GROKKING THROUGH THE LENS OF MINIMUM-NORM INTERPOLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Grokking shows that fitting the training data and learning the underlying signal can occur at very different stages. However, existing theories offer limited quantitative insight into how this delayed generalization depends on inductive bias and signal structure. Our work addresses the gap by developing a statistical theory that characterizes how regularization geometry and signal sparsity govern generalization near interpolation. In particular, we focus on the prototypical setting of high-dimensional regression and identify regimes in which sparsity-promoting regularization makes exact interpolation much more accurate than approximate fitting. In strongly overparameterized noiseless problems, we prove a zero–one generalization law and construct a family of convex norms whose interpolators transition from the trivial risk of the all-zero predictor to exact recovery, while keeping the training error equal to 0. Furthermore, when feature dimension and sample size are proportional, we provide a precise characterization of training and generalization errors along ℓ_r-regularization paths. This in turn allows us to quantify the generalization gain that remains near interpolation: we show that this gain increases as the norm becomes more sparsity-promoting and as the target becomes sparser, with a sharp drop in generalization reached for noiseless data and ℓ_1 regularization. Experiments on diagonal linear networks and transformers trained on modular arithmetic demonstrate the generality of our theoretical predictions. Finally, beyond grokking, our work reveals a statistical instability in minimum-norm interpolation: small perturbations in the regularization strength can lead to drastically different generalization, while preserving small training error.

1 INTRODUCTION

Overparameterized models can fit their training data exactly, even in the presence of noise, and nevertheless generalize well (Belkin et al., 2018; Muthukumar et al., 2020). This phenomenon, known as benign overfitting, has been rigorously studied for minimum-norm interpolators (MNIs) in linear regression (Bartlett et al., 2020; Tsigler and Bartlett, 2023) and reproducing kernel Hilbert spaces (Liang and Rakhlin, 2020; Liang et al., 2020), clarifying conditions under which interpolation is compatible with accurate prediction. Grokking provides an extreme manifestation of this compatibility between interpolation and generalization: a model attains near-zero training error yet generalizes poorly for many additional training steps; then, the test error drops sharply while the training error remains close to zero (Power et al., 2022; Nanda et al., 2023). Initially observed in small algorithmic problems, the phenomenon has been reported in image, language, and molecular prediction tasks (Liu et al., 2023; Humayun et al., 2024). This raises the question of which properties of the data and the learning procedure allow solutions with almost indistinguishable training errors to have dramatically different generalization.

Existing theoretical work explains grokking by analyzing linear estimators and classification (Levi et al., 2024; Beck et al., 2025), transitions between lazy and rich training (Kumar et al., 2024; Lyu et al., 2024; Mohamadi et al., 2024), and slow norm minimization under weight decay (Boursier et al., 2025). However, turning these explanations into an end-to-end guarantee requires proving early fitting, a prolonged period of poor generalization, and a final improvement in the test error. Xu et al. (2026) establish such guarantees, together with quantitative bounds on generalization delay. Their analysis nevertheless concerns over-parameterized linear regression with ℓ_2 regularization.

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.

This restriction is significant since nonlinear parameterization fundamentally changes the bias of training. For example, in diagonal linear networks, the scale and shape of initialization determine an implicit regularizer that is a hybrid between ridge ℓ_2-like and sparsity-promoting ℓ_1-like geometries (Woodworth et al., 2020; Azulay et al., 2021). Furthermore, existing analyses of gradient flow in diagonal linear networks with vanishing initialization establish convergence to the ℓ_1-MNI (Pesme and Flammarion, 2023) and connect training dynamics to ℓ_1-regularization paths (Berthier, 2023; 2025). Recently, Tikeng Notsawo et al. (2025) have connected grokking to sparsity- and low-rank-promoting regularization. These results establish the importance of matching regularization to the structure of the target, but leave open how this match governs the sharpness of generalization near interpolation. In particular, a quantitative account of how much generalization can improve for a small reduction in training error, and how this improvement varies with the regularizing norm and signal sparsity, is still missing.

To address this gap, we develop a statistical theory that studies generalization along regularization paths induced by different norms, and we compare near-interpolating solutions with the corresponding MNI. More precisely, we consider a linear model and the regularized square-root loss, which allows the optimum to be an exact interpolator when the regularization is below a critical threshold (see (2) for details). Our results establish when approaching exact interpolation produces a sharp drop in generalization error, and show how this behavior depends on the geometry of the norm and the sparsity of the problem. These theoretical predictions then shed light into late-stage generalization under non-linear training of diagonal linear networks and transformer models. We summarize our main contributions as follows:

- **A zero–one law for generalization, under exact interpolation.** When the number of features significantly exceeds the sample size, we construct a family of convex ℓ_1–ℓ_∞ hybrid norms for noiseless sparse regression. As the norm geometry varies, we show that the generalization performance exhibits a zero–one transition between exact recovery and trivial error, while every minimizer keeps interpolating exactly (Theorem 1). This gives a sharp separation between fitting the training data and recovering the underlying signal.
- **Quantifying the roles of norm geometry and sparsity.** When the number of features scales proportionally to the sample size, we use the convex Gaussian min–max theorem (CGMT) to characterize training and generalization error of ℓ_r-regularized square-root regression (Lemma 2), as well as the generalization error of ℓ_r-MNIs (Lemma 3). Using this characterization, we provide quantitative conditions leading to a sharp drop in generalization near interpolation. In particular, we show that this drop is more prominent *(i)* as the norm passes from ℓ_2 to ℓ_1 (Theorem 2), with the most extreme example for ℓ_1 regularization and noiseless data (Theorem 3), and *(ii)* as the regression problem becomes sparser (Theorem 4).
- **Evidence from training neural networks.** We test these predictions in two-layer diagonal linear networks, whose nonlinear parameterization induces an initialization-dependent implicit bias. The results support the predicted ordering of late-stage generalization gains: stronger sparsity-promoting bias and sparser targets yield larger improvements near interpolation. Experiments with transformers trained on modular arithmetic also show larger gains in generalization at smaller output scales, qualitatively consistent with this picture (Figure 2).

Beyond grokking, our results highlight a surprising statistical instability of the regularized square-root loss around its corresponding MNI: small perturbations of the norm geometry yield a zero–one generalization transition while preserving an exact fit (Theorem 1) and, along the ℓ_1 regularization path, generalization changes much faster than training error near interpolation (Theorem 3).

2 Related work

Grokking, implicit bias, and regularization. Initialization scale, layer imbalance, and target or output rescaling shape feature learning (Chizat et al., 2019; Saxe et al., 2014; Domine et al., 2025; Jarvis et al., 2025; Anguita et al., 2026; Kumar et al., 2024; Nam et al., 2025). In particular, Kunin et al. (2024); Kumar et al. (2024) show that unbalanced initializations and output rescaling can accelerate feature learning and reduce grokking delays in transformers. Data availability also influences learning regime and grokking profile: a "Goldilocks zone" allows eventual structured generalization while retaining an initial memorization phase (Liu et al., 2022; Kumar et al., 2024). For modular

addition, Mohamadi et al. (2024) establish a sample-complexity separation between permutation-equivariant kernel methods and quadratic networks. Early oscillations in the training loss have been used to predict grokking (Tikeng Notsawo et al., 2023), while Jeffares et al. (2024) connect its emergence empirically to benign overfitting. Related work has also focused on sparse subnetworks (Merrill et al., 2023), norm-efficient circuits (Varma et al., 2023), and low-rank weights (Yunis et al., 2024). Focusing on regularization, Tikeng Notsawo et al. (2025) analyze an initial loss-minimization phase followed by a slower dynamics driven by regularization, with applications to sparse recovery regularized via the ℓ_1 norm and matrix factorization regularized via the nuclear norm. ℓ_1 and ℓ_2 regularization are compared by Žunkovič and Ilievski (2024) in the context of linear classification. Most recently, Xu et al. (2026) quantify how initialization and weight decay affect the generalization delay for ridge regression. Our work goes beyond ridge regularization: motivated by explicit characterizations of implicit bias (Woodworth et al., 2020; Azulay et al., 2021), we study regularization paths in high-dimensional regression and quantify how generalization gains near interpolation depend on norm geometry and signal sparsity.

