THE SAMPLE COMPLEXITY OF ONE REWEIGHTED LEAST-SQUARES STEP

Anonymous authors
Paper under double-blind review

ABSTRACT

Recursive feature machines learn features by repeatedly refitting a predictor and reweighting its inputs. Their linear version generalizes iteratively reweighted least squares, which also underlies FOCUSS: each step fits the data by weighted least squares and reweights each coordinate by a power of its fitted value. Existing analyses describe the limit of this iteration, draw fresh data for every step, or hold only asymptotically; with fewer samples than features, the finite-sample cost of a step that reuses the data has been open. We bound it for the first step, which starts from the minimum-norm estimate $a = X^\dagger y$ and uses weights $|a_i|^q$. For a noiseless Gaussian design with n features and a fixed k-sparse target with equal magnitudes, $[n(k\log(n/\delta))^q]^{1/(q+1)}$ samples suffice for fixed relative error, and a high-probability lower bound matches the powers of n and k. Yet a already ranks the support correctly with order $k\log n$ samples. Between these scales, which differ by a factor polynomial in n, one step finds the relevant coordinates but does not use them: the many inactive coordinates, each with a small weight, together fit the observations at a lower weighted cost than the target. The gap persists for every ridge penalty and weight floor and closes when the power grows logarithmically in n. We extend the lower bound to unequal magnitudes and prove consistency under Gaussian noise with an oracle penalty. Experiments on the same simulated trials measure the gap and its cause, and show that a second step removes it.

1 INTRODUCTION

Recursive feature machines perform feature learning explicitly: they alternate between fitting a predictor and reweighting the inputs by the average gradient outer product of that predictor, a quantity proposed to capture how neural networks learn features (Radhakrishnan et al., 2025). For linear predictors on sparse problems, the method (lin-RFM) generalizes iteratively reweighted least squares (IRLS): each step fits the data by weighted least squares and then reweights each coordinate by a power of its fitted value. The same step drives FOCUSS and reweighted least-squares algorithms for diagonal linear networks (Gorodnitsky & Rao, 1997; Daubechies et al., 2010; Kaushik et al., 2024). One such step is therefore both a classical sparse-recovery update and the simplest model of a feature-learning step.

What one step learns from finite data matters for these methods. Reweighting schemes for diagonal networks reach low test error within a few steps (Kaushik et al., 2024), analyses of several steps apply one-step guarantees inductively (Zhu et al., 2025), and single-step analyses also describe feature learning in two-layer networks (Ba et al., 2022). Yet, with fewer samples than features, the sample cost of even the first reweighting step has not been determined when the step reuses the data. The iteration-wise theory of Kaushik et al. (2024) draws a fresh batch for every step and lists finite-sample guarantees and the setting without fresh batches as open directions.

The question is not which sparse estimator to use. For the Gaussian designs studied here, Lasso recovers the support of a k-sparse target from order $k\log n$ samples (Wainwright, 2009), and basis pursuit has comparable guarantees (Chen et al., 2001; Candès & Tao, 2005; Candès et al., 2006). We do not propose one reweighting step as a competitor. We ask what the step itself extracts from the data: whether it finds the relevant coordinates, and whether it turns them into an accurate fit. Guarantees for Lasso do not answer these questions about a feature-learning procedure.

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.

We study the first step from the minimum-norm estimate $a = X^\dagger y$, with weights $|a_i|^q$ and the same observations in both solves. Special cases are one update of Basic FOCUSS ($q = 1$), General FOCUSS ($q = 2$), unsmoothed ℓ_1-IRLS ($q = 1/2$), and lin-RFM with a power feature map (any q; Section 2). For a fixed k-sparse target with equal magnitudes and a Gaussian design, a already ranks the support correctly with order $k \log n$ samples (Theorem 1); it is the high-dimensional ordinary least-squares projection (HOLP) screening statistic (Wang & Leng, 2016; Wang et al., 2015). Keeping its k largest coordinates and refitting recovers the target exactly. The reweighting step needs more samples. For every fixed $q > 0$, the sample size needed for constant relative error scales as $n^{1/(q+1)} k^{q/(q+1)}$, with high-probability upper and lower bounds that match up to a logarithmic factor (Theorems 3 and 5). Between $k \log n$ and this scale, which differ by a factor polynomial in n, the step has found the relevant coordinates but does not use them: a ranks the support correctly, yet the reweighted fit keeps relative error at least 0.6 with high probability (Corollary 8). Each inactive coordinate receives a small weight, but the inactive coordinates are so many that together they fit the observations at a lower weighted cost than the target. Ridge penalties and weight floors do not change this. Powers that grow logarithmically in n remove the gap, and so does a second step in our experiments. Squared coefficient error equals excess prediction risk for isotropic test covariates, so the conclusions also hold for prediction.

Contributions.

- Matching upper and lower bounds on the sample cost of one step for every fixed power, with explicit dependence on the power (Theorems 3 and 5, Corollary 4).
- A separation between correct ranking and accurate estimation on the same data, for every ridge penalty and weight floor, with a computable lower bound that explains it (Proposition 6, Corollaries 7 and 8).
- Extensions to unequal magnitudes, Gaussian noise, and approximately sparse targets, and a correlated design where a fails to rank the support (Section 4).
- Same-trial experiments on ranking, error, and the lower bound, including a second step and sparse baselines (Section 5).

1.1 Relation to Existing Analyses

IRLS and one-step estimation. FOCUSS and smoothed IRLS have convergence theory (Gorodnitsky & Rao, 1997; Daubechies et al., 2010; Kümmerle et al., 2021). Broken adaptive ridge analyzes the limit of iterated quadratic penalties built from an initial estimate (Dai et al., 2018). One-step and multi-step statistical guarantees are also established for adaptive or reweighted ℓ_1 procedures (Zou, 2006; Zou & Li, 2008; Zhang, 2010; Fan et al., 2014). Our second-stage objective is quadratic and its output is generally dense. Hadamard parametrizations connect sparse penalties to smooth optimization (Hoff, 2017; Poon & Peyré, 2021); early-stopped gradient methods have a separate implicit-sparsity theory (Li et al., 2021). None of these results gives the sample cost of the first update.

Feature priming. Warmuth & Amid (2023) propose feature priming and pose an online adversarial-regret problem. Warmuth et al. (2025, Appendix D.4) analyze a two-stage regularized fit on reused observations for an overdetermined orthogonal design and a one-sparse target. Xu et al. (2026) study priming with power weights in the online setting; their lower bound also relies on interpolation through many low-weight nuisance coordinates. Our results concern batch estimation with an underdetermined Gaussian design, general sparsity and powers, and high-probability coefficient error. They complement the online results and do not resolve the regret problem of Warmuth & Amid (2023).

Iteration-wise random-design theory. Teramoto et al. (2021) use the replica method to study iterative reweighted Lasso. Kaushik et al. (2024) rigorously analyze successive quadratic updates for diagonal networks in a proportional limit, with independent batches and positive regularization. Zhu et al. (2025) prove asymptotic detection and learning guarantees for derivative-reweighted kernel machines on sparse polynomials, with fresh samples at each update. Reused-data analyses include iterated empirical-risk minimization (Cui & Lu, 2026) and finite-sample state evolution for algorithms with first-order and strongly convex–concave saddle-point updates (Celentano et al., 2025). Our

results treat the ridgeless, coordinate-weighted update directly, allow m/n to tend to zero, and give explicit upper and lower sample scales in terms of sparsity, dimension, power, and confidence, together with their separation from support screening.

2 Estimator and Statistical Setting

Let $X \in \mathbb{R}^{m\times n}$, $y = X\beta^\star$, and $S = \operatorname{supp}(\beta^\star)$ with $|S| = k$. Write X_S for the active columns and $X^\dagger$ for the Moore–Penrose pseudoinverse. Define

$$a = X^\dagger y, \qquad D_q = \operatorname{diag}(|a_1|^q, \dots, |a_n|^q), \qquad \widehat{\beta}_q = D_q(XD_q)^\dagger y. \tag{1}$$

We set $D_0 = I$. For positive weights and feasible observations,

$$\widehat{\beta}_q = \operatorname*{arg\,min}_{X\beta=y} \sum_{i=1}^{n} \frac{\beta_i^2}{|a_i|^{2q}}. \tag{2}$$

All coordinates of a are nonzero almost surely for a fixed nonzero target and an underdetermined Gaussian design. When weights can be zero, equation (1) remains the definition and the fit is confined to $\operatorname{range}(D_q)$.

Exact algorithmic correspondences. Each correspondence starts at a and uses the same data. Basic FOCUSS maps a to $A(XA)^\dagger y$, where $A = \operatorname{diag}(a)$ (Gorodnitsky & Rao, 1997, eq. (6)). Writing $A = D_1 J$ with a diagonal sign matrix gives $(XD_1J)^\dagger = J(XD_1)^\dagger$, so the update is $\widehat{\beta}_1$. For positive integer l, General FOCUSS uses entrywise powers a_i^l; the same sign cancellation gives $\widehat{\beta}_l$, including $q = 2$ (Gorodnitsky & Rao, 1997, eq. (8)). For $0 < p \le 1$, unsmoothed ℓ_p-IRLS minimizes $\sum_i \beta_i^2/|a_i|^{2-p}$ subject to $X\beta = y$ (Daubechies et al., 2010). Thus $q = 1 - p/2$, and $p = 1$ gives $q = 1/2$. Sparse lin-RFM starts with unit feature scales and updates them by $\phi(a_i^2)$ (Radhakrishnan et al., 2025, arXiv version, Appendix B, Algorithm 4). For $\phi(s) = s^\alpha$, its next coefficient vector is $\widehat{\beta}_{2\alpha}$. This covers every $q > 0$. We use no convergence guarantee for the smoothed or iterated methods.

Notation and assumptions. In Section 3, the entries of X are independent $N(0,1)$, the target is fixed independently of X, and there is no observation noise. Put

$$\beta_{\min} = \min_{i\in S} |\beta_i^\star|, \qquad \mu = \frac{\|\beta^\star\|_2^2}{\beta_{\min}^2} \ge k, \qquad E_q = \frac{\left\|\widehat{\beta}_q - \beta^\star\right\|_2}{\|\beta^\star\|_2}.$$

For the lower bound, define $k_{\text{eff}} = \|\beta^\star\|_2^2 \,/\, \|\beta^\star\|_\infty^2$. Thus $1 \le k_{\text{eff}} \le k \le \mu$, with equality throughout the last three quantities for equal-magnitude signals. In the noiseless results, probability is over X for the fixed target. Constants denoted by C, c are universal; a subscript q permits dependence on the fixed power. Norms of matrices are operator norms unless specified otherwise. For an independent x with $\mathbb{E}[xx^\top] = I$, the conditional excess prediction risk is

$$\mathbb{E}_x[(x^\top\widehat{\beta}_q - x^\top\beta^\star)^2 \mid X] = \left\|\widehat{\beta}_q - \beta^\star\right\|_2^2. \tag{3}$$

Top-k selection followed by a least-squares refit is a separate estimator, denoted by $\widehat{\beta}_{\text{hard}}$. Its set of fitted columns is $\operatorname{Top}_k(|a|)$. We do not identify it with a limit of equation (1) as $q \to \infty$.

3 Sample Complexity of One Update

We first distinguish coordinate ranking from coefficient accuracy. Since $a = P\beta^\star$ with $P = X^\dagger X$ a Haar-distributed rank-m projector,

$$\mathbb{E}_X \frac{\|a - \beta^\star\|_2^2}{\|\beta^\star\|_2^2} = 1 - \frac{m}{n}. \tag{4}$$

The identity follows from $P^2 = P$ and $\mathbb{E}P = (m/n)I$. Inactive coordinates contain only the fluctuation term:

$$a_i = \frac{m}{n}\beta_i^\star + e_i^\top(P - \tfrac{m}{n}I)\beta^\star, \qquad a_j = e_j^\top(P - \tfrac{m}{n}I)\beta^\star \quad (j \notin S).$$

Theorem 1 (Ranking by the minimum-norm estimate). *Let $0 < \delta < 1$ and $t = \log(8n/\delta)$. If $Ct \le m \le n/2$, then, with probability at least $1 - \delta$,*

$$\max_{j \notin S} |a_j| \le C \left\|\beta^\star\right\|_2 \frac{\sqrt{mt}}{n}, \tag{5}$$

$$\min_{i \in S} |a_i| \ge \frac{m}{n}\beta_{\min} - C \left\|\beta^\star\right\|_2 \frac{\sqrt{mt}}{n}. \tag{6}$$

Consequently, $m \ge C\mu t$ suffices for $\mathrm{Top}_k(|a|) = S$. On that event, $\widehat{\beta}_{\mathrm{hard}} = \beta^\star$ whenever X_S has full column rank.

The estimate a is the HOLP screening statistic (Wang & Leng, 2016). For isotropic designs and equal magnitudes, Wang et al. (2015, Theorem 6) give a screening condition of order $k^2 \log n$ that holds simultaneously over a class of signals. Theorem 1 instead fixes the target; this gives order $\mu \log n$ and the coordinate gap used below. With known k, hard selection uses the ordering directly, whereas finite-power reweighting keeps every inactive column.

Proposition 2 (From a weight ratio to coefficient error). *Let $p_i > 0$, $D = \mathrm{diag}(p)$, $y = X_S\beta^\star_S$, and X_S have full column rank. Define*

$$\widehat{\beta}(p) = \operatorname*{arg\,min}_{X\beta = y} \left\|D^{-1}\beta\right\|_2^2, \quad \rho = \frac{\max_{j \notin S} p_j}{\min_{i \in S} p_i}, \quad L = \left\|X_S^\dagger X_{S^c}\right\|_{\mathrm{op}}.$$

Then $\left\|\widehat{\beta}(p)_{S^c}\right\|_2 \le \rho \left\|\beta^\star\right\|_2$ and

$$\left\|\widehat{\beta}(p) - \beta^\star\right\|_2 \le \sqrt{1 + L^2}\, \rho \left\|\beta^\star\right\|_2. \tag{7}$$

Proof. Feasibility of $\beta^\star$ gives $\left\|D^{-1}\widehat{\beta}(p)\right\|_2 \le \left\|D^{-1}\beta^\star\right\|_2$. This bounds the inactive norm by $\rho \left\|\beta^\star\right\|_2$. Feasibility also gives $\widehat{\beta}(p)_S - \beta^\star_S = -X_S^\dagger X_{S^c}\widehat{\beta}(p)_{S^c}$. The active and inactive errors occupy disjoint coordinates, which proves equation (7). □

The ratio ρ bounds the norm of the fit outside the true support. The factor L bounds the error that this part of the fit causes on the active coordinates while the observations stay fixed. For Gaussian designs, L is at most order $\sqrt{n/m}$ with high probability once m exceeds order $k + \log(1/\delta)$. To make this upper bound small, the weight ratio must also offset the factor $\sqrt{n/m}$.

Theorem 3 (One-update upper bound). *Fix $q > 0$ and $0 < \delta < 1$. If $C\mu \log(8n/\delta) \le m \le n/2$, then, with probability at least $1 - \delta$,*

$$E_q \le C_q \sqrt{\frac{n}{m}} \left(\frac{\mu \log(8n/\delta)}{m}\right)^{q/2}. \tag{8}$$

In the same range of m, relative error at most $\varepsilon \in (0, 1)$ is guaranteed when

$$m \ge C_q \varepsilon^{-2/(q+1)} [n\{\mu \log(8n/\delta)\}^q]^{1/(q+1)}. \tag{9}$$

Proof mechanism. Theorem 1 gives $\rho \le C_q(\mu \log(8n/\delta)/m)^{q/2}$ for the weights $|a_i|^q$. Proposition 2 and a bound on L for Gaussian designs then give equation (8). The deterministic step permits the weights to depend on all of X and y; no independent second batch is introduced. Appendix A proves the concentration bounds and keeps their confidence-range conditions explicit.

The proof gives C_q explicitly, so the power may grow with the dimension.

