]		[1.4, 13.1]	[−1.0, 5.5]
PRISM, three quarters†	2.92–2.97	−18.7	4.9	10.2	2.07	5.6	0.0
		[−27.0, −12.6]	[2.5, 7.5]	[7.2, 13.2]	[1.60, 3.25]	[4.0, 7.3]	[−1.0, 1.1]
Community Alignment, all	11.45	−146.9	−0.8	−0.3	–	0.5	20.8
		[−172.5, −119.2]	[−7.0, −0.2]	[−10.3, 1.7]		[−4.5, 2.6]	[17.6, 24.7]
PRISM, all	10.69–11.02	−77.6	0.2	0.0	–	−0.1	4.3
		[−88.0, −68.6]	[−0.8, 1.0]	[−1.2, 0.9]		[−0.7, 0.2]	[3.4, 5.2]
Reward model labels the whole proxy; sentence features							
PRISM, 80 gold participants	1.15	−14.6	10.0	12.4	1.24	2.7	1.9
		[−30.4, −2.7]	[4.0, 15.5]	[6.1, 18.5]	[1.07, 2.15]	[1.3, 4.2]	[0.6, 3.2]
Rule along the most curved direction relabels part of another population's proxy; sentence features							
PRISM, 20%	1.53–1.60	−12.4	14.4	21.1	1.46	7.8	3.7
			[10.9, 19.5]	[16.9, 26.1]	[1.31, 1.70]	[6.4, 9.4]	[2.8, 4.8]
Community Alignment, 35%	1.65	2.5	10.8	13.1	1.21	2.6	1.9
			[5.4, 16.5]	[7.0, 19.4]	[1.05, 1.46]	[0.7, 4.6]	[0.9, 2.9]
Injected rule along the most curved direction‡; sentence features							
PRISM, 8 components	1.68	−20.6	15.3	19.4	1.27	4.9	3.4
			[10.3, 20.7]	[13.7, 25.3]	[1.16, 1.45]	[3.5, 6.5]	[2.2, 4.4]
PRISM, 16 components	2.18	−39.6	8.3	12.3	1.48	4.4	5.9
			[5.2, 11.0]	[7.8, 16.0]	[1.31, 1.94]	[3.0, 5.8]	[4.6, 7.3]
PRISM, 32 components	2.78	−15.5	5.9	10.8	1.83	5.3	5.7
			[3.5, 8.2]	[7.8, 14.0]	[1.56, 2.68]	[3.9, 6.6]	[4.8, 6.9]
Community Alignment, 16 components	1.73	2.1	10.8	15.5	1.44	5.3	2.1
			[8.0, 13.6]	[12.6, 18.9]	[1.32, 1.63]	[4.4, 6.5]	[1.6, 2.6]
Reward model relabels half of a proxy from the gold population‡; its own representation							
HH-RLHF, $N_1 = 1000$		7.9	17.0	26.8	1.58	11.8	8.1
		[−6.6, 21.8]	[3.3, 28.8]	[12.3, 39.4]	[1.24, 3.26]*	[6.2, 14.6]*	[4.3, 10.5]*

1.13 $[1.09, 1.18]$ times as much as WE; its error is 8.1% $[4.5, 12.1]$ and 3.5% $[2.6, 4.4]$ below that of WE, against 7.9–8.7% and 0.3–5.1% predicted. With three quarters relabeled, the improvement of WE falls to 4.9% on PRISM and to -1.5% on Community Alignment, while THE keeps 10.2% and 4.8% (ratio 2.07 on PRISM). In the sentence features the bias of the reward model is flatter (curvature 0.94 to 1.27 on PRISM, 0.87 to 0.91 on Community Alignment), and, as predicted, the ratio is 1.00 to 1.08 on PRISM and THE's gain over WE is -0.1% to 0.4% on Community Alignment. With the reward model alone, design (iv), pooling is 14.6% worse than gold-only while WE and THE improve on it by 10.0% and 12.4% (ratio 1.24); from 3,000 to about 24,000 relabeled comparisons, the improvement of THE grows from 9.4% to 12.4% and that of WE from 7.5% to 10.0%, while that of pooling falls from 0.9% to -14.6%. On HH-RLHF, design (v), the error of THE is 11.8% below that of WE (percentile joint interval $[6.2, 14.6]$; 9.6% predicted), and the improvements over gold-only are inflated by the gold-population labels in the proxy.

