AN ASYMPTOTIC ONE-SAMPLE SURROGATE FOR NON-GAUSSIAN ERM WITH GENERAL CONVEX REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study convex empirical risk minimization (ERM) with non-Gaussian features and nonsmooth penalties through a surrogate driven by a single Gaussian vector. The surrogate preserves the penalty and calibrates its coefficients using the actual joint law of the feature–label model. We prove that this population calibration exists and is unique. We conjecture a conditional Gaussian prediction principle for fresh scores when the data satisfy a Poincaré inequality. Conditional on this conjecture, standard concentration conditions on the data and regularity conditions on the losses imply agreement between the ERM minimizer and its surrogate in normalized means, prediction variances, and conditional fresh-score laws averaged over the fresh feature. The result covers general uniformly strongly convex penalties and, under projection anti-concentration, pure Lasso with regularization strengths bounded away from zero and infinity. Transfer of the surrogate's support statistics to the trained estimator remains open, although it is strongly supported by simulations. The core proof relies on a cavity Hamilton–Jacobi argument that compares the main statistics of the two minimizers without differentiating the penalty.

1 INTRODUCTION

We study high-dimensional convex empirical risk minimization:

$$\widehat{W} := \widehat{W}((x_i, y_i)_{i\in[n]}) = \arg\min_w \left\{ \rho(w) + \sum_{i=1}^{n} L_{y_i}(x_i^\top w/\sqrt{n}) \right\}. \tag{1}$$

The training pairs (x_i, y_i) are independent, with $x_i \in \mathbb{R}^p$ and $p/n \to c \in (0, \infty)$. Features $(x_i)_{i\in[n]}$ may be non-Gaussian with dependent coordinates. The loss is smooth and convex; the finite convex penalty $\rho : \mathbb{R}^p \to \mathbb{R}$ may be nonsmooth, including Lasso, which can set coefficients exactly to zero.

Gaussian theory and its universality extensions give precise descriptions under additional structure. They do not justify a general Gaussian replacement with matching covariances: the joint feature–label law can affect learning errors.

Inspired by earlier surrogate approaches (Loureiro et al., 2021a; Celentano et al., 2023), we introduce a surrogate designed to approximate the prediction behavior of $\widehat{W}$. With $\Sigma_x := \mathbb{E}\left[xx^\top\right]$ the feature second-moment matrix, define

$$W_*(G) := \arg\min_w \left\{ \rho(w) + \frac{\nu_*}{2} w^\top \Sigma_x w + \beta_* G^\top w - \sqrt{n}\, \eta_*^\top w \right\}, \qquad G \sim N(0, \Sigma_x). \tag{2}$$

The coefficients $\nu_*, \beta_* > 0$ and $\eta_* \in \mathbb{R}^p$ are the *calibration parameters*. For the given penalty, they are calibrated from the actual joint law of (x, y) and the loss by solving the fixed-point equations (4)–(5).

We call $W_*(G)$ a *one-sample surrogate* because, once calibrated, it replaces a minimization problem depending on n training pairs by one whose only random input is a single p-dimensional Gaussian vector G. The law of (x, y) remains encoded in Σ_x and the calibration parameters. This substantially simplifies the random optimization problem while preserving the original penalty ρ.

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.

Our empirical comparison is conditional on Gaussian approximation of fresh scores over training randomness (Conjecture 1). For any square-integrable random vector W, let $C_W := \mathbb{E}\left[(W - \mathbb{E}[W])(W - \mathbb{E}[W])^\top\right]$ be its covariance over all randomness defining W. Under the assumptions of Theorem 2, the normalized means and prediction variances of $\hat{W}$ and $W_*(G)$ agree asymptotically. Corollaries 3 and 4 give averaged fresh-score laws and prediction risks. Coordinate laws, empirical support transfer and support recovery remain open.

The true feature law enters through the introduction of a score surrogate during calibration

$$U_* := x^\top \mu_* + \alpha_* Z, \qquad Z \sim N(0,1), \quad Z \perp (x, y),$$

where $\mu_* := \mathbb{E}_G[W_*]/\sqrt{n}$ and $\alpha_*^2 := \mathrm{tr}(\Sigma_x C_{W_*})/n$. As noted in Yaakoubi et al. (2026), the law of $x^\top \mu_*$ might be non Gaussian, which gives a clear characterization of Gaussian universality.

Related work. Approximate message passing (AMP) and the convex Gaussian min–max theorem (CGMT) give precise descriptions of sparse estimators in Gaussian models (Bayati and Montanari, 2012; Thrampoulidis et al., 2018). Celentano et al. (2023, equations (7)–(11), Theorems 5 and 8) give a closely related covariance-weighted Gaussian-sequence surrogate and model-size guarantees for squared-loss Lasso in Gaussian linear models, including correlated designs. These descriptions do not directly cover the non-Gaussian feature laws considered here.

Non-Gaussian sparse estimation is also established in structured settings. Han and Shen (2023, Theorem 3.8, equations (3.8)–(3.10)) obtain coefficient and support-size descriptions for squared-loss Lasso with independent sub-Gaussian design entries, under their signal, noise and tuning conditions. Other extensions cover random-feature least squares with nonsmooth penalties and block-dependent linear models (Bosch et al., 2023; Lahiry and Sur, 2024). Esmaili Mallory et al. (2025) study ridge logistic regression with specified dependence among observations. Our training pairs remain independent, while coordinates within each feature vector may be dependent.

The vector proximal description of Loureiro et al. (2021a) is proved for jointly Gaussian teacher–student covariates. Their Conjecture 1 proposes a broader Gaussian replacement; their Section 3.4 and Pesce et al. (2023); Mai and Liao (2025) show its limitations. Keeping the actual feature law in $x^\top \mu_* + \alpha_* Z$ appears in the smooth-penalty analysis of Yaakoubi et al. (2026); related decompositions occur in the ridge factor models of Mai and Liao (2025). Here the original nonsmooth penalty is retained as well.

Montanari and Saeed (2022, Assumption 5, equations (11)–(12), Theorems 1–3) prove universality under Gaussian approximation of $x^\top u$ for deterministic parameters u in an admissible set, together with further conditions. That approximation concerns feature randomness. In Conjecture 1, the fresh feature x is fixed and only the centered fluctuation induced by training is approximated by a Gaussian. The mean retains the true feature law.

Proof strategy and organization. We interpolate between the empirical objective and the surrogate using $\mathcal{E}$ defined in (14). A leave-one-out argument and the score comparison (21) in Conjecture 1 give a Hamilton–Jacobi equation for the expected minimum. Its derivatives encode the prediction moments, which we transport without differentiating the penalty. Section 2 states the results. Section 3 provides illustration of the agreement between ERM minimizer and one-sample surrogate mainly through the behavior of their support $S := \{i \in [p] : W_i \neq 0\}$. Section 4 explains the Hamilton–Jacobi equation used to prove our main result and the appendices contain the details.

2 Main theoretical results

2.1 Definition and well-posedness of the surrogate

Fix n, p and a feature–label law $(x, y) \in \mathbb{R}^p \times \mathcal{Y}$, where $\mathcal{Y}$ is the label space, with finite second moment $\Sigma_x := \mathbb{E}\left[xx^\top\right]$. We use $|\cdot|$ for the Euclidean norm of a vector and the absolute value of a scalar. The population system defines the surrogate from this law, the loss, and the penalty; Theorem 2 later compares it with the trained estimator for pure Lasso and general uniformly strongly convex penalties. We use the Moreau-envelope notation of Thrampoulidis et al. (2018). For a convex

function $\omega : \mathbb{R}^d \to \mathbb{R}$ and $\tau > 0$, write

$$e_\omega(u;\tau) := \min_{z\in\mathbb{R}^d} \left\{ \omega(z) + \frac{|u-z|^2}{2\tau} \right\}. \tag{3}$$

The minimizing point is $\mathrm{prox}_{\tau\omega}(u)$, and $e_\omega(u;0) := \omega(u)$ is the right-limit extension. For the scalar losses, set $\mathrm{prox}_{0L_y}(u) := u$. In (4)–(5), loss expectations use the actual joint law of (x, y) and an independent $Z \sim N(0,1)$; penalty expectations use an independent $G \sim N(0, \Sigma_x)$. For a vector $\mu \in \mathbb{R}^p$ and $\alpha, \kappa \geq 0$, define the loss coefficients by (4), taking β to be the nonnegative square root. The population system is considered on the branch $\beta, \nu > 0$:

$$\begin{gathered} U := x^\top \mu + \alpha Z, \qquad \zeta := \mathrm{prox}_{\kappa L_y}(U), \qquad \xi := L_y'(\zeta), \\ \beta^2 := \mathbb{E}\left[\xi^2\right], \qquad \nu := \mathbb{E}\left[\frac{L_y''(\zeta)}{1+\kappa L_y''(\zeta)}\right], \qquad \eta := \nu\Sigma_x\mu - \mathbb{E}\left[x\xi\right]. \end{gathered} \tag{4}$$

$$W(G) := \arg\min_w \{\rho(w) + \nu w^\top \Sigma_x w/2 + \beta G^\top w - \sqrt{n}\,\eta^\top w\}, \tag{5a}$$

$$\mu = \frac{\mathbb{E}_G\left[W(G)\right]}{\sqrt{n}}, \qquad \alpha^2 = \frac{1}{n}\,\mathrm{tr}(\Sigma_x C_W)\,, \qquad \kappa = -\frac{\mathbb{E}_G\left[G^\top W(G)\right]}{\beta n}. \tag{5b}$$

Equations (4)–(5) contain auxiliary Gaussian expectations and the population law, but no realization of the training sample.

Theorem 1 (Existence and uniqueness of the population solution)**.** *Fix n, p. Assume $\mathbb{E}\left[|x|^2\right] < \infty$ and $\Sigma_x = \mathbb{E}\left[xx^\top\right] \succ 0$. Suppose $\rho : \mathbb{R}^p \to \mathbb{R}$ is convex with $0 \in \partial\rho(0)$, and every L_y is C^2 with*

$$|L_y'(u)| \leq L_1, \qquad 0 < L_y''(u) \leq L_2, \qquad u \in \mathbb{R}.$$

Then (4)–(5) *has at most one finite solution with $\alpha, \kappa \geq 0$ and $\beta, \nu > 0$; any such solution has $\alpha, \kappa > 0$.*

If, in addition, the losses are nonnegative with $\mathbb{E}\left[L_y(0)\right] < \infty$, and

$$\arg\min_{w\in\mathbb{R}^p} \rho(w) = \{0\}, \qquad \mathbb{E}\left[L_y'(0)^2\right] > 0, \tag{6}$$

then this solution exists.

Theorem 1 is proved in Appendix A. Under the existence assumptions of Theorem 1, denote the unique solution by $(\mu_*, \alpha_*, \beta_*, \nu_*, \kappa_*, \eta_*)$ and its optimizer by $W_*(G)$, as in (2). The penalty and loss conditions in Assumption 1 imply the existence assumptions of Theorem 1. Positivity at each finite size does not imply the uniform positive-variance condition in Theorem 2. Lemma A.1 in Appendix A.3 gives an equivalent min-max formulation.

2.2 Asymptotic setting and assumptions

For comparison with the trained estimator, consider a sequence with

$$n \longrightarrow \infty, \qquad p = p(n), \qquad \frac{p}{n} \longrightarrow c \in (0, \infty).$$

For each n, the training pairs $(x_i, y_i)_{i=1}^n$ are independent copies of (x, y). The population law, loss, penalty, and all derived quantities may depend on n. We suppress this dependence, including for Σ_x, $\widehat{W}$, and the population order parameters. Independence is required within each training sample; no coupling of different sample sizes is assumed. We assume

$$\underline{c}_x I_p \preceq \Sigma_x \preceq K I_p, \qquad \underline{c}_x > 0. \tag{7}$$

Enlarging K if necessary, also $p/n \leq K$ for all sufficiently large n. All constants in (7), the eventual bound $p/n \leq K$, and the loss, data-Poincaré, and penalty/design conditions in Assumptions 1–3, are uniform along the sequence unless dependence on a fixed auxiliary parameter is stated explicitly. Write $\|A\|_{\mathrm{hs}} := (\mathrm{tr}(A^\top A))^{1/2}$ for the Hilbert–Schmidt norm and $\|\cdot\|_{L^q}$ for a moment norm. Feature coordinates may be dependent and $\mathbb{E}\left[x\right]$ need not vanish. We study the estimator $\widehat{W}$ in (1); the score normalization remains $x^\top w/\sqrt{n}$ after thinning or deletion. The one-sample surrogate uses the same dimension and population law as the training problem. Our conclusions compare these size-matched objects; they do not require population coefficients to have dimension-independent limits.

Assumption 1 (Penalty and loss)**.** *The penalty $\rho : \mathbb{R}^p \to \mathbb{R}$ is convex, with* $\arg\min_{w\in\mathbb{R}^p} \rho(w) = \{0\}$. *The losses are nonnegative convex C^4 functions, with*

$$\sup_{y,u} |L_y^{(j)}(u)| \le L_j \quad (j = 1, 2, 3, 4), \qquad 0 \le L_y(0) \le L_0 \quad \textit{almost surely}.$$

For every $M < \infty$, $c_L(M) := \inf_{y,\,|u|\le M} L_y''(u) > 0$, and $\mathbb{E}\left[L_y'(0)^2\right] > 0$ at every problem size. The constants $L_0, \dots, L_4$ and $c_L(M)$, for each fixed M, are uniform along the sequence.

Binary logistic loss satisfies the loss conditions in Assumption 1, with $L_y(0) = \log 2$ and $\mathbb{E}\left[L_y'(0)^2\right] = 1/4$; squared loss has unbounded slope and needs additional moment estimates that we do not provide here for simplicity. Theorem 1 retains its broader finite-size assumptions. Theorems 1 and 5 allow general convex penalties; Theorem 2 covers pure Lasso and uniformly strongly convex penalties.

Assumption 2 (Poincaré inequality for the data)**.** *The joint law of (x, y) satisfies, for a constant $C_{\mathrm{P}} < \infty$ independent of dimension, for all bounded test functions F continuously differentiable in their continuous arguments*

$$\mathrm{Var}_{x,y}(F(x,y)) \le C_{\mathrm{P}} \mathbb{E}_{x,y}\left[|\nabla_x F(x,y)|^2 + \Gamma_y F(x,y)\right]. \tag{8}$$

where, for Euclidean labels, $\Gamma_y F(x,y) := |\nabla_y F(x,y)|^2$. For a fixed finite label space, use $\Gamma_y F(x,y) := \frac{1}{2}\sum_{y'\in\mathcal{Y}} |F(x,y') - F(x,y)|^2$.

The finite-label definition uses differences, so binary labels are included without differentiating them. Condition (8) concerns the joint law; it requires neither independence of x and y nor centered or independent feature coordinates. Applying it to functions of x alone gives, by Lemma B.1,

$$\mathbb{E}_x\left[\left|x^\top A x - \mathrm{tr}(\Sigma_x A)\right|^2\right] \le K_Q^2 \|A\|_{\mathrm{hs}}^2, \qquad K_Q^2 := 4C_{\mathrm{P}}K, \tag{9}$$

for every deterministic symmetric A.

Assumption 3 (Penalty and design alternatives)**.** *One of the following holds:*

1. Pure Lasso with projection anti-concentration. *For constants $0 < \lambda_- \le \lambda_+ < \infty$ independent of dimension,*

$$\rho(w) = \lambda\|w\|_1, \qquad 0 < \lambda_- \le \lambda \le \lambda_+ < \infty. \tag{10}$$

   The strength λ may depend on size, and there is a constant $A_{\mathrm{ac}} < \infty$, independent of size and direction, such that

$$\mathbb{P}\Big(|x^\top u| \le r\sqrt{u^\top \Sigma_x u}\Big) \le A_{\mathrm{ac}} r, \qquad u \ne 0, \quad 0 < r < 1. \tag{11}$$

2. Uniform strong convexity. *The penalty satisfies*

$$w \longmapsto \rho(w) - \frac{\gamma_{\mathrm{sc}}}{2}|w|^2 \quad \textit{is convex}, \qquad \gamma_{\mathrm{sc}} > 0 \quad \textit{independent of } n. \tag{12}$$

Anti-concentration is required only in the pure-Lasso alternative, with no smallness restriction on A_{ac}. The strongly convex alternative allows nonsmooth and nonseparable penalties, with their minimum at zero as in Assumption 1. It includes the elastic penalty

$$\rho(w) = \lambda\|w\|_1 + \rho_0(w), \qquad \rho_0 \in C^2(\mathbb{R}^p), \qquad \nabla^2\rho_0(w) \succeq \gamma_{\mathrm{sc}} I_p \quad \text{for all } w, \tag{13}$$

with $\lambda, \gamma_{\mathrm{sc}} > 0$ independent of n. The remainder may depend on size and be nonseparable; no upper Hessian bound is required.

2.3 Perturbed optimizers and prediction quantities

For $t \in [0,1]$, let $B_i(t) \sim \mathrm{Bernoulli}(t)$ retain observations independently, and write $B_i := B_i(t)$. The selectors, training pairs, and probe $G \sim N(0, \Sigma_x)$ are mutually independent. For $v, b > 0$ and $h \in \mathbb{R}^p$, define

$$\mathcal{E}_{t,v,b,h}(w) := \rho(w) + \sum_{i=1}^n B_i(t) L_{y_i}(x_i^\top w/\sqrt{n}) + \frac{v}{2} w^\top \Sigma_x w + \sqrt{2b}\, G^\top w - \sqrt{n}\, h^\top w. \tag{14}$$

$$W_{t,v,b,h} := \arg\min_{w} \mathcal{E}_{t,v,b,h}(w). \tag{15}$$

The parameters v, b, h add quadratic curvature, Gaussian noise, and a linear field to the retained empirical objective. At $t = 0$, no observation is retained, and the choice $(v, b, h) = (\nu_*, \beta_*^2/2, \eta_*)$ gives exactly the surrogate $W_*(G)$ in (2). At $t = 1$, every observation is retained. Extending (14) to $(v, b, h) = (0, 0, 0)$ recovers (1) and its selected minimizer $\widehat{W}$.

The Hamilton–Jacobi equation (34) in Section 4, together with (35), determines the path

$$(v(t), b(t), h(t)) := (1 - t)(\nu_*, \beta_*^2/2, \eta_*), \tag{16}$$

that allows to prove Theorem 2 below (more details are provided in Section 4).

For any square-integrable coefficient vector W, write

$$\mu := \mathbb{E}[W]/\sqrt{n}, \qquad \alpha^2 := \operatorname{tr}(\Sigma_x C_W)/n. \tag{17}$$

Here μ is the normalized mean, and α^2 is the prediction variance over the randomness defining W, averaged over a fresh feature x. For an interior optimizer, define

$$\kappa := \frac{1}{n\sqrt{n}} \mathbb{E}\left[\operatorname{tr}(\Sigma_x D_h W)\right]. \tag{18}$$

Here $D_h W$ is the Jacobian of the optimizer with respect to h, and κ is the average score response to the physical linear field $\sqrt{n}\, h$. This Jacobian exists almost everywhere because positive quadratic curvature makes the solution map Lipschitz; the Gaussian field avoids its exceptional set (Lemma B.2).

We introduce a fixed interior domain $\mathcal{D}$: on $\mathcal{D}$, t ranges over a closed interval inside $(0, 1)$, v, b range over compact subintervals of $(0, \infty)$, and $|h|$ is bounded, with all bounds independent of n. Whenever finite differences are taken, slightly wider bounds specify an enlarged domain. We use comparison on every such domain, a sufficient formulation stronger than imposing it only on the particular domains traversed by the proof.

To compare score laws along the interpolation, we first specify the relevant conditional law and distance. For a real random variable u and a random vector x, write $\mathcal{L}(u \mid x)$ for a version of the regular conditional law of u given x, characterized by

$$\int_{\mathbb{R}} f(a)\, \mathcal{L}(u \mid x)(\mathrm{d}a) = \mathbb{E}[f(u) \mid x] \quad \text{almost surely} \tag{19}$$

for every bounded Borel function $f : \mathbb{R} \to \mathbb{R}$. For Borel probability laws P, Q on $\mathbb{R}$ with finite first moments, define their 1-Wasserstein distance by

$$\mathsf{W}_1(P, Q) := \inf_{\pi \in \Pi(P,Q)} \int_{\mathbb{R}^2} |a - b|\, \pi(\mathrm{d}a, \mathrm{d}b) = \sup_{\mathrm{Lip}(f) \le 1} \left| \int_{\mathbb{R}} f\, \mathrm{d}P - \int_{\mathbb{R}} f\, \mathrm{d}Q \right|, \tag{20}$$

where $\Pi(P, Q)$ is the set of couplings of P and Q, and the supremum is over all 1-Lipschitz functions $f : \mathbb{R} \to \mathbb{R}$.

The following conjecture 1 is needed to close the Hamilton–Jacobi argument, and preliminary numerical experiments appear consistent with it.

Conjecture 1 (Conditional Gaussian prediction under Poincaré concentration)**.** *Along a sequence in which the training pairs are independent copies of (x, y), $p/n \to c \in (0, \infty)$, and $\underline{c}_x I_p \preceq \Sigma_x \preceq K I_p$, suppose Assumptions 1 and 2 hold. Let x be an independent fresh feature. Let W be either the full optimizer $W_{t,v,b,h}$ or its unforced cavity optimizer, and use the corresponding moments μ, α^2 from* (17)*. For every fixed interior domain $\mathcal{D}$ introduced in Section 2.3,*

$$\sup_{(t,v,b,h) \in \mathcal{D}} \mathbb{E}_x \left[\mathsf{W}_1\left(\mathcal{L}\left(\frac{x^\top W}{\sqrt{n}} \,\middle|\, x \right), N(x^\top \mu, \alpha^2) \right) \right] \longrightarrow 0. \tag{21}$$

A stronger Gaussian approximation is available in a smoother setting: Louart (2026, Assumptions 4–7, Theorem 2.1) proves a standardized central limit theorem for nondegenerate deterministic projections of a regularized M-estimator under columnwise Poincaré concentration, a smooth uniformly convex loss, a strongly convex quadratic penalty, and additional higher-derivative bounds. Such a conclusion cannot hold for the nonsmooth penalties considered here. Proposition G.5 in Appendix G.4 gives a pure-Lasso coordinate whose distance from its moment-matched Gaussian stays positive even for Gaussian data and positive probes.

2.4 PREDICTION MOMENTS AND CONSEQUENCES

Theorem 2 (Prediction moments for Lasso and strongly convex penalties)**.** *Along a sequence in which the training pairs are independent copies of (x, y), $p/n \to c \in (0, \infty)$, and $\underline{c}_x I_p \preceq \Sigma_x \preceq K I_p$, suppose Assumptions 1–3 and Conjecture 1 hold, and assume:*

$$\sup_n \left\{ \frac{1}{n} \mathbb{E}_G \left[|W_*|^2 \right] + \kappa_* \right\} < \infty, \qquad \liminf_{n \to \infty} \alpha_*^2 > 0. \tag{22}$$

Then

$$\left| \frac{\mathbb{E}\left[\widehat{W}\right]}{\sqrt{n}} - \mu_* \right| + \left| \frac{1}{n} \operatorname{tr}\left(\Sigma_x C_{\widehat{W}}\right) - \alpha_*^2 \right| \longrightarrow 0. \tag{23}$$

Theorem 2 identifies the empirical optimizer's normalized mean and prediction variance with those of the population surrogate. It does not identify the full covariance or empirical support.

Fresh prediction and classification.

Corollary 3 (Asymptotic prediction law)**.** *Under the assumptions of Theorem 2, for an independent fresh pair (x, y) and for every measurable family of uniformly Lipschitz tests f_y for which both expectations in* (24) *are finite,*

$$\left| \mathbb{E}\left[f_y(x^\top \widehat{W}/\sqrt{n}) \right] - \mathbb{E}\left[f_y(x^\top \mu_* + \alpha_* Z) \right] \right| \longrightarrow 0, \tag{24}$$

where $Z \sim N(0, 1)$ is independent of (x, y). The Lipschitz bound is common along the sequence.

The conditional law fixes the fresh feature x and averages training randomness. The Gaussian comparison law is evaluated at the current problem size; neither its signal projection nor its variance is required to have a limit.

Corollary 4 (Asymptotic binary classification error)**.** *Under the assumptions of Theorem 2, let (x, y) be an independent fresh pair with $y \in \{-1, 1\}$ and set $\widehat{y}(x) := \operatorname{sign}_+(x^\top \widehat{W}/\sqrt{n})$, where $\operatorname{sign}_+(z) := 1$ for $z \geq 0$ and -1 otherwise. With Φ_{N} the standard normal distribution function,*

$$\left| \mathbb{P}\{\widehat{y}(x) \neq y\} - \mathbb{E}_{x,y}\left[\Phi_{\mathrm{N}}\left(-\frac{y x^\top \mu_*}{\alpha_*} \right) \right] \right| \longrightarrow 0. \tag{25}$$

The probability averages the training sample and the fresh pair.

The condition $\liminf_n \alpha_*^2 > 0$ in (22) bounds the comparison Gaussian's density, so no additional margin or fitted-score density assumption is needed. No uniform conclusion is asserted as that variance vanishes. Corollaries 3 and 4 are proved in Appendix F.