Corollary 4 (Powers that grow with the dimension). *There are universal constants C_1, C_2 such that, under the conditions of Theorem 3, with probability at least $1 - \delta$, simultaneously for every $q > 0$,*

$$E_q \le C_1 \sqrt{\frac{n}{m}} \left(\frac{C_2\, \mu \log(8n/\delta)}{m}\right)^{q/2}.$$

Hence, if $m \ge 4C_2\, \mu \log(8n/\delta)$, then $E_q \le \varepsilon$ for every $q \ge \log_2\{C_1\sqrt{n/m}/\varepsilon\}$.

A fixed power pays a polynomial factor in the sample size. A power of order $\log(n/m) + \log(1/\varepsilon)$ reaches accuracy ε at the screening scale $\mu \log n$, at a numerical price: the weight condition number $(\max_i |a_i| / \min_i |a_i|)^q$ grows with q. Theorem 5 does not cover such powers, since its constant c_q decays exponentially in q.

Theorem 5 (One-update lower bound)**.** *Fix $q > 0$. There are universal constants $C_0, C, c > 0$ and a constant $c_q > 0$ depending only on q with the following property. Let $\beta^\star \neq 0$ be any fixed k-sparse target and put $N = n - k$. If*

$$k_{\text{eff}} \leq m, \qquad N \geq C_0 m, \qquad m^{q+1} \leq c_q N k_{\text{eff}}^q, \tag{10}$$

then $\mathbb{P}_X\{E_q \geq 0.6\} \geq 1 - Ce^{-cm}$.

Why soft weights can fail. Normalize $\|\beta^\star\|_2 = 1$. A constant fraction of the inactive coordinates of a have magnitude at least order $\sqrt{m}/n$. Their columns can interpolate y with norm of order $\sqrt{m/n}$. Meanwhile, on a set carrying most of the target's squared norm, the coordinates of a have magnitude at most order $m/(n\sqrt{k_{\text{eff}}})$. Below the scale in equation (10), this interpolant has such a low weighted cost that optimality forces the fitted norm on this active set below $1/4$. The target's norm there exceeds $\sqrt{3}/2$. The selected inactive block depends on a; Appendix B proves a singular-value bound simultaneously for all large blocks to justify this choice.

The argument has an exact deterministic form, valid for any positive weights and any ridge penalty.

Proposition 6 (A lower bound from interpolation by inactive coordinates)**.** *Let $p_i > 0$, $D = \operatorname{diag}(p)$, and $W_{\text{in}} = \min\{\|D^{-1}\beta\|_2 : X\beta = y,\ \beta_S = 0\}$, with $W_{\text{in}} = \infty$ if no such β exists. For $\lambda > 0$, let $\widehat{\beta}_\lambda$ minimize $\|X\beta - y\|_2^2 + \lambda \|D^{-1}\beta\|_2^2$, and let $\widehat{\beta}_0$ be the weighted minimum-norm interpolant. Then, for every $\lambda \geq 0$ and $T \subset S$,*

$$\left\|\widehat{\beta}_\lambda - \beta^\star\right\|_2 \geq \|\beta^\star_T\|_2 - \Big(\max_{i \in T} p_i\Big) W_{\text{in}}.$$

Proof. Let β^0 attain W_{in}. Since $X\beta^0 = y$, comparing objective values gives $\left\|D^{-1}\widehat{\beta}_\lambda\right\|_2 \leq W_{\text{in}}$ for every $\lambda \geq 0$. Hence $\left\|(\widehat{\beta}_\lambda)_T\right\|_2 \leq \max_{i \in T} p_i \, W_{\text{in}}$, and $\left\|\widehat{\beta}_\lambda - \beta^\star\right\|_2 \geq \|\beta^\star_T\|_2 - \left\|(\widehat{\beta}_\lambda)_T\right\|_2$. □

The proof of Theorem 5 bounds W_{in} by the cost of an interpolant supported on a block of inactive columns. Because the proposition holds for every λ and every choice of positive weights, the same event gives more.

Corollary 7 (Ridge penalties and weight floors)**.** *Under the conditions of Theorem 5, with probability at least $1 - Ce^{-cm}$, the following holds for all $\lambda \geq 0$ and $\omega \geq 0$ at once: with weights $p_i = \max\{|a_i|, \omega\}^q$, the fit $\widehat{\beta}_\lambda$ of Proposition 6 has relative error at least 0.6. This includes penalties and floors chosen from the data.*

For equal magnitudes, the bounds have the same powers of n and k and differ by a logarithmic factor. For unequal magnitudes, the upper bound depends on μ and the lower bound on k_{eff}; their

Table 1: Sample scales for equal-magnitude targets, fixed target accuracy, and fixed confidence. Finite-power bounds hold with high probability in their stated dimensional ranges. The lower column concerns coefficient error of the stated estimator.

Estimator	Sufficient scale	Lower-bound scale
Minimum norm	n [(4), mean square]	n [(4), mean square]
$q = 1/2$ (first ℓ_1-IRLS update)	$n^{2/3}(k \log n)^{1/3}$	$n^{2/3}k^{1/3}$
$q = 1$ (first Basic FOCUSS update)	$(nk \log n)^{1/2}$	$(nk)^{1/2}$
$q = 2$ (first General FOCUSS update)	$n^{1/3}(k \log n)^{2/3}$	$n^{1/3}k^{2/3}$
Any fixed $q > 0$	$[n(k \log n)^q]^{1/(q+1)}$	$[nk^q]^{1/(q+1)}$
Top-k selection from a and refit	$k \log n$, exact recovery	—

Finite-power rows: Theorems 3 and 5. Hard selection: Theorem 1. These are scaling statements, with constants depending on q and accuracy.

dependence on the coefficient profile need not match. Appendix C restates the sample exponent in terms of growth rates of $\mu \log n$ and m.

3.1 Correct Screening Can Coexist with Failed Estimation

Corollary 8 (Separation between screening and one soft update)**.** *Let $0 < \delta < 1$. Under the dimensional conditions of Theorem 5, if $m \geq C\mu \log(8n/\delta)$, then*

$$\mathbb{P}_X\{\mathrm{Top}_k(|a|) = S,\ \widehat{\beta}_{\mathrm{hard}} = \beta^\star,\ E_q \geq 0.6\} \geq 1 - \delta - Ce^{-cm}. \tag{11}$$

The probability is at least $1 - 2\delta$ if also $m \geq c^{-1} \log(C/\delta)$.

Proof. The sample condition implies $m \geq k$, and $N \geq C_0 m$ implies $m \leq n/2$. Apply Theorem 1 with failure probability δ and intersect its event with that of Theorem 5. Gaussian X_S has full column rank almost surely. A union bound proves the claim without an independence assumption. □

For equal magnitudes and fixed confidence, this range is nonempty when n/k is sufficiently large relative to $(\log n)^{q+1}$. For $q = 1$, it extends from order $k \log n$ to order $\sqrt{nk}$. Along sequences within this range with $m/(k \log n) \to \infty$, the inactive-to-active weight ratio tends to zero while the coefficient error remains bounded below. The number of available inactive columns matters even when their individual weights are small. Later updates can change the weight distribution; hard selection removes those columns immediately.

The output itself ranks the support correctly once $\left\|\widehat{\beta}_q - \beta^\star\right\|_2 < \beta_{\min}/2$; with the upper bound, this gives the conservative scale $kn^{1/(q+1)}(\log n)^{q/(q+1)}$ for equal magnitudes. In experiments the output ranks better than this suggests: at $(n, k, m) = (512, 8, 136)$ its top-k set is correct in all 100 trials for each tested power, versus 65 for a, and refitting on that set is then exact. A finite-sample explanation remains open.

4 Noise, Approximate Sparsity, and Design Dependence

The same observations can also define a regularized fit. Write $\beta^\star = b + r$, where b is supported on S, and $y = X\beta^\star + \xi = Xb + h$ with $h = Xr + \xi$. For positive weights p, including weights computed from y, define

$$\widehat{\beta}_{p,\lambda} = \arg\min_{\beta} \{\|X\beta - y\|_2^2 + \lambda \left\|D^{-1}\beta\right\|_2^2\}, \quad D = \mathrm{diag}(p), \quad \lambda > 0. \tag{12}$$

Proposition 9 (Regularized comparison bound)**.** *Assume X_S has full column rank. Put $s = \sigma_{\min}(X_S)$, $\kappa = \sigma_{\max}(X_S)/s$, $L = \left\|X_S^\dagger X_{S^c}\right\|_{\mathrm{op}}$, $p_S = \min_{i \in S} p_i$, $p_T = \max_{j \notin S} p_j$, and*

$$B_\lambda = \left(\|h\|_2^2 + \lambda \left\|D_S^{-1} b_S\right\|_2^2\right)^{1/2}.$$

Then

$$\left\|\widehat{\beta}_{p,\lambda} - \beta^\star\right\|_2 \leq \sqrt{1 + \kappa^2 L^2}\frac{p_T B_\lambda}{\sqrt{\lambda}} + \kappa \left\|X_S^\dagger h\right\|_2 + \frac{\lambda \|b\|_2}{s^2 p_S^2} + \|r\|_2. \tag{13}$$

The four terms control the fit outside the support, the noise projected onto the active columns, the bias from the penalty, and the dense part of the target. The projected term permits a statistical noise bound even though the weights depend on the same noisy observations. Appendix C proves the proposition from the active normal equations.

Theorem 10 (Gaussian noise with oracle regularization)**.** *Let $\beta^\star \neq 0$ be fixed and k-sparse, let X have independent $N(0,1)$ entries, and let $\xi \sim N(0, \sigma^2 I_m)$ be independent of X, with $\sigma > 0$. Fix*

$q > 0$ and $0 < \delta < 1$. Put $B = \|\beta^\star\|_2$, $\nu = \sigma/B$, $\mu_\sigma = (B+\sigma)^2/\beta_{\min}^2$, and $\ell = \log(32n/\delta)$. Suppose $C\mu_\sigma \ell \le m \le n/2$. Use $p_i = |(X^\dagger y)_i|^q$ without a floor, and define

$$A = \sqrt{\frac{n}{m}} \left(\frac{\mu_\sigma \ell}{m}\right)^{q/2}, \qquad p_0 = \left(\frac{m\beta_{\min}}{2n}\right)^q, \qquad \lambda = m p_0^2 (A\nu)^{2/3}.$$

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

$$\frac{\left\|\widehat{\beta}_{p,\lambda} - \beta^\star\right\|_2}{B} \le C_q \left[A + (A\nu)^{2/3} + \nu\sqrt{\frac{k + \log(1/\delta)}{m}}\right]. \tag{14}$$

For fixed signal-to-noise ratio and confidence, the bound tends to zero when $A \to 0$ and $k/m \to 0$. It keeps the noiseless polynomial sample scale for equal magnitudes, with a weaker dependence on accuracy. The penalty depends on unknown signal and noise parameters; validation tuning is evaluated separately in the experiments. Corollary 14 extends the stochastic bound to a fixed dense tail, with explicit dependence on its total norm and largest coordinate. Appendix C also bounds the effect of a weight floor.

For non-Gaussian or correlated features, both deterministic propositions remain valid. A stochastic extension needs, for the design at hand, a bound on the coordinates of a and control of the active singular values and of L. Bounded covariance condition number alone does not ensure successful screening: Proposition 15 gives a unit-diagonal Gaussian covariance for which an inactive coordinate of a outranks both active coordinates with probability tending to one. Whitening the design can mix the sparse coordinates. We test other designs empirically.

5 Experiments

Methods share each trial's target, design, and observations. Appendix D gives all grids, trial counts, tuning rules, and uncertainty summaries.

Screening and estimation on the same trials. Figure 1 measures the joint event that the top-k set of a is correct and the reweighted fit has relative error at least 0.25. At $(n,k) = (16384, 8)$, this event occurs for $q = 1$ in 99, 100, and 100 of 100 trials at $m = 384, 512, 768$; median errors are $0.727, 0.594, 0.379$. At $(65536, 4)$ and $q = 2$, the event occurs in at least 95 trials at each of $m = 160, 192, 224, 256, 320$. The refit's own top-$k$ set is correct in every trial at these highlighted settings. Correct ranking, even of the output, can therefore coexist with large coefficient error. The threshold 0.25 is chosen for display; with the threshold 0.6 of Theorem 5, the joint event occurs in 99 of 100 trials at $m = 384$ for $q = 1$, and in 96 and 86 trials at $m = 160$ and 192 for $q = 2$.

The mechanism, and what a second update changes. Proposition 6 can be evaluated in simulation, where S is known. On 40 of the Figure 1 datasets at each of six sample sizes ($n = 16384$, $k = 8$), we compute the ratio $R = W_{\text{in}}/\left\|D^{-1}\beta^\star\right\|_2$ between the smallest weighted cost of an interpolant supported on inactive coordinates and the weighted cost of the target (Appendix D.3). In all 219 trials where one update leaves error at least 0.5, $R < 1$; in all 137 trials with error at most 0.2, $R > 1$. The first update also improves the ranking: in each of the 374 trials where a ranks the support correctly, the ratio $\max_{j \notin S} |\widehat{\beta}_j| / \min_{i \in S} |\widehat{\beta}_i|$ for the output is smaller than the same ratio for a. A second update on the same data, with weights $\max\{|\widehat{\beta}_i|, 10^{-3}\left\|\widehat{\beta}\right\|_\infty\}^q$, raises R in all 480 trial–power pairs and above one in every trial with $m \ge 256$. At $q = 1$ and $m = 384$, the median of R rises from 0.62 to 6.5, and the median error falls from 0.73 to 0.026.

One update and sparse baselines. Figure 2 uses $(n,k) = (512, 8)$, a random support, random signs, and unit-norm equal-magnitude targets. At $m = 136$, median errors for minimum norm and $q = 1/2, 1, 2$ are 0.855, 0.553, 0.272, and 0.036. Top-k selection from a followed by least squares receives the true k and is exact when the selection is correct. Basis pursuit provides the sparse-recovery benchmark. Its median is below the plotting floor even at $m = 48$, but its error exceeds 10^{-3} in 23 of those 100 trials; the support panel shows these failures.

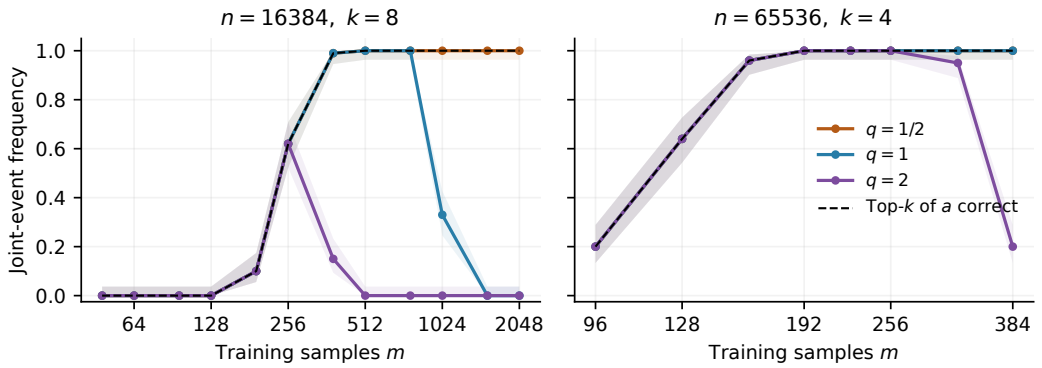


Figure 1: Same-trial screening–estimation separation. Solid curves show the frequency of $\{\mathrm{Top}_k(|a|) = S,\ E_q \geq 0.25\}$; the dashed curve shows how often the top-k set of a is correct. Shading gives pointwise Wilson 95% intervals from 100 trials. In the right panel, the $q = 1/2$ and $q = 1$ curves coincide with the dashed curve. Appendix D.2 reports the complete grids, including settings with no high-frequency separation at $q = 2$.