A sparse shift in real data. A registered design tests the ℓ_1 correction (5) where the shift is sparse. The features of HH-RLHF pairs are the difference in the logarithm of one plus the word count and $d - 1$ principal components of the sentence features with their linear length component removed, all standardized, for $d = 16$, 32 and 64; the gold samples have $N_1 = 300$, 1000 or 3000 pairs, and the

proxy is the other 20,029 pairs of the pool, of which a rule that prefers the shorter answer relabels a fifth, or a rule that prefers the longer answer 35%, the rest keeping human labels of the gold population; 60 seeds, joint intervals. The rules were chosen because, in the training data, replacing the length coordinate of the fit to the mixed labels by that of the fit to the human labels removes most of their bias (41.2° to 19.7° and 24.7° to 18.3° at $d = 64$). With the shorter-answer rule, the ℓ_1 correction reduces the error of WE by 32.1% to 47.8% in all nine cells, the lower ends of the intervals above 25%, and the gold-only error by 39% to 53%, a figure inflated by the gold-population labels in the proxy; pooling is worse than gold-only in seven cells. With the longer-answer rule its gain over WE is 1.2% to 38.3%, larger for larger N_1 and smaller for larger d, and seven of the nine intervals exclude zero. The corrections change 1.7 to 9.6 coordinates on average and include the length coordinate in 90% to 100% of the fits. When the reward model or Qwen2.5-1.5B-Instruct labels the whole proxy, the length coordinate carries little of the bias (12% and 6% of the reward model's for $d = 16$ and 64), and the correction is up to 10.3% worse than WE, significantly in six of twelve cells ($d = 16$, 64, $N_1 = 300$, 1000, 20 seeds, intervals over seeds).

The ℓ_1 correction. With the reward model relabeling half of the proxy in its own representation, we also ran, as an exploratory analysis, the ℓ_1 correction (5), with its penalty and ridge chosen by the oracle rule. On HH-RLHF, in design (v) with $N_1 = 500$, 1000 and 3000 gold pairs, its gain over WE is 4.0% $[0.3, 7.4]$, 5.9% $[-0.3, 12.1]$ and 3.7% $[-1.8, 10.1]$, against 8.4% $[4.4, 13.6]$, 11.8% $[7.5, 15.9]$ and 9.1% $[3.8, 14.8]$ for THE (intervals of the joint bootstrap). With a quarter of the proxy relabeled it is -0.8%, -3.3% and 6.4%, and with all of it -0.2%, 2.0% and 1.2% (30 seeds, intervals over seeds; only -3.3% and 6.4% exclude zero). On Community Alignment it is -1.4%, 1.0%, -3.6% and -2.4% for 10, 20, 40 and 80 gold annotators, against 9.4%, 11.2%, 8.9% and 3.0% for THE, and zero lies in every interval even over the 30 seeds alone. In the training data, the fit to the partly relabeled proxy is 29.0° (HH-RLHF) and 36.6° (Community Alignment) from the fit to the same pairs' human labels; correcting its five largest of 16 coordinates leaves 14.5° and 23.6°, and correcting eight leaves 7.6° and 17.1°. Neither shift is sparse, and the one on Community Alignment is less so.

Rule-based and injected judges. Along the most curved principal direction, a rule on another population's proxy gives ratios of 1.46 (PRISM) and 1.21 (Community Alignment), and along the flattest direction 0.99 to 1.02. Injected into a proxy from the gold population, it gives ratios of 1.27, 1.48 and 1.83 with 8, 16 and 32 components of PRISM, as the curvature of the bias direction grows from 1.68 to 2.18 and 2.78 times the mean, and 1.44 on Community Alignment (curvature 1.73), while pooling is up to 40% worse than gold-only on PRISM; a random direction gives 1.00 to 1.13, and the flattest direction or no bias 0.98 to 1.01. THE's gain over WE stays between 4.4% and 5.3% as the curvature grows: the ratio rises because the improvement of WE falls, to 5.9% with 32 components.