Precise asymptotics for regularized regression and minimum-norm interpolators. A rich line of work has studied high-dimensional regularized estimators and MNIs via different techniques. Approximate message passing yields exact asymptotics for Lasso (Bayati and Montanari, 2012), robust M-estimation (Donoho and Montanari, 2016), ℓ_1-MNI (Li and Wei, 2021) and ℓ_r-regularized least squares (Weng et al., 2018). In particular, Weng et al. (2018) study the effects of norm exponent and sparsity on the optimally tuned risk in the low-noise limit; in contrast, we focus on the generalization gain near interpolation at a fixed noise level. Kur et al. (2024); Kur and Bizeul (2026) relate the generalization of MNIs to 2-uniform convexity via the local theory of Banach spaces. Zhou et al. (2024) derive uniform bounds connecting population risk to training error and model complexity, including for near-interpolating predictors. Our strategy builds on Gordon's Gaussian comparison inequalities and the convex Gaussian min–max theorem (CGMT) (Gordon, 1985; Stojnic, 2013; Thrampoulidis et al., 2015b). These methods provide a characterization of the square-root Lasso and its generalizations (Oymak et al., 2013; Thrampoulidis et al., 2015a), convex regularized M-estimators (Thrampoulidis et al., 2018), as well as uniform distributional guarantees for the Lasso (Miolane and Montanari, 2021). Gaussian comparison methods also establish the generalization error of interpolators minimizing separable convex penalties (Varma and Hassibi, 2025) and yield sharp rates for ℓ_r-MNIs (Wang et al., 2022; Donhauser et al., 2022). Our contribution is a unified characterization of ℓ_r-MNIs and ℓ_r-regularized regression with the square-root loss, which we then use to understand the conditions causing a sharp drop in the generalization error near interpolation.

3 Preliminaries

For $a, b \in \mathbb{R}^p$ and $r \geq 1$, we write $\langle a, b\rangle := \sum_{j=1}^p a_j b_j$ and $\|b\|_r := \left(\sum_{j=1}^p |b_j|^r\right)^{1/r}$. We consider a linear regression model

$$y = X\beta_* + \varepsilon, \tag{1}$$

where $\beta_* \in \mathbb{R}^p$ is the vector of regression coefficients, $X \in \mathbb{R}^{n\times p}$ is the design matrix with independent $N(0, 1/n)$ entries, $y \in \mathbb{R}^n$ is the vector of observations, and the noise vector $\varepsilon \sim N(0, \sigma^2 I_n)$ is independent of X. We study estimators optimizing the ℓ_r-regularized square-root loss

$$\widehat{\beta}_{\kappa,r} \in \operatorname*{arg\,min}_{b\in\mathbb{R}^p} \left\{ \|y - Xb\|_2 + \frac{\kappa\lambda(n,p)}{\sqrt{n}} \|b\|_r \right\}. \tag{2}$$

For $r = 1$, this estimator is the square-root Lasso (Belloni et al., 2011). We note that, when the coefficient $\kappa\lambda(n,p)$ in front of the ℓ_r regularization is smaller than a critical value (see e.g. Theorem 4.2 in Friedlander and Tseng (2008)), $\widehat{\beta}_{\kappa,r}$ coincides with the ℓ_r-MNI given by

$$\widehat{\beta}_r \in \operatorname*{arg\,min}_{b\in\mathbb{R}^p} \{\|b\|_r : Xb = y\}. \tag{3}$$

For convenience, we express this coefficient as $\kappa\lambda(n,p)$, so the critical value of the regularization leading to interpolation is achieved when κ is a universal constant that does not scale with n, p. The precise form of $\lambda(n,p)$ depends on the scaling between n and p, and it is clarified later.

We measure the generalization error of an estimator via its mean squared error (MSE). Thus, the training and generalization error of an estimator β are respectively given by

$$\mathcal{E}(\beta) := \frac{1}{\sqrt{n}}\|y - X\beta\|_2, \qquad \mathsf{MSE}(\beta) := \frac{1}{p}\|\beta - \beta_*\|_2^2. \tag{4}$$

Let $0_p \in \mathbb{R}^p$ be the all-0 vector and $\mathsf{MSE}(0_p)$ its corresponding MSE. We denote normalized MSE and training error by $\mathsf{MSE}_s(\beta) := \mathsf{MSE}(\beta)/\mathsf{MSE}(0_p)$ and $\mathcal{E}_s(\beta) := \mathcal{E}(\beta)/\mathcal{E}(0_p)$. Throughout the paper, $c, C \in (0,1)$ are absolute constants and $c(\alpha), C(\alpha)$ constants that only depends on α. We use $O(\cdot), o(\cdot), \Omega(\cdot), \omega(\cdot)$ for asymptotic notation in n, p, and $\mathrm{conv}\{\cdot\}$ denotes the convex hull operator.

4 Main Results

Our results cover two asymptotic regimes. In the strongly overparameterized regime ($p \gg n$), Section 4.1 establishes that, while the training error remains negligible, the MSE exhibits a zero-one transition. In the proportional regime ($p \asymp n$), Section 4.2 gives precise characterizations of ℓ_r-regularized estimators and minimum-norm interpolators. Building on these characterizations, Section 4.3 then shows that smaller norm exponents and sparser regression problems yield sharper generalization gains near interpolation.

4.1 A zero–one law for ℓ_1 regularization when $p \gg n$

We consider the noiseless Gaussian model in (1), with $y = X\beta_*$ and

$$n, p \to \infty, \qquad p/n \longrightarrow \infty, \qquad \log p = o(n). \tag{5}$$

Set $L := 2\log(p/n)$ and $\lambda(n,p) = \sqrt{L}$ in (2). Set $\nu = \nu_{n,p} > 0$ with $\nu^2 = L - \log L + O(1)$ and deterministic $\eta = \eta_{n,p} \in (0,1)$ such that

$$k_* := \left\lfloor (1-\eta)\frac{n}{\nu^2} \right\rfloor \approx \frac{n}{2\log(p/n)}, \qquad \eta = o(1), \qquad L^{-1} + \frac{L}{n} = o(\eta). \tag{6}$$

Let $\beta_* \in \mathbb{R}^p$ be any fixed k_* sparse vector (i.e. $\|\beta_*\|_0 = k_*$) with $+1$ or -1 entries. We remark that k_* is also known as the leading weak-recovery sparsity scale (Donoho and Tanner, 2009).

Our first result shows a discontinuity in the MSE along the regularization path in the noiseless case.