3 PRACTICAL ILLUSTRATIONS

Section 3.1 presents the surrogate for Lasso and elastic regularization, including coordinate-wise formulas when the feature covariance is diagonal in the penalty's coordinates. Section 3.2 uses correlated block designs and evaluates the surrogate, comparing its support predictions with ERM under non-Gaussian and covariance-matched Gaussian feature laws.

3.1 LASSO AND ELASTIC REGULARIZATION

Take

$$\rho(w) := \lambda_1 \|w\|_1 + \frac{\lambda_2}{2} |w|^2, \qquad \|w\|_1 := \sum_{j=1}^p |w_j|, \qquad |w| := \left(\sum_{j=1}^p w_j^2 \right)^{1/2}.$$

Thus $\|w\|_1$ is the ℓ_1 norm, $\lambda_1 > 0$ is the Lasso strength, and $\lambda_2 \geq 0$ is the ridge strength. The same penalty notation is used in Appendix G.2. For $\Sigma_x = \operatorname{diag}(c_1, \ldots, c_p)$ in the penalty's coordinates,

the $c_j > 0$ are its diagonal entries. Put $\Gamma_{*,j} := \nu_* c_j + \lambda_2$ and $q_{*,j} := \sqrt{n}\,\eta_{*,j} - \beta_* \sqrt{c_j} Z_j$, with independent $Z_j \sim N(0,1)$. Then

$$W_{*,j} = \frac{\mathrm{soft}(q_{*,j}, \lambda_1)}{\Gamma_{*,j}}, \qquad \mathrm{soft}(z, \lambda_1) := \mathrm{sign}(z)(|z| - \lambda_1)_+, \tag{26}$$

where $a_+ := \max\{a, 0\}$ and $\mathrm{sign}(0) := 0$. The moment equations then become

$$\sqrt{n}\mu_{*,j} = \frac{\mathbb{E}\left[\mathrm{soft}(q_{*,j}, \lambda_1)\right]}{\Gamma_{*,j}}, \quad \alpha_*^2 = \frac{1}{n}\sum_j \frac{c_j \,\mathrm{Var}(\mathrm{soft}(q_{*,j}, \lambda_1))}{\Gamma_{*,j}^2}, \quad \kappa_* = \frac{1}{n}\sum_j \frac{c_j \mathbb{P}(|q_{*,j}| > \lambda_1)}{\Gamma_{*,j}}. \tag{27}$$

The loss supplies surrogate curvature $\nu_* c_j > 0$ even for pure Lasso ($\lambda_2 = 0$). Pure Lasso is covered by Theorem 2 when (10) and (11) hold. For $\inf_n \lambda_2 > 0$, elastic regularization satisfies (12) and uses Theorem 2 without anti-concentration or a Lasso threshold, also for size-dependent λ_1. Vanishing positive ridge does not satisfy this strongly convex branch. Diagonal second moments do not imply independent feature coordinates; only the one-sample surrogate separates. When Σ_x is not diagonal in the penalty's coordinates, the one-sample surrogate remains a coupled convex optimization. Active-set sensitivity gives another interpretation of its response; Corollary G.3 gives (138). At the population solution, it also gives $\nu_* \kappa_* = \mathbb{E}_G\left[|\,\mathrm{supp}\, W_*|\right]/n$ for pure Lasso. This surrogate identity does not transfer training support counts.

3.2 Non-Gaussian support recovery

Theorem 2 establishes prediction moments, but does not establish coordinate-wise support transfer. We test the calibrated surrogate's support predictions against finite-sample ERM, and compare non-Gaussian and Gaussian feature laws with the same mean and covariance.

With $(n, p) = (200, 400)$, the features comprise ten independent correlated blocks of ten coordinates and 300 independent standard Gaussian singletons. Within each block the correlation is 0.95. Figure 1(a) compares the marginal density of a representative non-Gaussian block coordinate with its matched Gaussian counterpart. In each design, the response is

$$y = \tanh(x^\top \theta_0 + \varepsilon), \tag{28}$$

where ε is a centered noise. The teacher θ_0 has twenty active block coordinates and five active singleton coordinates, with all the remaining coordinates set to zero. We use pseudo-Huber loss with linear fitted score and the elastic penalty $\rho(w) = \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2/2$. Under $\theta = w/\sqrt{n}$, the coefficient of $\|\theta\|_1$ in the mean-loss objective is $\lambda_1/\sqrt{n}$. Appendix H gives the complete sampling model and numerical protocol.

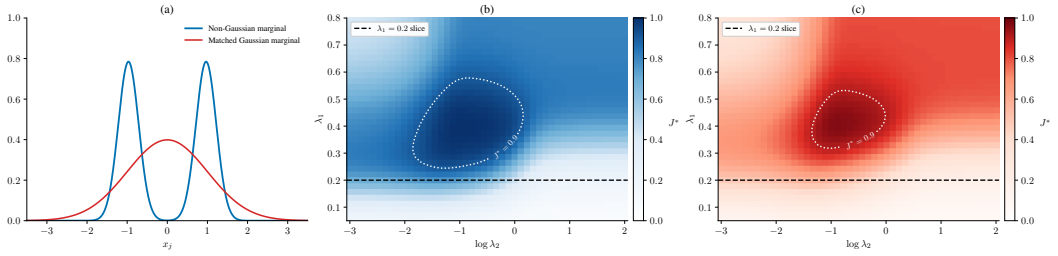

Figure 1: Feature marginals and predicted support recovery. (a) Exact block-coordinate marginal densities for the non-Gaussian design (blue solid) and matched Gaussian design (red solid); both have mean zero and variance one, and the full designs share the same covariance. (b,c) Surrogate-predicted Jaccard J^* on the 36×46 penalty grid for the non-Gaussian and matched Gaussian designs, respectively. White dotted contours mark $J^* = 0.9$, and horizontal dashed lines mark the $\lambda_1 = 0.2$ slice.

To measure recovery against the teacher, we introduce Jaccard similarity. For finite sets $\mathcal{A}, \mathcal{B}$ with nonempty union, $J(\mathcal{A}, \mathcal{B}) := \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|}$. Let $S_0 = \mathrm{supp}(\theta_0)$, introduce its predicted surrogate Jaccard by

$$J^* = \mathbb{E}_G[J(\mathrm{supp}(W_*(G)), S_0)]. \tag{29}$$

Figure 1(b,c) shows that moderate regularization improves support recovery, while the close agreement between surrogate and empirical Jaccard—with maximum grid discrepancies of 0.74 and 0.72 percentage points for the non-Gaussian and matched Gaussian designs, respectively—supports the accuracy of the surrogate's recovery predictions in this experiment.

Index the sets of active block coordinates and active singleton coordinates by $S_{0,b}$ and $S_{0,i}$ respectively. Figure 2(a) compares the non-Gaussian and covariance-matched Gaussian designs along the $\lambda_1 = 0.2$ slice. In both designs, increasing λ_2 initially improves block-signal recovery and overall Jaccard similarity; at larger λ_2, declining singleton-signal recovery and increasing false selections reduce Jaccard despite high block-signal recovery.

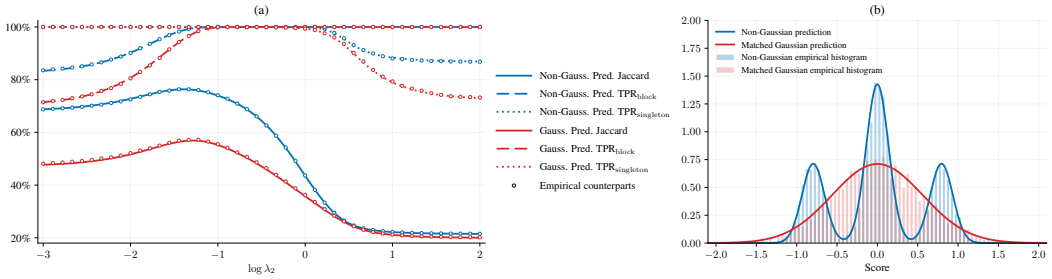

Figure 2: Support recovery and fresh-score distributions for the non-Gaussian (blue) and covariance-matched Gaussian (red) designs. (a) Support recovery along the $\lambda_1 = 0.2$ slice. For any random support S, the true-positive rate for block signals $\mathrm{TPR}_{\mathrm{block}} = \frac{|S \cap S_{0,b}|}{|S_{0,b}|}$, and that for singleton signal $\mathrm{TPR}_{\mathrm{singleton}} = \frac{|S \cap S_{0,i}|}{|S_{0,i}|}$. With $S = \mathrm{supp}(W_*(G))$, the surrogate mean Jaccard is $J^* = \mathbb{E}_G \left[\frac{|S_{0,b}|\,\mathrm{TPR}_{\mathrm{block}} + |S_{0,i}|\,\mathrm{TPR}_{\mathrm{singleton}}}{|S_0| + |S \setminus S_0|} \right]$. Lines show surrogate expectations: solid for Jaccard, long dashes for $\mathrm{TPR}_{\mathrm{block}}$, and dotted for $\mathrm{TPR}_{\mathrm{singleton}}$. "Pred." denotes surrogate predictions; "Gauss." denotes the covariance-matched Gaussian design. Open circles show empirical means of the corresponding metrics over independent training datasets per design. (b) Fresh-score densities at $(\lambda_1, \lambda_2) = (0.2, 0.1)$. Solid curves show the predicted laws of $x^\top \mu_* + \alpha_* Z$ under each design's own calibration, where $Z \sim N(0,1)$ is independent of x. The blue and red histograms show empirical scores $x^\top \widehat{W}/\sqrt{n}$ for the non-Gaussian and matched Gaussian designs. Writing $\mu_* = m_{b*} \mathbf{1}_{S_{0,b}} + m_{i*} \mathbf{1}_{S_{0,i}}$, where $\mathbf{1}_S$ is the coordinate-indicator vector of S, the calibrated values $(m_{b*}, m_{i*}, \alpha_*)$ are approximately $(0.04111, 0.04924, 0.02859)$ for the non-Gaussian design and $(0.03961, 0.05192, 0.04698)$ for the matched Gaussian design.

These experiments provide strong numerical evidence for the surrogate's support-recovery statistics predictions in the tested setting. The substantial recovery differences between non-Gaussian and covariance-matched Gaussian designs highlight the importance of retaining the actual feature–response law in the calibration. Establishing theoretical transfer of support statistics from the surrogate to ERM remains future work.

Figure 2(b) examines fresh-score distributions for Non-Gaussian and Gaussian designs. Corollary 3 gives the corresponding asymptotic prediction law under its assumptions: $x^\top \mu_* + \alpha_* Z$, with $Z \sim N(0,1)$ independent of x. The empirical histograms for both designs closely follow their respective predicted densities. Although the designs have the same feature covariance, their separately calibrated score laws differ, with an unstandardized W_1 distance of approximately 0.10994.

4 Hamilton–Jacobi mechanism

4.1 Pseudo-equation

Write the expected normalized minimum of the objective formula (14) as

$$\phi(t, v, b, h) := \frac{1}{n} \mathbb{E} \left[\min_w \mathcal{E}_{t,v,b,h}(w) \right]. \tag{30}$$

The parameter values $(t, v, b, h) = (1, 0, 0, 0)$ give the empirical problem, while $(0, \nu_*, \beta_*^2/2, \eta_*)$ give the calibrated surrogate. The expected minimum ϕ records the prediction quantities through the envelope identities

$$\mu = -\nabla_h \phi, \qquad \alpha^2 = 2\partial_v \phi - \mu^\top \Sigma_x \mu, \qquad \kappa = -\partial_b \phi. \tag{31}$$

Thus the deterministic linear, quadratic, and Gaussian perturbations reveal the mean, prediction variance, and response. Positive probes give the required curvature without differentiating the penalty; the identities are proved in Lemma B.5. We keep $v, b > 0$ during the comparison: $v = 0, b > 0$ can destroy Lasso coercivity.

Let W_- be the optimizer with one observation omitted, and let Δ be the increase in the minimum objective when that independent observation is reinserted. Lemma B.6 gives $\partial_t \phi = \mathbb{E}[\Delta]$. Proposition C.4, using Lemmas C.1 and C.2, replaces Δ by the envelope $e_{L_y}(u; \kappa_-)$ from (3), where $u := x^\top W_- / \sqrt{n}$. Proposition C.5 then applies (21) in Conjecture 1 to replace the fresh fluctuation by a Gaussian with deterministic averaged variance. Proposition C.6 replaces the cavity moments by full-sample moments. The actual feature–label law remains in the resulting Hamiltonian:

$$\mathcal{H}(\mu, \alpha, \kappa) := \mathbb{E}_{x,y,Z}\left[e_{L_y}(x^\top \mu + \alpha Z; \kappa)\right], \tag{32}$$

$$\widetilde{\mathcal{H}}(\delta) := \mathcal{H}\left(-\delta_h, \sqrt{2\delta_v - \delta_h^\top \Sigma_x \delta_h}, -\delta_b\right). \tag{33}$$

Here $\delta := (\delta_v, \delta_b, \delta_h)$ is a spatial gradient; $\widetilde{\mathcal{H}}$ is evaluated on its nonnegative-variance, nonnegative-response domain.

Theorem 5 (Asymptotic interior value equation)**.** *Along a sequence in which the training pairs are independent copies of (x, y), suppose $p/n \to c \in (0, \infty)$, $\underline{c}_x I_p \preceq \Sigma_x \preceq K I_p$, and Assumptions 1 and 2 hold. Suppose also that* (21) *in Conjecture 1 holds for full and unforced cavity problems. On every fixed interior domain $\mathcal{D}$ introduced in Section 2.3,*

$$\sup_{\mathcal{D}} \left|\partial_t \phi - \widetilde{\mathcal{H}}(\nabla_{v,b,h} \phi)\right| \longrightarrow 0. \tag{34}$$

Convergence in (34) *does not use terminal quadratic growth or a population calibration.*

4.2 Choice of the correct path that preserves statistics

Write $\nabla := \nabla_{v,b,h}$. By (31), fixed $\nabla\phi$ means fixed (μ, α^2, κ). The gradient $\nabla_\delta \widetilde{\mathcal{H}}$ is taken with respect to $\delta = (\delta_v, \delta_b, \delta_h)$; evaluated at $\delta = \nabla\phi(t, v, b, h)$, Lemma B.8 gives

$$\nabla_\delta \widetilde{\mathcal{H}}(\nabla\phi(t, v, b, h)) = (\nu, \beta^2/2, \eta), \qquad \nu\alpha = \mathbb{E}[Z\xi]. \tag{35}$$

These coefficients use the current moments in (4), before identification with the starred population solution. First suppose ϕ is smooth and satisfies the exact equation $\partial_t \phi = \widetilde{\mathcal{H}}(\nabla\phi)$. Along a path, the chain rule gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} \nabla\phi(t, v(t), b(t), h(t)) &= \nabla\widetilde{\mathcal{H}}(\nabla\phi) + \nabla^2\phi\,(\dot{v}, \dot{b}, \dot{h}) \\ &= \nabla^2\phi\left(\nabla_\delta \widetilde{\mathcal{H}}(\nabla\phi) + (\dot{v}, \dot{b}, \dot{h})\right).\end{aligned}$$

By (35), choosing $(\dot{v}, \dot{b}, \dot{h}) = -(\nu, \beta^2/2, \eta)$ cancels the two terms, so $\mathrm{d}(\nabla\phi)/\mathrm{d}t = 0$. Increasing t adds observations while reducing the probes and preserving the moments.

5 Final discussion

Two directions remain for future work. First, proving Conjecture 1 for nonsmooth penalties would remove the conjectural input to the prediction comparisons, under the remaining stated assumptions. Second, the numerical agreement of support statistics calls for a rigorous transfer from the surrogate to ERM. Prediction moments and fresh-score laws alone do not establish this: support selection is discontinuous at zero. A possible route is to combine finer coordinate comparisons with control of selection margins near the penalty threshold. Such results would clarify when the surrogate's accurate recovery predictions extend beyond the tested setting.

CODE AVAILABILITY.

The code used to reproduce the numerical experiments is available at `https://anonymous.4open.science/r/ERM_non_smooth_penalty-2609/`.

AI USE STATEMENT

Generative AI tools assisted with mathematical development, literature search, drafting including proof writing, and code and verification checks. The authors are responsible for reviewing the mathematical claims, proofs and cited sources, and for the final manuscript.

REPRODUCIBILITY STATEMENT

The appendices prove the stated asymptotic comparisons using convex response calculus, value comparisons, calibrated transport and cutoff coupling. Theorem 2 states its feature, conditional prediction, population and branch-specific penalty conditions explicitly. Lemma E.4 proves localization; Propositions E.2 and E.6 prove cutoff mean-square comparison. The source package includes compilation instructions, a proof-status ledger, a numerical protocol, and executable checks of the diagonal proximal formulas. For the support-recovery experiment, Appendix H specifies the sampling law, penalty grid, replicate ERM protocol, and numerical fixed-point validation.

REFERENCES

Mohsen Bayati and Andrea Montanari. The LASSO risk for Gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2012. doi: 10.1109/TIT.2011.2174612. URL `https://arxiv.org/abs/1008.2581`.

David Bosch, Ashkan Panahi, Ayca Ozcelikkale, and Devdatt Dubhashi. Random features model with general convex regularization: A fine grained analysis with precise asymptotic learning curves. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 11371–11414, 2023. URL `https://proceedings.mlr.press/v206/bosch23a.html`.

Matthew Esmaili Mallory, Kevin Han Huang, and Morgane Austern. Universality of high-dimensional logistic regression and a novel CGMT under dependence with applications to data augmentation. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 1799–1918, 2025. URL `https://proceedings.mlr.press/v291/esmaili-mallory25a.html`.

Qiyang Han and Yandi Shen. Universality of regularized regression estimators in high dimensions. *The Annals of Statistics*, 51(4):1799–1823, 2023. doi: 10.1214/23-AOS2309. URL `https://arxiv.org/abs/2206.07936`.

Samriddha Lahiry and Pragya Sur. Universality in block dependent linear models with applications to nonparametric regression. arXiv:2401.00344, 2024. URL `https://arxiv.org/abs/2401.00344`.

A. S. Lewis and S. Zhang. Partial smoothness, tilt stability, and generalized hessians. *SIAM Journal on Optimization*, 23(1):74–94, 2013. doi: 10.1137/110852103. URL `https://epubs.siam.org/doi/10.1137/110852103`.

A. S. Lewis, Jingwei Liang, and Tonghua Tian. Partial smoothness and constant rank. *SIAM Journal on Optimization*, 32(1):276–291, 2022. doi: 10.1137/19M1237909. URL `https://epubs.siam.org/doi/10.1137/19M1237909`.

Cosme Louart. A central limit theorem for regularized M-estimators, 2026. URL `https://arxiv.org/abs/2608.15034v1`.

Bruno Loureiro, Cédric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. In *Advances in Neural Information Processing Systems*, volume 34, 2021a. URL https://proceedings.neurips.cc/paper/2021/hash/9704a4fc48ae88598dcbdcdf57f3fdef-Abstract.html.

Andrea Montanari and Basil N. Saeed. Universality of empirical risk minimization. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4310–4312, 2022. URL https://proceedings.mlr.press/v178/montanari22a.html. Full version: https://arxiv.org/abs/2202.08832v2.

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized M-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018. doi: 10.1109/TIT.2018.2840720. URL https://arxiv.org/abs/1601.06233.

Ryan J. Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012. doi: 10.1214/12-AOS1003. URL https://arxiv.org/abs/1111.0653.

Chiheb Yaakoubi, Cosme Louart, Malik Tiomoko, and Zhenyu Liao. Characterization of Gaussian universality breakdown in high-dimensional empirical risk minimization. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2604.03146. Public manuscript: arXiv:2604.03146.

Pierre C. Bellec and Cun-Hui Zhang. Second order Stein: SURE for SURE and other applications in high-dimensional inference. Preprint, 2018. URL https://arxiv.org/abs/1811.04121.

Pierre C. Bellec and Takuya Koriyama. Existence of solutions to the nonlinear equations characterizing the precise error of M-estimators. Preprint, 2023. URL https://arxiv.org/abs/2312.13254.

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.

Michael Celentano, Andrea Montanari, and Yuting Wei. The Lasso with general Gaussian designs with applications to hypothesis testing. *The Annals of Statistics*, 51(5):2194–2220, 2023. doi: 10.1214/23-AOS2327. URL https://arxiv.org/abs/2007.13716.

Luca Pesce, Florent Krzakala, Bruno Loureiro, and Ludovic Stephan. Are Gaussian data all you need? The extents and limits of universality in high-dimensional generalized linear estimation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 27680–27708. PMLR, 2023. URL https://proceedings.mlr.press/v202/pesce23a.html.

Xiaoyi Mai and Zhenyu Liao. The breakdown of Gaussian universality in classification of high-dimensional linear factor mixtures. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/154743e7e9688cf77db5ee75807bda82-Abstract-Conference.html.

Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. arXiv:1312.3580, 2013. URL https://arxiv.org/abs/1312.3580.

A Population well-posedness and coefficient bounds

Theorem 1 follows from one convex program. For existence, we prove attainment, use its Karush–Kuhn–Tucker (KKT) conditions to construct the population coefficients, and verify their required positivity. A feasible midpoint and strict convexity of the scalar loss then give uniqueness. We subsequently verify the Moreau characterization in Lemma A.1, record the moment identities in Lemma A.3, and prove the coefficient bounds in Proposition A.4. Throughout, $\rho : \mathbb{R}^p \to \mathbb{R}$ is finite and convex, $\Sigma_x = \mathbb{E}\left[xx^\top\right] \succ 0$, and the probability spaces of $G \sim N(0, \Sigma_x)$ and (x, y, Z) are separate, with Z independent of (x, y). The centered loss bound permits the proof without assuming integrable $L_y(0)$: $|L_y(\zeta) - L_y(0)| \le L_1|\zeta|$.

A.1 Existence

Proof of existence in Theorem 1. Fix n, p. The variables in the auxiliary minimization problem are exactly

$$(W, \zeta, \alpha) \in L^2(G; \mathbb{R}^p) \times L^2(x, y, Z) \times \mathbb{R}.$$

Thus W is a measurable function of G, whereas ζ is a measurable function of (x, y, Z). For each W, define

$$\mu := \frac{\mathbb{E}_G[W]}{\sqrt{n}}, \qquad C_W := \mathbb{E}_G\left[(W - \mathbb{E}_G[W])(W - \mathbb{E}_G[W])^\top\right].$$

In particular, μ and C_W are determined by W and are not additional variables. The variance constraint in (36) forces $\alpha \ge 0$. Consider

$$\begin{aligned} \text{minimize} \quad & \frac{1}{n}\mathbb{E}_G[\rho(W) - \rho(0)] + \mathbb{E}[L_y(\zeta) - L_y(0)] \\ \text{subject to} \quad & \sqrt{\operatorname{tr}(\Sigma_x C_W)/n} \le \alpha, \\ & \|x^\top\mu + \alpha Z - \zeta\|_{L^2} + \frac{1}{n}\mathbb{E}_G\left[G^\top W\right] \le 0. \end{aligned} \tag{36}$$

The constraints are convex and continuous; $(W, \zeta, \alpha) = (0, 0, 0)$ is feasible with value zero. This Hilbert-space formulation is related to Bellec and Koriyama (2023). The coefficients β, ν, κ and η are not primal variables of (36). We first prove that the program attains its minimum. Slater's condition then supplies one nonnegative Lagrange multiplier for each constraint. After proving the quantities needed for division are positive, we rescale the multipliers and use the KKT stationarity and complementary-slackness conditions to recover (4)–(5).

Attainment. Uniqueness of the penalty minimum gives $c_\rho := \min_{|w|=1}\{\rho(w) - \rho(0)\} > 0$ and $\rho(w) - \rho(0) \ge c_\rho|w| - c_\rho$. On a feasible objective sublevel of height M, nonnegative loss therefore gives

$$\mathbb{E}_G[|W|]/\sqrt{n} + |\mu| \le C_M, \qquad \|W\|_{L^2}/\sqrt{n} \le C_M + \alpha/\sqrt{\lambda_{\min}(\Sigma_x)}.$$

The only remaining issue is to bound α. Define $d_L := \mathbb{E}\left[L_y'(1) - L_y'(-1)\right]/2 > 0$ and $\varepsilon_R := \|(G/\sqrt{n})\mathbf{1}_{\{|G|/\sqrt{n} > R\}}\|_{L^2} \to 0$. Symmetry of Z, convexity, and Gaussian truncation give

$$\begin{aligned} \mathbb{E}[L_y(\alpha Z)] &\ge \mathbb{E}[L_y(0)] + d_L(\alpha\mathbb{E}[|Z|] - 1), \\ -\frac{1}{n}\mathbb{E}_G\left[G^\top W\right] &\le RC_M + \varepsilon_R(C_M + \alpha/\sqrt{\lambda_{\min}(\Sigma_x)}). \end{aligned}$$

By feasibility and the L_1-Lipschitz loss,

$$\mathbb{E}[L_y(\zeta)] \ge \left(d_L\mathbb{E}[|Z|] - \frac{L_1\varepsilon_R}{\sqrt{\lambda_{\min}(\Sigma_x)}}\right)\alpha - C_{M,R}.$$

Choose R so that the coefficient is positive. This bounds α, then $\|W\|_{L^2}$ and, by the residual constraint in (36), $\|\zeta\|_{L^2}$. All bounds are at fixed size. The feasible set is weakly closed; Fatou and convexity make the penalty integral weakly lower semicontinuous, and the centered loss integral is convex and Lipschitz on L^2. Weak compactness therefore gives a minimizer.