Implementable rule. Under the implementable rule, THE's gain over WE remains significant with the reward model on half of another population's proxy and with the rule-based and the injected judges, at 39% to 135% of its size under the oracle rule, and with the reward model alone on PRISM (1.9% $[0.6, 3.2]$). It vanishes when the reward model labels three quarters of the proxy in its own representation (0.0% on PRISM, 2.5% $[-1.0, 5.5]$ on Community Alignment). With three quarters or all of the proxy relabeled there, WE and THE both fall below gold-only under this rule, because the rule does not fall back to gold-only when the proxy is nearly useless: by 10.6% and 10.6% on PRISM and by 20.0% and 16.9% on Community Alignment at three quarters, and by 44.5% to 98.4% with the whole proxy, where THE loses less than WE.

Large bias. When the reward model labels the whole proxy in its own representation, the proxy direction is 57° from the gold direction, and under the oracle rule WE and THE are within 0.8% of gold-only and within 0.5% of each other. This is the regime of Theorem 4 in which the bias is several gold-only errors and no weight can remove much of the gold-only error.

Predictions from training data. The predictions order the cells of the designs with another population's proxy well, with rank correlations of 0.69 to 0.83 under the oracle rule, and match the reward-model judge on Community Alignment, but they overstate the gains of the injected judges, which reach 0.28 to 0.40 of the prediction (0.5 to 0.75 with an exploratory de-noised angle, which

removes the reference noise by comparing the estimate with references fitted to two halves of the held-out units), and understate those on HH-RLHF. Two designs on Community Alignment contradict the curvature criterion under the oracle rule: along the answer-length direction (curvature 1.42) and with Qwen2.5-1.5B-Instruct (Qwen Team, 2024) as the judge (agreement with the annotators 44%, curvature 1.57), THE and WE improve on gold-only by the same amount. THE is ahead only under the implementable rule, by 2.9% and by 9–24%, and when the language model labels half of the proxy or more, both methods fall below gold-only under that rule. Without a bias THE loses 0.2–1% to WE, the cost of tuning one more hyperparameter.

The smoothed sign loss. To check the loss of the theory on real labels, the reward-model designs of Community Alignment, PRISM and HH-RLHF were rerun with the smoothed sign loss, with the same draws, seeds and references, a bandwidth $c\,s\,n_k^{-1/5}$ for each sample with $c = 1$ and s the scale of the data, weights and penalties on a grid extended where choices reached its edge, one warm-started path of fits, and the oracle rule. The value $c = 1$ was fixed after a first run, in which smaller bandwidths led different starting points to local minima 30° to 52° apart. With half of the proxy relabeled, the judge pulls the target of the sign loss almost onto its own direction: on the proxy pairs, the fit to the mixed labels is 72° from the fit to the human labels, 88–92% of the way to the fit to the judge's labels, against 29° to 37° and 35–45% for the logistic likelihood, because the judge's labels are deterministic while the human labels agree with any direction on about 60% of the pairs. The proxy then barely helps: WE and THE improve on gold-only by 3.5% and 6.3% on Community Alignment, by 0.7% and 0.9% on PRISM, and by -1.0% to -0.1% and -2.7% to -1.2% on HH-RLHF. We therefore added judge shares of 10%, 15%, 20% and 25%, registering each before its run, at which the fit to the mixed labels is 23° to 54° from the human fit on Community Alignment. There WE improves on gold-only by 28–33% on Community Alignment, and the error of THE is 4.2%, 5.1%, 6.0% and 4.3% below that of WE, with joint intervals $[1.4, 7.8]$, $[1.9, 8.4]$, $[1.7, 10.0]$ and $[0.4, 9.1]$ and THE better in 97–99.5% of the joint draws; on PRISM the gain is 0.8–1.5%, and on HH-RLHF, whose proxy otherwise comes from the gold population, it lies between -1.8% and 2.7% in the twelve cells ($N_1 = 500$, 1000 and 3000), 0.1–0.6% on average (intervals over seeds for PRISM and HH-RLHF). Aligned by the size of the proxy bias, the gains match those of the logistic likelihood on the same draws of Community Alignment: 4.2% at a bias of 23°, against 2.9% at 15° and 8.1% at 37°. Without a judge, the sign loss improves on gold-only by 22.6% (WE) and 23.0% (THE) on Community Alignment, as the logistic likelihood does. The sign estimates are less precise, with gold-only errors 10° to 18° larger than the logistic ones.

K.6 Asymptotic constants beyond the proved regimes

These computations evaluate the asymptotic constants of Proposition 2; they are not simulations of the estimators.