Lemma 1 (Zero–one generalization)**.** *Under the assumptions above, fix $\kappa > 0$, $\kappa \neq 1$, and consider (2) with $r = 1$ and $\sigma = 0$. Then, the minimizer $\widehat{\beta}_{\kappa,1}$ satisfies*

$$\left(\mathsf{MSE}_s(\widehat{\beta}_{\kappa,1}), \mathcal{E}_s(\widehat{\beta}_{\kappa,1})\right) \xrightarrow{\mathbb{P}} \left(\mathbf{1}_{\{\kappa>1\}}, \mathbf{1}_{\{\kappa>1\}}\right). \tag{7}$$

In words, (7) shows that a discontinuity in both training and generalization error is possible despite the convexity of the norm. The proof combines Gaussian comparison inequalities with the connection between symmetric Gaussian polytopes $Q_{n,p} := \mathrm{conv}\{\pm X_j : 1 \le j \le p\} \subset \mathbb{R}^n$ and ℓ_1-MNI (Donoho and Tanner, 2005; Kur and Pathak, 2026). The Minkowski functional of $Q_{n,p}$ gives the minimum ℓ_1 budget needed to interpolate the data, and a sharp estimate of its spherical mean guides the calibration of the sparsity k_*; see (26). For $\kappa < 1$, this calibration leaves a strict margin in Gordon's comparison theorem, ensuring that every nonzero perturbation of the ground truth increases the objective in (2), hence yielding unique exact recovery. For $\kappa > 1$, a clipping argument and properties of the Gaussian distribution give an ℓ_1–ℓ_2 norm comparison for $\widehat{\beta}_{\kappa,1}$. Then, a restricted-isometry argument and a comparison of the objective in (2) at $\widehat{\beta}_{\kappa,1}$ and at the zero estimator show that the normalized ℓ_2 energy of $\widehat{\beta}_{\kappa,1}$ vanishes.

Zero–one generalization with zero training error. The preceding lemma does not separate fitting from generalization: training and generalization error undergo the same transition. To keep exact interpolation while generalization exhibits a zero–one law, we slightly modify the ℓ_1 norm. Specifically, we replace ℓ_1 by its infimal convolution with a scaled ℓ_∞ norm:

$$N_\kappa(b) := \inf_{u+v=b} \left\{ \|u\|_1 + \frac{p}{\kappa}\sqrt{\frac{2}{\pi L}}\,\|v\|_\infty \right\}. \tag{8}$$

Its unit ball is $\operatorname{conv}\left\{B_1^p, \frac{\kappa}{p}\sqrt{\pi L/2}\, B_\infty^p\right\}$, and its geometric distance to the ℓ_1-norm is within a factor $O_\kappa(\sqrt{L})$, as

$$\frac{1}{\kappa}\sqrt{\frac{2}{\pi L}}\,\|b\|_1 \le N_\kappa(b) \le \|b\|_1, \qquad \forall b \in \mathbb{R}^p. \tag{9}$$

Leveraging this norm allows us to prove the following extreme example of grokking.

Theorem 1 (Zero–one generalization at interpolation)**.** *Under the same assumptions, fix $\kappa > 0$, $\kappa \neq 1$, and let*

$$\widetilde{\beta}_\kappa = \operatorname*{arg\,min}_{b\in\mathbb{R}^p}\left\{\|y - Xb\|_2 + \frac{\kappa\sqrt{L}}{2\sqrt{n}}\, N_\kappa(b)\right\}. \tag{10}$$

Then, it holds

$$\left(\mathsf{MSE}_s(\widetilde{\beta}_\kappa), \mathcal{E}_s(\widetilde{\beta}_\kappa)\right) \xrightarrow{\mathbb{P}} \left(\mathbf{1}_{\{\kappa>1\}}, 0\right). \tag{11}$$

The idea is that the solution $\widetilde{\beta}_\kappa$ in (10) is *(i)* k_*-sparse when $\kappa < 1$, and *(ii)* dense (all entries have roughly the same size) and interpolating when $\kappa > 1$. The result is illustrated in the left plot of Figure 1, which clearly displays how the normalized MSE jumps from 0 to 1 while the training error remains extremely small. Full proofs, as well as sketches, of the results in this subsection appear in Appendix A, where we also discuss the technical novelty of our approach (see Remarks 1-2 therein).

4.2 Precise characterization of estimators when $p \asymp n$

We next study the regime in which the sample size n and the feature dimension p grow at the same rate. Our aim is to characterize both training error and MSE for the family of ℓ_r regularizers, and Lemmas 2-3 below give a scalar description of these two regimes in terms of a pair of fixed point equations. This provides the starting point to establish under what conditions the MSE exhibits a sharp drop once the training error is already close to 0, as done in Section 4.3.

More formally, throughout this subsection and Section 4.3, we consider the model in (1) and assume $n, p \to \infty$ with $\delta_p = n/p \to \delta \in (0, \infty)$. We set $\lambda(n, p) = p^{1-1/r}$ in (2), so that $\widehat{\beta}_{\kappa,r}$ coincides with the MNI $\widehat{\beta}_r$ in (3) when κ is smaller than a universal constant independent of n, p. The empirical distribution of β_* is assumed to converge to P_B in W_q, with $q = \max\{2, r\}$. The scalar description given by Lemmas 2-3 is expressed through

$$f_{\tau,\alpha} := \tau\eta_{\alpha,r}(Z), \qquad Z = Z(\tau, B) := \frac{B}{\tau} + G, \qquad B \sim P_B, \quad G \sim N(0, 1), \tag{12}$$

where B and G are independent and the proximal operator $\eta_{\alpha,r}(z)$ is given by $\arg\min_{u\in\mathbb{R}}\left\{\frac{1}{2}(u - z)^2 + \frac{\alpha}{r}|u|^r\right\}$ (for $r = 1$, $\eta_{\alpha,1}(z) := \operatorname{sign}(z)(|z| - \alpha)_+$). We also define

$$H_r(\tau, \alpha) := \mathbb{E}\left[\left(\eta_{\alpha,r}(Z) - \frac{B}{\tau}\right)^2\right], \quad D_r(\tau, \alpha) := \mathbb{E}\left[\eta'_{\alpha,r}(Z)\right], \quad U_r(\tau, \alpha) := \mathbb{E}\left[|\eta_{\alpha,r}(Z))|^r\right], \tag{13}$$

with $D_1(\tau, \alpha) = \mathbb{P}(|Z| > \alpha)$. The scalar estimator in (12) applies shrinkage to a signal coordinate observed with effective Gaussian noise of standard deviation τ. The functions in (13) summarize its behavior: $\tau^2 H_r(\tau, \alpha) = \mathbb{E}(f_{\tau,\alpha} - B)^2$ is the MSE, $\tau^r U_r(\tau, \alpha) = \mathbb{E}|f_{\tau,\alpha}|^r$ is the r-th absolute moment and, as we will see, $D_r(\tau, \alpha)$ determines the training error $\mathcal{E}(\widehat{\beta}_{\kappa,r})$ via the expression $\tau(1 - D_r(\tau, \alpha)/\delta)$. The parameters τ and α will be identified via a set of two fixed point equations.

Regularized regression with square-root loss. We start by characterizing the estimator in (2) when its training error $\mathcal{E}(\widehat{\beta}_{\kappa,r})$ has a strictly positive limit.

Lemma 2 (Precise characterization for nonzero training error)**.** *Fix $\kappa > 0$. Let $\tau_\kappa, \alpha_\kappa > 0$ satisfy*

$$\tau_\kappa^2 = \sigma^2 + \frac{\tau_\kappa^2}{\delta} H_r(\tau_\kappa, \alpha_\kappa), \qquad \kappa = \begin{cases} \alpha_\kappa, & r = 1, \\ \alpha_\kappa U_r(\tau_\kappa, \alpha_\kappa)^{(r-1)/r}, & 1 < r < \infty. \end{cases} \tag{14}$$

Define

$$R_\kappa := \tau_\kappa\left(1 - \frac{D_r(\tau_\kappa, \alpha_\kappa)}{\delta}\right). \tag{15}$$

If $R_\kappa > 0$, then, almost surely, the minimizer $\widehat{\beta}_{\kappa,r}$ in (2) satisfies

$$\mathcal{E}(\widehat{\beta}_{\kappa,r}) \to R_\kappa, \qquad \mathsf{MSE}(\widehat{\beta}_{\kappa,r}) \to \tau_\kappa^2 H_r(\tau_\kappa, \alpha_\kappa) = \delta(\tau_\kappa^2 - \sigma^2). \tag{16}$$

In words, the first equation in (14) determines the effective noise level from the noise variance and the MSE. The second equation calibrates the shrinkage parameter α_κ to the regularization strength κ. Solving for $(\tau_\kappa, \alpha_\kappa)$ determines both limiting training error and the MSE via (16).