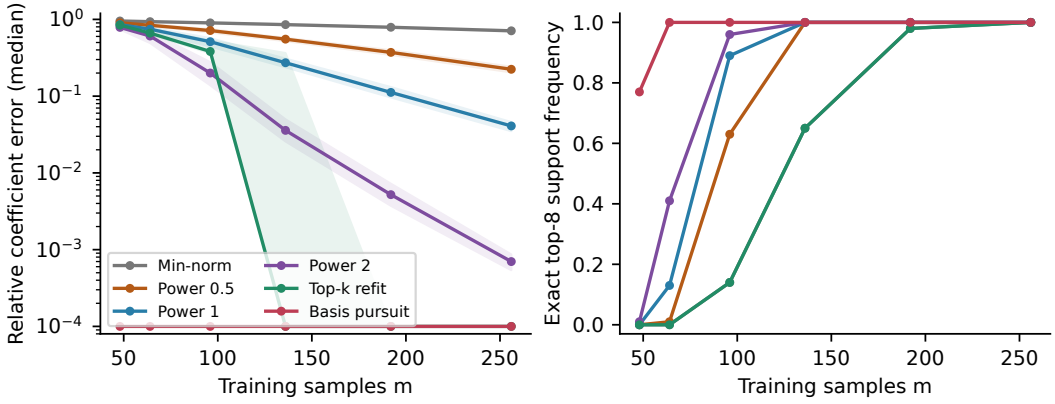


Figure 2: One-update coefficient error and exact top-k support frequency for iid Gaussian designs. Shading denotes interquartile ranges over 100 trials. Errors below 10^{-4} are plotted at that floor; the saved records keep the exact values. The right panel concerns ranking the output, including dense estimates.

Further updates and unequal coefficients. At $m = 96$, the $q = 1$ median error falls from 0.512 after one update to 0.047 after two and below 10^{-4} after five, using a relative weight floor of 10^{-3} after the first step. The gain persists for floors 10^{-4} and 10^{-2}, although the final accuracy changes (Table 15). For geometric active magnitudes, increasing the largest-to-smallest ratio from 1 to 8 raises the $q = 2$ median error at $m = 136$ from 0.037 to 0.184. Hard selection is also sensitive to weak coefficients: at ratio 8 and $m = 256$, its median error is 0.143, versus 0.010 for $q = 2$, which has lower error in 86 of 100 paired trials. The estimate a ranks the support correctly in only two trials at that setting. For $q = 1/2$ median error decreases as the magnitude ratio grows; for $q = 1$ it changes by at most about 36%, far less than the factor $(\mu/k)^{1/2}$ in the upper bound, which may therefore be loose in μ.

Alternative designs. Table 4 in Appendix D fixes $(n, k, m) = (512, 8, 136)$. It compares Rademacher, Laplace, and correlated Gaussian (AR) designs. Reweighted fits reduce coefficient error for all of them, hard selection is sensitive to correlation, and basis pursuit remains accurate. Theory for these designs remains open.

Noise and approximate sparsity. Table 5 in Appendix D compares regularized reweighted fits with ridge, Lasso, adaptive Lasso, and top-k refitting. Each regularized method selects its penalty

Table 2: Fixed-noise underdetermined sweep: mean excess test risk $\pm$ one standard error, multiplied by 10^3, over 60 paired trials per column. All cases use iid Gaussian features, $n = 2048$, $k = 8$, and $\sigma = 0.15$. The dense tail has norm 0.1. Penalty grids, validation size, and test size match Table 5. Top-k refitting receives the true k.

	Sparse target			Dense-tail target		
Method	$m = 256$	$m = 512$	$m = 1024$	$m = 256$	$m = 512$	$m = 1024$
Ridge	878.38 ± 2.87	755.65 ± 3.66	524.22 ± 3.07	888.11 ± 3.80	763.93 ± 3.56	526.73 ± 2.81
$q = 1/2$	492.70 ± 5.51	173.44 ± 2.66	40.01 ± 0.57	515.54 ± 6.40	184.59 ± 2.86	49.37 ± 0.60
$q = 1$	171.96 ± 4.95	28.37 ± 0.69	4.20 ± 0.08	198.67 ± 5.38	39.84 ± 0.84	13.43 ± 0.12
$q = 2$	21.74 ± 1.14	2.05 ± 0.09	0.34 ± 0.02	38.01 ± 1.49	12.03 ± 0.12	10.33 ± 0.06
Top-k refit	18.07 ± 5.77	0.38 ± 0.03	0.16 ± 0.01	28.86 ± 5.89	10.52 ± 0.05	10.31 ± 0.04
Lasso	7.13 ± 0.29	3.04 ± 0.11	1.51 ± 0.04	20.16 ± 0.42	13.92 ± 0.17	11.65 ± 0.08
Adaptive Lasso	2.10 ± 0.11	0.68 ± 0.04	0.29 ± 0.02	13.22 ± 0.20	10.96 ± 0.08	10.58 ± 0.08

on 256 independent validation rows, which are not counted in m. A dense tail of norm 0.1 tests approximate sparsity. Reweighted fits improve on ridge in all five settings. Lasso and adaptive Lasso have lower mean risk in the four underdetermined cases. In the overdetermined case, $q = 2$ has about one quarter of Lasso's risk and lies between top-k refitting and adaptive Lasso. Table 7 reports paired differences.

Fixed noise as the sample size grows. At $(n, k, \sigma) = (2048, 8, 0.15)$, the validation-tuned $q = 2$ median error for sparse targets falls from 0.141 to 0.046 to 0.018 as m grows from 256 to 512 to 1024. Table 2 reports sparse and dense-tail targets. Lasso has lower mean risk at $m = 256$; $q = 2$ has lower mean risk at $m = 512$ and 1024 for both targets. Adaptive Lasso has lower mean risk in five of the six settings; $q = 2$ has lower risk in the dense-tail case at $m = 1024$. Appendix D.6 gives quartiles and paired differences. With the oracle penalty of Theorem 10, the $q = 2$ median error at the same dimensions falls from 0.138 to 0.043 to 0.016 (40 separate trials per setting; Appendix D.7). Appendix D.8 adds an overdetermined grid with shrinking noise.

6 Discussion and Limitations

One reweighted update lowers the sample scale of minimum-norm least squares from n to $[n(k \log n)^q]^{1/(q+1)}$, but it does not reach the $k \log n$ scale of hard selection, even when a already ranks the support correctly. For a feature-learning procedure such as lin-RFM, finding the relevant coordinates and using them are therefore different tasks with different sample costs. The gap comes from the many inactive columns that the soft weights keep: the lower bound shows when their combined weighted cost prevents accurate recovery, and neither a ridge penalty nor a weight floor removes it. A power growing like $\log(n/m)$ removes it at a numerical cost, and so does a second update in our measurements: the first output separates the support by a wider margin than a does, and with the second weights, fitting the observations with inactive coordinates alone costs more than the target.

The stochastic results assume a Gaussian design and a target fixed independently of it. For unequal magnitudes, the upper bound depends on μ and the lower bound on k_{eff}; these parameters can differ substantially. For equal magnitudes, a logarithmic gap remains. The noisy theorem uses an oracle penalty and gives neither a minimax rate nor a guarantee for validation-selected regularization. The experiments are synthetic. The comparison with sparse penalties depends on sample size and signal profile, as Tables 5 and 2 show, and a fixed number of solves does not imply a runtime advantage, since large powers make the weights ill-conditioned.

The main open question is a finite-sample theory for several updates on the same data. By Proposition 2, a two-step guarantee reduces to a high-probability bound on the ratio $\max_{j \notin S} |\widehat{\beta}_j| / \min_{i \in S} |\widehat{\beta}_i|$ for the first output, whose measured behavior Appendix D.3 reports. Other open questions are ranking guarantees for a under correlated designs, removal of the logarithmic gap, and matching dependence on unequal magnitudes.

Reproducibility Statement

Proofs are in Appendices A–C. Appendix D specifies every experimental grid, selection rule, and numerical tolerance. The accompanying code includes the seeds, trial-level outputs, environment versions, and scripts that generate the figures and tables. All numerical results in this manuscript come from those saved runs.

Ethics Statement

This work studies synthetic sparse linear regression. It uses no personal data or human participants. Its guarantees apply to the stated mathematical model; they do not establish the reliability of feature selection in scientific or high-stakes applications.

AI Use Statement

Generative AI tools (ChatGPT, Codex, and Claude) assist with literature search, formulation and refinement of mathematical claims, proof development and checking, experimental design, code implementation, interpretation of results, and manuscript editing. The released pseudorandom code generates all synthetic data, and every reported number comes from a recorded computation. The authors are responsible for the final content, including AI-assisted claims, text, and code.

References

Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 37932–37946, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/f7e7fabd73b3df96c54a320862afcb78-Abstract-Conference.html.

Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005. doi: 10.1109/TIT.2005.858979. URL https://doi.org/10.1109/TIT.2005.858979.

Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006. doi: 10.1002/cpa.20124. URL https://doi.org/10.1002/cpa.20124.

Michael Celentano, Chen Cheng, Ashwin Pananjady, and Kabir Aladin Verchand. State evolution beyond first-order methods I: Rigorous predictions and finite-sample guarantees. *arXiv preprint arXiv:2507.19611*, 2025. URL https://arxiv.org/abs/2507.19611.

Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001. doi: 10.1137/S003614450037906X. URL https://doi.org/10.1137/S003614450037906X.

Hugo Cui and Yue M. Lu. Asymptotic theory of iterated empirical risk minimization, with applications to active learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.23031. arXiv:2601.23031.

Linlin Dai, Kani Chen, Zhihua Sun, Zhenqiu Liu, and Gang Li. Broken adaptive ridge regression and its asymptotic properties. *Journal of Multivariate Analysis*, 168:334–351, 2018. doi: 10.1016/j.jmva.2018.08.007. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6430210/.

Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C. Sinan Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 2010. doi: 10.1002/cpa.20303. URL https://doi.org/10.1002/cpa.20303.

Jianqing Fan, Lingzhou Xue, and Hui Zou. Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics*, 42(3):819–849, 2014. doi: 10.1214/13-AOS1198. URL https://arxiv.org/abs/1210.5992.

Irina F. Gorodnitsky and Bhaskar D. Rao. Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, 45(3):600–616, 1997. doi: 10.1109/78.558475. URL https://doi.org/10.1109/78.558475.

Peter D. Hoff. Lasso, fractional norm and structured sparse estimation using a Hadamard product parametrization. *Computational Statistics & Data Analysis*, 115:186–198, 2017. doi: 10.1016/j.csda.2017.06.007. URL https://doi.org/10.1016/j.csda.2017.06.007.

Chiraag Kaushik, Justin Romberg, and Vidya Muthukumar. Precise asymptotics of reweighted least-squares algorithms for linear diagonal networks. In *Advances in Neural Information Processing Systems*, volume 37, pp. 123755–123786, 2024. doi: 10.52202/079017-3933. URL https://arxiv.org/abs/2406.02769.

Christian Kümmerle, Claudio Mayrink Verdun, and Dominik Stöger. Iteratively reweighted least squares for basis pursuit with global linear convergence rate. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2873–2886, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/16bda725ae44af3bb9316f416bd13b1b-Abstract.html.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395. URL https://doi.org/10.1214/aos/1015957395.

Jiangyuan Li, Thanh V. Nguyen, Chinmay Hegde, and Raymond K. W. Wong. Implicit sparse regularization: The impact of depth and early stopping. In *Advances in Neural Information Processing Systems*, volume 34, pp. 28298–28309, 2021. URL https://arxiv.org/abs/2108.05574.

Clarice Poon and Gabriel Peyré. Smooth bilevel programming for sparse regularization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1543–1555, 2021. URL https://arxiv.org/abs/2106.01429.

Adityanarayanan Radhakrishnan, Mikhail Belkin, and Dmitriy Drusvyatskiy. Linear recursive feature machines provably recover low-rank matrices. *Proceedings of the National Academy of Sciences*, 122(13):e2411325122, 2025. doi: 10.1073/pnas.2411325122. URL https://arxiv.org/abs/2401.04553. Sparse-regression algorithm in Appendix B of arXiv:2401.04553v1.

Galen Reeves, Jiaming Xu, and Ilias Zadik. The all-or-nothing phenomenon in sparse linear regression. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2652–2663, 2019. URL https://proceedings.mlr.press/v99/reeves19a.html.

Chihiro Teramoto, Kazushi Mimura, and Kazunori Iwata. Typical performance of iterative reweighted Lasso for compressed sensing. In *29th European Signal Processing Conference (EUSIPCO)*, pp. 1940–1944, 2021. doi: 10.23919/EUSIPCO54536.2021.9616196. URL https://eurasip.org/Proceedings/Eusipco/Eusipco2021/pdfs/0001940.pdf.

Sara van de Geer, Peter Bühlmann, and Shuheng Zhou. The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electronic Journal of Statistics*, 5:688–749, 2011. doi: 10.1214/11-EJS624. URL https://doi.org/10.1214/11-EJS624.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, pp. 210–268. Cambridge University Press, 2012. doi: 10.1017/CBO9780511794308.006. URL https://arxiv.org/abs/1011.3027.

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55 (5):2183–2202, 2009. doi: 10.1109/TIT.2009.2016018. URL https://doi.org/10.1109/TIT.2009.2016018.

Xiangyu Wang and Chenlei Leng. High-dimensional ordinary least-squares projection for screening variables. *Journal of the Royal Statistical Society: Series B*, 78(3):589–611, 2016. doi: 10.1111/rssb.12127. URL https://doi.org/10.1111/rssb.12127.

Xiangyu Wang, Chenlei Leng, and David B. Dunson. On the consistency theory of high dimensional variable screening. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/540ae6b0f6ac6e155062f3dd4f0b2b01-Abstract.html.

Xiangyu Wang, Chenlei Leng, and Tom Boot. Wang and Leng (2016), high-dimensional ordinary least-squares projection for screening variables, journal of the royal statistical society series b, 78, 589–611. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4):880–881, 2021. doi: 10.1111/rssb.12427. URL https://doi.org/10.1111/rssb.12427. Correction.

Manfred K. Warmuth and Ehsan Amid. Open problem: Learning sparse linear concepts by priming the features. In *Proceedings of the Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 5937–5942, 2023. URL https://proceedings.mlr.press/v195/warmuth23a.html.

Manfred K. Warmuth, Wojciech Kotłowski, Matt Jones, and Ehsan Amid. How rotation invariant algorithms are fooled by noise on sparse targets. In *Proceedings of the 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pp. 1316–1360, 2025. URL https://proceedings.mlr.press/v272/warmuth25a.html.

Huibo Xu, Shi Fu, Qixin Zhang, and Dacheng Tao. Feature priming in online linear regression: Sparse-regret lower bounds and tight coordinatewise rates. *arXiv preprint arXiv:2608.17573*, 2026. URL https://arxiv.org/abs/2608.17573.

Tong Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11:1081–1107, 2010. URL https://www.jmlr.org/papers/v11/zhang10a.html.

Peng Zhao and Bin Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006. URL https://www.jmlr.org/papers/v7/zhao06a.html.

Libin Zhu, Damek Davis, Dmitriy Drusvyatskiy, and Maryam Fazel. Iteratively reweighted kernel machines efficiently learn sparse functions. *arXiv preprint arXiv:2505.08277*, 2025. URL https://arxiv.org/abs/2505.08277.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. doi: 10.1198/016214506000000735. URL https://doi.org/10.1198/016214506000000735.

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533, 2008. doi: 10.1214/009053607000000802. URL https://arxiv.org/abs/0808.1012.

A Projector Bounds and the Upper Theorem

The screening argument below is self-contained. It uses the HOLP statistic but not the screening lemmas of Wang & Leng (2016), one of which is amended by Wang et al. (2021).