Random designs. 180 designs: tangent dimension 2, 3, 5 or 8; eigenvalue ratio of H from 4 to 256; noise covariance V with the eigenvectors of H and eigenvalue ratio from 1.5 to 256, in reverse order to H (low-noise curved directions), in the same order, or in random order, 60 designs each; $N_2/N_1 \in \{1, 3, 10, 30\}$; an offset along a random eigenvector of size 0 to 0.5 times the gold-only error; one bandwidth for both samples; seven bias levels from $r = 0.25$ to 16. Each cell uses 40,000 antithetic draws of the bias direction, and an improvement counts as strict when $(C_{\mathrm{gold}} - C_{\mathrm{THE}})/(C_{\mathrm{gold}} - C_{\mathrm{WE}}) > 1.001$. In reverse order, THE improves strictly on the best weight in every cell with $r \geq 1$, in 97% of the cells with $r = 0.5$ and in 68% with $r = 0.25$, where the remaining improvements, of order r^4, are below the numerical resolution; the median improvement ratio rises from 1.03 at $r = 0.5$ to 1.97 at $r = 16$. In the same order, THE improves on the best weight in no cell. In random order, it improves strictly in 164 of the 202 cells in which a large finite penalty at the best weight helps, and in none of the 218 cells in which it does not help; with the 420 cells in the same order, no (ρ, λ) improves on the best weight in any of the 638 cells where a large finite penalty does not. The closed-form criterion (10) agrees in sign with this derivative under the bandwidth conditions of the theorems; with a fixed offset and bandwidth the two can differ, and the counts above use the derivative.

Separately tuned bandwidths. With bandwidths $h_k = c_k N^{-1/5}$, the offset of each sample scales as c_k^2 and its variance as $N/(c_k N_k)$, and each estimator optimizes both bandwidths and its own

hyperparameters. In 36 cells (a two-level spectrum with five curved and two flat directions and curvature ratio 256, with $V \propto H^{-1}$ or $V \propto H^{-1/2}$, and a two-dimensional reverse-order design with curvature ratio 64 and noise ratio 2; $N_2/N_1 \in \{3, 10\}$; r from 0.5 to 16), THE improves strictly on the best weight in every cell; for instance, with $V \propto H^{-1}$, $N_2 = 10N_1$ and $r = 4$, its gain over the best weight is 8.1% with separate bandwidths and 6.3% with a common one.

Large bias. With H and V sharing eigenvectors in reverse order, a common bandwidth and a uniform bias direction ($N_2 = 10N_1$, a two-level spectrum with curvature ratio 256), the ratio of the improvements over gold-only $(C_{\text{gold}} - C_{\text{THE}})/(C_{\text{gold}} - C_{\text{WE}})$ at $r = 1, 4, 16, 64$ is 1.86, 4.38, 5.24, 5.31 for $V \propto H^{-1}$; 1.44, 2.73, 3.07, 3.10 for $V \propto H^{-1/2}$; and 1.03, 1.11, 1.12, 1.12 when the eigenvalues of V differ by a factor 2. The constants of all methods tend to the gold-only constant as r grows, but the improvement of THE stays a fixed multiple of that of the best weight, as Theorem 28 states; in the last design the two-step correction keeps only 0.67 times the improvement of re-weighting, as in Proposition 9.

L Additional related work

This appendix expands the related-work paragraph of Section 1.

Preference learning for alignment. Reinforcement learning from human feedback fits a reward model to comparisons under a Bradley–Terry model or its multi-way extension and then optimizes a policy against it (Christiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022a). Direct preference optimization maximizes the same likelihood with the reward written through the policy (Rafailov et al., 2023), and Azar et al. (2024) place it in a general family of preference objectives and show that the Bradley–Terry form can make it overfit when preferences are deterministic. Zhu et al. (2023) show that the maximum likelihood estimator converges under the Bradley–Terry–Luce and Plackett–Luce models with linear rewards and study pessimistic policy learning built on it. Casper et al. (2023) survey open problems of this pipeline, including disagreement among annotators, and Gao et al. (2023) show that optimizing a proxy reward model too far lowers the score of the gold reward model that labeled its training data. The fitted model in these works is a single Bradley–Terry model for all comparisons. In our model each comparison is answered by an independently drawn user with a random utility vector, and such a fit can be inconsistent for the direction of the population center (Aouad et al., 2025).