The argument uses the CGMT framework to compare (2) with an auxiliary optimization involving independent Gaussian vectors. We first show that its population counterpart has the unique minimizer $f_{\tau_\kappa,\alpha_\kappa}$, characterized by the fixed-point equations (14). We then prove the stability of the minimizers of the auxiliary problem. This in turn allows us to use the CGMT to transfer their asymptotic behavior to the original estimator, thus yielding (16). The full proof is deferred to Appendix B, with an overview in Appendix B.1. Beyond the limits in (16), the argument also gives convergence in W_2 of the joint empirical distribution of $(\widehat{\beta}_{\kappa,r}, \beta_*)$, which implies convergence of the empirical averages of functions with quadratic growth, see Remark 3 in Appendix B.

Minimum-norm interpolation. For $\delta < 1$, we then provide a similar characterization for the minimum-ℓ_r-norm solution that fits the observations exactly.

Lemma 3 (Precise characterization for the MNI)**.** *Let $\delta \in (0, 1)$. Suppose that $\tau_{\mathrm{int}}, \alpha_{\mathrm{int}} > 0$ satisfy*

$$\tau_{\mathrm{int}}^2 = \sigma^2 + \frac{\tau_{\mathrm{int}}^2}{\delta} H_r(\tau_{\mathrm{int}}, \alpha_{\mathrm{int}}), \qquad D_r(\tau_{\mathrm{int}}, \alpha_{\mathrm{int}}) = \delta. \tag{17}$$

Then, almost surely, the minimum-norm interpolator $\widehat{\beta}_r$ in (3) satisfies

$$\mathsf{MSE}(\widehat{\beta}_r) \to \tau_{\mathrm{int}}^2 H_r(\tau_{\mathrm{int}}, \alpha_{\mathrm{int}}) = \delta\big(\tau_{\mathrm{int}}^2 - \sigma^2\big). \tag{18}$$

We note that the first scalar equation in (17) coincides with the one in (14); the second scalar equation in (17) corresponds to setting $R_\kappa = 0$ in (15), enforcing exact interpolation in the characterization. The proof first shows that a sufficiently small positive regularization parameter enforces minimum-norm interpolation, and then adapts the CGMT analysis of Lemma 2 to the case $D_r = \delta$. The full argument is in Appendix C.

Relation to previous work. Specializing Theorem 3.1 of Thrampoulidis et al. (2018) to the square-root loss and ℓ_r regularization gives the limit in probability of $\mathsf{MSE}(\widehat{\beta}_{\kappa,r})$. Lemma 2 additionally establishes the limit R_κ of the normalized training error, which is needed for our analysis of the interpolation threshold in Section 4.3. Furthermore, it gives almost-sure convergence and the joint empirical W_2 limit discussed in Remark 3.

For minimum-norm interpolation, Lemma 3 extends Li and Wei (2021) which is restricted to $r = 1$. For $r \geq 1$, specializing Theorem 3.1 of Varma and Hassibi (2025) to isotropic features and the potential $\psi(t) = |t|^r/r$ gives the limit in probability of $\mathsf{MSE}(\widehat{\beta}_r)$ for β_* having i.i.d. coordinates. Lemma 3 relaxes this i.i.d. assumption, and it gives almost-sure convergence.

We finally note that the application of results by Thrampoulidis et al. (2018); Varma and Hassibi (2025) still requires showing uniqueness of some parameters involved in the characterization. Our analysis directly gives uniqueness, after assuming existence of the solution of the fixed point equations; furthermore, existence at and just above the interpolation threshold follows from Lemma 24 (Appendix D.1) when $\sigma > 0$, $\delta \in (0, 1)$ and $r \in (1, 2)$.

4.3 Generalization near interpolation: norm geometry and sparsity

We now use the characterization in Section 4.2 to study how norm geometry (captured by the parameter r) and sparsity of β_* affect generalization near interpolation. We focus on comparing the decrease in MSE and training error, establishing conditions under which the MSE decreases more rapidly having fixed the training error.

More formally, consider $\delta \in (0, 1)$ and let $\overline{R}_r(\kappa)$ denote the limiting normalized training error of a global minimizer: $\overline{R}_r(\kappa)$ is either 0 or it equals $R_\kappa > 0$ in (15). Since $\overline{R}_r(\kappa)$ is nondecreasing in κ (Lemma 23 in Appendix D.1), we define the critical regularization leading to interpolation as

$$\kappa_r^\star := \sup\{\kappa \geq 0 : \overline{R}_r(\kappa) = 0\} = \inf\{\kappa \geq 0 : \overline{R}_r(\kappa) > 0\}. \tag{19}$$

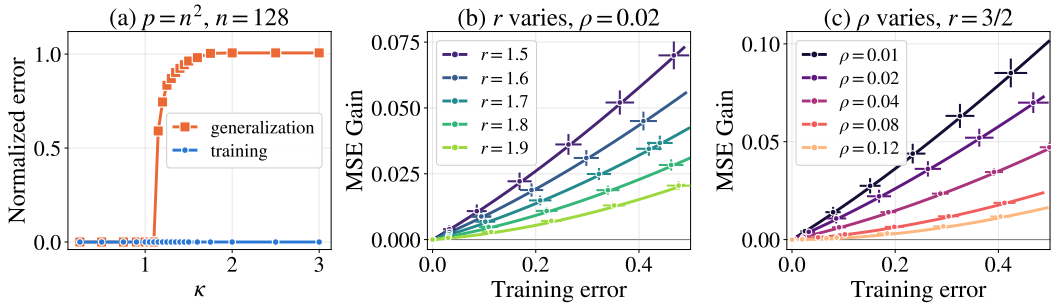

Figure 1: Training and generalization error (MSE) for regression regularized with the convex $\ell_1 - \ell_\infty$ norm in (8) (Left) and with ℓ_r norms (Center, Right). *Left:* We pick β_* to be a k_*-sparse vector with uniformly distributed ± 1 entries, consider the estimator in (10) and plot normalized training (blue) and generalization (orange) errors as a function of regularization strength κ. We report the mean $\pm$ one standard error over 10 seeds. The generalization error exhibits a zero–one transition around $\kappa = 1$, while the training error remains close to 0 (Theorem 1). *Center, Right:* We pick B as in (21) with W a Rademacher random variable, $\delta = 0.2$, $\sigma = 0.1$, $n = 1000$ and $p = 5000$. We consider the estimator $\widehat{\beta}_{\kappa,r}$ in (2) and plot MSE gain $\overline{\mathsf{MSE}}_r(\kappa) - \overline{\mathsf{MSE}}_r(\kappa_r^\star)$ as a function of training error $\mathcal{E}(\widehat{\beta}_{\kappa,r})$. Solid curves are theoretical predictions from Lemma 2, and dots are obtained by solving regularized regression numerically (mean $\pm$ one standard error over 20 seeds). In the central plot, for a fixed $\rho = 0.02$, the curves flatten as r increases (Theorem 2). In the right plot, for a fixed $r = 3/2$, the curves flatten as ρ increases (Theorem 4).