Lemma 11 (Coordinate concentration with an explicit range)**.** *Let P be a Haar-distributed rank-m orthogonal projector in $\mathbb{R}^n$ and let b be fixed. There are universal constants $C, c > 0$ such that, for $1 \le t \le cm$ and $m \le n/2$,*

$$\mathbb{P}\left\{\max_{i\le n} |e_i^\top (P - \tfrac{m}{n} I)b| > C \left\|b\right\|_2 \frac{\sqrt{mt}}{n}\right\} \le 6ne^{-t}.$$

Proof. By homogeneity, take $\left\|b\right\|_2 = 1$. For a fixed coordinate i, decompose $b = \alpha e_i + \sqrt{1-\alpha^2}v$ with $v \perp e_i$, $\left\|v\right\|_2 = 1$. Put $H = e_i^\top P e_i$. Rotational invariance gives

$$H \sim \mathrm{Beta}(m/2, (n-m)/2), \qquad Pe_i = He_i + \sqrt{H(1-H)}\,\Theta,$$

where, conditionally on H, Θ is uniform on the unit sphere of $e_i^\perp$.

To bound H, write $H = A/(A+B)$ with independent $A \sim \chi^2_m$ and $B \sim \chi^2_{n-m}$. For $V \sim \chi^2_d$, the inequalities of Laurent & Massart (2000) give

$$\mathbb{P}\{V - d > 2\sqrt{dt} + 2t\} \le e^{-t}, \qquad \mathbb{P}\{d - V > 2\sqrt{dt}\} \le e^{-t}.$$

Applying them to A and $A+B$ gives, outside a set of probability $4e^{-t}$,

$$|H - m/n| \le C\frac{\sqrt{mt} + t}{n}, \qquad H \le C\frac{m+t}{n},$$

provided $t \le cn$. Indeed, $A+B \ge n/2$ for a sufficiently small c, and $|nA - m(A+B)| \le n|A-m| + m|A+B-n|$. The additional condition $t \le cm$ reduces the two bounds to $|H - m/n| \le C\sqrt{mt}/n$ and $H \le Cm/n$.

The spherical-cap bound gives $|\Theta^\top v| \le C\sqrt{t/n}$ except with conditional probability $2e^{-t}$. Thus

$$e_i^\top (P - \tfrac{m}{n} I)b = \alpha(H - m/n) + \sqrt{1-\alpha^2}\sqrt{H(1-H)}\,\Theta^\top v$$

has magnitude at most $C\sqrt{mt}/n$ outside a set of probability $6e^{-t}$. A union bound over i proves the result. The confidence-range condition is essential to this simplified form; for unrestricted t, the t/n contribution cannot be discarded. □

Proof of Theorem 1. For Gaussian X, $a = P\beta^\star$ with P as in Lemma 11. Choose $t = \log(8n/\delta)$ and increase the universal constant in $m \ge Ct$ to meet the lemma's range. The inactive coordinates have no deterministic $(m/n)\beta^\star$ contribution, giving equation (5). The reverse triangle inequality gives equation (6). If $m \ge C\mu t$ with a sufficiently large constant, the fluctuation bound is at most $(m/(3n))\beta_{\min}$. Every active magnitude then strictly exceeds every inactive magnitude. On correct selection, $X_S^\dagger y = \beta^\star_S$ whenever X_S has full column rank. □

Lemma 12 (The operator $X_S^\dagger X_{S^c}$ for Gaussian designs)**.** *Let S be fixed with $|S| = k$. For $u = \sqrt{2\log(4/\delta)}$ and $\sqrt{m} > \sqrt{k} + u$, with probability at least $1-\delta$,*

$$\left\|X_S^\dagger X_{S^c}\right\|_{\mathrm{op}} \le \frac{\sqrt{n-k} + \sqrt{k} + u}{\sqrt{m} - \sqrt{k} - u}.$$

In particular, if $m \ge C(k + \log(4/\delta))$, this is at most $C\sqrt{n/m}$ and X_S has full column rank.

Proof. Condition on X_S. Each column of $X_S^\dagger X_{S^c}$ is an independent Gaussian vector with covariance $(X_S^\top X_S)^{-1}$. Hence the conditional distribution is that of $(X_S^\top X_S)^{-1/2} G$ for an independent $k \times (n-k)$ standard Gaussian matrix G. Gaussian singular-value inequalities give

$$\left\|G\right\|_{\mathrm{op}} \le \sqrt{n-k} + \sqrt{k} + u, \qquad \sigma_{\min}(X_S) \ge \sqrt{m} - \sqrt{k} - u,$$

with combined failure probability at most $2e^{-u^2/2} \le \delta$. These one-sided bounds follow from Gordon's inequality and Gaussian concentration (Vershynin, 2012, Theorem 5.32 and Proposition 5.34). Multiplication gives the claim. Enlarging C makes the denominator a fixed fraction of $\sqrt{m}$. □

Proof of Theorem 3. Apply Theorem 1 and Lemma 12 with failure probability $\delta/2$ each. The logarithm $\log(16n/\delta)$ is at most a universal constant times $\log(8n/\delta)$. Since $\mu \geq k$, the assumed sample lower bound implies the condition of Lemma 12. On the ranking event, after increasing that lower-bound constant,

$$\min_{i\in S} |a_i| \geq \frac{m}{2n}\beta_{\min}, \qquad \max_{j\notin S} |a_j| \leq C \left\|\beta^\star\right\|_2 \frac{\sqrt{m\log(8n/\delta)}}{n}.$$

Raising their ratio to q gives

$$\rho \leq C_q \left(\frac{\mu \log(8n/\delta)}{m}\right)^{q/2}.$$

No independence between this event and the event of Lemma 12 is required. Proposition 2, Lemma 12, and $m \leq n/2$ imply equation (8). Solving its right-hand side for m proves equation (9), subject to the theorem's dimensional range. □

Proof of Corollary 4. The event in the proof of Theorem 3 does not depend on q. On it, the ratio $r = \max_{j\notin S}|a_j| / \min_{i\in S}|a_i|$ satisfies $r^2 \leq 4C^2\mu\log(8n/\delta)/m$, where C is the constant in the two displayed bounds on a, and $\left\|X_S^\dagger X_{S^c}\right\|_{\text{op}} \leq C\sqrt{n/m}$. For $p_i = |a_i|^q$, the weight ratio of Proposition 2 is $\rho = r^q$ for every $q > 0$, so $E_q \leq \sqrt{1+L^2}\, r^q$. Since $m \leq n/2$, $\sqrt{1+L^2} \leq C_1\sqrt{n/m}$. This proves the display with $C_2 = 4C^2$. If $m \geq 4C_2\mu\log(8n/\delta)$, then $r \leq 1/2$ and $E_q \leq C_1\sqrt{n/m}\, 2^{-q}$, which is at most ε for the stated powers. □

B Proof of the Lower Bound

The inactive columns chosen by the lower-bound proof depend on the same observations used in the refit. The following uniform statement avoids treating that selection as independent.

Lemma 13 (All sufficiently large Gaussian column blocks)**.** *Fix $\alpha \in (0,1)$. There are constants $C_\alpha, s_\alpha, c_\alpha > 0$ such that, for a standard Gaussian matrix $G \in \mathbb{R}^{m\times N}$ with $N \geq C_\alpha m$, with probability at least $1 - 2e^{-c_\alpha N}$,*

$$\sigma_{\min}(G_B) \geq s_\alpha\sqrt{|B|} \quad \textit{for every } B \subset [N] \textit{ with } |B| \geq \alpha N.$$

Here $\sigma_{\min}(G_B) = \inf_{\|v\|_2=1} \left\|G_B^\top v\right\|_2$ is the smallest of its m row singular values.

Proof. Write g_j for column j and set $t = \alpha/8$. For a fixed unit vector v, let $J(v)$ count the columns with $|g_j^\top v| < t$. Then $J(v)$ is binomial with success probability $p \leq \sqrt{2/\pi}\, t \leq \alpha/8$. Exponential Markov at $\log 4$ gives

$$\mathbb{P}\{J(v) \geq \alpha N/2\} \leq \exp\{3Np - (\log 4)\alpha N/2\} \leq e^{-\alpha N/4}.$$

On the complementary event, every B with $|B| \geq \alpha N$ contains at least $|B|/2$ columns with $|g_j^\top v| \geq t$. Hence $\left\|G_B^\top v\right\|_2 \geq t\sqrt{|B|/2}$ simultaneously for all such B.

Take a deterministic ϵ-net of the unit sphere, where $\epsilon = t\sqrt{\alpha}/(6\sqrt{2})$. Its size is at most $(1 + 96\sqrt{2}/\alpha^{3/2})^m$. If

$$C_\alpha = \max\left\{1, \frac{8}{\alpha}\log\left(1 + \frac{96\sqrt{2}}{\alpha^{3/2}}\right)\right\},$$

then $N \geq C_\alpha m$ makes the count bound hold at every net point outside probability $e^{-\alpha N/8}$. Also, $\|G\|_{\text{op}} \leq 3\sqrt{N}$ outside probability $e^{-N/2}$: apply the Gaussian expected-norm bound and one-sided concentration with deviation $\sqrt{N}$ (Vershynin, 2012, Theorem 5.32 and Proposition 5.34).

On both events, approximate any unit vector v by a net point v_0. Uniformly over the permitted blocks,

$$\left\|G_B^\top v\right\|_2 \geq t\sqrt{|B|/2} - 3\sqrt{N}\,\epsilon \geq \frac{t}{2\sqrt{2}}\sqrt{|B|}.$$

Thus $s_\alpha = \alpha/(16\sqrt{2})$ and $c_\alpha = \alpha/8$ suffice, with total failure probability at most $2e^{-c_\alpha N}$. No union over column subsets is needed. □

Proof of Theorem 5. Rescaling $\beta^\star$ by a nonzero scalar rescales $\widehat{\beta}_q$ by the same scalar: the common factor in D_q cancels in $D_q(XD_q)^\dagger$, and y scales with the target. Hence E_q is unchanged. Assume $\|\beta^\star\|_2 = 1$, write $u = \beta^\star$, and put $P = X^\dagger X$. Then $\|u\|_\infty = 1/\sqrt{k_{\text{eff}}}$. Haar invariance gives

$$Pu = Zu + \sqrt{Z(1-Z)}\, w, \qquad Z = \|Pu\|_2^2 \sim \text{Beta}(m/2, (n-m)/2), \tag{15}$$

where w is uniform on the unit sphere of $u^\perp$. Write $Z \overset{d}{=} A/(A+B)$ with independent $A \sim \chi^2_m$ and $B \sim \chi^2_{n-m}$. Chi-square tail bounds applied to A and $A+B$ give

$$\frac{m}{4n} \le Z \le \frac{4m}{n} \le \frac{1}{2}, \tag{16}$$

outside probability $4e^{-m/16}$. The last inequality holds after choosing $C_0 \ge 8$ in $N \ge C_0 m$.

We use two properties of w. Represent it as $(I - uu^\top)g/\sqrt{R}$, where $g \sim N(0, I_n)$ and $R = \|(I - uu^\top)g\|_2^2 \sim \chi^2_{n-1}$. With failure probability at most $2e^{-(n-1)/16}$, $(n-1)/2 \le R \le 2(n-1)$.

First, for $j \notin S$ the numerator is g_j. Since $\mathbb{P}\{|g_j| \ge 1/2\} > 0.6$, a binomial tail bound gives at least $0.3N$ such coordinates outside probability $e^{-N/14}$. They form a set $G_0 \subset S^c$ on which

$$|w_j| \ge \frac{1}{4\sqrt{n}}.$$

Set $\alpha = 0.15$, so $|G_0| \ge 2\alpha N$.

Second, a set of active coordinates with controlled $|w_i|$ carries most of the target's squared norm. Indeed,

$$\sum_{i \in S} u_i^2 \{(I - uu^\top)g\}_i^2 \le 2 \sum_{i \in S} u_i^2 g_i^2 + 2(u^\top g)^2 \sum_{i \in S} u_i^4.$$

Here $\sum_i u_i^4 \le 1/k_{\text{eff}}$. The weighted chi-square tail bound (Laurent & Massart, 2000) gives

$$\sum_i u_i^2 g_i^2 \le 1 + 2\sqrt{m/k_{\text{eff}}} + 2m/k_{\text{eff}} \le 5m/k_{\text{eff}}$$

outside probability e^{-m}, using $m \ge k_{\text{eff}}$. Also $(u^\top g)^2 \le 2m$ outside probability $2e^{-m}$. On their intersection with the lower bound for R,

$$Q := \sum_{i \in S} u_i^2 w_i^2 \le \frac{56m}{k_{\text{eff}} n}.$$

Thus $T = \{i \in S : w_i^2 \le 224m/(k_{\text{eff}} n)\}$ satisfies $\sum_{i \in S \setminus T} u_i^2 \le 1/4$ and $\|u_T\|_2 \ge \sqrt{3}/2$.

Intersect these events with equation (16). Equation (15) yields

$$|a_j| \ge b_0 \frac{\sqrt{m}}{n} \qquad (j \in G_0), \quad b_0 = 1/(8\sqrt{2}), \tag{17}$$

$$|a_i| \le A_0 \frac{m}{n\sqrt{k_{\text{eff}}}} \qquad (i \in T), \quad A_0 = 34. \tag{18}$$

For the second inequality, $Z|u_i| + \sqrt{Z}|w_i| \le (4 + 2\sqrt{224})m/(n\sqrt{k_{\text{eff}}})$.

Apply Lemma 13 with $\alpha = 0.15$ to X_{S^c}. Choose $C_0 \ge \max\{C_\alpha, 1/\alpha, 8\}$, so every block of size at least αN has at least m columns and the lemma applies. Also $y = Xu \sim N(0, I_m)$, giving $\|y\|_2 \le 2\sqrt{m}$ outside probability $e^{-m/2}$. These events need not be independent of a.

Choose $B \subset G_0$ with $|B| = \lceil \alpha N \rceil$ and define β^B supported on B by $\beta^B_B = X_B^\dagger y$. The simultaneous block bound permits this data-dependent choice and gives

$$X\beta^B = y, \qquad \|\beta^B\|_2 \le \frac{2}{s_\alpha \sqrt{\alpha}} \sqrt{\frac{m}{N}}.$$

Set $W(\beta) = \|D_q^{-1}\beta\|_2$. Equation (17) implies

$$W(\beta^B) \le \frac{2}{s_\alpha \sqrt{\alpha}} \sqrt{\frac{m}{N}} \left(\frac{n}{b_0 \sqrt{m}} \right)^q. \tag{19}$$

All coordinates of a are nonzero almost surely. Thus $\widehat{\beta}_q$ minimizes W among interpolants and $W(\widehat{\beta}_q) \le W(\beta^B)$. Equations (18)–(19) give

$$\left\|(\widehat{\beta}_q)_T\right\|_2 \le \max_{i\in T} |a_i|^q W(\beta^B) \le K_q \left(\frac{m^{q+1}}{N k_{\text{eff}}^q}\right)^{1/2}, \qquad K_q = \frac{2}{s_\alpha \sqrt{\alpha}} \left(\frac{A_0}{b_0}\right)^q. \tag{20}$$

Choose $c_q = 1/(16K_q^2)$. Under equation (10), $\left\|(\widehat{\beta}_q)_T\right\|_2 \le 1/4$, so

$$E_q \ge \|u_T\|_2 - \left\|(\widehat{\beta}_q)_T\right\|_2 \ge \frac{\sqrt{3}}{2} - \frac{1}{4} > 0.6.$$

The failure probabilities are $4e^{-m/16}$, $2e^{-(n-1)/16}$, $e^{-N/14}$, $3e^{-m}$, $2e^{-c_\alpha N}$, and $e^{-m/2}$. Since $N \ge m$ and $n - 1 \ge m$, their sum is at most Ce^{-cm}. A union bound completes the proof. □

Proof of Corollary 7. Work on the event in the proof of Theorem 5, which involves neither λ nor ω. Rescaling all weights by a common positive factor only rescales λ, so the weights need not be normalized. Put $b = b_0\sqrt{m}/n$ and $A = A_0 m/(n\sqrt{k_{\text{eff}}})$. Equations (17)–(18) give $p_j \ge \max\{b, \omega\}^q$ for $j \in B$ and $p_i \le \max\{A, \omega\}^q$ for $i \in T$. Because $m \ge k_{\text{eff}}$ and $A_0 > b_0$, we have $A \ge b$, hence $\max\{A, \omega\}/\max\{b, \omega\} \le A/b$. The block interpolant β^B is an inactive interpolant, so $W_{\text{in}} \le \left\|D^{-1}\beta^B\right\|_2 \le \left\|\beta^B\right\|_2 / \min_{j\in B} p_j$. Proposition 6 with this T therefore gives, for every $\lambda \ge 0$,

$$E \ge \|u_T\|_2 - \left(\frac{A}{b}\right)^q \left\|\beta^B\right\|_2 \ge \frac{\sqrt{3}}{2} - K_q \left(\frac{m^{q+1}}{N k_{\text{eff}}^q}\right)^{1/2} \ge \frac{\sqrt{3}}{2} - \frac{1}{4} > 0.6,$$

using the bound on $\left\|\beta^B\right\|_2$ and the choice of c_q in the proof of Theorem 5. □

Relation to information-theoretic transitions. Reeves et al. (2019) study an all-or-nothing transition for a noisy binary sparse model. Their impossibility result concerns all estimators in an asymptotic regime. Theorem 5 gives a high-probability coefficient-error event for one specified noiseless estimator. Hard selection can succeed on the same data by Corollary 8; the failure is specific to the finite-power quadratic objective.