KKT multipliers and positivity. Let $\mathcal{F}(W, \zeta, \alpha)$ denote the objective in (36), and write its two constraint functions as

$$c_{\mathrm{var}}(W, \alpha) := \sqrt{\mathrm{tr}(\Sigma_x C_W)/n} - \alpha, \qquad c_{\mathrm{res}}(W, \zeta, \alpha) := \|x^\top \mu + \alpha Z - \zeta\|_{L^2} + \frac{1}{n}\mathbb{E}_G\left[G^\top W\right].$$

Slater's condition (a strictly feasible point with finite objective) holds for $W := -G/(1 + |G|)$, $\alpha > \sqrt{\|\Sigma_x\|_{\mathrm{op}}/n}$ and $\zeta := \alpha Z$: $\mu = 0$ and $\mathbb{E}_G\left[G^\top W\right] < 0$. Thus the KKT theorem supplies a multiplier $\tau_\nu \geq 0$ for $c_{\mathrm{var}} \leq 0$ and a multiplier $\beta \geq 0$ for $c_{\mathrm{res}} \leq 0$. Concretely, the minimizing triple is stationary for the Lagrangian

$$\begin{aligned} \mathcal{K}(W, \zeta, \alpha; \tau_\nu, \beta) &:= \mathcal{F}(W, \zeta, \alpha) + \tau_\nu c_{\mathrm{var}}(W, \alpha) + \beta c_{\mathrm{res}}(W, \zeta, \alpha), \\ 0 &\in \partial_{(W,\zeta,\alpha)} \mathcal{K}(W, \zeta, \alpha; \tau_\nu, \beta). \end{aligned} \tag{37}$$

Complementary slackness means

$$\tau_\nu c_{\mathrm{var}}(W, \alpha) = 0, \qquad \beta c_{\mathrm{res}}(W, \zeta, \alpha) = 0. \tag{38}$$

At a minimizer, set $U := x^\top \mu + \alpha Z$, $\tau_\kappa := \|U - \zeta\|_{L^2}$ and $\xi := L_y'(\zeta)$. Notice that τ_κ is the residual norm, not a Lagrange multiplier. Choose $s_{\mathrm{res}} \in \partial\|\cdot\|_{L^2}(U - \zeta)$; then $\|s_{\mathrm{res}}\|_{L^2} \leq 1$, with $s_{\mathrm{res}} = (U - \zeta)/\tau_\kappa$ when $\tau_\kappa > 0$. The ζ- and α-stationarity conditions for (37) are, respectively,

$$\xi = \beta s_{\mathrm{res}}, \qquad \tau_\nu = \beta\mathbb{E}\left[Z s_{\mathrm{res}}\right] = \mathbb{E}\left[Z\xi\right]. \tag{39}$$

In particular, $\|\xi\|_{L^2} \leq \beta$, and if $\tau_\kappa > 0$, then $\xi = \beta(U - \zeta)/\tau_\kappa$. Equation (38) gives $c_{\mathrm{res}}(W, \zeta, \alpha) = 0$ if $\beta > 0$ and $c_{\mathrm{var}}(W, \alpha) = 0$ if $\tau_\nu > 0$. When $\tau_\nu = 0$, stationarity in W is

$$-\sqrt{n}\,\mathbb{E}\left[x\xi\right] - \beta G \in \partial\rho(W). \tag{40}$$

When $\tau_\nu, \alpha > 0$, it is

$$-\frac{\tau_\nu}{\alpha}\Sigma_x(W - \mathbb{E}_G\left[W\right]) - \sqrt{n}\,\mathbb{E}\left[x\xi\right] - \beta G \in \partial\rho(W). \tag{41}$$

The convex sum rule applies because the constraints are continuous; testing its integral subgradient inequality on measurable subsets gives these pointwise inclusions.

Positivity follows before any division. If $\beta = 0$, then $\xi = \tau_\nu = 0$, $W = 0$, and feasibility gives $\zeta = \alpha Z$, contradicting $\|L_y'(\alpha Z)\|_{L^2} > 0$ by strict increase when $\alpha > 0$ and (6) when $\alpha = 0$. If $\alpha = 0$, then $W = \sqrt{n}\mu$, $\zeta = x^\top \mu$ and $\tau_\nu = 0$. Equation (40) is impossible: $\partial\rho(\sqrt{n}\mu)$ is bounded because ρ is finite, whereas βG is nondegenerate Gaussian. Finally, if $\tau_\kappa = 0$, Gaussian integration by parts yields $\tau_\nu = \alpha\mathbb{E}\left[L_y''(U)\right] > 0$ and $\mathbb{E}_G\left[G^\top W\right] = 0$. Jensen and (41) contradict this:

$$0 \leq \frac{\mathbb{E}_G\left[\rho(W)\right] - \rho(\mathbb{E}_G\left[W\right])}{n} \leq -\frac{\tau_\nu}{\alpha}\frac{\mathrm{tr}(\Sigma_x C_W)}{n} - \frac{\beta}{n}\mathbb{E}_G\left[G^\top W\right] = -\tau_\nu\alpha < 0.$$

Identification. Now set $\kappa := \tau_\kappa/\beta$. Equation (39) and scalar Gaussian integration by parts give

$$\zeta = \mathrm{prox}_{\kappa L_y}(U), \qquad \mathbb{E}\left[\xi^2\right] = \beta^2, \qquad \tau_\nu = \alpha\mathbb{E}\left[\frac{L_y''(\zeta)}{1 + \kappa L_y''(\zeta)}\right] = \alpha\nu > 0.$$

The equalities $c_{\mathrm{var}}(W, \alpha) = c_{\mathrm{res}}(W, \zeta, \alpha) = 0$ give, respectively, the variance and Gaussian-correlation equations in (5). With $\eta := \nu\Sigma_x\mu - \mathbb{E}\left[x\xi\right]$, (41) identifies the unique optimizer in (5a). All coefficients are finite. Comparing with zero gives

$$\frac{1}{n}\mathbb{E}_G\left[\rho(W) - \rho(0)\right] + \mathbb{E}\left[L_y(\zeta)\right] \leq \mathbb{E}\left[L_y(0)\right]. \tag{42}$$

□

A.2 Uniqueness

Proof of uniqueness in Theorem 1. Take solutions $(\mu, \alpha, \kappa, \nu, \beta, \eta)$ and $(\mu', \alpha', \kappa', \nu', \beta', \eta')$; primes refer to the second solution, coupled through the same G and (x, y, Z). We show that both minimize (36). Their feasible midpoint then identifies ζ; (4)–(5) then identify the remaining coefficients. The

uniqueness proof uses only $\mathbb{E}\left[|x|^2\right] < \infty$, $\Sigma_x \succ 0$, a finite convex ρ with $0 \in \partial\rho(0)$, and the C^2 loss bounds $|L'_y| \le L_1$ and $0 < L''_y \le L_2$ from Theorem 1.

For either solution, the surrogate inclusion gives

$$\nu\lambda_{\min}(\Sigma_x)|W| \le |\sqrt{n}\,\eta - \beta G|, \qquad 0 \le \rho(W) - \rho(0) \le (\sqrt{n}\,\eta - \nu\Sigma_x W - \beta G)^\top W.$$

Thus $W \in L^2$ and $\rho(W) \in L^1$. If $\alpha = 0$, the variance identity in (5) makes W deterministic. The surrogate optimality inclusion would then require the bounded set $\partial\rho(W)$ of the finite convex function ρ to contain the nondegenerate Gaussian term βG, which is impossible. Jensen gives $0 \le \{\mathbb{E}_G[\rho(W)] - \rho(\mathbb{E}_G[W])\}/n \le -\nu\alpha^2 + \beta^2\kappa$, so $\kappa > 0$. Put $\tau_\kappa := \beta\kappa$ and $\tau_\nu := \alpha\nu$. Both constraints in (36) are tight, $\tau_\kappa = \|U - \zeta\|_{L^2}$, $\xi = \beta(U-\zeta)/\tau_\kappa$, $\tau_\nu = \mathbb{E}[Z\xi]$, and

$$0 \in \partial\rho(W) + \frac{\tau_\nu}{\alpha}\Sigma_x(W - \mathbb{E}_G[W]) + \sqrt{n}\,\mathbb{E}[x\xi] + \beta G. \tag{43}$$

For any feasible $(\widetilde{W}, \widetilde{\alpha}, \widetilde{\zeta})$ of finite objective value, put $\widetilde{\mu} := \mathbb{E}_G\left[\widetilde{W}\right]/\sqrt{n}$. The convex subgradient inequalities give explicitly

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_G\left[\rho(\widetilde{W}) - \rho(W)\right] + \mathbb{E}\left[L_y(\widetilde{\zeta}) - L_y(\zeta)\right] \\
&\quad \ge -\tau_\nu\left(\sqrt{\mathrm{tr}(\Sigma_x C_{\widetilde{W}})/n} - \widetilde{\alpha}\right) \\
&\qquad - \beta\left(\|x^\top\widetilde{\mu} + \widetilde{\alpha}Z - \widetilde{\zeta}\|_{L^2} + \frac{1}{n}\mathbb{E}_G\left[G^\top\widetilde{W}\right]\right) \ge 0.
\end{aligned}$$

Hence both solutions minimize (36). At their feasible midpoint, the sum of the two nonnegative Jensen gaps is zero; in particular,

$$\mathbb{E}\left[\frac{L_y(\zeta) + L_y(\zeta')}{2} - L_y\left(\frac{\zeta+\zeta'}{2}\right)\right] = 0 \quad\Longrightarrow\quad \zeta = \zeta', \quad \xi = \xi', \quad \beta = \beta'. \tag{44}$$

The Jensen gap in (44) is integrable without an assumption on $L_y(0)$. Subtracting $U - \zeta = \kappa\xi$, which follows from (4), for the two solutions gives

$$x^\top(\mu - \mu') + (\alpha - \alpha')Z = (\kappa - \kappa')\xi.$$

Conditionally on almost every (x, y), continuity extends this identity to all Z. Boundedness of ξ forces $\alpha = \alpha'$; its strict increase in Z, since $\alpha > 0$, then forces $\kappa = \kappa'$. Finally $\Sigma_x \succ 0$ gives $\mu = \mu'$, (4) gives $\nu = \nu'$ and $\eta = \eta'$, and strict convexity in (5a) gives $W = W'$ almost surely. □

A.3 Moreau variational formulation

Under the existence assumptions of Theorem 1, for $\alpha, \kappa, \beta, \nu > 0$ and $\mu, \eta \in \mathbb{R}^p$, define

$$\begin{aligned}
\mathcal{J}(\mu, \alpha, \kappa; \beta, \nu, \eta) := \frac{\beta^2\kappa}{2} + \mathbb{E}_{x,y,Z}\left[e_{L_y}(x^\top\mu + \alpha Z; \kappa)\right] - \frac{\nu\alpha^2}{2} \\
+ \frac{1}{n}\mathbb{E}_G\left[e_{\rho\circ\Sigma_x^{-1/2}}\left(\frac{\Sigma_x^{-1/2}(\sqrt{n}\,\eta - \beta G)}{\nu}; \frac{1}{\nu}\right)\right] \\
- \frac{|\Sigma_x^{-1/2}\eta - \nu\Sigma_x^{1/2}\mu|^2}{2\nu} - \frac{\beta^2 p}{2n\nu}.
\end{aligned} \tag{45}$$

Here $(\rho \circ \Sigma_x^{-1/2})(z) := \rho(\Sigma_x^{-1/2}z)$, with symmetric positive-definite matrix powers. The population solution is the stationary point of this objective and realizes

$$\inf_{\substack{\mu\in\mathbb{R}^p \\ \alpha>0}} \sup_{\substack{\beta,\nu>0 \\ \eta\in\mathbb{R}^p}} \inf_{\kappa>0} \mathcal{J}(\mu, \alpha, \kappa; \beta, \nu, \eta). \tag{46}$$

Lemma A.1 proves this characterization by maximizing over η and rescaling two scalars to obtain a convex–concave problem. Its first-order equations are (4)–(5). Expected loss and penalty envelopes also enter the Gaussian linear-design analysis of Thrampoulidis et al. (2018).

In Lemma A.1 and its proof, $\mathcal{J}$ and $\widetilde{\mathcal{J}}$ use the centered loss $e_{L_y}(u;\kappa) - L_y(0)$, whose absolute value is at most $L_1|u| + \kappa L_1^2/2$. Restoring $\mathbb{E}[L_y(0)]$ when finite recovers the value of (45).

Lemma A.1 (Variational representation of the population equations)**.** *Under the uniqueness assumptions of Theorem 1, the positive stationary points of $\mathcal{J}$ are exactly the population solutions. Define*

$$\tau_\kappa := \beta\kappa, \qquad \tau_\nu := \alpha\nu,$$
$$\widetilde{\mathcal{J}}(\mu, \alpha, \tau_\kappa; \beta, \tau_\nu) := \sup_{\eta \in \mathbb{R}^p} \mathcal{J}\left(\mu, \alpha, \frac{\tau_\kappa}{\beta}; \beta, \frac{\tau_\nu}{\alpha}, \eta\right). \tag{47}$$

For positive scalar arguments, $\widetilde{\mathcal{J}}$ is finite, convex in $(\mu, \alpha, \tau_\kappa)$ and concave in (β, τ_ν). Its positive saddles correspond exactly to population solutions. Under the existence assumptions of Theorem 1, the solution also realizes (46)*.*

Proof. Maximizing over η prescribes the surrogate mean. The resulting quadratic perspectives give convexity and concavity; their lifted stationarity equations identify the saddle. We finally verify the specific nesting in (46).

Completing the square in the penalty envelope gives

$$\begin{aligned}\mathcal{J} = \frac{\beta^2\kappa}{2} + \mathbb{E}\left[e_{L_y}(x^\top\mu + \alpha Z; \kappa) - L_y(0)\right] - \frac{\nu}{2}(\alpha^2 + \mu^\top\Sigma_x\mu) + \eta^\top\mu \\ + \frac{1}{n}\mathbb{E}_G\left[\min_w \left\{\rho(w) + \frac{\nu}{2}w^\top\Sigma_x w + (\beta G - \sqrt{n}\,\eta)^\top w\right\}\right].\end{aligned} \tag{48}$$

The bound $\nu\lambda_{\min}(\Sigma_x)|W| \le |\sqrt{n}\,\eta - \beta G|$ justifies envelope differentiation. The derivatives in η, ν, β give the mean, variance and Gaussian-correlation equations in (5); those in κ, α, μ give (4), using $\mathbb{E}[Z\xi] = \alpha\nu$ from scalar Gaussian integration by parts.

For fixed $\nu, \beta > 0$, minimize $\mathbb{E}_G\left[\rho(W) + \nu W^\top\Sigma_x W/2 + \beta G^\top W\right]$ with prescribed mean $\mathbb{E}_G[W] = \sqrt{n}\mu$. The constant choice of W gives finiteness from above; the positive quadratic term gives a finite lower bound and bounded L^2 sublevels, hence weak attainment. The value is finite and convex in the prescribed mean. A supporting hyperplane at that mean supplies a field $\sqrt{n}\,\eta$ for which the constrained minimizer also minimizes the unconstrained integral minus $\sqrt{n}\,\eta^\top\mathbb{E}_G[W]$. Pointwise minimization identifies (5a); the bound $\nu\lambda_{\min}(\Sigma_x)|W| \le |\sqrt{n}\,\eta - \beta G|$ and the surrogate subgradient inequality give $W \in L^2$ and integrable $\rho(W)$. Thus the field supremum is attained. After this maximization, the rescaling in (47) gives

$$\begin{aligned}\widetilde{\mathcal{J}} = \frac{\beta\tau_\kappa - \tau_\nu\alpha}{2} + \mathbb{E}\left[e_{L_y}\left(x^\top\mu + \alpha Z; \frac{\tau_\kappa}{\beta}\right) - L_y(0)\right] \\ + \min_{\substack{W \in L^2(G;\mathbb{R}^p) \\ \mathbb{E}_G[W] = \sqrt{n}\mu}} \frac{1}{n}\mathbb{E}_G\left[\rho(W) + \frac{\tau_\nu}{2\alpha}(W - \sqrt{n}\mu)^\top\Sigma_x(W - \sqrt{n}\mu) + \beta G^\top W\right].\end{aligned} \tag{49}$$

Expand the loss envelope as a minimum over ζ with quadratic term $\beta(x^\top\mu + \alpha Z - \zeta)^2/(2\tau_\kappa)$. Both quadratic-over-positive-scalar terms are convex in the minimizing variables, and the mean constraint is affine. Partial minimization gives convexity. Before minimization over W and ζ, this lifted form of (49) is affine in (β, τ_ν); partial minimization therefore gives concavity in these two variables.

At a root, substitute $\mu = \mathbb{E}_G[W]/\sqrt{n}$ and minimize jointly in $W, \zeta, \alpha, \tau_\kappa$. The stationary equations are (43), $\xi = \beta(U - \zeta)/\tau_\kappa$, $\tau_\nu = \mathbb{E}[Z\xi]$, and $\tau_\kappa^2 = \|U - \zeta\|_{L^2}^2$. Thus convexity gives, for all positive $(\widetilde{\mu}, \widetilde{\alpha}, \widetilde{\tau}_\kappa)$,

$$\widetilde{\mathcal{J}}(\mu, \alpha, \tau_\kappa; \beta, \tau_\nu) \le \widetilde{\mathcal{J}}(\widetilde{\mu}, \widetilde{\alpha}, \widetilde{\tau}_\kappa; \beta, \tau_\nu). \tag{50}$$

At the root's minimizing W, ζ, the coefficients of both dual variables vanish:

$$\frac{\mathrm{tr}(\Sigma_x C_W)}{2n\alpha} - \frac{\alpha}{2} = 0, \qquad \frac{\tau_\kappa}{2} + \frac{\|U - \zeta\|_{L^2}^2}{2\tau_\kappa} + \frac{1}{n}\mathbb{E}_G\left[G^\top W\right] = 0.$$

Changing the dual variables leaves this value unchanged; minimizing in W, ζ can only decrease it. Hence, for all positive $(\widetilde{\beta}, \widetilde{\tau}_\nu)$,

$$\widetilde{\mathcal{J}}(\mu, \alpha, \tau_\kappa; \widetilde{\beta}, \widetilde{\tau}_\nu) \le \widetilde{\mathcal{J}}(\mu, \alpha, \tau_\kappa; \beta, \tau_\nu). \tag{51}$$

Conversely, at a positive saddle the unique W, ζ minima and their local L^2 bounds justify envelope differentiation. It gives

$$\mathbb{E}\left[\xi^2\right] = \beta^2, \quad \operatorname{tr}(\Sigma_x C_W)/n = \alpha^2, \quad \mathbb{E}_G\left[G^\top W\right]/n = -\tau_\kappa, \quad \mathbb{E}\left[Z\xi\right] = \tau_\nu. \tag{52}$$

The lifted W stationarity is (43): after substituting the mean, all terms except the penalty integral are continuously Fréchet differentiable on the product Hilbert space, so the convex sum rule applies without global finiteness of that integral on L^2. With $\kappa := \tau_\kappa/\beta$, $\nu := \tau_\nu/\alpha$ and $\eta := \nu\Sigma_x\mu - \mathbb{E}\left[x\xi\right]$, (52) gives (4)–(5).

Under the existence assumptions of Theorem 1, restore $\mathbb{E}\left[L_y(0)\right]$ in both objectives and write $\mathcal{J}_*$ for the root value. The η and κ terms are additively separated, so their finite supremum and infimum commute. Inequalities (50) and (51) imply

$$\sup_{\beta,\nu>0,\,\eta} \inf_{\kappa>0} \mathcal{J}(\mu,\alpha,\kappa;\beta,\nu,\eta) \geq \inf_{\tau_\kappa>0} \widetilde{\mathcal{J}}(\mu,\alpha,\tau_\kappa;\beta_*,\tau_{\nu,*}) \geq \mathcal{J}_*,$$
$$\sup_{\beta,\nu>0,\,\eta} \inf_{\kappa>0} \mathcal{J}(\mu_*,\alpha_*,\kappa;\beta,\nu,\eta) \leq \sup_{\beta,\tau_\nu>0} \widetilde{\mathcal{J}}(\mu_*,\alpha_*,\tau_{\kappa,*};\beta,\tau_\nu) = \mathcal{J}_*.$$

The lower bound fixes the dual variables at $(\beta_*, \tau_{\nu,*})$, equivalently choosing $\nu = \tau_{\nu,*}/\alpha$ for each α. The upper bound fixes the primal variables at $(\mu_*, \alpha_*, \tau_{\kappa,*})$, equivalently choosing $\kappa = \tau_{\kappa,*}/\beta$ for each β. This proves (46) without interchanging κ with the coefficient supremum. Substituting $U - \zeta = \kappa\xi$ gives $\mathcal{J}_* = \mathbb{E}_G\left[\rho(W_*)\right]/n + \mathbb{E}\left[L_y(\zeta_*)\right]$. If $\mathbb{E}\left[L_y(0)\right] \leq L_0$, (42) also gives

$$\rho(w) - \rho(0) \geq \gamma_0|w|^2/2 \implies \mathbb{E}_G\left[|W_*|^2\right]/n \leq 2L_0/\gamma_0,$$
$$\rho(w) = \lambda\|w\|_1 \implies \lambda\mathbb{E}_G\left[\|W_*\|_1\right]/n \leq L_0.$$

The Lasso bound alone does not control normalized second moments. □

Proposition A.2 (The score-nondegeneracy condition cannot be omitted)**.** *Let $\rho(w) := \lambda\|w\|_1$, $\lambda > 0$, and $L_y(u) := \log\cosh u$. If $\mathbb{E}\left[|x|^2\right] < \infty$ and $\Sigma_x \succ 0$, then $\widehat{W} = 0$ for every sample, but the population equations have no finite solution with $\beta, \nu > 0$. Here $L_y'(0) = 0$ almost surely.*

Proof. Loss and penalty are nonnegative and vanish together only at zero. At a supposed root, multiply the surrogate inclusion by W and average:

$$0 \leq \frac{\lambda}{n}\mathbb{E}_G\left[\|W\|_1\right] = -\mu^\top\mathbb{E}\left[x\xi\right] - \nu\alpha^2 + \kappa\beta^2 = -\mathbb{E}\left[\zeta\tanh\zeta\right] \leq 0.$$

Hence $\zeta = 0$ and $\beta^2 = \mathbb{E}\left[\xi^2\right] = 0$, a contradiction. □

A.4 Population coefficient bounds

Lemma A.3 (Population moment and Gaussian-correlation identities)**.** *At any finite population root satisfying $\underline{c}_x I_p \preceq \Sigma_x \preceq K I_p$,*

$$\mathbb{E}_G\left[|W_*|^2\right]/n = |\mu_*|^2 + \operatorname{tr} C_{W_*}/n, \qquad \underline{c}_x \operatorname{tr} C_{W_*}/n \leq \alpha_*^2 \leq K \operatorname{tr} C_{W_*}/n, \tag{53}$$

$$\beta_*\kappa_* = -\frac{1}{n}\mathbb{E}_G\left[G^\top(W_* - \sqrt{n}\mu_*)\right] \leq \sqrt{p/n}\,\alpha_*. \tag{54}$$

Thus bounded normalized surrogate second moments are equivalent to bounded $|\mu_*|^2 + \alpha_*^2$.

Proof. Use the mean–covariance decomposition and $\underline{c}_x I \preceq \Sigma_x \preceq KI$ for (53). For (54), center W_* and apply Cauchy–Schwarz in the Σ_x^{-1} and Σ_x metrics, using $\mathbb{E}_G\left[G^\top\Sigma_x^{-1}G\right] = p$. □

Proposition A.4 gives uniform coefficient and surrogate-moment bounds on the region

$$\mathcal{Q} := \{q = (\mu, \alpha^2, \kappa) : |\mu| \leq M,\ 0 \leq \alpha^2 \leq \alpha_+^2,\ 0 \leq \kappa \leq K_\kappa\}, \qquad \alpha \geq 0. \tag{55}$$

Use $\|q\|_1 := |\mu| + |\alpha^2| + |\kappa|$, also for differences. For $q \in \mathcal{Q}$, compute $\zeta_q, \xi_q, \nu_q, \beta_q, \eta_q$ by (4) and W_q by (5a), with independent $G \sim N(0, \Sigma_x)$. No fixed-point equation is imposed on q.