For $\kappa > \kappa_r^\star$, we denote the limiting MSE as $\overline{\mathsf{MSE}}_r(\kappa) := \delta\{\tau_\kappa^2 - \sigma^2\}$ and define

$$S_r := \lim_{\kappa \downarrow \kappa_r^\star} \frac{\frac{d}{d\kappa}\overline{\mathsf{MSE}}_r(\kappa)}{\frac{d}{d\kappa}R_\kappa}. \tag{20}$$

The quantity S_r measures how much generalization changes for a given change in training error: $S_r > 0$ means that an exact interpolation outperforms approximate one, and a larger value of S_r is evidence of more significant grokking. Throughout this section, we focus on the sparse family of regression problems where the empirical distribution of β_* approaches the law of

$$B = \frac{\xi_\rho W}{\sqrt{\rho}}, \qquad \xi_\rho \sim \text{Bernoulli}(\rho), \tag{21}$$

where ξ_ρ and W are independent. This choice is rather broad, and it includes Rademacher, Gaussian, bounded, and finite-mixture random variables.

Role of norm geometry. The result below shows that, for sparse regression problems, grokking is more pronounced as the regularization passes from ℓ_2 to ℓ_1, i.e., S_r is decreasing in r.

Theorem 2 (Smaller r implies more grokking)**.** *Fix $1 < r_- < r_+ < 2$ and $\sigma \geq 0$. For a fixed $\lambda > 0$, let B be given by (21) with $\rho = \lambda\delta$. Assume $\mathbb{E}W^2 = 1$ and $\mathbb{E}|W|^{2/(r_- - 1)} < \infty$. Then, there exist $\lambda_0 := \lambda_0(r_-, r_+, W) > 0$, $\delta_0 := \delta_0(r_-, r_+, W, \lambda, \sigma) > 0$ such that, whenever $\lambda \in (0, \lambda_0)$ and $\delta \in (0, \delta_0)$, S_r is strictly positive and decreasing in r on (r_-, r_+).*

In words, as $\delta \to 0$ (i.e., as the model gets more over-parameterized) and for a sufficiently sparse problem (i.e., $\rho \lesssim \delta$), exact interpolation generalizes better than approximate fitting, and smaller norm exponents yield a sharper decrease in the MSE. Remarkably, the result holds regardless of the amount of noise σ in the data. We note that $\rho/\delta = \lambda$ is small, so the number of nonzero entries in β_* is small compared to n, but the relative sparsity ρ does not vanish in n ($\rho = \Theta(1)$). The result is illustrated in the central plot of Figure 1, and its proof is deferred to Appendix D.2.

We next show that, for noiseless data ($\sigma = 0$) and ℓ_1 regularization ($r = 1$), the MSE improvement for a given change in the training error may be unbounded. This result aligns with Lemma 1 above, albeit in a different regime of n, p, and it is proved in Appendix D.3.

Theorem 3 (Unbounded improvement in MSE for ℓ_1 regularization). *Fix $0 < \rho < 1$, and let B be given by (21). Assume that $\mathbb{E}W^2 = 1$ and, as $t \downarrow 0$,*

$$\mathbb{P}(|W| \le t) = o(t^2). \tag{22}$$

Set $r = 1$, $\sigma = 0$, and

$$\delta = \delta_c(\rho) := \min_{\alpha \ge 0} \left\{ \rho(1+\alpha^2) + (1-\rho)\mathbb{E}(|G| - \alpha)_+^2 \right\}, \qquad G \sim N(0,1). \tag{23}$$

Let $\alpha_c > 0$ be the minimizer of (23). Then $\kappa_1^\star = \alpha_c$, and the limit in (20) satisfies $S_1 = +\infty$.

In words, the limiting MSE and training error both tend to zero as $\kappa \downarrow \kappa_1^\star$, but the training error does so at a faster rate. Hence, near interpolation, a small reduction in training error yields a much larger reduction in MSE. This results in delayed generalization captured by an unbounded S_1. Theorem 3 identifies two sufficient conditions leading to such phenomenon: *(i)* the nonzero part of the signal β_* has little mass around 0, which follows from the fact that W does not satisfy even a mild small ball condition as (22) entails; *(ii)* the sample size equals the critical value $\delta_c(\rho)$ in (23), which corresponds to the classical weak recovery threshold for noiseless ℓ_1 minimization with Gaussian designs (Donoho and Tanner, 2009; Amelunxen et al., 2014). Remarkably, the situation is quite different in the presence of label noise: when $\sigma > 0$, the ℓ_1-MNI has a worse MSE than an approximate interpolator, i.e., $S_1 < 0$ (see Proposition 3 in Appendix D.4 for the precise statement and proof).

We finally note that exact interpolation hurts generalization for ℓ_2 regularization: for $\sigma > 0$, allowing positive training error reduces the MSE, i.e., $S_2 < 0$; furthermore, $S_2 = 0$ in the noiseless case (see Proposition 4 in Appendix D.5 for the precise statement and proof). To reconcile this result with the analysis of grokking for ridge regularization in Xu et al. (2026), note that in their mechanism, generalization is delayed by the slow decay of an initialization component irrelevant for the training data. In contrast, our results identify when regularization geometry supports such a delay.

Role of sparsity. We next fix any ℓ_r norm and show that grokking is more pronounced as the regression problem becomes sparser.

Theorem 4 (More sparsity implies more grokking). *Fix $1 < r < 2$ and $\sigma \ge 0$. For $0 < \rho < 1$, let B be given by (21) with $\rho = \lambda\delta$. Assume $\mathbb{E}W^2 = 1$ and $\mathbb{E}|W|^{r/(r-1)} < \infty$. For any $0 < \lambda_- < \lambda_+ < \infty$, there exists $\delta_0 := \delta_0(r, W, \lambda_-, \lambda_+, \sigma) > 0$ such that, whenever $\delta \in (0, \delta_0)$, S_r is strictly decreasing in ρ on $[\lambda_-\delta, \lambda_+\delta]$. Moreover, there exists $\lambda_c = \lambda_c(r, W) > 0$ such that, if $\lambda_+ < \lambda_c$, then δ_0 may be chosen so that $S_r > 0$ throughout this interval.*

The result is illustrated in the right plot of Figure 1, and its proof is in Appendix D.6. We finally note that, if there is no structure in the regression problem (i.e., B is Gaussian), then exact interpolation hurts generalization (i.e., $S_r < 0$ for every $1 \le r < 2$), see Proposition 5 in Appendix D.7 for the precise statement and proof.

5 Numerical results

We next ask whether the orderings predicted by Theorems 2 and 4 persist beyond regression with an explicit regularization. The initialization scale is known to shape the implicit bias of gradient-based training (Gunasekar et al., 2018; Azulay et al., 2021; Kunin et al., 2024; Kumar et al., 2024), so it can be used to tune the geometry of the effective regularizer. We use it first in two-layer diagonal linear networks (DLNs) (Azulay et al., 2021), where the implicit regularizer is known in closed form. We then use it in transformers trained on modular arithmetic (Kumar et al., 2024).