C Perturbations and Scope

Proof of Proposition 9. Evaluate the objective in equation (12) at b. Since $y = Xb + h$, optimality gives

$$\left\|X\widehat{\beta}_{p,\lambda} - y\right\|_2^2 + \lambda \left\|D^{-1}\widehat{\beta}_{p,\lambda}\right\|_2^2 \le B_\lambda^2.$$

Thus, for $v = (\widehat{\beta}_{p,\lambda})_{S^c}$, $\|v\|_2 \le p_T B_\lambda/\sqrt{\lambda}$. Put $G = X_S^\top X_S$, $R = \lambda D_S^{-2}$, and $M = (G + R)^{-1}G$. The active normal equations imply

$$(\widehat{\beta}_{p,\lambda})_S - b_S = -MX_S^\dagger X_{S^c} v + MX_S^\dagger h - (G + R)^{-1} R b_S.$$

Since

$$M = G^{-1/2}(I + G^{-1/2} R G^{-1/2})^{-1} G^{1/2}, \qquad \|M\|_{\text{op}} \le \kappa,$$

the vector with active part $-MX_S^\dagger X_{S^c} v$ and inactive part v has norm at most $\sqrt{1 + \kappa^2 L^2}\,\|v\|_2$. The remaining active terms have norms at most $\kappa \left\|X_S^\dagger h\right\|_2$ and $\lambda \|b\|_2 / (s^2 p_S^2)$. Apply the triangle inequality and add $\|r\|_2$. The argument is deterministic and allows arbitrary dependence of D on (X, y). □

Proof of Theorem 10. We establish simultaneous events for the design, the noisy estimate $X^\dagger y$, and the projected noise. All constants below are independent of the target, dimensions, noise level, and confidence; q enters only when coordinate magnitudes are raised to a power. Put $P = X^\dagger X$.

Gaussian singular-value bounds and $m \le n/2$ give $\sigma_{\min}(X) \ge c\sqrt{n}$ outside probability $\delta/16$, after increasing the constant in the assumed sample lower bound. Here the smallest singular value is the

smallest of the m positive singular values. The beta law of each P_{ii} and a coordinate union bound give

$$\max_i P_{ii} \le Cm/n$$

outside probability $\delta/16$. For example, apply the chi-square bounds from Lemma 11 at level $\log(64n/\delta) \le cm$. If $X = U\Sigma V^\top$ is a thin singular-value decomposition, then

$$\left\|e_i^\top X^\dagger\right\|_2^2 = e_i^\top V\Sigma^{-2}V^\top e_i \le P_{ii}/\sigma_{\min}(X)^2 \le Cm/n^2.$$

Conditionally on X, each $(X^\dagger\xi)_i$ is Gaussian with variance at most $C\sigma^2 m/n^2$ on these design events. Another coordinate union bound yields

$$\left\|X^\dagger\xi\right\|_\infty \le C\sigma\frac{\sqrt{m\ell}}{n} \tag{21}$$

with additional failure probability at most $\delta/8$.

Lemma 11 applied to the fixed $\beta^\star$, combined with equation (21) and $y = X\beta^\star + \xi$, now gives

$$\min_{i\in S}|(X^\dagger y)_i| \ge \frac{m\beta_{\min}}{2n}, \qquad \max_{j\notin S}|(X^\dagger y)_j| \le C(B+\sigma)\frac{\sqrt{m\ell}}{n}.$$

The active bound uses $m \ge C\mu_\sigma\ell$ with a sufficiently large universal constant. Its additional failure probability is at most $6ne^{-\ell} = 3\delta/16$. Consequently, with $r_0 = (\mu_\sigma\ell/m)^{q/2}$,

$$p_S \ge p_0, \qquad p_T \le C_q p_0 r_0. \tag{22}$$

Because $\mu_\sigma \ge k$, Gaussian singular-value inequalities and Lemma 12 also give

$$s \ge c\sqrt{m}, \qquad \kappa \le C, \qquad L \le C\sqrt{n/m}$$

with combined failure probability at most $\delta/4$. Gaussian norm concentration gives $\|\xi\|_2 \le C\sigma\sqrt{m}$ outside probability $\delta/8$. Finally, conditionally on X_S, the vector $X_S^\dagger\xi$ has law $\sigma(X_S^\top X_S)^{-1/2}g$, where $g \sim N(0, I_k)$. Hence, on the active singular-value event,

$$\left\|X_S^\dagger\xi\right\|_2 \le C\sigma\sqrt{\frac{k+\log(1/\delta)}{m}}$$

with additional failure probability at most $\delta/8$. The sum of the allocated failure probabilities is below δ. Conditioning only establishes individual events; it does not make the weights independent of the noise.

Apply Proposition 9 with $b = \beta^\star$, $h = \xi$, and $r = 0$. For any $t > 0$, set $\lambda = mp_0^2 t$. Then

$$B_\lambda \le \|\xi\|_2 + \sqrt{\lambda}B/p_S, \qquad \frac{\lambda}{s^2p_S^2} \le Ct.$$

Equations (13) and (22) therefore imply

$$\frac{\left\|\widehat{\beta}_{p,\lambda} - \beta^\star\right\|_2}{B} \le C_q\left[A + \frac{A\nu}{\sqrt{t}} + \nu\sqrt{\frac{k+\log(1/\delta)}{m}} + t\right].$$

The choice $t = (A\nu)^{2/3}$ balances the second and fourth terms and proves the theorem. All coordinates of $X^\dagger y$ are nonzero almost surely under the Gaussian model, so the weights are positive. □

Corollary 14 (A fixed approximately sparse target)**.** *Let $\beta^\star = b + r$ be fixed, where $S = \operatorname{supp}(b)$ has size k, $b \ne 0$, and $r_S = 0$. Let X and ξ have the Gaussian distributions of Theorem 10, allowing $\sigma = 0$. Put $B = \|b\|_2$, $\tau = \|r\|_2$, and assume $\tau^2 + \sigma^2 > 0$. Define*

$$b_{\min} = \min_{i\in S}|b_i|, \quad \mu_\sigma = \frac{(\|\beta^\star\|_2 + \sigma)^2}{b_{\min}^2}, \quad \nu = \frac{\sqrt{\tau^2+\sigma^2}}{B}, \quad \zeta = \frac{\|r\|_\infty}{b_{\min}}, \quad \ell = \log(32n/\delta).$$

For fixed $q > 0$ and $0 < \delta < 1$, suppose $C\mu_\sigma\ell \le m \le n/2$. Set

$$A = \sqrt{\frac{n}{m}}\left(\zeta + \sqrt{\frac{\mu_\sigma\ell}{m}}\right)^q, \qquad p_0 = \left(\frac{mb_{\min}}{2n}\right)^q, \qquad \lambda = mp_0^2(A\nu)^{2/3}.$$

For $p_i = |(X^\dagger y)_i|^q$, with probability at least $1-\delta$,

$$\frac{\left\|\widehat{\beta}_{p,\lambda} - \beta^\star\right\|_2}{B} \le C_q \left[A + (A\nu)^{2/3} + \nu\sqrt{\frac{k+\log(1/\delta)}{m}}\right] + \frac{\tau}{B}. \tag{23}$$

Proof. Apply the projector argument in Theorem 10 to the full fixed vector $\beta^\star$. On S, its entries equal those of b. On S^c, the deterministic part of $X^\dagger y$ is $(m/n)r$. The same events for the design and for $X^\dagger y$ therefore give

$$\min_{i\in S} |(X^\dagger y)_i| \ge \frac{m b_{\min}}{2n}, \qquad \max_{j\notin S} |(X^\dagger y)_j| \le \frac{m}{n}\|r\|_\infty + C(\|\beta^\star\|_2 + \sigma)\frac{\sqrt{m\ell}}{n}.$$

Thus $p_S \ge p_0$ and

$$p_T \le C_q p_0 \left(\zeta + \sqrt{\mu_\sigma \ell/m}\right)^q.$$

Since $r_S = 0$ and r is fixed, $h = Xr + \xi$ has law $N(0, (\tau^2+\sigma^2)I_m)$ and is independent of X_S. Gaussian concentration yields

$$\|h\|_2 \le CB\nu\sqrt{m}, \qquad \left\|X_S^\dagger h\right\|_2 \le CB\nu\sqrt{\frac{k+\log(1/\delta)}{m}}$$

on the same probability allocations as in Theorem 10. The latter bound uses conditioning on X_S only. Dependence between h, X_{S^c}, and the weights is allowed. Also $\mu_\sigma \ge k$, so the bounds on the active singular values and on L apply. Substitute these bounds in Proposition 9, put $\lambda = mp_0^2 t$, and choose $t = (A\nu)^{2/3}$. The extra term τ/B comes from comparison with the sparse part b. The same union bound gives total failure probability below δ. □

The tail affects the guarantee through both its total norm τ and its largest coordinate relative to $b_{\min}$. A small tail norm alone does not ensure that the weights separate S from S^c.

Proposition 15 (A correlated design where screening fails)**.** *Let $m \to \infty$ and $m/n \to 0$. Let the rows of X be independent $N(0,\Sigma_n)$ vectors, where*

$$\Sigma_n = \mathrm{diag}(\Sigma_3, I_{n-3}), \qquad \Sigma_3 = \begin{pmatrix} 1 & 0 & 3/5 \\ 0 & 1 & 3/5 \\ 3/5 & 3/5 & 1 \end{pmatrix}, \qquad \beta^\star = (1,1,0,\dots,0)/\sqrt{2}.$$

The covariance has unit diagonal and condition number less than 13. For $y = X\beta^\star$ and $a = X^\dagger y$,

$$\frac{n}{m}(a_1,a_2,a_3) \longrightarrow \frac{1}{\sqrt{2}}(1,1,6/5) \quad \textit{in probability.}$$

Consequently, $\mathbb{P}\{\mathrm{Top}_2(|a|) = \{1,2\}\} \to 0$. This applies, for example, when $m = \lfloor\sqrt{n}\rfloor$.

Proof. The eigenvalues of Σ_3 are 1 and $1 \pm 3\sqrt{2}/5$, which gives the condition-number bound. Write $X = [U\ V]$, where U contains the first three columns. The entries of V are independent standard Gaussians. Gaussian singular-value bounds give $\left\|VV^\top/n - I_m\right\|_{\mathrm{op}} \to 0$ in probability. Also

$$\left\|UU^\top/n\right\|_{\mathrm{op}} \le \|U\|_F^2/n = O_{\mathbb{P}}(m/n) \to 0.$$

Thus $H = XX^\top/n$ satisfies $\left\|H^{-1} - I_m\right\|_{\mathrm{op}} \to 0$ in probability. For $j = 1,2,3$, $\|x_j\|_2 \|y\|_2/m = O_{\mathbb{P}}(1)$, so

$$\frac{n}{m}a_j = \frac{1}{m}x_j^\top H^{-1} y = \frac{1}{m}x_j^\top y + o_{\mathbb{P}}(1).$$

The law of large numbers gives the displayed limit from the first three coordinates of $\Sigma_n\beta^\star$. The third limit strictly exceeds the first two, so the top-two set of a is wrong with probability tending to one. □

This example rules out a covariance extension based only on condition number. It does not show that the reweighted fit fails, since its coordinate ordering can differ from that of a.

Perturbation of a and weight floors. Let $a^0 = X^\dagger Xb$, suppose its inactive magnitudes are at most U_{in} and its active magnitudes are at least $A_{\text{act}} > 0$, and put $\eta = \left\|X^\dagger h\right\|_\infty$. If $\eta < A_{\text{act}}$, the stabilized weights $p_i = \max\{|(X^\dagger y)_i|, \omega\}^q$ with floor $\omega \geq 0$ satisfy

$$\frac{\max_{j \notin S} p_j}{\min_{i \in S} p_i} \leq \left(\frac{\max\{U_{\text{in}} + \eta, \omega\}}{\max\{A_{\text{act}} - \eta, \omega\}} \right)^q . \tag{24}$$

Thus a large floor can erase the coordinate contrast. The perturbation $h = Xr + \xi$ includes both noise and a dense target tail; its effect depends on the actual design through $X^\dagger h$.

Proof of equation (24). The identity $X^\dagger y = a^0 + X^\dagger h$ implies inactive magnitudes at most $U_{\text{in}} + \eta$ and active magnitudes at least $A_{\text{act}} - \eta > 0$. Taking the maximum with ω, dividing, and raising to q gives the claim. No independence is needed. □

Zero coordinates of a. For $0 < m < n$ and Gaussian X, all coordinates of a are nonzero almost surely for a fixed nonzero target. Positivity need not hold simultaneously for all sparse targets. For a projector P with $0 < P_{jj} < 1$, the identity $\sum_\ell P_{j\ell}^2 = P_{jj}$ gives an $\ell \neq j$ with $P_{j\ell} \neq 0$. Set $b_j = 1$, $b_\ell = -P_{jj}/P_{j\ell}$, and all other entries to zero. Then $(Pb)_j = 0$ although $b_j \neq 0$, and a positive-power update deletes that coordinate. This example explains the role of the fixed-target quantifier and of weight floors; it supplies no constant relative-error lower bound uniform over dimensions.

Dense refits and top-k recovery. If all weights are positive, $y \neq 0$, and X_{S^c} has row rank m, the weighted interpolant cannot be supported entirely on S. Its optimality equations give $\widehat{\beta} = D^2 X^\top \nu$. If $\widehat{\beta}_{S^c} = 0$, then $X_{S^c}^\top \nu = 0$, hence $\nu = 0$, contradicting $X\widehat{\beta} = y \neq 0$. Gaussian X_{S^c} has that rank almost surely when $n - k \geq m$.

For any estimate $\widetilde{\beta}$, if $\left\|\widetilde{\beta} - \beta^\star\right\|_2 < \beta_{\min}/2$, then its active magnitudes exceed $\beta_{\min}/2$ and its inactive magnitudes are less than $\beta_{\min}/2$. Thus its top k coordinates equal S. Substituting equation (8) and $\mu = k$ yields the sufficient scale stated in Section 3.1. It is a global-norm conversion, with no associated lower bound for thresholded support recovery.

For comparison, Lasso sign consistency is governed by an irrepresentable condition under the regularity assumptions of Zhao & Yu (2006). For isotropic Gaussian designs, its sharp asymptotic sample threshold is $2k\log(n-k)$ to leading order, subject to appropriate penalty and minimum-signal conditions (Wainwright, 2009). For equal magnitudes, our noiseless ranking guarantee for a has the same order $k \log n$, whereas accurate coefficients from one soft update require the larger scale of Theorem 5. Thresholded and adaptive Lasso have separate guarantees (van de Geer et al., 2011); the lower bound here concerns the unthresholded reweighted least-squares estimate.