Proposition A.4 (Uniform coefficient and surrogate-moment bounds)**.** *Suppose $\underline{c}_x I_p \preceq \Sigma_x \preceq K I_p$, $p/n \le K$, and Assumption 1 holds. Uniformly for $q \in \mathcal{Q}$,*

$$\nu_0 \le \nu_q \le L_2, \qquad 0 \le \beta_q \le L_1, \qquad \beta_q \ge \nu_q \alpha, \qquad |\mathbb{E}\left[x\xi_q\right]| + |\eta_q| \le C.$$

The surrogate has a unique minimizer, including when $\beta_q = 0$, and $\mathbb{E}_G\left[|W_q|^2\right]/n \le C$. The constants $\nu_0 > 0, C$ depend only on $\underline{c}_x, K, L_1, L_2$, the parameters M, α_+, K_κ defining (55)*, and the lower-curvature function $c_L(\cdot)$ in Assumption 1. In particular, $\alpha_* \ge \underline{\alpha} > 0$ implies $\beta_* \ge \nu_0 \underline{\alpha} > 0$ whenever $q_* \in \mathcal{Q}$.*

Proof. Since $\mathbb{E}\left[U^2\right] \le KM^2 + \alpha_+^2$, choose a fixed B with $\mathbb{P}(|U| \le B) \ge 1/2$. The equation $\zeta_q + \kappa L_y'(\zeta_q) = U$, obtained from (4), gives $|\zeta_q| \le |U| + K_\kappa L_1$, hence

$$\nu_q \ge \frac{c_L(B + K_\kappa L_1)}{2(1 + K_\kappa L_2)} =: \nu_0 > 0, \qquad \mathbb{E}\left[Z\xi_q\right] = \alpha\nu_q \le \beta_q.$$

Scalar Gaussian integration by parts gives $\mathbb{E}\left[Z\xi_q\right] = \alpha\nu_q$, and Cauchy–Schwarz gives $\alpha\nu_q \le \beta_q$, also when $\alpha = 0$. For every unit vector u, $|u^\top \mathbb{E}\left[x\xi_q\right]| \le \sqrt{K} L_1$, which bounds η_q. Finally, positive curvature and monotonicity against $0 \in \partial\rho(0)$ give a unique optimizer with $\nu_0 \underline{c}_x |W_q| \le |\sqrt{n}\,\eta_q - \beta_q G|$. Taking second moments and using $\mathrm{tr}(\Sigma_x)/n \le Kp/n$ proves $\mathbb{E}_G\left[|W_q|^2\right]/n \le C$. □

B Convex response calculus and Hamiltonian identities

B.1 Positive-probe response and moments

Positive quadratic curvature makes the optimizer Lipschitz in its linear field, even when ρ is nonsmooth. Lemmas B.3, B.4, and B.5 give the moment bounds and value derivatives used in Appendices C–D. In Lemmas B.3, B.5, and B.6, v, b stay in fixed positive bounded intervals and $|h|$ is bounded; finite differences use a slightly enlarged domain. The calculus permits $t = 0, 1$ and requires only $0 \in \partial\rho(0)$, not loss-score nondegeneracy.

Lemma B.1 (Feature moment bounds)**.** *Suppose Assumption 2 holds, $\Sigma_x \preceq K I_p$, and $p = O(n)$. Then* (9) *holds with $K_Q^2 = 4C_{\mathrm{P}}K$. Put $K_4 := K_Q^2 + K^2$. For deterministic symmetric A and $u \in \mathbb{R}^p$,*

$$\mathbb{E}_x\left[|x^\top Ax - \mathrm{tr}(\Sigma_x A)|\right] \le K_Q \|A\|_{\mathrm{hs}}, \qquad \mathbb{E}_x\left[|u^\top x|^4\right] \le K_4 |u|^4, \qquad \mathbb{E}_x\left[|x|^4\right] \le K_4 p^2.$$

Consequently, $\sup_n \mathbb{E}_x\left[(1 + |x|^2/n)\mathbf{1}_{\{|x| > R\sqrt{n}\}}\right] \to 0$ *as* $R \to \infty$.

Proof. First prove (9) without assuming fourth moments. Set $Q := x^\top Ax$, which is integrable by the feature second moment. Choose smooth bounded truncations T_m with $|T_m'| \le 1$, $|T_m(s)| \le |s|$, and $T_m(s) \to s$. Since $T_m(Q)$ depends only on x, its label energy vanishes, and (8) gives

$$\mathrm{Var}_x(T_m(Q)) \le 4C_{\mathrm{P}}\mathbb{E}_x\left[|Ax|^2\right] = 4C_{\mathrm{P}}\,\mathrm{tr}(A^2\Sigma_x) \le 4C_{\mathrm{P}}K\|A\|_{\mathrm{hs}}^2.$$

Dominated convergence of the means and Fatou's lemma for the centered squares give (9). Cauchy–Schwarz applied to (9) gives $\mathbb{E}_x\left[|x^\top Ax - \mathrm{tr}(\Sigma_x A)|\right] \le K_Q\|A\|_{\mathrm{hs}}$. The choice $A = uu^\top$ gives $\mathbb{E}_x\left[|u^\top x|^4\right] \le K_Q^2|u|^4 + (u^\top\Sigma_x u)^2$. Summing coordinate fourth moments by Cauchy–Schwarz gives $\mathbb{E}_x\left[|x|^4\right] \le K_4 p^2$; Markov's inequality bounds the tail by $C(R^{-4} + R^{-2})$. □

Lemma B.2 (Well-posedness and response under positive curvature)**.** *Suppose Assumption 1 holds and $\Sigma_x \succeq \underline{c}_x I_p$. Fix observations and selectors and put*

$$f_v(w) := \rho(w) + \sum_i B_i L_{y_i}(x_i^\top w/\sqrt{n}) + \frac{v}{2} w^\top \Sigma_x w, \qquad \ell := \sqrt{n}\,h - \sqrt{2b}\,G.$$

For $v > 0$, $W(\ell) := \arg\min_w\{f_v(w) - \ell^\top w\}$ exists uniquely and

$$v\underline{c}_x |W(\ell) - W(\ell')|^2 \le (\ell - \ell')^\top (W(\ell) - W(\ell')), \qquad |W(\ell) - W(\ell')| \le \frac{|\ell - \ell'|}{v\underline{c}_x}. \tag{56}$$

Its almost-everywhere normalized field response $J := D_h W/\sqrt{n} = D_\ell W$ is symmetric with $0 \preceq J \preceq (v\underline{c}_x)^{-1} I_p$; for $b > 0$ it exists at the Gaussian-perturbed field almost surely.

Proof. The positive quadratic term makes every linear tilt coercive (tending to $+\infty$ as $|w| \to \infty$) and $v\underline{c}_x$-strongly convex. Thus the minimizer exists uniquely, and strong monotonicity of $\ell \in \partial f_v(W(\ell))$ gives (56). Since $W = \nabla f_v^*$, its almost-everywhere Jacobian satisfies $0 \preceq D_\ell W \preceq (v\underline{c}_x)^{-1} I_p$. The nondegenerate Gaussian field avoids its null exceptional set; Fubini also justifies differentiation along almost every scalar force. □

Lemma B.3 (Second moments on fixed interior domains)**.** *Let $\mathcal{D}$ be a fixed interior domain introduced in Section 2.3. For $(t, v, b, h) \in \mathcal{D}$, let $W := W_{t,v,b,h}$ be the optimizer in* (15)*, and let W_- be the optimizer obtained from* (14) *by deleting only $B_j(t)L_{y_j}(x_j^\top w/\sqrt{n})$ for one specified $j \in [n]$. All other selectors and the same Gaussian probe are retained, and no additional linear force is added. Suppose Assumption 1 holds, $\underline{c}_x I_p \preceq \Sigma_x \preceq K I_p$, and $p/n \le K$. Then*

$$\sup_{(t,v,b,h)\in\mathcal{D}} (\|W\|_{L^2} + \|W_-\|_{L^2}) \le B_{\mathcal{D}}\sqrt{n}, \qquad \|U\|_{L^2} := \left(\mathbb{E}\left[|U|^2\right]\right)^{1/2}. \tag{57}$$

The expectation is over the training sample, the selectors, and the Gaussian probe, and the bound is uniform in the omitted index. Bound (57) *also holds when t is set to 0 or 1, provided v, b, h remain within the coordinate bounds defining $\mathcal{D}$. The constant $B_{\mathcal{D}}$ depends only on $\underline{c}_x, K, L_1$ and those coordinate bounds.*

Proof. Fix $(t, v, b, h) \in \mathcal{D}$. For $z_i := B_i L'_{y_i}(0)x_i$, monotonicity against zero gives

$$v\underline{c}_x|W| \le \left| n^{-1/2}\sum_i z_i + \sqrt{2b}\,G - \sqrt{n}\,h \right|, \qquad \mathbb{E}\left[|z_i|^2\right] \le Cn, \quad |\mathbb{E}\left[z_i\right]| \le L_1\sqrt{K}. \tag{58}$$

Independence yields $\mathbb{E}\left[|n^{-1/2}\sum_i z_i|^2\right] \le Cn$. Together with $\mathbb{E}\left[|G|^2\right] = \operatorname{tr}\Sigma_x = O(n)$, this proves (57). For W_-, the same argument omits the corresponding summand z_i. The constants are uniform over $\mathcal{D}$ and the omitted index. Since the proof uses only $0 \le B_i(t) \le 1$, it also applies at $t = 0$ and $t = 1$ under the stated coordinate bounds. □

Lemma B.4 (Interior fourth moments)**.** *Let $\mathcal{D}$ be a fixed interior domain introduced in Section 2.3, and let W and W_- be the optimizers defined in Lemma B.3. Along a sequence in which the training pairs are independent copies of (x, y), suppose $p/n \to c \in (0, \infty)$, $\underline{c}_x I_p \preceq \Sigma_x \preceq K I_p$, and Assumptions 1 and 2 hold. Then there is $C_{\mathcal{D}} < \infty$, independent of the omitted index, such that*

$$\sup_{(t,v,b,h)\in\mathcal{D}} \frac{\mathbb{E}\left[|W|^4\right] + \mathbb{E}\left[|W_-|^4\right]}{n^2} \le C_{\mathcal{D}}. \tag{59}$$

The estimate remains valid after omitting any fixed number of observations, with the original denominator n, and when t is set to 0 or 1 under the coordinate bounds for v, b, h defining $\mathcal{D}$. It uses no prediction comparison or additional quadratic-form assumption.

Proof. For $z_i := B_i L'_{y_i}(0)x_i$, Lemma B.1 gives $\mathbb{E}\left[|z_i|^4\right] \le Cn^2$ and $|\mathbb{E}\left[z_i\right]| \le L_1\sqrt{K}$. The fourth-moment inequality for independent centered vectors gives

$$\mathbb{E}\left[\left|\sum_i (z_i - \mathbb{E}\left[z_i\right])\right|^4\right] \le \sum_i \mathbb{E}\left[|z_i - \mathbb{E}\left[z_i\right]|^4\right] + 3\left(\sum_i \mathbb{E}\left[|z_i - \mathbb{E}\left[z_i\right]|^2\right]\right)^2.$$

Thus $\mathbb{E}\left[|n^{-1/2}\sum_i z_i|^4\right] \le Cn^2$. Equation (58), $\mathbb{E}\left[|G|^4\right] = O(n^2)$, and $n^2|h|^4 = O(n^2)$ prove (59). Omitting any fixed number of summands gives the cavity claim. The argument is uniform over $\mathcal{D}$ and is unchanged at $t = 0$ and $t = 1$ under the stated coordinate bounds. □

B.2 Value and Hamiltonian derivatives

We collect the envelope identities that turn value gradients into moments. Gaussian integration by parts in the proof of Lemma B.5 concerns the actual probe G; Conjecture 1 enters only in Proposition C.5. Define the Hamilton–Jacobi coordinates and Hamiltonian by

$$z := (v, b, h), \qquad \pi := \nabla_z \phi, \tag{60}$$

$$H_*(\pi) := \widetilde{\mathcal{H}}(\pi) = \mathbb{E}\left[e_{L_y}\left(-x^\top \pi_h + \sqrt{2\pi_v - \pi_h^\top \Sigma_x \pi_h}\, Z; -\pi_b\right)\right], \tag{61}$$

on $-\pi_b \geq 0$ and $2\pi_v \geq \pi_h^\top \Sigma_x \pi_h$.

Lemma B.5 (Envelope and Gaussian-correlation identities)**.** *Under the assumptions of Lemma B.3,*

$$\mu = -\nabla_h \phi, \qquad \alpha^2 = 2\partial_v \phi - \mu^\top \Sigma_x \mu, \qquad \kappa = -\partial_b \phi = -\frac{\mathbb{E}\left[G^\top W\right]}{n\sqrt{2b}}. \tag{62}$$

Equivalently,

$$\mu = -\pi_h, \qquad \alpha^2 = 2\pi_v - \pi_h^\top \Sigma_x \pi_h, \qquad \kappa = -\pi_b, \qquad \partial_b \phi = \frac{\mathbb{E}\left[G^\top W\right]}{n\sqrt{2b}}. \tag{63}$$

Moreover $\nabla H_*(\pi) = (\nu, \beta^2/2, \eta)$, with continuous boundary extension, and the defect r defined in (83) *satisfies $\phi_t = H_*(\nabla_z \phi) + r$.*

Proof. The unique-minimizer envelope theorem gives

$$\nabla_h \phi = -\frac{\mathbb{E}[W]}{\sqrt{n}}, \qquad \partial_v \phi = \frac{\mathbb{E}\left[W^\top \Sigma_x W\right]}{2n}, \qquad \partial_b \phi = \frac{\mathbb{E}\left[G^\top W\right]}{n\sqrt{2b}}.$$

Estimate (57) provides local domination. Since $D_G W = -\sqrt{2b} J$, Gaussian integration by parts gives $\mathbb{E}\left[G^\top W\right] = -\sqrt{2b}\mathbb{E}\left[\mathrm{tr}(\Sigma_x J)\right]$. Centering the quadratic moment proves (62) and (63); Lemmas B.8 and B.10 give the Hamiltonian gradient and its extension. □

Lemma B.6 (Exact Bernoulli differentiation)**.** *Under the assumptions of Lemma B.3, let Δ be the minimum-value increment from reinserting one independent observation into its unforced cavity. Then $\partial_t \phi = \mathbb{E}[\Delta]$.*

Proof. For $F(B) := \min_w \mathcal{E}(w)$, differentiating the finite product Bernoulli law gives

$$\partial_t \phi = \frac{1}{n}\sum_i \mathbb{E}\left[F(B^{i,1}) - F(B^{i,0})\right] = \mathbb{E}[\Delta],$$

where $B^{i,j}$ fixes selector i to j. Estimate (57), together with $\mathcal{E}(W) \leq \mathcal{E}(0)$, gives the required integrability; exchangeability gives $n^{-1}\sum_i \mathbb{E}\left[F(B^{i,1}) - F(B^{i,0})\right] = \mathbb{E}[\Delta]$. □

Lemma B.7 (Moreau derivatives and Gaussian averaging)**.** *Under the loss conditions of Assumption 1, put $\zeta := \mathrm{prox}_{\kappa L_y}(u)$ and $\xi := L_y'(\zeta)$. For $\kappa \geq 0$, with right derivatives at zero,*

$$\partial_u e_{L_y} = \xi, \qquad \partial_{uu} e_{L_y} = \frac{L_y''(\zeta)}{1 + \kappa L_y''(\zeta)}, \qquad \partial_\kappa e_{L_y} = -\frac{\xi^2}{2}. \tag{64}$$

Their respective absolute bounds are $L_1, L_2, L_1^2/2$. In particular,

$$\left|\mathbb{E}_Z\left[e_{L_y}(z_0 + \alpha_1 Z; \kappa) - e_{L_y}(z_0 + \alpha_2 Z; \kappa)\right]\right| \leq \frac{L_2}{2}|\alpha_1^2 - \alpha_2^2|. \tag{65}$$

Proof. Differentiate $\zeta + \kappa L_y'(\zeta) = u$ and apply the envelope theorem. For any $f \in C^2(\mathbb{R})$ with bounded first and second derivatives, differentiation and Gaussian integration by parts give, for $\alpha > 0$,

$$\partial_{\alpha^2}\mathbb{E}_Z[f(z_0 + \alpha Z)] = \frac{1}{2\alpha}\mathbb{E}_Z[Z f'(z_0 + \alpha Z)] = \frac{1}{2}\mathbb{E}_Z[f''(z_0 + \alpha Z)], \qquad Z \sim N(0,1). \tag{66}$$

The formulas for $\partial_{\alpha^2}\mathbb{E}_Z[f(z_0 + \alpha Z)]$ in (66) extend to $\alpha = 0$ by continuity, using the right variance derivative. Applying (66) to $f(u) := e_{L_y}(u; \kappa)$ and using (64) gives (65). □

Lemma B.8 (Hamiltonian coefficient identities)**.** *If $\mathbb{E}\left[|x|^2\right] < \infty$ and the loss derivative bounds in Assumption 1 hold,* (32)–(33) *satisfy* (35)*, with right variance derivatives at $\alpha = 0$.*

Proof. Equations (64) and (66), with the definitions (4), give

$$\mathrm{d}\mathcal{H} = \mathbb{E}[x\xi]^\top \mathrm{d}\mu + \frac{\nu}{2}\mathrm{d}(\alpha^2) - \frac{\beta^2}{2}\mathrm{d}\kappa.$$

For $\delta = (\delta_v, \delta_b, \delta_h) = \nabla_{v,b,h}\phi$, (31) gives $\mathrm{d}\mu = -\mathrm{d}\delta_h$, $\mathrm{d}\kappa = -\mathrm{d}\delta_b$, and $\mathrm{d}(\alpha^2) = 2\mathrm{d}\delta_v + 2\mu^\top \Sigma_x \mathrm{d}\delta_h$. Thus $\mathrm{d}\widetilde{\mathcal{H}} = \nu \mathrm{d}\delta_v + (\beta^2/2)\mathrm{d}\delta_b + (\nu\Sigma_x\mu - \mathbb{E}[x\xi])^\top \mathrm{d}\delta_h$. Scalar Gaussian integration by parts also gives $\mathbb{E}[Z\xi] = \nu\alpha$. □

B.3 Coupling and coefficient stability

For U independent of a fresh pair (x, y), define its prediction error by

$$\mathcal{G}(U) := \mathbb{E}_x \left[\mathsf{W}_1\big(\mathcal{L}\big(x^\top U/\sqrt{n} \mid x\big), N\big(x^\top \mathbb{E}[U]/\sqrt{n}, \operatorname{tr}(\Sigma_x C_U)/n\big)\big)\right].$$

Proposition B.9 transfers $\mathcal{G}(U)$ and the moments from positive probes to the endpoint.

Proposition B.9 (Coupling bounds for moments and fresh scores)**.** *Assume $\Sigma_x \preceq K I_p$. For coupled square-integrable U, W jointly independent of (x, y),*

$$|\mathcal{G}(W) - \mathcal{G}(U)| \le C_{\mathrm{gc}} \frac{\|W - U\|_{L^2}}{\sqrt{n}}, \qquad C_{\mathrm{gc}} := \left(2 + \sqrt{2/\pi}\right) \sqrt{\|\Sigma_x\|_{\mathrm{op}}}. \tag{67}$$

Writing $d := \|W - U\|_{L^2}/\sqrt{n}$, one also has

$$\begin{gathered} |\mathbb{E}[W - U]|/\sqrt{n} \le d, \qquad \mathbb{E}_x\left[\mathsf{W}_1\big(\mathcal{L}(x^\top W/\sqrt{n} \mid x), \mathcal{L}(x^\top U/\sqrt{n} \mid x)\big)\right] \le \sqrt{K} d, \\ \frac{\|C_W - C_U\|_{\mathrm{hs}} + |\operatorname{tr}(\Sigma_x[C_W - C_U])|}{n} \le (1 + K) d \frac{\|W\|_{L^2} + \|U\|_{L^2}}{\sqrt{n}}. \end{gathered} \tag{68}$$

Proof. Couple both scores using (U, W) and both comparison Gaussians using one standard normal. The mean and standard-deviation differences are bounded by

$$\frac{|\mathbb{E}[W - U]|}{\sqrt{n}} \le d, \qquad \left|\sqrt{\operatorname{tr}(\Sigma_x C_W)/n} - \sqrt{\operatorname{tr}(\Sigma_x C_U)/n}\right| \le \sqrt{\|\Sigma_x\|_{\mathrm{op}}} d. \tag{69}$$

The standard-deviation bound in (69) follows from the reverse triangle inequality in centered weighted L^2. Averaging the score difference and applying the Wasserstein triangle inequality gives (67). Expanding the difference of the centered outer products and using $\|ab^\top\|_{\mathrm{hs}} = |a||b|$ and Cauchy–Schwarz gives (68). □

Lemma B.10 (Stability of the population coefficients)**.** *Assume the loss derivative bounds in Assumption 1 and $\Sigma_x \preceq K I_p$. For every $M < \infty$, let (ν, β, η) and (ν', β', η') be the coefficients defined by* (4) *at (μ, α^2, κ) and $(\mu', (\alpha')^2, \kappa')$, respectively. If $|\mu|, |\mu'| \le M$, $0 \le \kappa, \kappa' \le M$, and $\alpha, \alpha' \ge 0$, then*

$$|\nu - \nu'| + |\beta^2 - (\beta')^2| + |\eta - \eta'| \le C\big(|\mu - \mu'| + |\alpha^2 - (\alpha')^2| + |\kappa - \kappa'|\big).$$

Here C depends only on M, K, L_1, L_2, L_3, L_4.

Proof. Put $\zeta(u, \kappa, y) := \operatorname{prox}_{\kappa L_y}(u)$, $\xi(u, \kappa, y) := L_y'(\zeta(u, \kappa, y))$, and $D := 1 + \kappa L_y''(\zeta) \ge 1$. Implicit differentiation gives

$$\begin{gathered} \partial_u \zeta = \frac{1}{D}, \qquad \partial_\kappa \zeta = -\frac{\xi}{D}, \\ \partial_u \xi = \frac{L_y''(\zeta)}{D}, \qquad \partial_u^2 \xi = \frac{L_y'''(\zeta)}{D^3}, \qquad \partial_u^3 \xi = \frac{L_y^{(4)}(\zeta)}{D^4} - \frac{3\kappa L_y'''(\zeta)^2}{D^5}, \\ \partial_\kappa \xi = -\frac{L_y''(\zeta)\xi}{D}, \qquad \partial_\kappa \partial_u \xi = -\frac{L_y'''(\zeta)\xi}{D^3} - \frac{L_y''(\zeta)^2}{D^2}. \end{gathered}$$

The κ-derivatives at $\kappa = 0$ are right derivatives. Assumption 1, $D \ge 1$, and $0 \le \kappa \le M$ show that

$$\sup_{u, y,\, 0 \le \kappa \le M} \big(|\xi| + |\partial_u \xi| + |\partial_u^2 \xi| + |\partial_u^3 \xi| + |\partial_\kappa \xi| + |\partial_\kappa \partial_u \xi|\big) \le C. \tag{70}$$

We now make the comparison in (4) explicit. Set

$$\Delta_\mu := \mu - \mu', \qquad \Delta_{\alpha^2} := \alpha^2 - (\alpha')^2, \qquad \Delta_\kappa := \kappa - \kappa',$$

and, for $s \in [0, 1]$, define

$$\begin{aligned} \mu_s &:= \mu' + s\Delta_\mu, \quad \alpha_s^2 := (\alpha')^2 + s\Delta_{\alpha^2}, \\ \kappa_s &:= \kappa' + s\Delta_\kappa, \quad U_s := x^\top \mu_s + \alpha_s Z, \end{aligned}$$

where $\alpha_s := \sqrt{\alpha_s^2}$. For a scalar function $f = f(u, \kappa, y)$ with bounded $\partial_u f$, $\partial_u^2 f$, and $\partial_\kappa f$, condition on (x, y) and apply (66) to the Gaussian variance α_s^2. The fundamental theorem of calculus gives

$$\begin{aligned}&\mathbb{E}\left[f(U_1, \kappa_1, y) - f(U_0, \kappa_0, y)\right] \\ &\quad = \int_0^1 \Big\{\mathbb{E}\left[(x^\top \Delta_\mu)\partial_u f(U_s, \kappa_s, y)\right] + \frac{\Delta_{\alpha^2}}{2}\mathbb{E}\left[\partial_u^2 f(U_s, \kappa_s, y)\right] + \Delta_\kappa \mathbb{E}\left[\partial_\kappa f(U_s, \kappa_s, y)\right]\Big\}\,\mathrm{d}s.\end{aligned} \tag{71}$$

At $\alpha_s = 0$, the variance term is the right derivative specified in (66).