Diagonal linear networks. We parameterize $\beta = u_+ \circ v_+ - u_- \circ v_-$ and run full-batch gradient descent on $\mathcal{E}(\beta)$ with initialization $u_{\pm,i}(0)^2 + v_{\pm,i}(0)^2 = \sqrt{k}/2$. Azulay et al. (2021) show that when gradient flow converges to a zero-loss solution $\hat{\beta}(\infty)$, this solution minimizes an initialization-dependent regularizer:

$$\hat{\beta}(\infty) = \arg\min_{X\beta = y} Q_k(\beta), \qquad Q_k(\beta) = \sum_{j=1}^{p} q_k(\beta_j), \qquad q_k(x) = \frac{1}{2}\int_0^x \operatorname{arcsinh}(2z/\sqrt{k})dz.$$

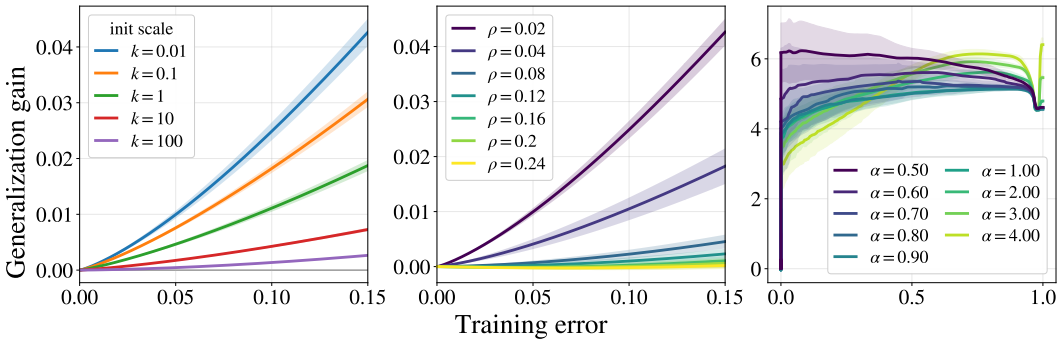

Figure 2: Residual generalization error as a function of the training error. **Diagonal linear networks** (Left, Center) are trained by gradient descent on the square-root loss ($\delta = 0.2$, $\sigma = 0.03$, $n = 2000$, $p = 10^4$; mean $\pm$ one standard deviation over 20 seeds). *Left:* at fixed $\rho = 0.02$, the gain in generalization near interpolation decreases as k increases, mirroring the decrease of S_r in r (Theorem 2). *Center:* at fixed $k = 0.01$, the gain in generalization near interpolation decreases as ρ increases (Theorem 4). **Transformers** (Right) are trained by AdamW on the cross-entropy loss for a modular arithmetic task (mean $\pm$ one standard deviation over 5 seeds). We regulate the output scale α controlling the learning regime and implicit bias. As expected, the gain in generalization near interpolation decreases with α. For implementation details, see Appendix E.

In words, as k goes from 0 to ∞, the initialization scale k moves Q_k continuously between ℓ_1-like and ℓ_2-like geometries. Furthermore, Figure 3 in Appendix E shows that the performance of the solution $\hat{\beta}(t)$ obtained via early stopping closely matches that of Q_k-regularized regression, upon tuning the regularization strength according to the stopping time t. Overall, this provides numerical evidence that training lies on a regularization path of the type analyzed in Section 4.3.

In the first two plots of Figure 2, we take B as in (21), with W a Rademacher random variable, and plot the residual generalization error ($\mathsf{MSE}(\hat{\beta}(t)) - \mathsf{MSE}(\hat{\beta}(\infty))$) as a function of the training error $\mathcal{E}(\beta(t))$. The experiment clearly shows that larger gains in MSE near interpolation are exhibited for small values of k (left plot) and of ρ (central plot). This is in agreement with Theorems 2 and 4, which predict a sharper drop in MSE for ℓ_1-like geometry (associated to small k) and sparse problems (associated to small ρ).

Transformer on a modular arithmetic task. We train a transformer on a modular arithmetic task following Power et al. (2022); Kumar et al. (2024), varying the output scale α to probe changes in the learning regime and implicit bias. The right plot of Figure 2 displays a decrease in late-stage generalization gains as α increases. This is qualitatively consistent with the predicted ordering that stronger sparsity-promoting bias yields larger gains near interpolation.

6 Conclusions and Perspectives

We studied why predictors with nearly indistinguishable training errors can have drastically different generalization errors. In the strongly overparameterized regime ($p \gg n$), we construct a family of norms whose MNIs exhibit a zero–one generalization transition as the norm geometry varies, while maintaining exact interpolation. In the proportional regime ($p \asymp n$), we quantify the generalization gain for a given change in training error near interpolation, and we show that this gain increases as the target becomes sparser and the norm becomes more sparsity-promoting. Experiments in diagonal networks and transformers trained on a modular arithmetic task support the predicted dependence on initialization and sparsity.

A natural next step is to quantify the correspondence observed between early-stopped diagonal networks and regression with their initialization-dependent implicit regularizer. This would help connect our analysis of regularization paths to finite-time training dynamics, distinguishing the magnitude of the generalization gain from the time required to realize it. For transformers, extending this perspective requires identifying the representation in which target sparsity is relevant and testing its role through controlled variations of task structure and output scale.

AI USE STATEMENT

We used generative AI tools to polish the writing, implement straightforward parts of the code, fill in a few proof details after we had identified the key methods and arguments, check proofs for correctness, and improve the coverage of related work. We did not use generative AI tools to develop the paper's core ideas, define the problem setting, or prove the main theoretical results. We take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the assistance of generative AI.

REFERENCES

Dennis Amelunxen, Martin Lotz, Michael B. McCoy, and Joel A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference*, 3(3):224–294, 2014.

Nicolas Anguita, Francesco Locatello, Andrew M Saxe, Marco Mondelli, Flavia Mancini, Samuel Lippl, and Clementine Domine. A theory of how pretraining shapes inductive bias in fine-tuning. In *International Conference on Machine Learning*. PMLR, 2026.

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *International Conference on Machine Learning*, pages 468–477. PMLR, 2021.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2012.

Alon Beck, Noam Levi, and Yohai Bar-Sinai. Grokking at the edge of linear separability. In *International Conference on Machine Learning*. PMLR, 2025.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2018.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Raphaël Berthier. Incremental learning in diagonal linear networks. *Journal of Machine Learning Research*, 24(171):1–26, 2023.

Raphaël Berthier. Diagonal linear networks and the lasso regularization path. *arXiv preprint arXiv:2509.18766*, 2025.

Etienne Boursier, Scott Pesme, and Radu-Alexandru Dragomir. A theoretical framework for grokking: Interpolation followed by Riemannian norm minimisation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012. doi: 10.1007/s10208-012-9135-7.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Clementine Domine, Nicolas Anguita, Alexandra M Proca, Lukas Braun, Daniel Kunin, Pedro Mediano, and Andrew Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. In *International Conference on Learning Representations*, volume 2025, pages 102485–102536, 2025.

Konstantin Donhauser, Nicolo Ruggeri, Stefan Stojanovic, and Fanny Yang. Fast rates for noisy interpolation require rethinking the effect of inductive bias. In *International Conference on Machine Learning*, pages 5397–5428. PMLR, 2022.

David Donoho and Andrea Montanari. High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166:935–969, 2016.

David L Donoho and Jared Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proceedings of the National Academy of Sciences*, 102(27):9452–9457, 2005.

David L. Donoho and Jared Tanner. Counting faces of randomly projected polytopes when the projection radically lowers dimension. *Journal of the American Mathematical Society*, 22(1): 1–53, 2009.

Michael P Friedlander and Paul Tseng. Exact regularization of convex programs. *SIAM Journal on Optimization*, 18(4):1326–1350, 2008.

Yehoram Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985. doi: 10.1007/BF02759761.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ahmed Imtiaz Humayun, Randall Balestriero, and Richard Baraniuk. Deep networks always grok and here is why. In *International Conference on Machine Learning*, pages 20722–20745. PMLR, 2024.

Devon Jarvis, Sebastian Lee, Clementine Domine, Andrew M. Saxe, and Stefano Sarao Mannelli. A theory of initialisation's impact on specialisation. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(11):114001, 2025.

Alan Jeffares, Alicia Curth, and Mihaela van der Schaar. Deep learning through a telescoping lens: A simple model provides empirical insights on grokking, gradient boosting & beyond. In *Advances in Neural Information Processing Systems*, volume 37, pages 123498–123533, 2024.