How the power enters. For fixed error and confidence, the sufficient scale in equation (9) is $n^{1/(q+1)} s^{q/(q+1)}$, where $s = \mu \log(8n/\delta)$. If $s = n^\gamma$ and $m = n^\theta$, the right-hand side of equation (8) is $C_q n^{[1+q\gamma-(q+1)\theta]/2}$. It tends to zero for fixed q when $\theta > (1 + q\gamma)/(q + 1)$, that is, when $q > (1 - \theta)/(\theta - \gamma)$ for $\gamma < \theta < 1$. Corollary 4 gives the version with constants explicit in q.

D Experimental Protocol and Additional Results

D.1 Designs, Targets, and Randomness

Master seed 20260925 generates the main Gaussian grid, the dimension/sparsity grid, the main noisy comparison, all design-comparison trials, and the floor-sensitivity check. Master seed 20260926 generates the heterogeneous-magnitude, shrinking-noise, fixed-noise, oracle-penalty, and large-dimension grids. Each case and trial receives a 64-bit seed derived by SHA-256 from its parameter dictionary, the name of its grid in the released code, and the master seed. The large-dimension suites use the first eight bytes, read big-endian, of the digest of compact sorted-key JSON with fields `experiment`, `k`, `m`, `master_seed`, `n`, and `trial`. The other suites record their seed derivation in the released scripts. Changing execution order does not change a trial. Methods within a trial use identical data. Unless stated otherwise, the target support is a uniform k-subset, its signs are

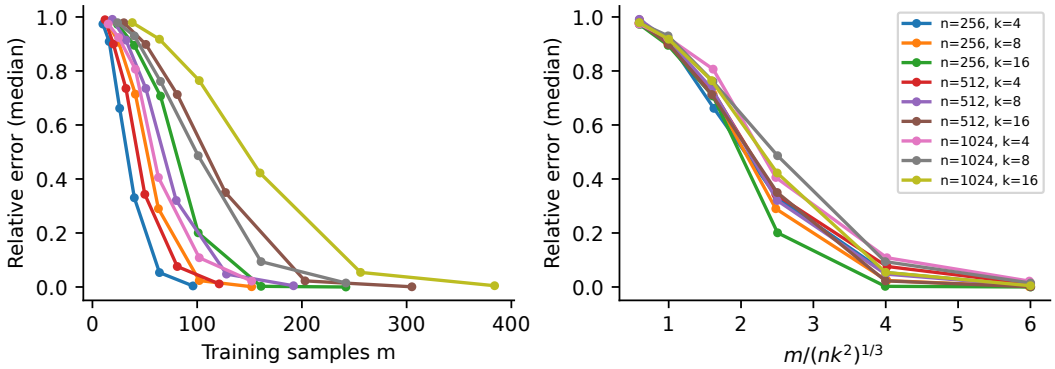


Figure 3: Dimension and sparsity dependence for $q = 2$, with 30 trials per point. Both panels display the same median errors. The right panel uses the polynomial scale from Theorem 5; it does not remove the confidence logarithm or finite-size effects in the upper bound.

independent Rademacher variables, and its nonzero magnitudes are $1/\sqrt{k}$. The theorem's constants are independent of the chosen support and signs, so drawing them independently of the design is compatible with its fixed-target quantifier.

Gaussian entries are $N(0, 1)$; Rademacher entries are uniform on $\{-1, 1\}$; Laplace entries have scale $1/\sqrt{2}$. AR rows are generated recursively from independent standard Gaussians using $x_j = \rho x_{j-1} + \sqrt{1-\rho^2} z_j$ and $x_1 = z_1$. Thus their covariance is exactly $\rho^{|i-j|}$ in distribution. Columns are not normalized or centered after generation. All models have zero population feature means and unit marginal variances.

The main noiseless grid is $m \in \{48, 64, 96, 136, 192, 256\}$ at $(n, k) = (512, 8)$, with 100 trials per value. The design comparison uses the same dimensions and $m = 136$, with 200 trials for Gaussian, Rademacher, Laplace, AR(0.60), and AR(0.85) designs. Each design has 200 trials with distinct seeds under one protocol; the first 100 Gaussian trials are the $m = 136$ trials of the main grid. The repeated-update comparison continues the main-grid trials at $m = 96$ and 136 for $q = 1/2$ and $q = 1$ through update 10, recording updates 1, 2, 5, and 10. The first update is unsmoothed. Later weights are $\max\{|\beta_i|, 10^{-3} \|\beta\|_\infty\}^q$.

The dimension/sparsity grid uses $n \in \{256, 512, 1024\}$, $k \in \{4, 8, 16\}$, and

$$m = \text{round}\{c(nk^2)^{1/3}\}, \qquad c \in \{0.6, 1, 1.6, 2.5, 4, 6\},$$

with 30 trials per case. All these sample sizes are below n. The broad grid includes cases outside the upper theorem's sufficient range; it is an empirical scaling check, without fitted transition constants. Figure 3 displays $q = 2$ against both the raw and rescaled sample counts. The curves need not coincide: the theorem has unknown constants, an upper/lower logarithmic gap, and an explicit coordinate-gap condition.

D.2 Direct Separation and Larger Dimensions

The same-trial event is computed by joining, for each trial, the support outcome of a and the error of the reweighted fit. Table 9 reports the main $n = 512$ grid with error threshold 0.1. The released summary also includes thresholds 0.05, 0.2, $1/(2\sqrt{2})$, and 0.6, together with conditional probabilities given correct screening. The displayed joint frequency uses all trials in its denominator; it is not conditional on successful screening.

Two larger equal-magnitude grids use 100 trials per setting and record every $q \in \{1/2, 1, 2\}$. The first has $k = 8$, $n \in \{4096, 16384\}$, and

$$m \in \{48, 64, 96, 128, 192, 256, 384, 512, 768, 1024, 1536, 2048\}.$$

The second has $k = 4$, $n \in \{16384, 65536\}$, and

$$m \in \{96, 128, 160, 192, 224, 256, 320, 384\}.$$

The second grid is a follow-up: the first has no setting where the frequency of a correct top-k set of a and the frequency of $E_2 \geq 0.25$ both reach 0.95. Both grids are reported in full. Figure 1 displays one dimension from each grid; Tables 10 and 11 report their frequencies and intervals. Table 12 gives the error distributions at the highlighted settings. The saved CSV summaries include both dimensions and thresholds 0.1, 0.25, and 0.6.

At $k = 8, n = 16384, m = 384$, the $q = 2$ joint frequency is only 0.15; at $k = 4, n = 65536$, it is at least 0.95 for five tested sample sizes. The larger grid therefore shows a separation that is narrow or absent at smaller n/k. These observations illustrate the mechanism and do not estimate the theorem's universal constants. At all highlighted settings, least squares on the k largest coordinates of a is exact whenever that selection is correct.

The large-dimension implementation computes a by Cholesky factorization of $XX^\top$, with an SVD fallback, and uses the same unfloored weighted interpolant. Across all large-dimension runs, the largest relative interpolation residual is below 1.1×10^{-13}. A separate implementation that computes a by SVD regenerates selected seeds and reproduces their errors and support outcomes.

D.3 The Mechanism and the Second Update

This diagnostic regenerates the large-dimension datasets with $n = 16384$, $k = 8$, equal magnitudes, trials 0–39, and $m \in \{192, 256, 384, 512, 768, 1024\}$ from their recorded seeds; their first-update errors equal the released values exactly. For $q \in \{1, 2\}$ and weights p, it computes W_{in} of Proposition 6 by weighted minimum-norm interpolation on the inactive columns, the lower bound of that proposition maximized over $T \subset S$, and the cost ratio $R = W_{\text{in}} / \left\|D^{-1}\beta^\star\right\|_2$. The second update reuses the observations and the weights $\max\{|\widehat{\beta}_i|, 10^{-3}\left\|\widehat{\beta}\right\|_\infty\}^q$ of the repeated-update experiments. Table 3 reports the results. In every trial, the lower bound does not exceed the error, and the second error does not exceed the bound of Proposition 2 computed with the second weights. An independent SVD implementation reproduces trial 0 at every sample size and power.

The following observation is exploratory; the protocol fixed before the run does not include it. In all 374 trials where a ranks the support correctly, the ratio $\max_{j \notin S} |\widehat{\beta}_j| / \min_{i \in S} |\widehat{\beta}_i|$ for the first output is at most the same ratio for a raised to the power $2q + 1$. A high-probability bound of this form would, through Proposition 2, give a two-step guarantee at a smaller sample scale than one step; we do not prove such a bound.

Table 3: The mechanism on the $n = 16384$, $k = 8$ datasets of Figure 1, 40 trials per row. "a" and "Output" count trials whose top-k set is correct for a and for the first output. "Prop. 6" counts trials where the lower bound of Proposition 6 is positive. R is the cost ratio for the first-update and second-update weights. The last column counts trials where Proposition 2, with the second weights, bounds the second error below one. Errors and ratios are medians.

q	m	a	$E^{(1)}$	Prop. 6	$R^{(1)}$	Output	$E^{(2)}$	$R^{(2)}$	Bound < 1
1	192	6	0.916	40	0.296	15	0.415	1.24	0
1	256	22	0.858	40	0.401	36	0.150	2.5	0
1	384	39	0.732	40	0.615	40	0.026	6.51	1
1	512	40	0.595	23	0.847	40	6.8×10^{-3}	12.9	39
1	768	40	0.386	0	1.3	40	1.1×10^{-3}	32	40
1	1024	40	0.245	0	1.82	40	2.8×10^{-4}	63.8	40
2	192	6	0.705	33	0.604	35	2.6×10^{-4}	84.5	6
2	256	22	0.494	5	1.02	40	1.9×10^{-6}	872	37
2	384	39	0.205	0	2.02	40	7.5×10^{-9}	1.3×10^{4}	40
2	512	40	0.085	0	3.37	40	3.8×10^{-10}	5.8×10^{4}	40
2	768	40	0.023	0	6.75	40	2.6×10^{-11}	2.0×10^{5}	40
2	1024	40	8.1×10^{-3}	0	11.6	40	1.6×10^{-11}	2.6×10^{5}	40

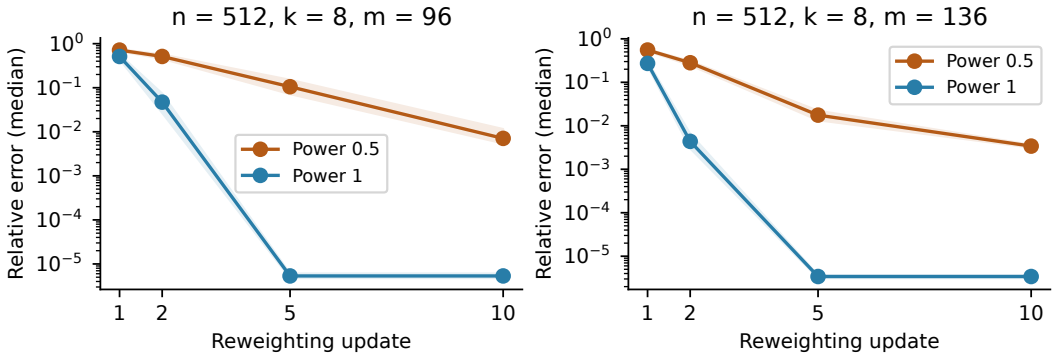


Figure 4: Repeated reweighting on the main-grid Gaussian data. The first update is unfloored. Updates 2–10 use $\max\{|\beta_i|, 10^{-3}\,\|\beta\|_\infty\}^q$. Shading denotes interquartile ranges over 100 trials.

D.4 Unequal Magnitudes and Further Updates

For the $n = 512, k = 8$ heterogeneous grid, active magnitudes form a geometric progression with largest-to-smallest ratio $d \in \{1, 2, 4, 8\}$ and are normalized to unit Euclidean norm. For each d and $m \in \{96, 136, 192, 256\}$, 100 independent trials compare all six noiseless methods. The resulting μ values are approximately $8, 17.7, 46.9, 141.6$; $k_{\text{eff}} = \mu/d^2$. Figure 5 and Tables 13–14 report every setting. For $d = 8, m = 256, q = 2$ has median error 0.010 and exact support frequency 0.90. The exact support frequency of a is 0.02. Coefficient error and recovery of the weakest coordinate can therefore behave differently.

A separate paired-profile grid uses $n = 4096, k = 8$, $m \in \{128, 192, 256, 384, 512, 768\}$, and 60 independent designs per sample size. On each design it uses the same support and signs for three unit-norm profiles: equal magnitudes ($\mu = 8$), geometric magnitudes chosen to give $\mu = 80$, and seven equal entries with one ten times smaller ($\mu = 701$). For $q = 1/2$ and $q = 1$, median errors remain close across profiles even as support recovery by a deteriorates. At $m = 768$, the top-k set of a is correct with frequencies 1.00, 0.32, and 0.00, respectively. The full paired records are included. This grid does not support interpreting the upper bound's $\mu^{q/2}$ factor as a necessary error multiplier.

Figure 4 shows later updates on the main-grid Gaussian trials. Table 15 repeats these trajectories with relative floors 10^{-4}, 10^{-3}, and 10^{-2} after the first step. All floors use the same datasets, so the comparison is paired. The gain from further updates persists across all three floors; the attainable error changes with the floor.

D.5 Noisy Models and Penalty Selection

Every noisy case uses $k = 8$. Training dimensions, distributions, and noise levels for the main comparison are given in Table 5. In that comparison, the noise standard deviation is σ per observation and does not shrink with m. Each trial has 256 independent validation rows and 2048 independent test rows drawn from the same feature distribution as training. Validation labels contain independent noise of the same standard deviation. Test risk is computed against $X_{\text{test}}\beta^\star$ without label noise. For the approximately sparse case, an independent Gaussian vector on S^c is normalized to norm 0.1 and added to the sparse signal. The full target then has squared norm 1.01. Prediction errors in the table are unnormalized excess risks for that full target; coefficient errors in the saved data are relative errors.

The minimum-norm estimate is $a = X^\dagger y$. For a reweighted ridge fit, form $p_i = \max\{|a_i|, 10^{-3}\,\|a\|_\infty\}^q$ and divide all weights by their maximum. For $q = 0$, all weights equal one and the method is ridge. Each method separately minimizes validation mean squared error over 22 logarithmically spaced values $\alpha \in [10^{-6}, 10]$ for

$$\frac{1}{m}\,\|X\beta - y\|_2^2 + \alpha\,\left\|D^{-1}\beta\right\|_2^2.$$

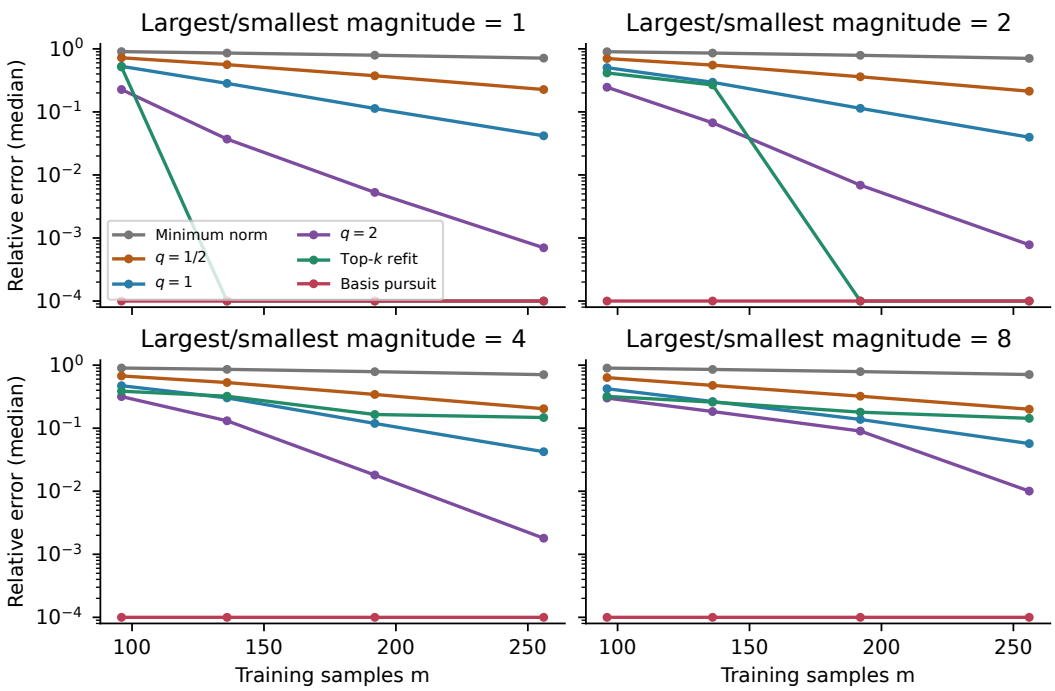


Figure 5: Median relative coefficient error for geometric active magnitudes, with 100 trials per setting. All methods use the same data within a trial. Errors below 10^{-4} are plotted at that floor; the tables and saved records keep the exact values.