The identity $\partial_u \xi = L_y''(\zeta)/(1 + \kappa L_y''(\zeta))$ rewrites the definitions of ν, ν', β^2, and $(\beta')^2$ in (4) as

$$\begin{aligned}\nu &= \mathbb{E}\left[\partial_u \xi(U_1, \kappa_1, y)\right], \qquad \nu' = \mathbb{E}\left[\partial_u \xi(U_0, \kappa_0, y)\right], \\ \beta^2 &= \mathbb{E}\left[\xi(U_1, \kappa_1, y)^2\right], \quad (\beta')^2 = \mathbb{E}\left[\xi(U_0, \kappa_0, y)^2\right].\end{aligned} \tag{72}$$

For compactness, write $g_s := g(U_s, \kappa_s, y)$ for any function g. Combining (72) with (71), first with $f = \partial_u \xi$ and then with $f = \xi^2$, yields the exact identities

$$\nu - \nu' = \int_0^1 \Big\{\mathbb{E}\left[(x^\top \Delta_\mu)(\partial_u^2 \xi)_s\right] + \frac{\Delta_{\alpha^2}}{2}\mathbb{E}\left[(\partial_u^3 \xi)_s\right] + \Delta_\kappa \mathbb{E}\left[(\partial_\kappa \partial_u \xi)_s\right]\Big\}\,\mathrm{d}s \tag{73}$$

and

$$\begin{aligned}\beta^2 - (\beta')^2 = \int_0^1 \Big\{&2\mathbb{E}\left[(x^\top \Delta_\mu)\xi_s(\partial_u \xi)_s\right] \\ &+ \Delta_{\alpha^2}\mathbb{E}\left[(\partial_u \xi)_s^2 + \xi_s(\partial_u^2 \xi)_s\right] + 2\Delta_\kappa \mathbb{E}\left[\xi_s(\partial_\kappa \xi)_s\right]\Big\}\,\mathrm{d}s.\end{aligned} \tag{74}$$

Since $\Sigma_x \preceq KI$, Cauchy–Schwarz gives $\mathbb{E}\left[|x^\top \Delta_\mu|\right] \le \sqrt{K}|\Delta_\mu|$. The derivative bounds (70), applied to (73)–(74), therefore imply

$$|\nu - \nu'| + |\beta^2 - (\beta')^2| \le C\big(|\Delta_\mu| + |\Delta_{\alpha^2}| + |\Delta_\kappa|\big). \tag{75}$$

It remains to compare $\mathbb{E}\left[x\xi\right]$, which enters the definition $\eta = \nu \Sigma_x \mu - \mathbb{E}\left[x\xi\right]$ in (4). Put

$$B := \mathbb{E}\left[x\xi(U_1, \kappa_1, y)\right], \qquad B' := \mathbb{E}\left[x\xi(U_0, \kappa_0, y)\right].$$

For every unit vector e, apply (71) to $f = \xi$ while retaining the factor $e^\top x$. This gives

$$\begin{aligned}e^\top (B - B') = \int_0^1 \Big\{&\mathbb{E}\left[(e^\top x)(x^\top \Delta_\mu)(\partial_u \xi)_s\right] \\ &+ \frac{\Delta_{\alpha^2}}{2}\mathbb{E}\left[(e^\top x)(\partial_u^2 \xi)_s\right] + \Delta_\kappa \mathbb{E}\left[(e^\top x)(\partial_\kappa \xi)_s\right]\Big\}\,\mathrm{d}s.\end{aligned} \tag{76}$$

Here

$$\mathbb{E}\left[|e^\top x|\,|x^\top \Delta_\mu|\right] \le K|\Delta_\mu|, \qquad \mathbb{E}\left[|e^\top x|\right] \le \sqrt{K}.$$

Taking the supremum over unit e in (76) shows that

$$|B - B'| \le C\big(|\Delta_\mu| + |\Delta_{\alpha^2}| + |\Delta_\kappa|\big). \tag{77}$$

Finally, the definition of η in (4) gives

$$\eta - \eta' = (\nu - \nu')\Sigma_x \mu + \nu' \Sigma_x \Delta_\mu - (B - B').$$

Use $|\mu| \le M$, $0 \le \nu' \le L_2$, $\|\Sigma_x\|_{\mathrm{op}} \le K$, (75), and (77) to prove the inequality in Lemma B.10. □

C Three-step proof of Theorem 5

C.1 Cavity reduction and the three signed errors

The proof follows the chain

insertion cost $\longrightarrow$ scalar Moreau envelope $\longrightarrow$ cavity Hamiltonian $\longrightarrow$ full Hamiltonian.

Only the scalar-envelope-to-cavity-Hamiltonian step uses Conjecture 1. Write $\varepsilon_{\mathrm{pred}}(\mathcal{D})$ for the supremum in (21), including the full/cavity family. The insertion-cost-to-scalar-envelope step uses

the scalar-response concentration derived from (9) in Lemma C.1 and its finite-force transfer in Lemma C.2.

Delete observation 1 and write Ψ_- for its background objective, W_- for its minimizer, and $\mu_-, \alpha_-^2, \kappa_-$ for its deterministic moments and response. The deleted pair (x, y) is independent of this background. Set

$$\Delta := \min_w\{\Psi_-(w) + L_y(x^\top w/\sqrt{n})\} - \min_w \Psi_-(w). \tag{78}$$

For $|r| \le L_1$, retain the cavity data and Gaussian probe used to define W_-, and set

$$W_-(r) := \arg\min_w\{\Psi_-(w) - rx^\top w/\sqrt{n}\},$$

$$q_x(r) := x^\top W_-(r)/\sqrt{n}, \qquad J_-(r) := \frac{1}{\sqrt{n}} D_h W_-(r), \qquad J_- := J_-(0).$$

The finite-force error used in Proposition C.4 is

$$\omega_{\mathrm{lin}} := \mathbb{E}\left[\sup_{|r|\le L_1} |q_x(r) - q_x(0) - \kappa_- r|\right]. \tag{79}$$

For cavity-to-full deletion, write $\omega_\mu := \sqrt{(\mu-\mu_-)^\top \Sigma_x (\mu-\mu_-)}$, $\omega_{\alpha^2} := |\alpha^2 - \alpha_-^2|$, and $\omega_\kappa := |\kappa - \kappa_-|$. With $u := q_x(0)$, define the three signed errors by

$$r^{(1)} := \mathbb{E}\left[\Delta - e_{L_y}(u;\kappa_-)\right], \tag{80}$$

$$r^{(2)} := \mathbb{E}\left[e_{L_y}(u;\kappa_-)\right] - \mathcal{H}(\mu_-, \alpha_-, \kappa_-), \tag{81}$$

$$r^{(3)} := \mathcal{H}(\mu_-, \alpha_-, \kappa_-) - \widetilde{\mathcal{H}}(\nabla\phi). \tag{82}$$

Summing (80)–(82) cancels the intermediate terms. Since $\mathbb{E}[\Delta] = \partial_t \phi$ by Lemma B.6, the value defect satisfies

$$r := \partial_t \phi - \widetilde{\mathcal{H}}(\nabla\phi) = \sum_{j=1}^{3} r^{(j)}. \tag{83}$$

C.2 Concentration and finite-force response

Throughout this section, $\mathcal{D}$ is a fixed interior domain introduced in Section 2.3. Finite differences are taken in a fixed enlargement with fixed positive lower bounds for v and b. All estimates are uniform on $\mathcal{D}$, and constants may depend on its compact ranges, the bound on $|h|$, and the margins of the enlarged domain. The response argument is

$$\text{minimum-value concentration} \implies \text{amplitude-derivative concentration}$$
$$\implies \text{response-trace concentration}.$$

The quadratic-form estimate (9) and independence yield (85); Lemma C.2 then transfers (85) to bounded forces.

Lemma C.1 (Scalar response from feature concentration)**.** *Under the assumptions of Lemma B.4, full and unforced cavity responses satisfy*

$$\sup_{\mathcal{D}} \|\mathrm{tr}(\Sigma_x J)/n - \kappa\|_{L^2} \longrightarrow 0. \tag{84}$$

If (9) *also holds, then for a fresh feature independent of J,*

$$\varepsilon_{\mathrm{resp}}(\mathcal{D}) := \sup_{\mathcal{D}} \mathbb{E}\left[|x^\top J x/n - \kappa|\right] \longrightarrow 0. \tag{85}$$

Each family uses its own deterministic κ. Neither prediction comparison nor projection anti-concentration is used.

Proof. We first concentrate a derivative of the random minimum by finite differences in the probe amplitude. A second-order Stein identity then identifies that derivative with the response trace, up to a vanishing error. Keep the observations, selectors and probe fixed, and put $\sigma := \sqrt{2b}$ and

$$\mathcal{V}(\sigma) := \frac{1}{n}\min_w \left\{\rho(w) + \sum_i B_i L_{y_i}(x_i^\top w/\sqrt{n}) + \frac{v}{2} w^\top \Sigma_x w + \sigma G^\top w - \sqrt{n}\, h^\top w\right\}.$$

Comparison with zero and the supporting loss hyperplanes bound $|\mathcal{V}(\sigma) - \rho(0)/n|$ by $L_0 + |n^{-1/2}\sum_i B_i L'_{y_i}(0)x_i + \sigma G - \sqrt{n}\,h|^2/(2v\underline{c}_x n)$, which has bounded L^2 norm by Lemma B.4. The insertion cost D_i satisfies

$$0 \le D_i \le L_0 + L_1|x_i^\top W_-|/\sqrt{n}, \qquad \mathbb{E}\left[D_i^2\right] \le C_{\mathcal{D}},$$

because the cavity W_- is independent of (x_i, y_i). Replacing one observation block costs two insertions into a common cavity. Efron–Stein and Gaussian Poincaré therefore give

$$\mathrm{Var}(\mathcal{V}(\sigma)) \le \frac{C_{\mathcal{D}}}{n} + \frac{\sigma^2}{n^2}\mathbb{E}\left[W^\top \Sigma_x W\right] \le \frac{C_{\mathcal{D}}}{n}.$$

The envelope theorem gives $\mathcal{V}'(\sigma) = G^\top W/n$; differentiating $\mathcal{V}'(\sigma) = G^\top W/n$ and using $D_G W = -\sigma J$ gives $\mathcal{V}''(\sigma) = -G^\top JG/n$ almost everywhere, with $\|\mathcal{V}''\|_{L^2} \le C_{\mathcal{D}}$. Comparing $\mathcal{V}'(\sigma)$ with an amplitude difference of length ℓ_{fd} inside the buffered domain gives

$$\left\| \frac{G^\top W - \mathbb{E}\left[G^\top W\right]}{n} \right\|_{L^2} \le C_{\mathcal{D}}\left(\ell_{\mathrm{fd}} + \frac{1}{\ell_{\mathrm{fd}}\sqrt{n}}\right) = o(1), \tag{86}$$

where $n \to \infty$ precedes $\ell_{\mathrm{fd}} \downarrow 0$.

Conditional on the data, $D_G W = -\sigma J$ and the second-order Gaussian Stein identity (Bellec and Zhang, 2018) gives

$$\mathbb{E}\left[(G^\top W + \sigma\,\mathrm{tr}(\Sigma_x J))^2\right] = \mathbb{E}\left[W^\top \Sigma_x W\right] + \sigma^2 \mathbb{E}\left[\mathrm{tr}((\Sigma_x^{1/2} J \Sigma_x^{1/2})^2)\right] \le C_{\mathcal{D}} n. \tag{87}$$

The response bound in Lemma B.2 justifies the identity (87) by Gaussian $W^{1,2}$ approximation. First-order integration by parts centers the left-hand side of (87). Since σ is bounded below on $\mathcal{D}$, (86) proves (84). Finally, condition on the unforced J, which is independent of x:

$$\mathbb{E}\left[|x^\top Jx/n - \kappa|\right] \le \frac{K_Q}{n}\mathbb{E}\left[\|J\|_{\mathrm{hs}}\right] + \|\mathrm{tr}(\Sigma_x J)/n - \kappa\|_{L^2} \le \frac{K_Q\sqrt{p}}{v\underline{c}_x n} + o(1).$$

This proves the scalar response; the quadratic-form estimate (9) is never conditioned on a response depending on x through a nonzero force. □

Lemma C.2 (From unforced to finite scalar response)**.** *Suppose the assumptions of Lemma B.4 and the quadratic-form estimate* (9) *hold. Then*

$$\sup_{\mathcal{D}} \omega_{\mathrm{lin}} \longrightarrow 0. \tag{88}$$

Proof. A bounded force is a translation of the actual Gaussian probe. Truncate only the fresh-feature norm so that this change of measure has bounded L^2 density, then integrate the response along the force interval. For $\sigma := \sqrt{2b}$, force r replaces G by $G - rx/(\sigma\sqrt{n})$; the translated Gaussian density is

$$\exp\left(-\frac{rx^\top \Sigma_x^{-1} G}{\sigma\sqrt{n}} - \frac{r^2 x^\top \Sigma_x^{-1} x}{2\sigma^2 n}\right).$$

On $A_R := \{|x| \le R\sqrt{n}\}$ its conditional L^2 norm is bounded uniformly for $|r| \le L_1$. Hence, for nonnegative measurable H,

$$\mathbb{E}\left[\mathbf{1}_{A_R} H\left(G - \frac{rx}{\sigma\sqrt{n}}\right)\right] \le C_{\mathcal{D},R}\left(\mathbb{E}\left[\mathbf{1}_{A_R} H(G)^2\right]\right)^{1/2}. \tag{89}$$

Here H may also depend on the fresh feature and cavity data, which are held fixed in the Gaussian change of measure. With $H(G) := |x^\top J_-(G)x/n - \kappa_-|$, (85) and $H \le C_{\mathcal{D}}(1 + R^2)$ on A_R give

$$\sup_{|r| \le L_1} \mathbb{E}\left[|x^\top J_-(r)x/n - \kappa_-|\right] \le C_{\mathcal{D},R}\sqrt{\varepsilon_{\mathrm{resp}}(\mathcal{D})} + C_{\mathcal{D}}\mathbb{E}_x\left[(1 + |x|^2/n)\mathbf{1}_{A_R^c}\right] = o(1).$$

The quantity $\mathbb{E}_x\left[(1 + |x|^2/n)\mathbf{1}_{A_R^c}\right]$ vanishes by the feature-tail conclusion of Lemma B.1, first taking $n \to \infty$ and then $R \to \infty$. Absolute continuity and Fubini give $q_x(r) - q_x(0) = \int_0^r x^\top J_-(s)x/n\,\mathrm{d}s$ almost surely. Thus

$$\omega_{\mathrm{lin}} \le \int_{-L_1}^{L_1} \mathbb{E}\left[|x^\top J_-(s)x/n - \kappa_-|\right]\,\mathrm{d}s = o(1).$$

The supremum in ω_{lin} also covers a random insertion force; no forced-optimizer prediction comparison is needed. □

Lemma C.3 (Normalized interior covariance)**.** *Under the assumptions of Lemma B.4, full and cavity minimizers satisfy*

$$\sup_{\mathcal{D}} \|C_W\|_{\mathrm{op}}/n \longrightarrow 0, \qquad \sup_{\mathcal{D}} \|C_W\|_{\mathrm{hs}}/n \longrightarrow 0. \tag{90}$$

Only the projection and feature-norm fourth-moment bounds in Lemma B.1 are needed, without prediction comparison or additional quadratic-form concentration.

Proof. We bound each directional variance by replacing one observation through an independent cavity. For deterministic $|u| = 1$, independence at zero force and the projection fourth-moment bound in Lemma B.1 give $\mathbb{E}\left[|u^\top J_- x|^4\right] \le C_{\mathcal{D}}$. The Gaussian-shift inequality (89) and the feature-tail conclusion of Lemma B.1 yield

$$\sup_{|r|\le L_1} \mathbb{E}\left[|u^\top J_-(r)x|^2/n\right] \le C_{\mathcal{D},R}/n + C_{\mathcal{D}}/R^2. \tag{91}$$

Indeed, apply (89) to $|u^\top J_- x|^2/n$ on A_R, and use $\|J_-(r)\|_{\mathrm{op}} \le (v\underline{c}_x)^{-1}$ on its complement. Insertion gives $W_+ = W_-(r_*)$ for $r_* := -L'_y(x^\top W_+/\sqrt{n}) \in [-L_1, L_1]$. Integrating along the entire force interval and applying Cauchy–Schwarz bounds $\mathbb{E}\left[|u^\top(W_+ - W_-)|^2\right]$ by $C_{\mathcal{D},R}/n + C_{\mathcal{D}}/R^2$, the right-hand side of (91). Efron–Stein over the observation blocks and Gaussian Poincaré then give

$$\mathrm{Var}(u^\top W) \le C_{\mathcal{D},R} + C_{\mathcal{D}} n/R^2 + 2b\mathbb{E}\left[u^\top J\Sigma_x J u\right] \le C'_{\mathcal{D},R} + C_{\mathcal{D}} n/R^2.$$

Divide by n, take $n \to \infty$ and then $R \to \infty$, uniformly in u. Finally, $\operatorname{tr} C_W \le C_{\mathcal{D}} n$ and $\|C_W\|_{\mathrm{hs}}^2 \le \|C_W\|_{\mathrm{op}} \operatorname{tr} C_W$ prove $\|C_W\|_{\mathrm{hs}}/n \to 0$ in (90). □

C.3 Bounding the three errors

Proposition C.4 (Scalar-response error)**.** *Under the assumptions of Lemma B.2, with $v, b > 0$, one has $|r^{(1)}| \le L_1\omega_{\mathrm{lin}}$. Under the additional assumptions of Lemma C.2, $L_1\omega_{\mathrm{lin}} \to 0$ uniformly on $\mathcal{D}$.*

Proof. For $F_x(r) := \min_w\{\Psi_-(w) - rx^\top w/\sqrt{n}\}$, the envelope identity $F'_x(r) = -q_x(r)$ gives

$$F_x(r) - F_x(0) = -ur - \kappa_- r^2/2 - \int_0^r (q_x(s) - u - \kappa_- s)\,\mathrm{d}s. \tag{92}$$

Fenchel duality, applicable since the loss is finite and continuous, gives

$$\Delta = \sup_z \left\{ \int_0^z q_x(-s)\,\mathrm{d}s - L_y^*(z) \right\}, \qquad e_{L_y}(u;\kappa_-) = \sup_z\{zu - \kappa_- z^2/2 - L_y^*(z)\}.$$

Both suprema use $\operatorname{dom} L_y^* \subset [-L_1, L_1]$. Their integrands differ by at most $L_1 \sup_{|r|\le L_1} |q_x(r) - u - \kappa_- r|$; take expectations. □

Proposition C.5 (Gaussian-comparison error)**.** *Suppose Assumption 1 and the comparison* (21) *in Conjecture 1 hold. Then $|r^{(2)}| \le L_1\varepsilon_{\mathrm{pred}}(\mathcal{D})$ uniformly on $\mathcal{D}$.*

Proof. Condition on the independent pair (x, y). Independence of W_- from this pair makes its score law depend only on x. Since $e_{L_y}(\cdot;\kappa_-)$ is L_1-Lipschitz uniformly in y, Wasserstein duality and averaging give

$$|r^{(2)}| \le L_1 \mathbb{E}_x\left[\mathsf{W}_1\left(\mathcal{L}(x^\top W_-/\sqrt{n} \mid x), N(x^\top \mu_-, \alpha_-^2)\right)\right].$$

The comparison (21) uses the deterministic variance α_-^2 appearing in the Hamiltonian. □

Proposition C.6 (Cavity-to-full error)**.** *Suppose Assumption 1 holds, $\underline{c}_x I_p \preceq \Sigma_x \preceq K I_p$, and $p/n \le K$. Let $\mathcal{D}$ be a fixed interior domain introduced in Section 2.3, and suppose its defining bounds remain valid on a fixed enlargement used for finite differences. Then, uniformly on $\mathcal{D}$,*

$$|r^{(3)}| \le L_1\omega_\mu + \frac{L_2}{2}\omega_{\alpha^2} + \frac{L_1^2}{2}\omega_\kappa. \tag{93}$$

All three discrepancies vanish uniformly, without prediction comparison or feature quadratic-form concentration.

Proof. Lemma B.7 bounds changes of the Hamiltonian in mean, variance and response by the three terms on the right-hand side of (93). Couple W, W_- using the same data and probe. Strong monotonicity of the cavity objective gives, when the deleted observation is present,

$$|W_+ - W_-| \le L_1|x|/(v\underline{c}_x\sqrt{n}). \tag{94}$$

Consequently $D := \|W - W_-\|_{L^2} \le C_{\mathcal{D}}$, $\omega_\mu \le \sqrt{K}D/\sqrt{n}$, and, by (68), $\omega_{\alpha^2} \le C_{\mathcal{D}}D/\sqrt{n}$. For the response, use the actual-probe correlation formula:

$$\omega_\kappa = \frac{|\mathbb{E}\left[G^\top(W - W_-)\right]|}{n\sqrt{2b}} \le \frac{\sqrt{\mathrm{tr}(\Sigma_x)/n}\, D}{\sqrt{2bn}}. \tag{95}$$

Since $D \le C_{\mathcal{D}}$, (68) and (95) show that $\omega_\mu, \omega_{\alpha^2}, \omega_\kappa \to 0$ uniformly on $\mathcal{D}$. □

Proof of Theorem 5. Propositions C.4–C.6 and Lemma C.2 give

$$\sup_{\mathcal{D}} |r^{(j)}| \longrightarrow 0 \qquad (j = 1, 2, 3). \tag{96}$$

Equation (83) yields

$$\sup_{\mathcal{D}} |\partial_t\phi - \widetilde{\mathcal{H}}(\nabla\phi)| \le \sum_{j=1}^{3} \sup_{\mathcal{D}} |r^{(j)}| \longrightarrow 0. \tag{97}$$

□

C.4 Responses needed for the Lasso support bound

The proof of Lemma E.3 needs a local active-coordinate formula and a response evaluated at the observation-selected insertion force. Controlling this response requires the change of variables in Lemma C.8; fixed-force estimates alone do not control the random insertion force.

Lemma C.7 (Local active-coordinate response)**.** *Let W minimize $\rho_0(w) + \lambda\|w\|_1 - \ell^\top w$, with ρ_0 convex and C^2 and $\lambda > 0$. Suppose its support S and active signs σ_S satisfy strict inactive slack $|\ell_j - \nabla_j\rho_0(W)| < \lambda$ for $j \notin S$, and $[\nabla^2\rho_0(W)]_{SS} \succ 0$. Then the global minimizer is unique and, locally in ℓ, has the same support and signs. It is continuously differentiable with*

$$D_\ell W = P_S([\nabla^2\rho_0(W)]_{SS})^{-1}P_S^\top, \tag{98}$$

where P_S embeds the active coordinates.

Proof. Apply the implicit-function theorem to $[\nabla\rho_0(W)]_S + \lambda\sigma_S = \ell_S$, $W_{S^c} = 0$. In a small neighborhood, signs, strict slack and restricted positivity persist, so this branch satisfies the full KKT conditions and is a global minimizer. To identify it as the unique optimizer, let a be its ℓ_1 subgradient. For any displacement d, convexity gives

$$\rho_0(W + d) + \lambda\|W + d\|_1 - \ell^\top(W + d) - \{\rho_0(W) + \lambda\|W\|_1 - \ell^\top W\} \ge \lambda\sum_{j\notin S}(1 - |a_j|)|d_j|.$$

Another minimizer must therefore have $d_{S^c} = 0$. Convexity would make its joining segment constant, contradicting $d_S^\top[\nabla^2\rho_0(W)]_{SS}d_S > 0$ unless $d = 0$. Differentiating the active KKT equations gives (98) (with zero derivative for empty support). The formula concerns local support constancy, as in partial-smoothness sensitivity (Lewis et al., 2022; Lewis and Zhang, 2013); finite forces may cross thresholds. □

Lemma C.8 (Response at the optimizing insertion force)**.** *Suppose the assumptions of Lemma C.2 hold. Let $W_+(G)$ include a fresh observation (x, y) and let $W_-(G)$ be its unforced cavity, with $G \sim N(0, \Sigma_x)$ and $\sigma := \sqrt{2b}$. Conditional on the cavity data and fresh pair, set*

$$u_-(z) := x^\top W_-(z)/\sqrt{n}, \qquad T(z) := z - \frac{L_y'(u_-(z))}{\sigma\sqrt{n}}x.$$

Then T is a bijection, and almost everywhere

$$\det DT(z) = 1 + L_y''(u_-(z))x^\top J_-(z)x/n \ge 1, \qquad W_+(T(z)) = W_-(z). \tag{99}$$

For $z = T^{-1}(G)$,

$$J_+(G) = J_-(z) - \frac{L_y''(u_-(z))}{1 + L_y''(u_-(z))x^\top J_-(z)x/n} \frac{J_-(z)xx^\top J_-(z)}{n}, \tag{100}$$

$$\sup_{\mathcal{D}} \mathbb{E}\left[|x^\top J_-(T^{-1}(G))x/n - \kappa_-|\right] \longrightarrow 0. \tag{101}$$

Proof. We absorb the inserted loss into the Gaussian field. The resulting map has a positive rank-one Jacobian, which controls its density and transfers the unforced response to the optimizing force. Substitution in the KKT inclusions gives $W_+(T(z)) = W_-(z)$, while

$$\nabla_z u_-(z) = -\sigma J_-(z)x/\sqrt{n}, \qquad DT(z) = I + L_y''(u_-(z))xx^\top J_-(z)/n.$$

The map preserves coordinates orthogonal to x. On every parallel line, monotonicity gives a longitudinal slope at least one, and the displacement is bounded by $L_1|x|/(\sigma\sqrt{n})$; hence it is onto and one-to-one. The Lipschitz dependence of u_- also controls transverse changes, so T and its inverse are Lipschitz for fixed x. The rank-one determinant formula proves (99).

Changing variables on each affine line parallel to x gives the density of $T^{-1}(G)$ relative to $N(0, \Sigma_x)$:

$$\Lambda_T(z) := \left[1 + L_y''(u_-(z))\frac{x^\top J_-(z)x}{n}\right] \exp\left(\frac{L_y'(u_-(z))x^\top \Sigma_x^{-1} z}{\sigma\sqrt{n}} - \frac{L_y'(u_-(z))^2 x^\top \Sigma_x^{-1} x}{2\sigma^2 n}\right).$$

On $A_R := \{|x| \le R\sqrt{n}\}$ the determinant is bounded, and $|L_y'| \le L_1$ gives a uniform conditional L^2 density bound using Gaussian exponential moments. Thus

$$\mathbb{E}\left[\mathbf{1}_{A_R} H(T^{-1}(G))\right] \le C_{\mathcal{D},R}\left(\mathbb{E}\left[\mathbf{1}_{A_R} H(G)^2\right]\right)^{1/2} \tag{102}$$

for nonnegative H, with its other arguments fixed conditionally. For $H(G) := |x^\top J_-(G)x/n - \kappa_-|$, (85) and $H \le C_{\mathcal{D}}(1 + R^2)$ on A_R bound the right side by $C_{\mathcal{D},R}\sqrt{\varepsilon_{\mathrm{resp}}(\mathcal{D})}$. The complement contributes at most $C_{\mathcal{D}}\mathbb{E}_x\left[(1 + |x|^2/n)\mathbf{1}_{A_R^c}\right]$. The feature-tail conclusion of Lemma B.1 shows that $\mathbb{E}_x[(1 + |x|^2/n)\mathbf{1}_{A_R^c}]$ vanishes when $n \to \infty$ is taken before $R \to \infty$, proving (101).