Tanishq Kumar, Blake Bordelon, Samuel Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *International Conference on Learning Representations*, volume 2024, pages 23010–23035, 2024.

Daniel Kunin, Allan Raventós, Clementine Domine, Feng Chen, David Klindt, Andrew Saxe, and Surya Ganguli. Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 81157–81203, 2024.

Gil Kur and Pierre Bizeul. Minimum norm interpolation via the local theory of banach spaces: The role of 2-uniform convexity. *arXiv preprint arXiv:2603.28956*, 2026.

Gil Kur and Reese Pathak. Minimum norm interpolation via the local theory of banach spaces: The role of gaussianity. *arXiv preprint arXiv:2607.07694*, 2026.

Gil Kur, Pedro Abdalla, Pierre Bizeul, and Fanny Yang. Minimum norm interpolation meets the local theory of banach spaces. In *International Conference on Machine Learning*. PMLR, 2024.

Gil Kur, Zong Shang, Paul Simanjuntak, Guillaume Lecué, and Reese Pathak. A new perspective on minimum-norm interpolation under gaussian covariates. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

Noam Levi, Alon Beck, and Yohai Bar-Sinai. Grokking in linear estimators–a solvable model that groks without understanding. In *International Conference on Learning Representations*, volume 2024, pages 32932–32948, 2024.

Yue Li and Yuting Wei. Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*, 2021.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR, 2020.

Christopher Liaw, Abbas Mehrabian, Yaniv Plan, and Roman Vershynin. A simple tool for bounding the deviation of random matrices on geometric sets. In Bo'az Klartag and Emanuel Milman, editors, *Geometric Aspects of Functional Analysis*, volume 2169 of *Lecture Notes in Mathematics*, pages 277–299. Springer, 2017. doi: 10.1007/978-3-319-45282-1_18.

Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 34651–34663, 2022.

Ziming Liu, Eric J. Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. In *International Conference on Learning Representations*, 2023.

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon Du, Jason Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *International Conference on Learning Representations*, volume 2024, pages 33897–33936, 2024.

William Merrill, Nikolaos Tsilivis, and Aman Shukla. A tale of two circuits: Grokking as competition of sparse and dense subnetworks. In *ICLR 2023 Workshop on Understanding Foundation Models*, 2023.

Léo Miolane and Andrea Montanari. The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4):2313–2335, 2021.

Mohamad Amin Mohamadi, Zhiyuan Li, Lei Wu, and Danica J Sutherland. Why do you grok? A theoretical analysis on grokking modular addition. In *International Conference on Machine Learning*, pages 35934–35967, 2024.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 67–83, 2020.

Yoonsoo Nam, Seok Hyeong Lee, Clementine Domine, Yeachan Park, Charles London, Wonyl Choi, Niclas Goring, and Seungjai Lee. Position: Solve layerwise linear models first to understand neural dynamical phenomena (neural collapse, emergence, lazy/rich regime, and grokking). In *International Conference on Machine Learning*, pages 81897–81929. PMLR, 2025.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.

Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. The squared-error of generalized LASSO: A precise analysis. In *51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1002–1009. IEEE, 2013.

Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. In *Advances in Neural Information Processing Systems*, volume 36, pages 7475–7505, 2023.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Andrew Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. LASSO with non-linear measurements is equivalent to one with linear measurements. In *Advances in Neural Information Processing Systems*, volume 28, 2015a.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. The gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*, 2015b.

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized M-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

Pascal Jr. Tikeng Notsawo, Hattie Zhou, Mohammad Pezeshki, Irina Rish, and Guillaume Dumas. Predicting grokking long before it happens: A look into the loss landscape of models which grok. *arXiv preprint arXiv:2306.13253*, 2023.

Pascal Jr. Tikeng Notsawo, Guillaume Dumas, and Guillaume Rabusseau. Grokking beyond the Euclidean norm of model parameters. In *International Conference on Machine Learning*, volume 267, pages 28552–28618. PMLR, 2025.

Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023.

Kanumuri Nithin Varma and Babak Hassibi. Optimal implicit bias in linear regression. *arXiv preprint arXiv:2506.17187*, 2025.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*, 2023.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press, 2012.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Bojan Žunkovič and Enej Ilievski. Grokking phase transitions in learning local rules with gradient descent. *Journal of Machine Learning Research*, 25(199):1–52, 2024.

Guillaume Wang, Konstantin Donhauser, and Fanny Yang. Tight bounds for minimum ℓ_1-norm interpolation of noisy data. In *International Conference on Artificial Intelligence and Statistics*, pages 10572–10602. PMLR, 2022.

Haolei Weng, Arian Maleki, and Le Zheng. Overcoming the limitations of phase transition by higher order analysis of regularization techniques. *The Annals of Statistics*, 46(6A):3099–3129, 2018.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

Mingyue Xu, Gal Vardi, and Itay Safran. To grok grokking: Provable grokking in ridge regression. In *International Conference on Machine Learning*, 2026.

David Yunis, Kumar Kshitij Patel, Samuel Wheeler, Pedro Savarese, Gal Vardi, Karen Livescu, Michael Maire, and Matthew R. Walter. Approaching deep learning through the spectral dynamics of weights. *arXiv preprint arXiv:2408.11804*, 2024.

Lijia Zhou, Frederic Koehler, Danica J Sutherland, and Nathan Srebro. Optimistic rates: A unifying theory for interpolation learning and regularization in linear regression. *ACM / IMS Journal of Data Science*, 1(2):1–51, 2024.

A Proofs of Lemma 1 and Theorem 1

A.1 Sketch of the proof of Lemma 1

Calibrating the ground truth. Let $S = \mathrm{supp}(\beta_*)$ and $R := \|y\|_2$. Recall that $L = 2\log(p/n)$ and that k_* is defined in (6), so $\|\beta_*\|_1 = \|\beta_*\|_2^2 = k_*$. We work under the assumptions of Section 4.1 and keep $\kappa > 0$ fixed. Denote the objective by

$$F_\kappa(b) := \|y - Xb\|_2 + \tau_\kappa \|b\|_1, \qquad \tau_\kappa := \kappa\sqrt{L/n}. \tag{24}$$

Define the symmetric Gaussian polytope

$$Q_{n,p} := XB_1^p = \mathrm{conv}\{\pm X_j : 1 \le j \le p\} \subset \mathbb{R}^n. \tag{25}$$

This is almost surely the unit ball of a norm $\|\cdot\|_{Q_{n,p}}$, since X has full row rank. The Minkowski functional of $Q_{n,p}$ satisfies

$$\|z\|_{Q_{n,p}} = \inf_{Xb=z} \|b\|_1, \qquad z \in \mathbb{R}^n. \tag{26}$$

Thus, F_κ trades residual length against the minimum ℓ_1 budget needed to produce a prediction. The mean spherical norm of $Q_{n,p}$ is

$$M(Q_{n,p}) := \int_{S^{n-1}} \|\omega\|_{Q_{n,p}} \, \mathrm{d}\sigma_{n-1}(\omega),$$

where σ_{n-1} is normalized surface measure on the unit sphere in $\mathbb{R}^n$. By (26), this is the average minimum ℓ_1 budget required to interpolate a unit response. In Kur and Pathak (2026) it is proven that [1]

$$\frac{n}{\left(\mathbb{E}_X M(Q_{n,p})\right)^2} = L - \log L + O(1).$$

This motives our deterministic calibration $\nu^2 = L - \log L + O(1)$. To see this, as $k_* = (1-\eta)n/\nu^2 + O(1)$ and $R/\sqrt{k_*} = 1 + O_{\mathbb{P}}(n^{-1/2})$, the calibration gives

$$\frac{R}{k_*} = \frac{\nu}{\sqrt{n}}\left[1 + \frac{\eta}{2} + O\left(\eta^2 + \frac{L}{n}\right) + O_{\mathbb{P}}(n^{-1/2})\right].$$

Consequently,

$$\frac{R}{k_*} = \frac{\nu}{\sqrt{n}}\left(1 + \frac{\eta}{2} + o_{\mathbb{P}}(\eta)\right), \qquad \frac{F_\kappa(\beta_*)}{F_\kappa(0_p)} = \frac{\tau_\kappa k_*}{R} \xrightarrow{\mathbb{P}} \kappa. \tag{27}$$

Thus, the planted representation produces slightly more response length per unit ℓ_1 budget than the calibration scale $\nu/\sqrt{n}$. The objective comparison points to exact recovery for $\kappa < 1$ and vanishing coefficient energy relative to the signal for $\kappa > 1$. However, comparing the ground truth with zero is not sufficient: competing predictors may mix planted and off-support coordinates. We control all such competitors, working around β_* below the transition and around zero above it.