Thus $\lambda = m\alpha$ in equation (12). A singular-value decomposition of XD evaluates the complete regularization path. The scale normalization of D is part of this protocol; a common rescaling is immaterial only for the unregularized estimator.

Lasso minimizes

$$(2m)^{-1} \|X\beta - y\|_2^2 + \alpha \|\beta\|_1 .$$

Its path uses 60 logarithmically spaced penalties from $\alpha_{\max} = \|X^\top y\|_\infty / m$ down to $10^{-4}\alpha_{\max}$. For adaptive Lasso, set

$$p_i = \frac{\max\{|a_i|, 10^{-3}\|a\|_\infty\}}{\|a\|_\infty},$$

solve the same Lasso path in the design XD, and return $D\widehat{\theta}$. Its penalty in original coordinates is $\alpha \sum_i |\beta_i|/p_i$, with its own $\alpha_{\max}$ computed on XD. This is an adaptive-Lasso baseline with a stated minimum-norm initialization; we do not invoke the oracle-property assumptions of Zou (2006).

Top-k refitting receives the true k and uses the k largest coordinates of a. It has no regularization parameter. All tuned methods share the validation observations and tune independently. No winner is selected using test error, and no method is dropped after evaluation.

D.6 Fixed-Noise Underdetermined Scaling

This grid fixes $n = 2048$, $k = 8$, and per-observation noise standard deviation $\sigma = 0.15$, and varies $m \in \{256, 512, 1024\}$. Each sample size has 60 independent sparse-target trials and 60 independent dense-tail trials, for 360 datasets. The tail is an independent Gaussian vector on S^c normalized to norm 0.1. The seven methods, 256-row validation set, 2048-row test set, penalty grids, and weight floor follow the preceding protocol. Every case is underdetermined, including $m = 1024 = n/2$.

Table 2 gives mean excess risk and standard errors. Table 18 gives relative coefficient-error medians and quartiles. At $q = 2$, median error decreases from 0.141 to 0.046 to 0.018 for the sparse target, and from 0.193 to 0.108 to 0.101 with the dense tail. The latter values approach the tail-to-target norm ratio $0.1/\sqrt{1.01} \simeq 0.100$; the dense tail is not recovered exactly.

Table 19 reports the paired comparisons with all three sparse baselines. At $m = 1024$ with a dense tail, the mean risk difference of $q = 2$ minus adaptive Lasso is $(-0.251 \pm 0.042) \times 10^{-3}$, whereas its difference from top-k refitting is $(0.018 \pm 0.038) \times 10^{-3}$. The uncertainty is across the 60 paired trials. These measurements concern validation-selected penalties, separately from the oracle choice in Theorem 10.

D.7 The Oracle Penalty at Fixed Noise

To evaluate the estimator specified in Theorem 10, we use $n = 2048$, $k = 8$, $\sigma = 0.15$, $q = 2$, and 40 independent Gaussian trials for each $m \in \{256, 512, 1024\}$. Targets have unit norm and equal magnitudes. The weights are unfloored and unnormalized, and $\lambda = mp_0^2(A\nu)^{2/3}$ is evaluated exactly as displayed in the theorem with $\delta = 0.1$. This choice uses the true $\beta_{\min}$ and σ.

Table 17 reports median relative errors and quartiles. For comparison, each trial also evaluates 25 penalties $\lambda 10^{j/2}$, $j = -12, \ldots, 12$, and selects the smallest true coefficient error on that grid. This second choice is an oracle benchmark, separate from validation tuning. Median error for the theorem's penalty decreases from 0.138 to 0.043 to 0.016. At $m = 512$ and 1024, its worst observed error ratio to the grid minimum is 1.09 and 1.04, respectively. These measurements evaluate the penalty formula at finite dimensions; they do not calibrate the theorem's sufficient sample or probability constants.

The oracle grid uses master seed 20260926. Each trial's seed is the first eight SHA-256 bytes, read big-endian, of compact sorted-key JSON with fields `experiment` (value `E4`), `m`, `master_seed`, and `trial`. Independent SVD-based computations of a and of the ridge fits reproduce all 120 trials and every value on their 25-point penalty paths.

D.8 Overdetermined Regression with Shrinking Noise

A separate Gaussian grid uses $n = 128, k = 8$, $m \in \{256, 640, 1280\}$, and per-observation noise standard deviation $\sigma = 0.15/\sqrt{m}$. Each setting has 100 independent trials. The targets have unit norm and equal active magnitudes. Validation and test sample sizes, penalty grids, and weight floors match the preceding protocol. We also include $q = 4$.

Table 16 reports median relative coefficient error. Table 20 reports normalized expected noisy test MSE, defined as

$$1 + \frac{\left\| X_{\text{test}}(\widehat{\beta} - \beta^\star) \right\|_2^2}{2048\,\sigma^2}.$$

This averages over hypothetical independent test-label noise while keeping the finite test design fixed. Its optimum is one; it differs from both coefficient error and excess risk. Standard errors are across the 100 independent trials. The $q = 2$ and $q = 4$ results agree to the shown precision in several cells; their full-precision records are distinct. This overdetermined, shrinking-noise grid is separate from the underdetermined model of Theorem 10.

D.9 Solvers and Diagnostics

Minimum-norm and restricted fits use a singular-value decomposition. A weighted interpolant is computed as $D^2X^\top(XD^2X^\top)^{-1}y$ using Cholesky factorization. If factorization fails or the interpolation residual exceeds $10^{-8}\max\{\|y\|_2, 1\}$, the solver falls back to a singular-value decomposition of XD. A relative interpolation residual above 10^{-7} terminates the run. No ridge jitter is added to a nominally noiseless fit. Basis pursuit solves the standard split-variable linear program with HiGHS and primal/dual feasibility tolerances 10^{-8}. Solver failure terminates the trial rather than returning a missing or zero error.

Lasso paths use coordinate descent with tolerance 10^{-7} and a maximum of 100000 iterations, retried with 1000000 iterations if needed. The run rejects a path that has an unresolved convergence warning or a dual gap above $10^{-7}\|y\|_2^2/m$. The algebra tests verify the FOCUSS sign cancellation, the $q = 1/2$ quadratic update, both deterministic error bounds on random problems, and the basis-pursuit residual. The experiments use NumPy 1.26.4, SciPy 1.13.1, scikit-learn 1.6.1, and one BLAS thread per process. The result metadata records versions and source/output hashes. These runs provide no comparison of algorithmic wall-clock efficiency.

D.10 Uncertainty and Full Comparisons

Error summaries use medians and interquartile ranges. Table 6 supplies these for every cell in Table 4, together with exact top-k support frequency and pointwise Wilson 95% intervals. Values below 10^{-8} are displayed using that threshold. For a selected set of size k, top-k precision, recall, and F1 all equal $|\widehat{S} \cap S|/k$; Table 8 reports their trial mean. Full summaries include F1 standard errors. For the main noisy table, standard errors are sample standard deviations of per-trial risks divided by $\sqrt{60}$. Table 7 uses per-trial differences between methods, preserving pairing. A negative difference favors $q = 2$. We make no multiple-comparison significance claims.

Table 4: Median relative coefficient error, 200 paired trials per design. Values below 10^{-8} are displayed as such. Table 6 gives quartiles and exact support frequencies; Table 8 gives mean top-k F1 scores.

Design	Min-norm	$q = 1/2$	$q = 1$	$q = 2$	Top-k refit	Basis pursuit
Gaussian	0.856	0.558	0.281	0.039	$< 10^{-8}$	$< 10^{-8}$
Rademacher	0.860	0.566	0.276	0.035	$< 10^{-8}$	$< 10^{-8}$
Laplace	0.855	0.554	0.274	0.037	$< 10^{-8}$	$< 10^{-8}$
AR(0.60)	0.861	0.576	0.293	0.039	0.364	$< 10^{-8}$
AR(0.85)	0.856	0.583	0.296	0.033	0.385	$< 10^{-8}$

Table 5: Mean excess test prediction risk $\pm$ one standard error, multiplied by 10^3, over 60 paired trials. Cases 1–4 have $(n, k, m) = (256, 8, 96)$; case 5 has $(128, 8, 256)$. Cases 1, 2, and 5 are iid Gaussian with $\sigma = 0.05, 0.15, 0.15$, respectively. Case 3 adds a norm-0.1 dense tail to case 2. Case 4 uses AR(0.60) features and $\sigma = 0.15$. Adaptive Lasso uses a as its initial estimate. Risk is measured against noiseless test responses.

Method	Case 1	Case 2	Case 3	Case 4	Case 5
Ridge	627.24 ± 5.91	632.25 ± 5.30	649.34 ± 6.81	441.64 ± 6.22	22.18 ± 0.49
$q = 1/2$	235.93 ± 8.52	261.88 ± 6.71	288.88 ± 9.45	191.34 ± 6.37	6.99 ± 0.20
$q = 1$	65.52 ± 5.07	106.90 ± 4.47	135.82 ± 7.50	82.84 ± 4.64	1.52 ± 0.08
$q = 2$	7.72 ± 1.03	28.87 ± 1.88	53.00 ± 4.99	25.18 ± 2.43	0.78 ± 0.05
Top-k refit	81.29 ± 12.37	106.90 ± 11.59	114.18 ± 11.91	101.93 ± 11.09	0.72 ± 0.04
Lasso	1.52 ± 0.09	13.96 ± 0.62	27.30 ± 1.05	13.25 ± 0.71	3.27 ± 0.15
Adaptive Lasso	0.55 ± 0.05	5.90 ± 0.37	18.53 ± 0.75	6.54 ± 1.09	0.85 ± 0.05

Table 6: Noiseless design comparison: quartiles of relative coefficient error and exact top-k support frequency with Wilson 95% intervals. All rows use 200 trials at $(n, k, m) = (512, 8, 136)$.

Design	Method	25th pct.	Median	75th pct.	Support freq.
Gaussian	Minimum norm	0.847	0.856	0.865	0.60 [0.53, 0.67]
Gaussian	$q = 1/2$	0.527	0.558	0.581	1.00 [0.98, 1.00]
Gaussian	$q = 1$	0.244	0.281	0.312	1.00 [0.98, 1.00]
Gaussian	$q = 2$	0.027	0.039	0.052	1.00 [0.98, 1.00]
Gaussian	Top-k refit	$< 10^{-8}$	$< 10^{-8}$	0.366	0.60 [0.53, 0.67]
Gaussian	Basis pursuit	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	1.00 [0.98, 1.00]
Rademacher	Minimum norm	0.847	0.860	0.868	0.68 [0.61, 0.74]
Rademacher	$q = 1/2$	0.524	0.566	0.595	0.99 [0.97, 1.00]
Rademacher	$q = 1$	0.235	0.276	0.319	1.00 [0.98, 1.00]
Rademacher	$q = 2$	0.025	0.035	0.051	1.00 [0.98, 1.00]
Rademacher	Top-k refit	$< 10^{-8}$	$< 10^{-8}$	0.364	0.68 [0.61, 0.74]
Rademacher	Basis pursuit	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	1.00 [0.98, 1.00]
Laplace	Minimum norm	0.844	0.855	0.865	0.55 [0.48, 0.62]
Laplace	$q = 1/2$	0.524	0.554	0.588	0.99 [0.96, 1.00]
Laplace	$q = 1$	0.239	0.274	0.329	1.00 [0.98, 1.00]
Laplace	$q = 2$	0.028	0.037	0.059	1.00 [0.98, 1.00]
Laplace	Top-k refit	$< 10^{-8}$	$< 10^{-8}$	0.365	0.55 [0.48, 0.62]
Laplace	Basis pursuit	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	1.00 [0.98, 1.00]
AR(0.60)	Minimum norm	0.850	0.861	0.868	0.34 [0.28, 0.41]
AR(0.60)	$q = 1/2$	0.542	0.576	0.602	0.97 [0.94, 0.99]
AR(0.60)	$q = 1$	0.256	0.293	0.330	1.00 [0.98, 1.00]
AR(0.60)	$q = 2$	0.028	0.039	0.051	0.99 [0.97, 1.00]
AR(0.60)	Top-k refit	$< 10^{-8}$	0.364	0.383	0.34 [0.28, 0.41]
AR(0.60)	Basis pursuit	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	1.00 [0.98, 1.00]
AR(0.85)	Minimum norm	0.847	0.856	0.867	0.17 [0.12, 0.23]
AR(0.85)	$q = 1/2$	0.558	0.583	0.614	0.94 [0.90, 0.97]
AR(0.85)	$q = 1$	0.259	0.296	0.334	1.00 [0.98, 1.00]
AR(0.85)	$q = 2$	0.024	0.033	0.044	1.00 [0.98, 1.00]
AR(0.85)	Top-k refit	0.363	0.385	0.528	0.17 [0.12, 0.23]
AR(0.85)	Basis pursuit	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$	1.00 [0.98, 1.00]

Table 7: Paired mean difference in excess test risk: reweighted ridge at $q = 2$ minus the stated baseline. Both numerical columns are multiplied by 10^3. Cases are defined in Table 5. Standard errors use per-trial differences over the 60 common seeds.

Case	Baseline	Mean difference	Standard error
1	Ridge	−619.52	5.45
1	Lasso	6.20	0.98
1	Adaptive Lasso	7.17	1.00
1	Top-k refit	−73.57	11.72
2	Ridge	−603.38	4.43
2	Lasso	14.90	1.62
2	Adaptive Lasso	22.97	1.72
2	Top-k refit	−78.04	10.69
3	Ridge	−596.33	5.49
3	Lasso	25.71	4.49
3	Adaptive Lasso	34.47	4.51
3	Top-k refit	−61.18	9.60
4	Ridge	−416.46	5.39
4	Lasso	11.93	1.94
4	Adaptive Lasso	18.64	1.89
4	Top-k refit	−76.75	10.39
5	Ridge	−21.40	0.48
5	Lasso	−2.49	0.12
5	Adaptive Lasso	−0.07	0.02
5	Top-k refit	0.06	0.02

Table 8: Mean top-k F1, 200 paired trials per design at $(n, k, m) = (512, 8, 136)$. The selected set always has size k, including for dense estimates.