Finally, the bi-Lipschitz map preserves null sets. Differentiating $W_+(G) = W_-(T^{-1}(G))$ almost everywhere and applying the rank-one inverse formula to DT proves (100). The change of variables (99) permits active-set changes along the insertion force. □

D Direct comparison and calibrated transport

We transport the population moments along $(v, b, h) = (1 - t)(\nu_*, \beta_*^2/2, \eta_*)$ to every fixed cutoff $t = 1 - \epsilon$. A quadratically penalized value increment measures the squared gradient discrepancy and satisfies a scalar differential inequality. A direct small-time coupling starts the transport proof without using Conjecture 1 at $t = 0$.

Recall $z = (v, b, h)$ and $\pi = \nabla_z \phi$ from (60). The gradient of ϕ determines the moments through (63).

Lemma D.1 (Spatial regularity at positive probes)**.** *Under the assumptions of Lemma B.3, fix $0 < v_- \le v \le v_+ < \infty$, $0 < b_- \le b \le b_+ < \infty$, and $|h| \le R_h < \infty$. Then $\nabla_z \phi$ is bounded and Lipschitz, with dimension-independent constants uniform in $t \in [0, 1]$. The value is jointly C^1 in time and space, with one-sided time derivatives at the endpoints. On bounded subsets of the convex physical domain*

$$\mathcal{P} := \{\pi : -\pi_b \ge 0, \quad 2\pi_v \ge \pi_h^\top \Sigma_x \pi_h\},$$

∇H_* is Lipschitz with dimension-independent constant L_H, and

$$|H_*(\pi) - H_*(\pi') - \nabla H_*(\pi')^\top(\pi - \pi')| \le \frac{L_H}{2}|\pi - \pi'|^2 \tag{103}$$

on each segment in the bounded subset. Boundary gradients mean their continuous extensions.

Proof. Couple optimizers at z, z' with identical observations, selectors and probe. For $v, v' \geq v_\circ > 0$, the monotonicity estimate (56) gives

$$\underline{c}_x v_\circ \frac{|W - W'|}{\sqrt{n}} \leq |h - h'| + |\sqrt{2b} - \sqrt{2b'}| \frac{|G|}{\sqrt{n}} + K|v - v'| \frac{|W'|}{\sqrt{n}}.$$

Lemma B.3 implies $\|W - W'\|_{L^2}/\sqrt{n} \leq C|z - z'|$. Equations (62) and (60) express the spatial-gradient components through $-\mathbb{E}[W]/\sqrt{n}$, $\mathbb{E}\left[W^\top \Sigma_x W\right]/(2n)$ and $\mathbb{E}\left[G^\top W\right]/(n\sqrt{2b})$. Applying the coupling bound to these quantities gives the spatial bounds. At each finite size, Bernoulli averaging is a finite sum over retained subsets with polynomial time weights. Each expected subset value is C^1 by the moment bound (57) and the optimizer Lipschitz estimate (56), giving joint C^1 regularity.

Lemma B.10 and the Lipschitz coordinate map (63) bound ∇H_*. The set $\mathcal{P}$ is the intersection of a half-space with the epigraph of a convex quadratic. Integrating the gradient along its segments gives (103); boundary segments follow by translating both endpoints by $\delta(e_v - e_b)$ and letting $\delta \downarrow 0$. □

Lemma D.2 (Gradient transport from a bounded value error)**.** *Fix π_* and set $z(t) := z(s) - (t - s)\nabla H_*(\pi_*)$ for $s \leq t \leq T$. Suppose ϕ is jointly C^1 on a neighborhood of $\{(t, z(t) + d) : s \leq t \leq T,\ |d| \leq r_0\}$, with one-sided time derivatives at the endpoints, spatial gradient bounded by M and L-Lipschitz, and $|\pi_*| \leq M$. Assume* (103) *holds on a convex set $\mathcal{P}_0$ containing π_* and $\nabla_z \phi(t, z(t) + d)$ for $s \leq t \leq T$ and $|d| \leq r_0$. If $\phi_t - H_*(\nabla_z \phi) = \mathcal{R}$ with $|\mathcal{R}| \leq e$ on the tube, then*

$$\sup_{s \leq t \leq T} |\nabla_z \phi(t, z(t)) - \pi_*|^2 \leq C\big(|\nabla_z \phi(s, z(s)) - \pi_*|^2 + e\big), \tag{104}$$

where C depends only on $M, L, L_H, r_0, T - s$.

Proof. We represent the squared gradient error by a maximum of value differences. Its envelope derivative involves only first derivatives of ϕ, so the value residual is never differentiated. Choose $A > L$ with $Ar_0 > 2M$, and define

$$\mathcal{E}(t) := \max_{|d| \leq r_0} \{\phi(t, z(t) + d) - \phi(t, z(t)) - \pi_*^\top d - A|d|^2/2\}.$$

The objective is $(A - L)$-strongly concave and has outward radial derivative at most $2Mr_0 - Ar_0^2 < 0$ on the boundary. Its unique maximizer d_t is therefore interior, continuous in t, and satisfies

$$\nabla_z \phi(t, z(t) + d_t) - \pi_* = A d_t, \qquad |d_t| \leq \frac{|\nabla_z \phi(t, z(t)) - \pi_*|}{A - L}.$$

Spatial Taylor bounds, using $d = (\nabla_z \phi(t, z(t)) - \pi_*)/(A + L)$ for the lower bound, give

$$\frac{|\nabla_z \phi(t, z(t)) - \pi_*|^2}{2(A + L)} \leq \mathcal{E}(t) \leq \frac{|\nabla_z \phi(t, z(t)) - \pi_*|^2}{2(A - L)}. \tag{105}$$

The test point lies in the ball because $2M < Ar_0$. The envelope theorem and $\dot{z} = -\nabla H_*(\pi_*)$ yield

$$\begin{aligned} \mathcal{E}'(t) = {} & H_*(\pi_* + A d_t) - H_*(\nabla_z \phi(t, z(t))) \\ & - \nabla H_*(\pi_*)^\top (\pi_* + A d_t - \nabla_z \phi(t, z(t))) \\ & + \mathcal{R}(t, z(t) + d_t) - \mathcal{R}(t, z(t)). \end{aligned}$$

Both Hamiltonian arguments are actual gradients in $\mathcal{P}_0$. Equation (103) around π_* therefore gives

$$\mathcal{E}'(t) \leq \frac{L_H}{2}(A^2 |d_t|^2 + |\nabla_z \phi(t, z(t)) - \pi_*|^2) + 2e \leq C_0 \mathcal{E}(t) + 2e,$$

where $C_0 := L_H(A + L)(1 + A^2/(A - L)^2)$. Gronwall's inequality and (105) prove (104). □

D.1 Forward transport from the exact population root

Lemma D.3 (Forward transport from the calibrated population solution)**.** *Along a sequence in which the training pairs are independent copies of (x, y), suppose $p/n \to c \in (0, \infty)$, $\underline{c}_x I_p \preceq \Sigma_x \preceq K I_p$, Assumptions 1 and 2 hold, the comparison* (21) *in Conjecture 1 holds, and* (22) *holds. For fixed $0 < \epsilon < 1$, set*

$$W^\epsilon := W_{1-\epsilon,\, \epsilon\nu_*,\, \epsilon\beta_*^2/2,\, \epsilon\eta_*}, \qquad q^\epsilon := (\mu^\epsilon, (\alpha^\epsilon)^2, \kappa^\epsilon). \tag{106}$$

The components of q^ϵ are the actual moments and response of W^ϵ defined in (17)–(18)*, and*

$$\|q^\epsilon - q_*\|_1 \longrightarrow 0, \qquad q_* = (\mu_*, \alpha_*^2, \kappa_*). \tag{107}$$

No penalty/design alternative from Assumption 3 is required.

Proof. A direct coupling controls the initial short time interval. We then use Theorem 5 and Lemma D.2 to transport the gradient along the remaining interval. Lemma A.3, Proposition A.4, and (22) give constants $0 < c < C < \infty$ such that, for all sufficiently large sizes,

$$|\mu_*| + \alpha_* + \kappa_* + \nu_* + \beta_* + |\eta_*| \le C, \qquad \min\{\nu_*, \beta_*\} \ge c.$$

Set

$$z_* := (\nu_*, \beta_*^2/2, \eta_*), \qquad z(t) := (1-t)z_*, \qquad \pi_* := \nabla_z \phi(0, z_*).$$

Equations (14) and (5a) identify the optimizer at zero observations as W_*, so $\nabla H_*(\pi_*) = z_*$ by Lemma B.5. For fixed ϵ, choose a tube of dimension-independent radius about this path on $[0, 1-\epsilon]$ with positive probe lower bounds. Lemma D.1 gives uniform M, L, L_H there; its gradients and π_* lie in a bounded convex subset of $\mathcal{P}$.

Couple W_{s,z_*} and W_* with the same probe for small $s > 0$. Strong convexity, objective comparison and nonnegative retained losses give

$$\frac{1}{n}\mathbb{E}\left[|W_{s,z_*} - W_*|^2\right] \le Cs\,\mathbb{E}_{x,y,G}\left[L_y(x^\top W_*/\sqrt{n})\right] \le Cs. \tag{108}$$

Independence of the fresh pair, the bound $\mathbb{E}_G\left[|W_*|^2\right]/n = O(1)$ in (22), and $L_y(u) \le L_0 + L_1|u|$ imply $\mathbb{E}_{x,y,G}[L_y(x^\top W_*/\sqrt{n})] \le C$, which proves the bound $\mathbb{E}[|W_{s,z_*} - W_*|^2]/n \le Cs$ in (108). Changing z_* to $z(s)$ costs $O(s)$ in normalized L^2 by Lemma D.1; (63) gives $|\nabla_z \phi(s, z(s)) - \pi_*|^2 \le Cs$. For fixed s, ϵ, Theorem 5 supplies a uniform value error tending to zero on the remaining tube. Hence Lemma D.2 gives

$$\limsup_{n\to\infty} |\nabla_z \phi(1-\epsilon, \epsilon z_*) - \pi_*|^2 \le C_\epsilon s.$$

Let $s \downarrow 0$ after the dimension limit and use the Lipschitz envelope map (63). The cutoff remains fixed throughout. □

E CUTOFF REMOVAL AND PROOF OF THEOREM 2

Lemma D.3 identifies the moments at the calibrated cutoff (106). Proposition E.2 removes the cutoff under uniform strong convexity. For pure Lasso, Lemma E.3 gives the strict support gap (112), Lemma E.4 gives the localization tail (115), and Lemma E.5, together with inactive-coordinate slack, gives the convexity-remainder lower bound (128). Combining it with (124), (126), and (127) proves (121), with the dimension limit taken before the cutoff limit.

Lemma E.1 (Existence and measurable endpoint selection)**.** *Suppose Assumption 1 holds together with one of the two alternatives in Assumption 3. The prescribed empirical optimizer exists and is measurable. It is unique under* (12)*; under* (10)*, the minimum-norm selection is unique.*

Proof. The continuous empirical objective F is coercive by (10) or (12). Under (10), its minimizer set is compact and convex. If $W^\dagger$ is its minimum-norm element, the unique minimizers $W_k := \arg\min_w\{F(w) + |w|^2/(2k)\}$ satisfy

$$|W_k| \le |W^\dagger|, \qquad F(W_k) - \min F \le |W^\dagger|^2/(2k),$$

so $W_k \to W^\dagger$. Strongly convex continuous objectives have measurable minimizers by approximation over a countable dense set; this proves measurability of each W_k, its limit, and the unique minimizer in the strongly convex alternative. □

Proposition E.2 (Calibrated-cutoff removal under strong convexity)**.** *Under the assumptions of Lemma D.3, suppose that ρ is strongly convex with dimension-independent modulus $\gamma_{\mathrm{sc}} > 0$. Couple $\widehat{W}$ and W^ϵ using the same observations. For all sufficiently large sizes and $0 < \epsilon < 1$,*

$$\frac{1}{n}\mathbb{E}\left[|\widehat{W} - W^\epsilon|^2\right] \le C\epsilon. \tag{109}$$

Proof. We restore the omitted observations and remove the probes separately. Let $I := I(1-\epsilon)$ and $\overline{W} := \arg\min_w F_I(w)$, where $F_I(w) := \rho(w) + \sum_{i\in I} L_{y_i}(x_i^\top w/\sqrt{n})$. Comparison with zero and strong convexity give

$$\frac{|\widehat{W}|^2}{n}, \ \frac{|\overline{W}|^2}{n} \le \frac{2L_0}{\gamma_{\mathrm{sc}}}, \qquad \frac{\gamma_{\mathrm{sc}}}{2}|\widehat{W} - \overline{W}|^2 \le \sum_{i\notin I} L_{y_i}(x_i^\top \overline{W}/\sqrt{n}). \tag{110}$$

Conditional on the selectors and retained observations, each omitted pair is independent of $\overline{W}$. Since $L_y(z) \le L_0 + L_1|z|$,

$$\frac{1}{n}\mathbb{E}\left[|\widehat{W} - \overline{W}|^2\right] \le \frac{2\epsilon}{\gamma_{\mathrm{sc}}}\left(L_0 + L_1\sqrt{2KL_0/\gamma_{\mathrm{sc}}}\right).$$

Strong monotonicity of ∂F_I compares the probed and unprobed retained minimizers:

$$|W^\epsilon - \overline{W}| \le \frac{|\epsilon\nu_*\Sigma_x\overline{W} + \sqrt{\epsilon}\,\beta_* G - \epsilon\sqrt{n}\,\eta_*|}{\gamma_{\mathrm{sc}} + \epsilon\nu_*\underline{c}_x}.$$

Proposition A.4, (110), and $\mathbb{E}_G\left[|G|^2\right] \le Kp$ bound the squared expectation divided by n by $C\epsilon$. The squared triangle inequality proves (109). □

E.1 The support gap

Lemma E.3 (An upper bound for the interior Lasso support)**.** *Along a sequence in which the training pairs are independent copies of (x, y), suppose $p/n \to c \in (0, \infty)$, $\underline{c}_x I_p \preceq \Sigma_x \preceq K I_p$, and Assumptions 1 and 2 hold. Let $\rho(w) = \lambda\|w\|_1$, $\lambda > 0$. For $W := W_{t,v,b,h}$ and its deterministic response κ, let $\mathcal{D}$ be a fixed interior domain introduced in Section 2.3, and suppose its defining bounds remain valid on a fixed enlargement used for finite differences. Then*

$$\sup_{\mathcal{D}} \mathbb{E}\left[\left(\frac{|\operatorname{supp}(W)|}{n} - v\kappa - t\frac{L_2\kappa}{1 + L_2\kappa}\right)_+\right] \to 0, \qquad u_+ := \max\{u, 0\}. \tag{111}$$

If the assumptions of Lemma D.3 also hold, there are $s_0 < 1$ and $\epsilon_1 > 0$ such that, for each fixed $0 < \epsilon < \epsilon_1$,

$$\mathbb{P}\{|\operatorname{supp}(W^\epsilon)| > s_0 n\} \to 0. \tag{112}$$

Proof. The active-coordinate trace identity reduces support size to scalar responses. These responses are evaluated at observation-selected insertion forces, so we use Lemma C.8. Write $S := \operatorname{supp}(W)$, $J := D_h W/\sqrt{n}$, $q_i := x_i^\top W/\sqrt{n}$, and $H := v\Sigma_x + n^{-1}\sum_i B_i(t)L''_{y_i}(q_i)x_i x_i^\top$. Inactive KKT inequalities are strict almost surely: conditional on G_{-j}, the optimizer constrained by $w_j = 0$ is independent of G_j, and its jth residual is affine with nonzero slope in a Gaussian of variance $1/(\Sigma_x^{-1})_{jj} > 0$. Lemma C.7 therefore gives $J = P_S H_{SS}^{-1} P_S^\top$, hence the samplewise identity

$$\frac{|S|}{n} = v\frac{\operatorname{tr}(\Sigma_x J)}{n} + \frac{1}{n}\sum_i B_i(t)L''_{y_i}(q_i)\frac{x_i^\top J x_i}{n}. \tag{113}$$

For $f(u) := L_2 u/(1 + L_2 u)$, the rank-one formula in Lemma C.8 bounds each summand by

$$L''_{y_i}(q_i)\frac{x_i^\top J x_i}{n} \le f\left(\frac{x_i^\top J_{-i}(-L'_{y_i}(q_i))x_i}{n}\right).$$

The matrix on the right depends on (x_i, y_i). Conditional on $B_i(t) = 1$, (101), the L_2-Lipschitz property of f, and (95) replace this term by $f(\kappa)$ with uniform $o_{L^1}(1)$ error. Also, Lemma C.1 gives $\operatorname{tr}(\Sigma_x J)/n = \kappa + o_{L^1}(1)$, while $n^{-1}\sum_i B_i(t) = t + o_{L^1}(1)$. Substitution proves (111) without a prediction CLT or anti-concentration.

Set $K_\kappa := 1 + \sup_n \kappa_*$, which is finite by (22). Lemma D.3 gives $\kappa^\epsilon \le K_\kappa$ eventually at each fixed cutoff. With $\delta := (1 + L_2 K_\kappa)^{-1}$ and $\nu_* \le L_2$,

$$\epsilon\nu_*\kappa^\epsilon + (1 - \epsilon)\frac{L_2\kappa^\epsilon}{1 + L_2\kappa^\epsilon} \le 1 - \delta + \epsilon L_2 K_\kappa \le 1 - \delta/2$$

for sufficiently small fixed ϵ. Markov's inequality in (111), with $s_0 := 1 - \delta/4$, proves (112). □

E.2 Localized loss curvature

Lemma E.4 supplies the fixed $O(\sqrt{n})$ ball and the exceptional-moment bound (115). Lemma E.5 supplies loss curvature on directions concentrated on fewer than n coordinates; its event is required only to have vanishing failure probability.

Lemma E.4 (Localization from a bounded Lasso root)**.** *Along a sequence in which the training pairs are independent copies of (x, y), suppose $p/n \to c \in (0, \infty)$, $\underline{c}_x I_p \preceq \Sigma_x \preceq K I_p$, and Assumptions 1 and 2 hold. Let $\rho(w) = \lambda\|w\|_1$, with $0 < \lambda_- \le \lambda \le \lambda_+ < \infty$, and suppose $\sup_n\{\mathbb{E}_G\left[|W_*|^2\right]/n + \kappa_*\} < \infty$. There are fixed $R, r_h > 0$ and events $\mathcal{L}_n$, depending only on observations and selectors, with $\mathbb{P}(\mathcal{L}_n^c) \le Ce^{-cn}$ on which, simultaneously for $t \in [1/2, 1]$, $v \ge 0$, and $|h| \le r_h$, the objective*

$$\lambda\|w\|_1 + \sum_{i \in I(t)} L_{y_i}(x_i^\top w/\sqrt{n}) + \frac{v}{2} w^\top \Sigma_x w - \sqrt{n}\, h^\top w \tag{114}$$

is coercive and all its minimizers have norm less than $R\sqrt{n}$. Here $I(t) := \{i : U_i \le t\}$ for independent uniform selectors U_i. Moreover,

$$\mathbb{E}\left[\frac{|\widehat{W}|^2}{n} \mathbf{1}_{\mathcal{L}_n^c}\right] \le Cne^{-cn} \to 0. \tag{115}$$

Proof. We first derive population linear growth from the surrogate KKT condition and local Moreau-loss curvature. A fourth-moment complexity bound and fixed-height clipping then transfer growth to the empirical objective on a fixed sphere.

Set $f(w) := \mathbb{E}_{x,y,Z}\left[e_{L_y}(x^\top w/\sqrt{n} + \alpha_* Z; \kappa_*)\right]$. The population equations supply $s(G) \in \partial\|W_*(G)\|_1$ with

$$\frac{\lambda}{n}\mathbb{E}_G\left[s(G)\right] = -\nabla f(\mathbb{E}_G\left[W_*\right]), \qquad \frac{\lambda}{n}\|w\|_1 \ge -\nabla f(\mathbb{E}_G\left[W_*\right])^\top w. \tag{116}$$

The averaged subgradient supplies this linear bound; it need not be a subgradient at the mean. On $|w - \mathbb{E}_G\left[W_*\right]| \le \sqrt{n}$,

$$u^\top \nabla^2 f(w) u = \frac{1}{n}\mathbb{E}\left[(x^\top u)^2 \frac{L_y''(\zeta)}{1 + \kappa_* L_y''(\zeta)}\right] \ge c_{\mathrm{lin}}|u|^2/n, \quad \zeta := \mathrm{prox}_{\kappa_* L_y}(x^\top w/\sqrt{n} + \alpha_* Z).$$

Indeed, the projection fourth-moment bound in Lemma B.1 gives $\mathbb{E}\left[(x^\top u)^2 \mathbf{1}_{\{|x^\top w/\sqrt{n}| > T\}}\right] \le C|u|^2/T$; choose fixed large T, restrict also to $|Z| \le 1$, and use local loss curvature. Integrating along rays from the mean, then using (116) and $f(w) \le \mathbb{E}\left[L_y(x^\top w/\sqrt{n})\right] + L_1\alpha_*\mathbb{E}\left[|Z|\right]$, gives

$$\mathbb{E}\left[L_y(x^\top w/\sqrt{n})\right] + \lambda\|w\|_1/n \ge c_{\mathrm{lin}}|w|/\sqrt{n} - C_{\mathrm{lin}}. \tag{117}$$

Because the penalty is nonnegative, (117) remains valid with $c_{\mathrm{lin}}/2$ when the loss is multiplied by $1/2$.

For $\mathcal{T}_{R,D} := \{w : |w| \le R\sqrt{n},\ \lambda\|w\|_1/n \le D\}$, the projection fourth moments imply uniformly bounded fourth moments of $n^{-1/2}\sum_i \varepsilon_i B_i(1/2)x_{i,j}$, where ε_i are independent Rademacher signs. Splitting its coordinates at height T yields the decisive complexity estimate

$$\mathbb{E}\left[\sup_{w \in \mathcal{T}_{R,D}} \left|\frac{1}{n\sqrt{n}} \sum_i \varepsilon_i B_i(1/2) x_i^\top w\right|\right] \le \inf_{T>0}\{DT/\lambda_- + CR/T\} \le C\sqrt{RD/\lambda_-}. \tag{118}$$

For $\ell_{H,y} := \min\{L_y, H\}$, symmetrization and contraction give the deviation bound in (118), up to constants; its clipping bias is $C(1 + R^2)/H$. Bounded differences adds at most one to the deviation except with probability $2e^{-2n/H^2}$.

Choose small r_h with $r_h + C\sqrt{r_h/\lambda_-} < c_{\mathrm{lin}}/8$ and set $D := L_0 + r_h R + 2$. On $|w| = R\sqrt{n}$, penalties exceeding nD make (114) larger than $n(L_0 + 1)$. For the remaining vectors, monotonicity of $I(t)$ and (117) give the normalized lower bound

$$c_{\mathrm{lin}}R/2 - C_{\mathrm{lin}} - r_h R - C\sqrt{R(L_0 + r_h R + 2)/\lambda_-} - C.$$

Choose fixed large R to make this exceed $L_0 + 1$, then fix H so that the clipping bias is at most one. The quadratic term is nonnegative and the value at zero is at most nL_0, so convexity along rays proves that (114) is coercive and that all its minimizers satisfy $|w| < R\sqrt{n}$. Finally, $\lambda_-\|\widehat{W}\|_1 \le nL_0$ on every sample; multiplication by $\mathbb{P}(\mathcal{L}_n^c) \le Ce^{-cn}$ proves (115). □

Lemma E.5 (Uniform curvature near coordinate spaces below the rank threshold)**.** *Along a sequence in which the training pairs are independent copies of (x, y), suppose $p/n \to c \in (0, \infty)$, $\underline{c}_x I_p \preceq \Sigma_x \preceq K I_p$, Assumptions 1 and 2 hold, and* (11) *holds. For fixed $0 \le s < 1$ and $R < \infty$, there are $\tau, \epsilon_0, \gamma_R > 0$ and events $\mathcal{A}_n$ with $\mathbb{P}(\mathcal{A}_n^c) \to 0$ such that*

$$\frac{1}{n} \sum_{i \in I} L''_{y_i}(x_i^\top w/\sqrt{n})(x_i^\top d)^2 \ge \gamma_R |d|^2 \tag{119}$$

simultaneously whenever $|S| \le sn$, $\|d_{S^c}\|_1 \le \tau\sqrt{n}|d|$, $|w| \le R\sqrt{n}$, and $|I^c| \le \epsilon_0 n$. The sets S and I may depend on the data.

Proof. An anti-concentration net argument gives many nonzero projections, uniformly over sparse directions. We extend it to small off-support ℓ_1 perturbations and discard rows with large scores.