Heterogeneous annotators. When annotators differ, a single Bradley–Terry fit over unrestricted utilities ranks the alternatives by their Borda count (Siththaranjan et al., 2024). Chakraborty et al. (2024) show that one reward cannot represent diverse subpopulations, learn a mixture of rewards and optimize the worst-off group; personalized methods learn several reward models or condition the reward on a user-specific latent variable (Park et al., 2024; Poddar et al., 2024), and pluralistic and social-choice approaches ask how diverse preferences should be represented or aggregated (Sorensen et al., 2024; Conitzer et al., 2024). In econometrics the same choice model is the random-coefficients, or mixed, logit (McFadden & Train, 2000), estimated by maximizing its likelihood under a parametric law for the coefficients; the result of Section 5 for a known Gaussian heterogeneity law uses this likelihood. The sign estimator leaves the law unknown and uses only its symmetry. Unlike the works above, we fix the target population and treat the other population as a biased proxy.

Cheaper and weaker supervision. AI feedback labels comparisons with a language model instead of people (Bai et al., 2022b; Lee et al., 2024), and weak-to-strong generalization trains a strong model on labels produced by a weaker one (Burns et al., 2024); in both, the labels need not agree with the target. Zhu et al. (2025) give sufficient conditions under which proxy preference data provably reduce the sample complexity of learning the target policy. Prediction-powered inference combines a small labeled sample with machine-learning predictions on a large unlabeled sample, and removes the bias of the predictions using the labeled units, on which both are observed (Angelopoulos et al., 2023). Our gold and proxy samples are unpaired: no comparison has both a gold and a proxy response, so the shift between the populations has to be learned from the two fitted directions rather than from paired differences.

Maximum score estimation. The sign estimator is the maximum score estimator of Manski (1975) applied to paired comparisons. Kim & Pollard (1990) derive its cube-root rate and its non-Gaussian limit, the maximizer of a Gaussian process with quadratic drift, and Horowitz (1992) replaces the indicator by a smooth kernel to obtain faster rates and asymptotic normality under additional smoothness. Mukherjee et al. (2021) study the estimator when the dimension grows with the sample size, slowly or beyond it, and give minimax rates up to a logarithmic factor; Chen & Lee (2018) select variables by a best-subset constraint on the score, and Feng et al. (2022) penalize a smoothed score by the ℓ_1 norm and obtain the rate $(s \log d/N)^{\beta/(2\beta+1)}$ for smoothness β, which for $\beta = 2$ matches the gold term $(s \log d/N_1)^{2/5}$ of our smoothed sparse result. All of these use one population.

Transfer learning with a biased source. Chen et al. (2015) fit linear regressions to a small target sample and a larger, possibly biased, auxiliary sample while penalizing the difference between the two models' predictions; THE uses a quadratic coupling of the same kind for the sign loss, and Section 4 identifies when a finite coupling improves on the best weighted fit. For smooth losses the coupling acts, to first order, as a matrix weight built from $X^\top X$ or the Fisher information, and it reduces to a scalar weight on the proxy loss when this matrix is a multiple of the identity (Lindley & Smith, 1972; Chen & Ibrahim, 2006). Chen et al. (2015) compare it in simulations with a plug-in scalar average of the two least-squares fits and prove, with such a data-driven average, that for Gaussian responses with five or more coefficients and ten or more error degrees of freedom, using the small sample alone is inadmissible at every bias size. In a Gaussian linear model with squared loss and an isotropic random bias, the coupled estimator with the right weight and penalty is the mixed estimator of Theil & Goldberger (1961), so there its advantage, averaged over the bias, over every scalar weight follows from classical arguments; the criterion of Section 4 also covers the smoothed sign loss with its smoothing offset and the angular error, with exact constants. Bastani (2021) fits the proxy data first and corrects this fit on the gold data with an ℓ_1 penalty; that paper proves an upper bound for this estimator and lower bounds for model averaging and weighted losses at their best tuning, which improve on the better single source by at most a constant factor. This comparison, an upper bound for the new estimator against lower bounds for the baselines at their best tuning, is the form used in this paper. Li et al. (2022) extend the correction to several sources with minimax rates, Tian & Feng (2023) extend it to generalized linear models, and Cai & Wei (2021) study transfer for nonparametric classification under posterior drift. These analyses use smooth losses or plug-in classifiers; the sign loss is discontinuous, and its population minimizer is identified only through the symmetry of the heterogeneity.