Design	Min-norm	$q = 1/2$	$q = 1$	$q = 2$	Top-k refit	Basis pursuit
Gaussian	0.947	1.000	1.000	1.000	0.947	1.000
Rademacher	0.959	0.999	1.000	1.000	0.959	1.000
Laplace	0.939	0.999	1.000	1.000	0.939	1.000
AR(0.60)	0.899	0.996	1.000	0.999	0.899	1.000
AR(0.85)	0.846	0.992	1.000	1.000	0.846	1.000

Table 9: Main Gaussian grid: number of trials in which the top-k set of a is correct, and frequency of the joint event $\{\mathrm{Top}_k(|a|) = S,\ E_q > 0.1\}$, with Wilson 95% intervals. All 100 trials enter each denominator; $(n, k) = (512, 8)$.

m	a correct	$q = 1/2$	$q = 1$	$q = 2$
48	0/100	0.00 [0.00, 0.04]	0.00 [0.00, 0.04]	0.00 [0.00, 0.04]
64	0/100	0.00 [0.00, 0.04]	0.00 [0.00, 0.04]	0.00 [0.00, 0.04]
96	14/100	0.14 [0.09, 0.22]	0.14 [0.09, 0.22]	0.05 [0.02, 0.11]
136	65/100	0.65 [0.55, 0.74]	0.65 [0.55, 0.74]	0.00 [0.00, 0.04]
192	98/100	0.98 [0.93, 0.99]	0.69 [0.59, 0.77]	0.00 [0.00, 0.04]
256	100/100	1.00 [0.96, 1.00]	0.00 [0.00, 0.04]	0.00 [0.00, 0.04]

Table 10: Large Gaussian grid at $(n, k) = (16384, 8)$: number of trials in which the top-k set of a is correct, and frequency of that event together with $E_q \geq 0.25$. Brackets give pointwise Wilson 95% intervals from 100 trials.

m	a correct	$q = 1/2$	$q = 1$	$q = 2$
48	0/100	0.00 [0.00, 0.04]	0.00 [0.00, 0.04]	0.00 [0.00, 0.04]
64	0/100	0.00 [0.00, 0.04]	0.00 [0.00, 0.04]	0.00 [0.00, 0.04]
96	0/100	0.00 [0.00, 0.04]	0.00 [0.00, 0.04]	0.00 [0.00, 0.04]
128	0/100	0.00 [0.00, 0.04]	0.00 [0.00, 0.04]	0.00 [0.00, 0.04]
192	10/100	0.10 [0.06, 0.17]	0.10 [0.06, 0.17]	0.10 [0.06, 0.17]
256	62/100	0.62 [0.52, 0.71]	0.62 [0.52, 0.71]	0.62 [0.52, 0.71]
384	99/100	0.99 [0.95, 1.00]	0.99 [0.95, 1.00]	0.15 [0.09, 0.23]
512	100/100	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	0.00 [0.00, 0.04]
768	100/100	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	0.00 [0.00, 0.04]
1024	100/100	1.00 [0.96, 1.00]	0.33 [0.25, 0.43]	0.00 [0.00, 0.04]
1536	100/100	1.00 [0.96, 1.00]	0.00 [0.00, 0.04]	0.00 [0.00, 0.04]
2048	100/100	1.00 [0.96, 1.00]	0.00 [0.00, 0.04]	0.00 [0.00, 0.04]

Table 11: Large Gaussian grid at $(n, k) = (65536, 4)$: number of trials in which the top-k set of a is correct, and frequency of that event together with $E_q \geq 0.25$. Brackets give pointwise Wilson 95% intervals from 100 trials.

m	a correct	$q = 1/2$	$q = 1$	$q = 2$
96	20/100	0.20 [0.13, 0.29]	0.20 [0.13, 0.29]	0.20 [0.13, 0.29]
128	64/100	0.64 [0.54, 0.73]	0.64 [0.54, 0.73]	0.64 [0.54, 0.73]
160	96/100	0.96 [0.90, 0.98]	0.96 [0.90, 0.98]	0.96 [0.90, 0.98]
192	100/100	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]
224	100/100	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]
256	100/100	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]
320	100/100	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	0.95 [0.89, 0.98]
384	100/100	1.00 [0.96, 1.00]	1.00 [0.96, 1.00]	0.20 [0.13, 0.29]

Table 12: Error distributions at the separation settings highlighted in the main text. Brackets give interquartile ranges. The joint event is a correct top-k set of a together with $E_q \geq 0.25$. Each row uses 100 trials.

n	k	m	q	Median E_q [IQR]	a correct	Joint event
16384	8	384	1	0.727 [0.707, 0.741]	99/100	99/100
16384	8	512	1	0.594 [0.575, 0.613]	100/100	100/100
16384	8	768	1	0.379 [0.360, 0.397]	100/100	100/100
65536	4	160	2	0.812 [0.777, 0.837]	96/100	96/100
65536	4	192	2	0.693 [0.654, 0.735]	100/100	100/100
65536	4	224	2	0.591 [0.558, 0.633]	100/100	100/100
65536	4	256	2	0.514 [0.465, 0.551]	100/100	100/100
65536	4	320	2	0.332 [0.296, 0.364]	100/100	95/100

Table 13: Median relative coefficient error for geometric active magnitudes. Here d is the largest-to-smallest magnitude ratio, $n = 512$, $k = 8$, and each setting has 100 paired trials.

d	m	μ	k_{eff}	Min-norm	$q = 1/2$	$q = 1$	$q = 2$	Top-k	BP
1	96	8.0	8.00	0.904	0.723	0.529	0.227	0.516	$< 10^{-8}$
1	136	8.0	8.00	0.859	0.562	0.283	0.037	$< 10^{-8}$	$< 10^{-8}$
1	192	8.0	8.00	0.793	0.375	0.113	0.005	$< 10^{-8}$	$< 10^{-8}$
1	256	8.0	8.00	0.713	0.226	0.042	0.001	$< 10^{-8}$	$< 10^{-8}$
2	96	17.7	4.42	0.901	0.703	0.504	0.246	0.415	$< 10^{-8}$
2	136	17.7	4.42	0.858	0.553	0.295	0.067	0.268	$< 10^{-8}$
2	192	17.7	4.42	0.790	0.361	0.114	0.007	$< 10^{-8}$	$< 10^{-8}$
2	256	17.7	4.42	0.706	0.213	0.040	0.001	$< 10^{-8}$	$< 10^{-8}$
4	96	46.9	2.93	0.901	0.675	0.472	0.317	0.386	$< 10^{-8}$
4	136	46.9	2.93	0.858	0.530	0.303	0.131	0.322	$< 10^{-8}$
4	192	46.9	2.93	0.788	0.343	0.119	0.018	0.165	$< 10^{-8}$
4	256	46.9	2.93	0.707	0.204	0.042	0.002	0.148	$< 10^{-8}$
8	96	141.6	2.21	0.899	0.635	0.423	0.301	0.319	$< 10^{-8}$
8	136	141.6	2.21	0.855	0.477	0.265	0.184	0.259	$< 10^{-8}$
8	192	141.6	2.21	0.790	0.322	0.138	0.090	0.180	$< 10^{-8}$
8	256	141.6	2.21	0.710	0.200	0.057	0.010	0.143	$< 10^{-8}$

Table 14: Exact top-k support frequency for the heterogeneous settings of Table 13, with 100 trials per row. Quartiles of coefficient error and Wilson support intervals are in the CSV summary.

d	m	Min-norm	$q = 1/2$	$q = 1$	$q = 2$	Top-k refit	Basis pursuit
1	96	0.08	0.51	0.87	0.96	0.08	1.00
1	136	0.61	0.99	1.00	1.00	0.61	1.00
1	192	0.96	1.00	1.00	1.00	0.96	1.00
1	256	1.00	1.00	1.00	1.00	1.00	1.00
2	96	0.02	0.25	0.71	0.80	0.02	1.00
2	136	0.11	0.86	0.98	0.97	0.11	1.00
2	192	0.63	1.00	1.00	1.00	0.63	1.00
2	256	0.98	1.00	1.00	1.00	0.98	1.00
4	96	0.00	0.01	0.24	0.42	0.00	1.00
4	136	0.00	0.41	0.73	0.74	0.00	1.00
4	192	0.11	0.95	0.99	0.97	0.11	1.00
4	256	0.40	1.00	1.00	1.00	0.40	1.00
8	96	0.00	0.00	0.02	0.02	0.00	1.00
8	136	0.00	0.04	0.29	0.36	0.00	1.00
8	192	0.00	0.44	0.72	0.69	0.00	1.00
8	256	0.02	0.92	0.91	0.90	0.02	1.00

Table 15: Weight-floor sensitivity: median relative coefficient error after each update, using the 100 main-grid trials per m at $(n, k) = (512, 8)$. The relative floor applies only after the first update.

m	Power	Floor	Update 1	Update 2	Update 5	Update 10
96	$q = 1/2$	10^{-4}	7.16×10^{-1}	5.13×10^{-1}	1.05×10^{-1}	3.85×10^{-3}
96	$q = 1/2$	10^{-3}	7.16×10^{-1}	5.13×10^{-1}	1.05×10^{-1}	7.09×10^{-3}
96	$q = 1/2$	10^{-2}	7.16×10^{-1}	5.14×10^{-1}	1.23×10^{-1}	5.34×10^{-2}
96	$q = 1$	10^{-4}	5.12×10^{-1}	4.71×10^{-2}	5.34×10^{-8}	5.34×10^{-8}
96	$q = 1$	10^{-3}	5.12×10^{-1}	4.71×10^{-2}	5.34×10^{-6}	5.34×10^{-6}
96	$q = 1$	10^{-2}	5.12×10^{-1}	4.99×10^{-2}	5.35×10^{-4}	5.35×10^{-4}
136	$q = 1/2$	10^{-4}	5.53×10^{-1}	2.82×10^{-1}	1.68×10^{-2}	3.64×10^{-4}
136	$q = 1/2$	10^{-3}	5.53×10^{-1}	2.82×10^{-1}	1.76×10^{-2}	3.41×10^{-3}
136	$q = 1/2$	10^{-2}	5.53×10^{-1}	2.84×10^{-1}	3.77×10^{-2}	3.36×10^{-2}
136	$q = 1$	10^{-4}	2.72×10^{-1}	4.38×10^{-3}	3.41×10^{-8}	3.41×10^{-8}
136	$q = 1$	10^{-3}	2.72×10^{-1}	4.38×10^{-3}	3.41×10^{-6}	3.41×10^{-6}
136	$q = 1$	10^{-2}	2.72×10^{-1}	5.34×10^{-3}	3.41×10^{-4}	3.41×10^{-4}

Table 16: Median relative coefficient error with $n = 128, k = 8$ and shrinking noise $\sigma = 0.15/\sqrt{m}$, from 100 paired trials per column.

Method	$m = 256$	$m = 640$	$m = 1280$
Ridge	0.00923	0.00293	0.00140
$q = 1/2$	0.00519	0.00186	0.00109
$q = 1$	0.00176	0.00071	0.00038
$q = 2$	0.00161	0.00064	0.00032
$q = 4$	0.00161	0.00064	0.00032
Top-k refit	0.00159	0.00062	0.00030
Lasso	0.00355	0.00127	0.00067
Adaptive Lasso	0.00164	0.00066	0.00034

Table 17: The displayed oracle penalty of Theorem 10, with $(n, k, \sigma, q, \delta) = (2048, 8, 0.15, 2, 0.1)$ and 40 trials per row. Brackets give interquartile ranges. The grid minimum uses true coefficient error on 25 penalties, including the displayed penalty. The last column is the largest per-trial error ratio to that grid minimum.

m	A	Oracle-formula error [IQR]	Grid-best median	Max. ratio
256	1.566	0.138 [0.121, 0.154]	0.138	1.53
512	0.554	0.043 [0.038, 0.047]	0.043	1.09
1024	0.196	0.016 [0.014, 0.019]	0.016	1.04

Table 18: Median relative coefficient error with interquartile ranges for the fixed-noise grid in Table 2. Each cell uses 60 trials. The error is divided by the full target norm, which is 1 for sparse targets and $\sqrt{1.01}$ for the dense-tail targets.

Tail norm	Method	$m = 256$	$m = 512$	$m = 1024$
0	Ridge	0.938 [0.935, 0.940]	0.872 [0.866, 0.876]	0.724 [0.713, 0.730]
0	$q = 1/2$	0.705 [0.690, 0.718]	0.418 [0.402, 0.435]	0.200 [0.191, 0.205]
0	$q = 1$	0.413 [0.392, 0.441]	0.169 [0.158, 0.180]	0.065 [0.062, 0.067]
0	$q = 2$	0.141 [0.125, 0.165]	0.046 [0.038, 0.050]	0.018 [0.016, 0.021]
0	Top-k refit	0.026 [0.022, 0.032]	0.018 [0.016, 0.021]	0.013 [0.010, 0.015]
0	Lasso	0.083 [0.076, 0.091]	0.056 [0.049, 0.058]	0.038 [0.036, 0.041]
0	Adaptive Lasso	0.045 [0.039, 0.052]	0.025 [0.020, 0.030]	0.016 [0.013, 0.019]
0.1	Ridge	0.940 [0.936, 0.944]	0.869 [0.865, 0.876]	0.724 [0.715, 0.732]
0.1	$q = 1/2$	0.716 [0.697, 0.739]	0.423 [0.409, 0.443]	0.221 [0.214, 0.228]
0.1	$q = 1$	0.441 [0.414, 0.476]	0.196 [0.185, 0.208]	0.115 [0.113, 0.118]
0.1	$q = 2$	0.193 [0.169, 0.205]	0.108 [0.106, 0.112]	0.101 [0.100, 0.101]
0.1	Top-k refit	0.105 [0.103, 0.108]	0.102 [0.101, 0.103]	0.101 [0.100, 0.101]
0.1	Lasso	0.141 [0.133, 0.148]	0.117 [0.114, 0.120]	0.107 [0.106, 0.109]
0.1	Adaptive Lasso	0.114 [0.110, 0.118]	0.104 [0.103, 0.105]	0.101 [0.101, 0.102]

Table 19: Paired risk comparisons for the fixed-noise grid. Differences are $q = 2$ minus the named baseline; means and standard errors are multiplied by 10^3. The last column counts trials in which $q = 2$ has lower risk, out of the 60 paired trials.

Tail norm	m	Baseline	Mean difference	Standard error	$q = 2$ lower
0	256	Lasso	14.604	0.968	0/60
0	256	Adaptive Lasso	19.635	1.078	0/60
0	256	Top-k refit	3.668	5.604	8/60
0	512	Lasso	−0.989	0.062	58/60
0	512	Adaptive Lasso	1.363	0.063	0/60
0	512	Top-k refit	1.669	0.080	0/60
0	1024	Lasso	−1.168	0.038	60/60
0	1024	Adaptive Lasso	0.058	0.015	9/60
0	1024	Top-k refit	0.180	0.018	0/60
0.1	256	Lasso	17.853	1.173	0/60
0.1	256	Adaptive Lasso	24.791	1.353	0/60
0.1	256	Top-k refit	9.146	5.325	8/60
0.1	512	Lasso	−1.895	0.087	60/60
0.1	512	Adaptive Lasso	1.067	0.080	2/60
0.1	512	Top-k refit	1.511	0.110	1/60
0.1	1024	Lasso	−1.315	0.070	57/60
0.1	1024	Adaptive Lasso	−0.251	0.042	57/60
0.1	1024	Top-k refit	0.018	0.038	37/60

Table 20: Normalized expected noisy test MSE, mean ± standard error, for the same 100-trial shrinking-noise grid as Table 16. The optimum is one. The metric is defined in Appendix D.8.

Method	$m = 256$	$m = 640$	$m = 1280$
Ridge	1.9922 ± 0.0204	1.2437 ± 0.0039	1.1127 ± 0.0015
$q = 1/2$	1.3093 ± 0.0078	1.1031 ± 0.0022	1.0696 ± 0.0014
$q = 1$	1.0370 ± 0.0019	1.0148 ± 0.0008	1.0101 ± 0.0008
$q = 2$	1.0324 ± 0.0018	1.0134 ± 0.0008	1.0081 ± 0.0007
$q = 4$	1.0325 ± 0.0018	1.0134 ± 0.0008	1.0081 ± 0.0007
Top-k refit	1.0316 ± 0.0017	1.0123 ± 0.0007	1.0061 ± 0.0004
Lasso	1.1464 ± 0.0054	1.0486 ± 0.0015	1.0278 ± 0.0010
Adaptive Lasso	1.0333 ± 0.0018	1.0140 ± 0.0008	1.0086 ± 0.0007