First, for a fixed $B < \infty$,

$$\mathbb{P}\left\{ \sup_{|u|=1} \frac{1}{n} \sum_i |x_i^\top u| > B \right\} \le Cn^{-1} + Ce^{-cn^{1/4}}. \tag{120}$$

To see this, truncate at $|x_i| \le n^{3/4}$. The discarded average $T_n := n^{-1} \sum_i |x_i| \mathbf{1}_{\{|x_i| > n^{3/4}\}}$ satisfies $\mathbb{E}\left[T_n^4\right] \le C/n$ by the projection fourth moments. For the truncated class, symmetrization and contraction bound the expected centered supremum by $4\sqrt{\mathrm{tr}(\Sigma_x)/n}$; its variance is at most K and envelope at most $2n^{3/4}$. Theorem 2.3 of Bousquet (2002) bounds a fixed excess with probability $Ce^{-cn^{1/4}}$. Markov's inequality for T_n proves (120).

On the event

$$\left\{ \sup_{|u|=1} \frac{1}{n} \sum_i |x_i^\top u| \le B \right\},$$

whose complement is bounded in (120), choose $s < s' < 1$ and $1 - 4\omega_{\mathrm{ac}} > s'$. Nets of radius $\delta := \omega_{\mathrm{ac}} r/(4B)$ over all coordinate unit spheres of dimension at most $s'n$ have total size at most $2^p(1 + 2/\delta)^{s'n}$. For a fixed net point, anti-concentration and independence bound the probability that at least $(1 - 4\omega_{\mathrm{ac}})n$ projections have magnitude at most $2r$ by $2^n(Cr)^{(1-4\omega_{\mathrm{ac}})n}$. Thus the union bound is

$$2^{p+n}(1 + 2/\delta)^{s'n}(Cr)^{(1-4\omega_{\mathrm{ac}})n} \le e^{-cn}$$

for fixed small r. The strict exponent inequality is why every $s < 1$ is allowed. The first-absolute-moment bound extends the net conclusion to all sparse unit vectors, each with at least $3\omega_{\mathrm{ac}} n$ projections exceeding r in magnitude. This proves the needed data-selected uniformity directly; see Koltchinskii and Mendelson (2013) for small-ball methods more generally.

For $|d| = 1$, augment S by the largest $(s' - s)n$ off-support coordinates. The remaining Euclidean norm is at most $C\tau/\sqrt{s' - s}$. Choose τ small: (120) then leaves at least $2\omega_{\mathrm{ac}} n$ projections of d with magnitude at least $r/4$. Uniformly in $|w| \le R\sqrt{n}$, the fraction of rows with $|x_i^\top w/\sqrt{n}| > M$ is at most BR/M. Choose $BR/M \le \omega_{\mathrm{ac}}/2$ and $\epsilon_0 \le \omega_{\mathrm{ac}}/2$. At least $\omega_{\mathrm{ac}} n$ rows remain with squared projection at least $r^2/16$ and loss curvature at least $c_L(M)$, proving (119). The failure probability is only required to vanish, as in (120). □

E.3 Curvature, subgradient slack, and cutoff removal

Proposition E.6 (Mean-square removal of the calibrated cutoff)**.** *Under the pure-Lasso branch of Theorem 2, couple $W^\epsilon := W_{1-\epsilon, \epsilon\nu_*, \epsilon\beta_*^2/2, \epsilon\eta_*}$ and the minimum-norm empirical minimizer $\widehat{W}$ using the same observations. For sufficiently small fixed $\epsilon > 0$,*

$$\limsup_{n \to \infty} \frac{1}{n} \mathbb{E}\left[|\widehat{W} - W^\epsilon|^2 \right] \le C\epsilon^{1/4}. \tag{121}$$

Proof. We combine the averaged upper bound (124) with the curvature-or-slack lower bound (128). The exceptional contributions are controlled by (126) and (127) before the cutoff is sent to zero. Put

$I := I(1-\epsilon)$, $v := \epsilon\nu_*$, $\sigma := \sqrt{\epsilon}\beta_*$, $h := \epsilon\eta_*$, $d := \widehat{W} - W^\epsilon$, and set $J^\epsilon := D_h W^\epsilon/\sqrt{n}$ and $\kappa^\epsilon := \mathbb{E}\left[\mathrm{tr}(\Sigma_x J^\epsilon)\right]/n$. Lemma D.3 and Proposition A.4 give

$$\sup_{0<\epsilon<\epsilon_1} \limsup_{n\to\infty} \left\{\mathbb{E}\left[|W^\epsilon|^2\right]/n + \kappa^\epsilon\right\} < \infty, \qquad \beta_* \geq \beta_- > 0, \tag{122}$$

with $\beta_* \leq L_1$, $\nu_* \leq L_2$, and $|\eta_*| \leq C$.

Let $F_I(w) := \lambda\|w\|_1 + \sum_{i\in I} L_{y_i}(x_i^\top w/\sqrt{n})$. Optimality gives the nonnegative remainder

$$\begin{aligned} \mathcal{D}_\epsilon &:= F_I(\widehat{W}) - F_I(W^\epsilon) - (\sqrt{n}\,h - v\Sigma_x W^\epsilon - \sigma G)^\top d \geq 0, \\ \mathcal{D}_\epsilon &\leq \sum_{i\notin I} L_{y_i}(x_i^\top W^\epsilon/\sqrt{n}) + (v\Sigma_x W^\epsilon + \sigma G - \sqrt{n}\,h)^\top d. \end{aligned} \tag{123}$$

Conditional on $i \notin I$, (x_i, y_i) is independent of W^ϵ, which has the unforced cavity law. Thus the omitted-loss expectation is bounded using $L_y(z) \leq L_0 + L_1|z|$, (94), and (122) by $C\epsilon n + o_n(n)$. Lemma E.4 bounds $\mathbb{E}\left[|\widehat{W}|^2\right]/n$, so the v, $\sqrt{n}\,h$ terms also contribute at most $C\epsilon n + o_n(n)$. The Gaussian term must keep its sign: $\widehat{W}$ is independent of G and $D_G W^\epsilon = -(\sigma/\sqrt{n})D_h W^\epsilon = -\sigma J^\epsilon$, giving

$$\frac{\sigma}{n}\mathbb{E}\left[G^\top d\right] = -\frac{\sigma}{n}\mathbb{E}\left[G^\top W^\epsilon\right] = \sigma^2\kappa^\epsilon = \epsilon\beta_*^2\kappa^\epsilon.$$

Consequently, before any event restriction,

$$\limsup_{n\to\infty} \mathbb{E}\left[\mathcal{D}_\epsilon\right]/n \leq C\epsilon. \tag{124}$$

Let $s^\epsilon \in \lambda\partial\|W^\epsilon\|_1$ be the actual optimality subgradient and put $S_a := \{j : |s_j^\epsilon| > \lambda - a\}$, $0 < a < \lambda_-$. Active coordinates belong to S_a; the additional coordinates obey

$$\mathbb{E}\left[\#\{j : W_j^\epsilon = 0,\ |s_j^\epsilon| > \lambda - a\}\right] \leq \frac{2pa}{\sigma\sqrt{2\pi\underline{c}_x}}. \tag{125}$$

Indeed, conditional on the data and G_{-j}, the optimizer constrained by $w_j = 0$ is independent of G_j. Its jth residual is affine with slope $-\sigma$; the two threshold intervals have total length $2a$, and $\mathrm{Var}(G_j \mid G_{-j}) \geq \underline{c}_x$. This conditional density bound proves (125), also for correlated probes. For $s_0 < s_1 < 1$ from Lemma E.3, Markov's inequality therefore gives

$$\limsup_{n\to\infty} \mathbb{P}\{|S_a| > s_1 n\} \leq Ca/\sqrt{\epsilon}. \tag{126}$$

Take $R, r_h, \mathcal{L}_n$ from Lemma E.4 and set $\mathcal{L}_{n,\epsilon} := \mathcal{L}_n \cap \{|h - \sigma G/\sqrt{n}| \leq r_h\}$. For small fixed cutoffs, Gaussian concentration gives exponential failure probability and $|\widehat{W}|, |W^\epsilon| \leq R\sqrt{n}$ on $\mathcal{L}_{n,\epsilon}$, with R independent of ϵ. On $\mathcal{L}_{n,\epsilon}^c$, $|\widehat{W}|^2/n \leq nL_0^2/\lambda_-^2$ and the fixed-cutoff fourth moment in Lemma B.4 imply

$$\mathbb{E}\left[\frac{|\widehat{W}|^2 + |W^\epsilon|^2}{n}\mathbf{1}_{\mathcal{L}_{n,\epsilon}^c}\right] \to 0. \tag{127}$$

The W^ϵ term in (127) uses Cauchy–Schwarz and Lemma B.4 at fixed ϵ.

Let $\mathcal{A}_n$ denote the event in Lemma E.5, evaluated at $s = s_1$ and at the value of R chosen in Lemma E.4. On $\mathcal{A}_n \cap \mathcal{L}_{n,\epsilon}$, when $|S_a| \leq s_1 n$ and $|I^c| \leq \epsilon_0 n$, there are two possibilities. If $\|d_{S_a^c}\|_1 \leq \tau\sqrt{n}|d|$, integrate the loss curvature along $W^\epsilon + ud$, $0 \leq u \leq 1$. Otherwise the inactive coordinates outside S_a contribute penalty slack at least $a\|d_{S_a^c}\|_1$. Hence

$$\mathcal{D}_\epsilon \geq \min\{\gamma_R/2, a\tau/(2R)\}|d|^2 \geq ca|d|^2. \tag{128}$$

The event $\mathcal{A}_n$ is uniform over the selected set and the score segment. Both $\mathbb{P}(\mathcal{A}_n^c)$ and $\mathbb{P}\{|I^c| > \epsilon_0 n\}$ vanish when $\epsilon < \epsilon_0$; on $\mathcal{L}_{n,\epsilon}$ their mean-square contributions vanish because $|d|^2/n \leq 4R^2$. Define

$$\mathcal{G}_{n,\epsilon,a} := \mathcal{A}_n \cap \mathcal{L}_{n,\epsilon} \cap \{|S_a| \leq s_1 n\} \cap \{|I^c| \leq \epsilon_0 n\}.$$

Using nonnegativity of $\mathcal{D}_\epsilon$ to apply (124) on $\mathcal{G}_{n,\epsilon,a}$, then (126) and (127) on $\mathcal{G}_{n,\epsilon,a}^c$, gives

$$\limsup_{n\to\infty} \frac{1}{n}\mathbb{E}\left[|d|^2\right] \leq C\left(\frac{\epsilon}{a} + \frac{a}{\sqrt{\epsilon}}\right).$$

After this dimension limit, choose $a := \epsilon^{3/4}$ to obtain (121). □

Proof of Theorem 2. Proposition E.6 in the pure-Lasso branch and Proposition E.2 in the strongly convex branch give

$$\lim_{\epsilon\downarrow 0}\limsup_{n\to\infty}\mathbb{E}\left[|\widehat{W}-W^{\epsilon}|^2\right]/n=0.$$

Lemma D.3 identifies the cutoff moments at each fixed ϵ. By (7) and (17),

$$\frac{1}{n}\mathbb{E}\left[|W^{\epsilon}|^2\right]=|\mu^{\epsilon}|^2+\frac{1}{n}\operatorname{tr}C_{W^{\epsilon}}\le|\mu^{\epsilon}|^2+\frac{(\alpha^{\epsilon})^2}{\underline{c}_x},$$

so the convergence of μ^{ϵ} and $(\alpha^{\epsilon})^2$ in (107) bounds the normalized cutoff second moment. The L^2 triangle inequality together with Proposition E.6 or Proposition E.2 also bounds $\mathbb{E}[|\widehat{W}|^2]/n$. Proposition B.9 therefore transfers the normalized mean and covariance trace, proving (23) as $\epsilon\downarrow 0$. □

F Prediction laws and classification

The cutoff coupling transfers scalar score laws and normalized covariance bounds to the empirical optimizer. Classification then follows by smoothing the threshold, using $\liminf_n \alpha_*^2>0$ from (22).

Proof of Corollary 3. Couple W^{ϵ} and $\widehat{W}$ as in (106), independently of the fresh pair (x,y). Write $D_{\epsilon}:=\|\widehat{W}-W^{\epsilon}\|_{L^2}/\sqrt{n}$, $\mu^{\epsilon}:=\mathbb{E}\left[W^{\epsilon}\right]/\sqrt{n}$ and $(\alpha^{\epsilon})^2:=\operatorname{tr}(\Sigma_x C_{W^{\epsilon}})/n$. Define

$$\mathfrak{W}:=\mathbb{E}_x\left[\mathsf{W}_1\Big(\mathcal{L}(x^{\top}\widehat{W}/\sqrt{n}\mid x),N(x^{\top}\mu_*,\alpha_*^2)\Big)\right],$$
$$e_{\mathrm{pred}}(\epsilon):=\mathbb{E}_x\left[\mathsf{W}_1\big(\mathcal{L}(x^{\top}W^{\epsilon}/\sqrt{n}\mid x),N(x^{\top}\mu^{\epsilon},(\alpha^{\epsilon})^2)\big)\right].$$

Proposition B.9 and a common standard Gaussian give

$$\mathfrak{W}\le\sqrt{K}D_{\epsilon}+e_{\mathrm{pred}}(\epsilon)+\sqrt{K}|\mu^{\epsilon}-\mu_*|+\sqrt{2/\pi}\sqrt{|(\alpha^{\epsilon})^2-\alpha_*^2|}. \tag{129}$$

Independence from the entire fresh pair makes these score laws unchanged if the fresh label is also fixed, even when y depends on x. At fixed $\epsilon>0$, the comparison (21) in Conjecture 1 gives $e_{\mathrm{pred}}(\epsilon)\to 0$, and Lemma D.3 gives convergence of the cutoff moments. Proposition E.6 or Proposition E.2 gives $\lim_{\epsilon\downarrow 0}\limsup_n D_{\epsilon}=0$. Taking $n\to\infty$ at fixed ϵ and then $\epsilon\downarrow 0$ in (129) proves

$$\mathbb{E}_x\left[\mathsf{W}_1\Big(\mathcal{L}(x^{\top}\widehat{W}/\sqrt{n}\mid x),N(x^{\top}\mu_*,\alpha_*^2)\Big)\right]\longrightarrow 0. \tag{130}$$

For each fixed (x,y), Wasserstein duality bounds the discrepancy for f_y by its common Lipschitz constant Λ times the score-law distance conditional on x. Averaging over (x,y) gives $\Lambda\mathfrak{W}$, proving (24). □

Proof of Corollary 4. The comparison Gaussian has density at most $M_{\mathrm{dens}}:=(\sqrt{2\pi}\,\underline{\alpha})^{-1}$. Upper and lower ramps of width $\ell_{\mathrm{ramp}}>0$ around a half-line boundary give, for any scalar laws P,Q with this density bound on Q,

$$|P(I)-Q(I)|\le\mathsf{W}_1(P,Q)/\ell_{\mathrm{ramp}}+M_{\mathrm{dens}}\ell_{\mathrm{ramp}}, \tag{131}$$

for every open or closed half-line I; continuity of Q handles either endpoint convention. With $\operatorname{sign}_+(0)=1$, the error half-lines are $(-\infty,0)$ for $y=1$ and $[0,\infty)$ for $y=-1$. Condition on (x,y) to select the error half-line. By independence from the fresh pair, the score laws still depend only on x, so averaging bounds the error discrepancy by $\mathfrak{W}/\ell_{\mathrm{ramp}}+M_{\mathrm{dens}}\ell_{\mathrm{ramp}}$. Corollary 3, followed by $\ell_{\mathrm{ramp}}\downarrow 0$, proves convergence. The Gaussian conditional error is $\Phi_{\mathrm{N}}(-yx^{\top}\mu_*/\alpha_*)$, giving (25), including possible fitted-score ties. □

Corollary F.1 (Boundary fresh-score comparison and covariance bound)**.** *Under the assumptions of Theorem 2, $\mathcal{G}(\widehat{W})\to 0$ and $\|C_{\widehat{W}}\|_{\mathrm{hs}}/n\to 0$, where $\mathcal{G}$ is defined in Proposition B.9.*

Proof. For the cutoff coupling (106), define

$$D_\epsilon := \frac{\|\widehat{W} - W^\epsilon\|_{L^2}}{\sqrt{n}}, \qquad \mu^\epsilon := \frac{\mathbb{E}[W^\epsilon]}{\sqrt{n}}, \qquad (\alpha^\epsilon)^2 := \frac{1}{n}\operatorname{tr}(\Sigma_x C_{W^\epsilon}),$$

and

$$e_{\mathrm{pred}}(\epsilon) := \mathbb{E}_x\left[\mathsf{W}_1\big(\mathcal{L}(x^\top W^\epsilon/\sqrt{n} \mid x), N(x^\top \mu^\epsilon, (\alpha^\epsilon)^2)\big)\right].$$

Proposition B.9 gives

$$\mathcal{G}(\widehat{W}) \le e_{\mathrm{pred}}(\epsilon) + CD_\epsilon, \qquad \frac{\|C_{\widehat{W}}\|_{\mathrm{hs}}}{n} \le CD_\epsilon + \frac{\|C_{W^\epsilon}\|_{\mathrm{hs}}}{n}. \tag{132}$$

The factor $(\|W^\epsilon\|_{L^2} + \|\widehat{W}\|_{L^2})/\sqrt{n}$ in (68) is bounded in the ordered limit by Lemma D.3 together with Proposition E.6 in the pure-Lasso case or Proposition E.2 in the strongly convex case. Lemma C.3 makes $\|C_{W^\epsilon}\|_{\mathrm{hs}}/n$ in (132) vanish at each fixed cutoff. Taking $n \to \infty$ and then $\epsilon \downarrow 0$ proves $\mathcal{G}(\widehat{W}) \to 0$ and $\|C_{\widehat{W}}\|_{\mathrm{hs}}/n \to 0$. □

G Penalty formulas and a directional counterexample

We first specialize the surrogate to quadratic penalties, then to Lasso and elastic net, and check consistency with Gaussian linear-model formulas. The last subsection distinguishes Gaussian fresh scores from Gaussian coefficient projections. The penalty formulas use a finite population solution $(\mu_*, \alpha_*, \kappa_*, \beta_*, \nu_*, \eta_*)$, with $\beta_*, \nu_* > 0$, and the surrogate $W_*(G)$. Write U_*, ζ_*, ξ_* for the quantities in (4) evaluated at this solution. Throughout, $D_\eta W_*$ denotes differentiation with respect to the linear field η, evaluated at η_*, with all other coefficients fixed; it does not differentiate the calibration equations. Transfer of prediction moments requires all assumptions of Theorem 2; support transfer remains open. The squared-loss calculations are algebraic consistency checks, since squared loss does not satisfy Assumption 1.

G.1 Quadratic penalties

For $\rho(w) := w^\top H w/2$, with $H \succeq 0$, put $Q_* := (\nu_*\Sigma_x + H)^{-1}$. Ridge regularization corresponds to $H = \lambda_2 I_p$. Completing the square gives

$$\begin{aligned} W_*(G) &= \sqrt{n}\, Q_*\eta_* - \beta_* Q_* G, \\ \mathbb{E}_G[W_*] &= \sqrt{n}\, Q_*\eta_*, \qquad C_{W_*} = \beta_*^2 Q_*\Sigma_x Q_*, \end{aligned} \tag{133}$$

and the population equations reduce to

$$\begin{aligned} H\mu_* + \mathbb{E}[x\xi_*] = 0, \qquad \kappa_* &= \frac{1}{n}\operatorname{tr}(\Sigma_x Q_*), \\ \alpha_*^2 &= \frac{\beta_*^2}{n}\operatorname{tr}(\Sigma_x Q_* \Sigma_x Q_*). \end{aligned} \tag{134}$$

These identities hold for $H \succeq 0$ whenever a population solution exists. For $H \succ 0$, existence follows from Theorem 1 under its loss assumptions. The strongly convex branch of Theorem 2 additionally requires a uniform positive lower bound on the eigenvalues of H.

In the variational reduction, the unstarred coefficients remain free variables. For $\alpha, \kappa, \beta, \nu > 0$ and $\mathbb{E}[L_y(0)] < \infty$, maximizing (45) in η gives $\eta = (\nu\Sigma_x + H)\mu$ and

$$\begin{aligned} \sup_{\eta \in \mathbb{R}^p} \mathcal{J}(\mu, \alpha, \kappa; \beta, \nu, \eta) = \frac{\beta^2\kappa}{2} + \mathbb{E}_{x,y,Z}[e_{L_y}(x^\top\mu + \alpha Z; \kappa)] + \frac{1}{2}\mu^\top H\mu \\ - \frac{\nu\alpha^2}{2} - \frac{\beta^2}{2n}\operatorname{tr}[\Sigma_x(\nu\Sigma_x + H)^{-1}]. \end{aligned} \tag{135}$$

The normalized penalty is $\rho(\sqrt{n}\,\cdot)/n$, whose value at μ is $\mu^\top H\mu/2$. For centered features, this reduction agrees with Yaakoubi et al. (2026, equation (4)), whose feature covariance then equals $\Sigma_x = \mathbb{E}[xx^\top]$. This algebraic identity does not interchange the extrema in (46).

Proposition G.1 (Quadratic surrogate and loss consistency)**.** *For $\rho(w) := w^\top H w/2$, $H \succeq 0$, every finite population solution satisfies* (133)–(134)*. For squared loss $L_y(u) := (u-y)^2/2$ and $y := x^\top\theta_0 + \varepsilon$, with deterministic θ_0 and independent centered noise of variance σ^2, put*

$$T_* := \frac{1}{n}\operatorname{tr}(\Sigma_x Q_* \Sigma_x Q_*), \qquad \Delta_* := (\mu_* - \theta_0)^\top \Sigma_x (\mu_* - \theta_0).$$

If $1 - \nu_*^2 T_* > 0$, then

$$\begin{aligned} \nu_* &= \frac{1}{1+\kappa_*}, \qquad \mu_* = \nu_* Q_* \Sigma_x \theta_0, \\ \frac{1}{2}\mathbb{E}[(x^\top\mu_* + \alpha_* Z - y)^2] &= \frac{\sigma^2 + \Delta_*}{2(1-\nu_*^2 T_*)}, \end{aligned} \tag{136}$$

where $Z \sim N(0,1)$ is independent of (x,y).

Proof. Completing the square gives (133) and $D_\eta W_* = \sqrt{n}\, Q_*$. Substitution into the mean, variance, response, and field equations gives (134). For arbitrary $\nu > 0$ and $\beta \geq 0$, the expected surrogate minimum is

$$-\frac{n}{2}\eta^\top(\nu\Sigma_x + H)^{-1}\eta - \frac{\beta^2}{2}\operatorname{tr}[(\nu\Sigma_x + H)^{-1}\Sigma_x].$$

Substituting this expression into (48), after restoring the loss constant, gives the maximizing field and (135), also at $\beta = 0$.

The relation $U_* - \zeta_* = \kappa_*\xi_*$ and Gaussian integration by parts give

$$\beta_*^2 = \frac{\mathbb{E}[(U_* - \zeta_*)^2]}{\kappa_*^2}, \qquad \nu_*\alpha_* = \frac{\mathbb{E}[Z(U_* - \zeta_*)]}{\kappa_*}.$$

Together with (134), these recover the quadratic equations of Yaakoubi et al. (2026, Theorem 4.4) for centered features. For squared loss,

$$\xi_* = \frac{U_* - y}{1+\kappa_*}, \qquad \eta_* = \nu_*\Sigma_x\theta_0, \qquad \beta_*^2 = \nu_*^2(\sigma^2 + \Delta_* + \alpha_*^2).$$

Combining these identities with $\alpha_*^2 = \beta_*^2 T_*$ proves (136). For centered features, the risk formula agrees with Yaakoubi et al. (2026, Corollary 5.1). □

G.2 Lasso and elastic-net formulas

Consider

$$\rho(w) := \lambda_1\|w\|_1 + \frac{\lambda_2}{2}\|w\|_2^2, \qquad \lambda_1 > 0, \quad \lambda_2 \geq 0,$$

where $\|w\|_2 = |w|$ in the notation of the main text. Pure Lasso corresponds to $\lambda_2 = 0$.

Proposition G.2 (Diagonal sparse surrogate and moment equations)**.** *For the above penalty and diagonal $\Sigma_x = \operatorname{diag}(c_1, \ldots, c_p) \succ 0$, the surrogate W_* and its moments satisfy* (26)–(27)*, evaluated at the starred population coefficients.*

Proof. Write $G_j = \sqrt{c_j}Z_j$, with independent $Z_j \sim N(0,1)$, and put

$$\Gamma_{*,j} := \nu_* c_j + \lambda_2, \qquad q_{*,j} := \sqrt{n}\,\eta_{*,j} - \beta_*\sqrt{c_j}Z_j.$$

Scalar optimality gives $q_{*,j} - \Gamma_{*,j}W_{*,j} \in \lambda_1\partial|W_{*,j}|$, hence

$$W_{*,j} = \frac{\operatorname{soft}(q_{*,j}, \lambda_1)}{\Gamma_{*,j}}.$$

Taking means and variances gives the equations for μ_* and α_*^2. Almost everywhere, $\partial_{\eta_j} W_{*,j} = \sqrt{n}\,\Gamma_{*,j}^{-1}\mathbf{1}_{\{|q_{*,j}|>\lambda_1\}}$. Gaussian integration by parts yields

$$\kappa_* = \frac{1}{n\sqrt{n}}\mathbb{E}_G[\operatorname{tr}(\Sigma_x D_\eta W_*)] = -\frac{1}{\beta_* n}\mathbb{E}_G[G^\top W_*], \tag{137}$$

which gives the response equation in (27). Independence here concerns the Gaussian probe coordinates; diagonal second moments do not imply independent feature coordinates. □

Corollary G.3 (Correlated sparse surrogate response)**.** *For the same penalty and a general $\Sigma_x \succ 0$, put $S_* := \mathrm{supp}(W_*(G))$, and let P_{S_*} embed its coordinates into $\mathbb{R}^p$. Almost surely,*

$$D_\eta W_* = \sqrt{n}\, P_{S_*} \big(\nu_*(\Sigma_x)_{S_*S_*} + \lambda_2 I_{|S_*|}\big)^{-1} P_{S_*}^\top. \tag{138}$$

For pure Lasso, $\nu_*\kappa_* = \mathbb{E}_G[|S_*|]/n$.

Proof. The optimizer is piecewise affine in $\sqrt{n}\,\eta_* - \beta_* G$, whose nondegenerate Gaussian law avoids its finitely many activation boundaries. Lemma C.7 gives (138), with zero derivative when S_* is empty. For $\lambda_2 = 0$, $\mathrm{tr}(\Sigma_x D_\eta W_*) = \sqrt{n}\,|S_*|/\nu_*$; (137) proves the support-size identity. At $\lambda_1 = 0$, (138) follows directly from (133) with $H = \lambda_2 I_p$, and the surrogate has full support almost surely. The identity is related to active-set degrees of freedom (Tibshirani and Taylor, 2012); it does not establish support transfer to $\widehat{W}$. □

G.3 Gaussian linear-model consistency

Proposition G.4 (Gaussian linear-model consistency)**.** *Let $L_y(u) := (u-y)^2/2$, $x \sim N(0, I_p)$, and $y := x^\top\theta_0 + \varepsilon$, with deterministic θ_0 and independent centered noise of variance σ^2. For $\rho(w) := \lambda_1\|w\|_1 + \lambda_2\|w\|_2^2/2$, with $\lambda_1 > 0$ and $\lambda_2 \geq 0$, put*

$$\tau_* := \frac{\beta_*}{\nu_*}, \qquad \mathsf{S}_*(z) := \frac{\mathrm{soft}(z, \lambda_1/\nu_*)}{1 + \lambda_2/\nu_*}.$$

Every finite population solution satisfies

$$\begin{aligned} \tau_*^2 &= \sigma^2 + \frac{1}{n}\sum_{j=1}^p \mathbb{E}_G[(W_{*,j} - \sqrt{n}\,\theta_{0,j})^2], \\ 1 - \nu_* &= \frac{1}{n}\sum_{j=1}^p \mathbb{E}_Z[\mathsf{S}_*'(\sqrt{n}\,\theta_{0,j} + \tau_* Z)]. \end{aligned} \tag{139}$$

Suppose λ_1, λ_2 are fixed, $\nu_* \to \nu_\infty > 0$, $\tau_* \to \tau_\infty > 0$, $p/n \to c$, and the empirical law of $\sqrt{n}\,\theta_{0,j}$ converges with second moments to the law of a random variable B. Writing $\mathsf{S}_\infty(z) := \mathrm{soft}(z, \lambda_1/\nu_\infty)/(1 + \lambda_2/\nu_\infty)$, one obtains

$$\begin{aligned} \tau_\infty^2 &= \sigma^2 + c\mathbb{E}[(\mathsf{S}_\infty(B + \tau_\infty Z) - B)^2], \\ \nu_\infty &= 1 - c\mathbb{E}[\mathsf{S}_\infty'(B + \tau_\infty Z)], \end{aligned} \tag{140}$$

where $Z \sim N(0,1)$ is independent of B.

Proof. For squared loss, $\xi_* = (U_* - y)/(1 + \kappa_*)$ and $\nu_* = (1+\kappa_*)^{-1}$. Since $\mathbb{E}[x\xi_*] = \nu_*(\mu_* - \theta_0)$, $\eta_* = \nu_*\theta_0$. The soft-threshold formula and Gaussian symmetry give

$$W_{*,j} \overset{\mathrm{law}}{=} \mathsf{S}_*(\sqrt{n}\,\theta_{0,j} + \tau_* Z).$$

Expanding $\mathbb{E}[(U_* - y)^2]$ and using $\mu_* = \mathbb{E}_G[W_*]/\sqrt{n}$ and $\alpha_*^2 = \mathrm{tr}(C_{W_*})/n$ gives the first equation in (139). Equation (137) gives

$$\nu_*\kappa_* = \frac{1}{n}\sum_{j=1}^p \mathbb{E}_Z[\mathsf{S}_*'(\sqrt{n}\,\theta_{0,j} + \tau_* Z)],$$

and $\nu_*\kappa_* = 1 - \nu_*$ gives the second equation. Gaussian averaging makes the response continuous and bounded, while the squared error has quadratic growth. The assumed convergence with second moments therefore gives (140). For pure Lasso, these are the state-evolution and threshold-calibration formulas of Bayati and Montanari (2012), with sample-to-dimension ratio $1/c$; the elastic-net map is the corresponding specialization of Thrampoulidis et al. (2018). These comparisons are algebraic. □

G.4 Why Gaussianity in every deterministic direction fails

The following example distinguishes the conditional fresh-score comparison (21) from a Gaussian approximation of every fixed coefficient projection. Its data are Gaussian, so the obstruction is not a lack of Poincaré concentration.

Proposition G.5 (A non-Gaussian coordinate under Gaussian data)**.** *Let $p = n$, let $x_i \sim N(0, I_n)$ be independent, and take $y_i \equiv 1$, binary logistic loss $L_y(u) := \log(1 + e^{-yu})$, and pure Lasso $\rho(w) := \lambda_1\|w\|_1$ with fixed $\lambda_1 > 0$ and $\lambda_2 = 0$. Set $W_t := W_{t,1,1/2,0}$ as in (15). The data law satisfies a dimension-independent Poincaré inequality, and there are constants $t_0 \in (0, 1)$ and $c_0 > 0$, independent of n, such that, for every coordinate j,*

$$\mathsf{W}_1(\mathcal{L}((W_{t_0})_j), N(\mathbb{E}[(W_{t_0})_j], \mathrm{Var}((W_{t_0})_j))) \geq c_0/2. \tag{141}$$

Nevertheless, for an independent fresh $x \sim N(0, I_n)$,

$$\begin{aligned} &\mathbb{E}_x\big[\mathsf{W}_1\big(\mathcal{L}(x^\top W_0/\sqrt{n} \mid x), N(0, s_{\lambda_1}^2)\big)\big] \longrightarrow 0, \\ &s_{\lambda_1}^2 := \mathbb{E}_Z[\mathrm{soft}(-Z, \lambda_1)^2], \qquad Z \sim N(0, 1). \end{aligned} \tag{142}$$

Proof. Here $\Sigma_x = I_n$. The Gaussian Poincaré inequality gives $\mathrm{Var}(F(x, 1)) \leq \mathbb{E}_x[|\nabla_x F(x, 1)|^2]$. Moreover, $\mathrm{Var}(x^\top Ax) = 2\|A\|_{\mathrm{hs}}^2$ for symmetric A, so (9) holds. At $t = 0$, the positive quadratic probe gives

$$\mathcal{E}_{0,1,1/2,0}(w) = \lambda_1\|w\|_1 + \tfrac{1}{2}\|w\|_2^2 + G^\top w, \qquad (W_0)_j = \mathrm{soft}(-G_j, \lambda_1), \quad G \sim N(0, I_n).$$

Thus the coordinates of W_0 are independent and symmetric, each with an atom at zero of mass $a_{\lambda_1} := \mathbb{P}(|Z| \leq \lambda_1) > 0$ and variance $s_{\lambda_1}^2 > 0$. Put $M_{\lambda_1} := (2\pi s_{\lambda_1}^2)^{-1/2}$ and let $N_{\lambda_1} \sim N(0, s_{\lambda_1}^2)$. The 1-Lipschitz test $f_d(z) := (d - |z|)_+$ satisfies

$$\mathbb{E}[f_d((W_0)_j)] \geq d\, a_{\lambda_1}, \qquad \mathbb{E}[f_d(N_{\lambda_1})] \leq M_{\lambda_1} d^2.$$

Choosing $d := a_{\lambda_1}/(2M_{\lambda_1})$ gives

$$\mathsf{W}_1(\mathcal{L}((W_0)_j), N(0, s_{\lambda_1}^2)) \geq c_0 := \frac{a_{\lambda_1}^2}{4M_{\lambda_1}} > 0. \tag{143}$$

Couple W_t and W_0 using the same Gaussian probe. Strong monotonicity gives

$$|W_t - W_0| \leq \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n B_i(t) x_i L'_{y_i}(x_i^\top W_0/\sqrt{n}) \right|.$$

Conditional on W_0, these summands are independent. Before the factor $1/\sqrt{n}$, each has second moment at most tn and mean of norm at most t: for every deterministic unit vector e,

$$\big|\mathbb{E}_{x,y}[(e^\top x) L'_y(x^\top W_0/\sqrt{n}) \mid W_0]\big| \leq (\mathbb{E}_x[(e^\top x)^2])^{1/2} = 1.$$

Consequently,

$$\frac{1}{n}\mathbb{E}[|W_t - W_0|^2] \leq t + \frac{n-1}{n} t^2 \leq t + t^2. \tag{144}$$

Permutation symmetry gives $\|(W_t)_j - (W_0)_j\|_{L^2} \leq \sqrt{t + t^2}$, and coordinate-sign symmetry gives $\mathbb{E}[(W_t)_j] = \mathbb{E}[(W_0)_j] = 0$. Coupling their moment-matched Gaussians with one standard normal and using the reverse triangle inequality in L^2 yields

$$\mathsf{W}_1(\mathcal{L}((W_t)_j), N(0, \mathrm{Var}((W_t)_j))) \geq c_0 - (1 + \sqrt{2/\pi})\sqrt{t + t^2}.$$

Choose

$$t_0 := \min\{1/4, c_0^2/[16(1 + \sqrt{2/\pi})^2]\} \in (0, 1/4].$$

This choice is independent of n, makes the subtracted term at most $c_0/2$, and lies in a fixed positive-probe interior domain. It proves (141).

For (142), note that $\mathbb{E}[W_0] = 0$ and $C_{W_0} = s_{\lambda_1}^2 I_n$. Conditional on fresh x, the score $x^\top W_0/\sqrt{n}$ is a sum of independent centered variables with variance $s_{\lambda_1}^2|x|^2/n$. Couple the coordinates of x

across dimensions as an infinite independent Gaussian sequence. Almost surely, $|x|^2/n \to 1$ and $\max_{j\le n} |x_j|/\sqrt{n} \to 0$. Together with the finite second moment of $\mathrm{soft}(-Z, \lambda_1)$, these properties give the conditional Lindeberg condition and hence weak convergence to $N(0, s^2_{\lambda_1})$. Bounded conditional second moments upgrade this to convergence in 1-Wasserstein distance. Finally,

$$\mathsf{W}_1(\mathcal{L}(x^\top W_0/\sqrt{n} \mid x), N(0, s^2_{\lambda_1})) \le s_{\lambda_1}|x|/\sqrt{n} + \sqrt{2/\pi}\, s_{\lambda_1}.$$

The right-hand side has uniformly bounded second moment over x, so uniform integrability permits averaging over x and proves (142). This zero-time calculation does not establish (21) at t_0. □

H SUPPORT-RECOVERY SIMULATION DETAILS

H.1 FEATURE LAWS, TEACHER, AND PENALTY NORMALIZATION

Let $n = 200$ and $p = 400$. For each observation and each of ten independent blocks, generate

$$Z_b^{(a)} = \frac{H_b + aS_b}{\sqrt{1+a^2}}, \qquad H_b \sim N(0,1), \quad \mathbb{P}(S_b = 1) = \mathbb{P}(S_b = -1) = \tfrac{1}{2}.$$

The non-Gaussian design uses $a = 8$; the matched Gaussian design uses $a = 0$. For coordinate $j = 1, \dots, 10$ in block b, set

$$x_{bj} = \sqrt{0.95}\, Z_b^{(a)} + \sqrt{0.05}\, E_{bj}, \qquad E_{bj} \sim N(0,1).$$

All primitive variables are independent across observations and blocks, and the E_{bj} are independent within blocks. The remaining 300 coordinates are independent standard Gaussians, independent of the blocks. Both designs have mean zero and covariance

$$\Sigma_x = \Sigma = \mathrm{diag}(C_{\mathrm{block}}, \dots, C_{\mathrm{block}}, I_{300}), \qquad C_{\mathrm{block}} = 0.05 I_{10} + 0.95 \mathbf{1}\mathbf{1}^\top.$$

For $a = 8$, the latent factor is a mixture of $N(-8/\sqrt{65}, 1/65)$ and $N(8/\sqrt{65}, 1/65)$ with equal weights. Figure 1(a) plots the exact marginal density of a block coordinate,

$$\tfrac{1}{2}N(-m, s_x^2) + \tfrac{1}{2}N(m, s_x^2), \qquad m = 8\sqrt{0.95/65}, \quad s_x^2 = 0.05 + 0.95/65,$$

and the exact $N(0,1)$ density. Here $m^2 + s_x^2 = 1$. Singleton coordinates have the same Gaussian marginal in both designs.

The first two blocks contain twenty signal coordinates, each with coefficient $b = \sqrt{0.9/(20 \cdot 9.55)} = \sqrt{0.9/191}$. Five singleton coordinates have coefficient $b_{\mathrm{iso}} = \sqrt{0.1/5}$; all other coefficients are zero. Thus $|S_0| = 25$ and $\theta_0^\top \Sigma \theta_0 = 1$, with block and singleton contributions 0.9 and 0.1. The coordinate groups are

$$S_{0,b} = \{1, \dots, 20\}, \qquad S_{0,i} = \{101, \dots, 105\},$$
$$S_{\setminus 0,b} = \{21, \dots, 100\}, \quad S_{\setminus 0,i} = \{106, \dots, 400\}.$$

For a selected support S, put $\mathrm{TP} = |S \cap S_0|$ and $\mathrm{FP} = |S \setminus S_0|$. Then $J(S, S_0) = \mathrm{TP}/(25 + \mathrm{FP})$, so Jaccard accounts for false selection as well as signal retention.

In both designs, use $y = \tanh(x^\top \theta_0 + \varepsilon)$ with independent $\varepsilon \sim N(0, 0.01^2)$. The pseudo-Huber loss and fitted objective are

$$L_y(u) = \delta_{\mathrm{H}}^2 \left(\sqrt{1 + \frac{(u-y)^2}{\delta_{\mathrm{H}}^2}} - 1 \right), \qquad \delta_{\mathrm{H}} = 0.2, \tag{145}$$

$$\widehat{\theta} = \arg\min_\theta \left\{ \frac{1}{n} \sum_{i=1}^n L_{y_i}(x_i^\top \theta) + \frac{\lambda_1}{\sqrt{n}} \|\theta\|_1 + \frac{\lambda_2}{2} \|\theta\|^2 \right\}. \tag{146}$$

No intercept or sample standardization is used. This objective is $1/n$ times the manuscript objective under $\widehat{\theta} = \widehat{W}/\sqrt{n}$. The nonlinear response transforms the teacher score, not the features; the support target remains that of θ_0.

Loss regularity. For fixed $\delta_{\mathrm{H}} = 0.2$, the loss is nonnegative, convex, and C^∞, with bounded derivatives of orders one through four. Since $|y| \le 1$,

$$|L_y'(u)| \le \delta_{\mathrm{H}}, \qquad L_y(0) \le \tfrac{1}{2},$$
$$\inf_{|y|\le 1,\, |u|\le M} L_y''(u) \ge \left(1 + ((M+1)/\delta_{\mathrm{H}})^2\right)^{-3/2} > 0.$$

The score at zero is nondegenerate because the teacher score has a nondegenerate continuous distribution. These facts verify the loss conditions of Assumption 1; they do not by themselves establish all hypotheses or the asymptotic population-moment regime of Theorem 2. The experiment is a finite-size empirical test of support predictions, for which a transfer theorem remains open.

H.2 Grid, ERM replication, and numerical validation

The two-dimensional grid is $\mathcal{G} = \mathcal{L}_1 \times \mathcal{L}_2$, with

$$\mathcal{L}_1 = \{0.05 + 0.75k/35 : k = 0, \dots, 35\}, \qquad \mathcal{L}_2 = \{10^{-3+k/9} : k = 0, \dots, 45\}.$$

There are $36 \times 46 = 1656$ points per design. The logarithmic spacing is $1/9$, up to $\log \lambda_2 = 2$. Figure 2(a) uses the same 46 ridge values at exactly $\lambda_1 = 0.2$. This value is a row of $\mathcal{G}$, so all slice predictions and ERM fits reuse that row. In total there are 3312 grid calibrations and no additional slice calibrations.

For each design we generate 500 independent datasets. A dataset is reused at every penalty pair, and the two designs share the primitive normal draws and signs within each replicate. The grid requires 1,656,000 ERM fits; the exact slice reuses these fits. Let $d \in \{\mathrm{NG}, \mathrm{G}\}$ index the non-Gaussian and covariance-matched Gaussian designs, respectively. For each $r = 1, \dots, 500$, let $\widehat{W}_d^{(r)}(\lambda_1, \lambda_2)$ denote the ERM estimator fitted to the rth independent training dataset from design d, with penalties (λ_1, λ_2). The empirical mean Jaccard is

$$\overline{J}_{500,d}(\lambda_1, \lambda_2) := \frac{1}{500} \sum_{r=1}^{500} J\Big(\mathrm{supp}\big(\widehat{W}_d^{(r)}(\lambda_1, \lambda_2)\big), S_0\Big).$$

This sample average estimates the expected Jaccard

$$\mathbb{E}_{\mathrm{training}}\Big[J\Big(\mathrm{supp}\big(\widehat{W}_d^{(1)}(\lambda_1, \lambda_2)\big), S_0\Big)\Big],$$

where the expectation is over the random training dataset for design d at the fixed penalty pair (λ_1, λ_2). The sample average uses individual Jaccard ratios, not a ratio of average counts. Similarly, J_d^* denotes the surrogate mean Jaccard J^* for that design. Numerical supports use $|\widehat{\theta}_j| > 10^{-8}$ for ERM and $|\theta_{*,j}| > 10^{-8}$ for the surrogate, where $\theta_* = W_*/\sqrt{n}$. The same definition is used throughout.

The optimizer uses accelerated proximal-gradient steps with backtracking, adaptive restart, and exact soft-thresholding for the ℓ_1 term. Neighboring penalty solutions provide warm starts. Across all fits, the maximum KKT residual is below 10^{-9} and the maximum absolute primal-dual gap is 4.64×10^{-15}. Eighteen representative fits and all 500 datasets at each design's largest grid discrepancy were refit with KKT tolerance 10^{-11}; the support metrics were unchanged. Increasing the selection threshold to 10^{-6} changes the ERM mean Jaccard by at most 0.497 percentage points over all grid and slice points.

For each plotted metric, open circles show the empirical mean over 500 training datasets; error bars are omitted. Reusing datasets induces dependence across penalty values.

The FPE computation evaluates (4)–(5) separately under each actual feature–response law. In particular, $\eta_* = \nu_* \Sigma \mu_* - \mathbb{E}[x\xi]$ is evaluated explicitly; the squared-loss identity $\eta_* = \nu_* \Sigma \theta_0$ is not imposed. Conditional on the sum of the two active block signs, the teacher score and surrogate input score are jointly Gaussian. The non-Gaussian loss expectation is a three-component mixture with weights $1/4, 1/2, 1/4$. The tanh response and the scalar pseudo-Huber proximal residual are evaluated inside the integrand. We use composite normal-weighted Gauss–Legendre quadrature in the teacher-score direction and initially 128-node Gauss–Hermite quadrature in the independent direction, checked at higher order.

The penalty channel uses 2^{12} antithetic scrambled-Sobol Gaussian probes for calibration; support evaluation uses a separate, independently scrambled ensemble of 2^{16} probes with the same antithetic construction. Singleton selection probabilities are analytic; independent block support-count distributions are combined to compute expected Jaccard. The maximum relative closure residual is 5.16×10^{-12}, and the maximum checked loss-quadrature discrepancy is 5.66×10^{-15}. Fourteen selected grid and slice points were recalibrated using new, independently scrambled antithetic Sobol Gaussian ensembles of 2^{14} probes for calibration and 2^{17} probes for evaluation, together with higher-order loss quadrature. Predicted Jaccard changed by at most 3.73×10^{-4}. These checks assess numerical sensitivity, not rigorous integration-error bounds.

H.3 Grid comparison and recovery profiles

For each design define the observed grid discrepancy

$$D_d = \max_{(\lambda_1,\lambda_2)\in\mathcal{G}} |J_d^*(\lambda_1,\lambda_2) - \overline{J}_{500,d}(\lambda_1,\lambda_2)|.$$

The maxima and their locations are

Design	$100D_d$ (pp)	λ_1	λ_2	J_d^*	$\overline{J}_{500,d}$
Non-Gaussian	0.74	0.735714	0.010000	0.528659	0.536080
Matched Gaussian	0.72	0.692857	0.001668	0.390684	0.397920

These quantities include Monte Carlo error from 500 datasets; selection of the maximum over the grid also affects their interpretation. They are not noise-free bounds on the surrogate approximation error.

The largest sampled between-design surrogate difference is $J_{\mathrm{NG}}^* - J_{\mathrm{G}}^* = 0.213799$, at $(\lambda_1,\lambda_2) = (0.221429, 0.00464159)$, where the predictions are 0.768126 and 0.554328. Along the exact $\lambda_1 = 0.2$ slice, the largest difference is 0.211278 at $\lambda_2 = 0.00599484$. Both designs attain their sampled slice maximum near $\lambda_2 = 0.04642$, with predicted Jaccard 0.7635 and 0.5691, respectively. These are exploratory grid summaries, not independently established tuning rules or global optima. The setting was selected through pilot experiments.

The declining right tail of Figure 2(a) reflects more than missed signals. In the non-Gaussian design, predicted mean false selections increase from about 7.79 at the slice maximum to 89.75 at $\lambda_2 = 100$, while $\mathrm{TPR}_{\mathrm{singleton}}$ falls from approximately one to 0.869. For the Gaussian design the corresponding false selection counts are 18.17 and 94.39, with $\mathrm{TPR}_{\mathrm{singleton}}$ falling to 0.730. $\mathrm{TPR}_{\mathrm{block}}$ is near one at the right endpoint in both designs. These support-count predictions explain why high block retention alone does not imply high Jaccard.

H.4 Fresh-score distributions and empirical histograms

Figure 2(b) uses the representative grid point $(\lambda_1,\lambda_2) = (0.2, 0.1)$, chosen as a moderate penalty. Write $\mu_{*,d}$ and $\alpha_{*,d}$ for the calibrated parameters of design $d \in \{\mathrm{NG}, \mathrm{G}\}$, and let x_d have that design's fresh-feature law. The predicted score is

$$u_d = x_d^\top \mu_{*,d} + \alpha_{*,d} Z, \qquad Z \sim N(0,1),$$

with Z independent of x_d. Its density is evaluated analytically: conditioning on the two active block signs gives a three-normal mixture with weights $1/4, 1/2, 1/4$ for the non-Gaussian design. The matched Gaussian score is normal, with mean zero and variance $\mu_{*,\mathrm{G}}^\top \Sigma_x \mu_{*,\mathrm{G}} + \alpha_{*,\mathrm{G}}^2$.

For each design d, each of the 5000 saved ERM fits $\widehat{W}_d^{(r)}(0.2, 0.1)$ is paired with one independently drawn fresh feature vector $x_d^{(r)}$, independent of its training dataset. Training datasets and fresh features are independent across replicates within each design. Each histogram contains the 5000 scores

$$\widehat{u}_{d,r} = \frac{(x_d^{(r)})^\top \widehat{W}_d^{(r)}(0.2, 0.1)}{\sqrt{n}}, \qquad r = 1, \ldots, 5000.$$

The first 500 fits reuse the grid experiment, and 4500 additional independent training datasets per design are fitted only at this penalty pair, using the same sampling model and solver tolerances. The

support-recovery averages in panel (a) still use 500 datasets per design. The saved coefficients are already $\widehat{\theta} = \widehat{W}/\sqrt{n}$, so no further normalization is applied when evaluating $x^\top \widehat{\theta}$. The two designs share primitive random draws within each replicate, as in the grid experiment. Each histogram has bin width 0.05, with heights equal to bin counts divided by the product of 5000 and the bin width. Bars occupy 80% of each bin for visual separation; normalization uses the full bin width. Each histogram observation thus uses a distinct fitted model and a distinct fresh feature. Neither the response transformation nor additional noise is applied to these fitted linear scores; training variability is represented by the fitted models.

Let F_d be the CDF of u_d. On the original score scale, we compute

$$\mathsf{W}_1\big(\mathcal{L}(u_{\mathrm{NG}}), \mathcal{L}(u_{\mathrm{G}})\big) = \int_{\mathbb{R}} |F_{\mathrm{NG}}(t) - F_{\mathrm{G}}(t)| \, \mathrm{d}t \simeq 0.10994428.$$

This compares the two predicted unconditional score laws, not an empirical-to-predicted discrepancy. Adaptive quadrature uses bounds extending twelve component standard deviations beyond every mixture mean; calculations at tolerances 10^{-10} and 10^{-12} agree within 10^{-8}. No score standardization is used.