Exact recovery for $\kappa < 1$. Let $s = (\beta_*)_S$, which is also the sign vector on the support. The directional derivative of the ℓ_1 norm at β_* is

$$D(h) := \langle s, h_S\rangle + \|h_{S^c}\|_1 = \sup_{z \in \partial\|\beta_*\|_1} \langle z, h\rangle. \tag{28}$$

The subgradient inequality therefore gives

$$F_\kappa(\beta_* + h) - F_\kappa(\beta_*) \ge \|Xh\|_2 + \tau_\kappa D(h).$$

The dilated subdifferential $\nu\partial\|\beta_*\|_1$ is a shifted coordinate cube, so its expected squared distance from a standard Gaussian vector can be computed explicitly. Specifically, (53) in the proof of Lemma 8 gives

$$\mathbb{E}\,\mathrm{dist}(g_p, \nu\partial\|\beta_*\|_1)^2 = k_*(1+\nu^2) + (p-k_*)h_0(\nu) \le n\left(1 - \frac{\eta}{2}\right),$$

[1] The cited work provides a sharper expansion. The coarser $O(1)$ remainder is sufficient for our argument, since $\eta L \to \infty$. Also note that the assumptions $L^{-1} + L/n = o(\eta)$ and $\eta = o(1)$ imply $L/n = o(\eta)$ and $n^{-1/2} = o(\eta)$.

for all sufficiently large dimensions, where h_0 is the Gaussian excess function in Lemma 5. The final inequality follows from the sparsity calibration of k_* and the bound $(p/n)h_0(\nu) = O(L^{-1}) = o(\eta)$. Thus, the expected squared distance lies below n by a margin of order ηn. By Jensen's inequality and (44), this yields

$$\mathbb{E}\|g_n\|_2 - \mathbb{E}\,\mathrm{dist}(g_p, \nu\partial\|\beta_*\|_1) \gtrsim \eta\sqrt{n}.$$

Applying Gordon's comparison with a deterministic offset (the first probability bound in Lemma 4, with $U = S^{n-1}$) to $\sqrt{n}\,X$ with offset $f = -\nu D$, as in the proof of Lemma 8, therefore gives, with probability at least $1 - e^{-cn\eta^2}$,

$$\|Xh\|_2 + \frac{\nu}{\sqrt{n}} D(h) \gtrsim \eta\|h\|_2 \qquad \text{for every } h \in \mathbb{R}^p. \tag{29}$$

Since $\alpha := \kappa\sqrt{L}/\nu \to \kappa < 1$, eventually $0 < \alpha < 1$, and

$$\|Xh\|_2 + \tau_\kappa D(h) = \alpha\left(\|Xh\|_2 + \frac{\nu}{\sqrt{n}} D(h)\right) + (1-\alpha)\|Xh\|_2 \gtrsim_\kappa \eta\|h\|_2.$$

Hence, every nonzero perturbation increases the objective. This proves unique exact recovery, $\widehat{\beta}_{\kappa,1} = \beta_*$, establishing the $\kappa < 1$ regime of the lemma.

The null side: $\kappa > 1$. Any improvement over zero must exploit unusually large initial correlations. Recall that $\tau_\kappa = \kappa\sqrt{L/n}$, and define

$$q := X^\mathsf{T}\frac{y}{R}, \qquad \gamma := \frac{1+\kappa}{2}\sqrt{L/n} < \tau_\kappa, \qquad Z_\gamma^2 := \sum_{j=1}^{p}(|q_j| - \gamma)_+^2. \tag{30}$$

Here q records the initial feature–response correlations: $q_j = \langle X_j, y\rangle/R$ measures the alignment of feature j with the response direction before fitting any coefficients. By Cauchy–Schwarz,

$$\|y - Xb\|_2 \geq \left\langle \frac{y}{R}, y - Xb \right\rangle = R - \langle q, b\rangle.$$

Thus, $F_\kappa(b) \leq R$ implies $\tau_\kappa\|b\|_1 \leq \langle q, b\rangle$. Clipping each coordinate of q to $[-\gamma, \gamma]$ decomposes q into a vector with ℓ_∞ norm at most γ and a remainder with ℓ_2 norm Z_γ. Hölder's and Cauchy–Schwarz inequalities therefore yield

$$F_\kappa(b) \leq R \quad\Longrightarrow\quad \tau_\kappa\|b\|_1 \leq \langle q, b\rangle \leq \gamma\|b\|_1 + Z_\gamma\|b\|_2. \tag{31}$$

In the proof of Lemma 9 below, Gaussianity gives $Z_\gamma \lesssim \sqrt{k_*/n}$ with probability tending to one; see (56). Subtracting $\gamma\|b\|_1$ in (31) and using $\tau_\kappa - \gamma = \frac{\kappa-1}{2}\sqrt{L/n}$ gives

$$\|b\|_1 \lesssim \frac{1}{\kappa-1}\sqrt{\frac{k_*}{L}}\,\|b\|_2 \lesssim_\kappa \frac{k_*}{\sqrt{n}}\|b\|_2 \qquad (F_\kappa(b) \leq R). \tag{32}$$

The last comparison uses $k_* L/n \to 1$. Thus, every nonzero vector with $F_\kappa(b) \leq R$ has effective sparsity $\|b\|_1^2/\|b\|_2^2$ bounded by $O_\kappa(k_*^2/n) = O_\kappa(n/L^2)$.

The extra factor of L below the scale n/L is important: by (32),

$$\frac{L}{n}\|b\|_1^2 \lesssim_\kappa \frac{k_*}{n}\|b\|_2^2 = o(\|b\|_2^2) \qquad (F_\kappa(b) \leq R).$$

Thus, the correction term in the global Euclidean lower bound (50), as well as $\tau_\kappa^2\|b\|_1^2$, can be absorbed into a fixed positive multiple of $\|b\|_2^2$. An effective-sparsity bound of only $O_\kappa(n/L)$ would give corrections of order $\|b\|_2^2$, without ensuring that their constants are small enough for absorption.

For every b with $F_\kappa(b) \leq R$, we have $\|y - Xb\|_2 \leq R - \tau_\kappa\|b\|_1$, with a nonnegative right-hand side. Squaring and applying the clipping bound gives

$$\begin{aligned}\|Xb\|_2^2 - \tau_\kappa^2\|b\|_1^2 &\leq 2R\big(\langle q, b\rangle - \tau_\kappa\|b\|_1\big) \\ &\leq 2RZ_\gamma\|b\|_2 \lesssim \frac{k_*}{\sqrt{n}}\|b\|_2,\end{aligned}$$

where we used $R \asymp \sqrt{k_*}$ on the same high-probability event. By (32) and the global lower bound in Lemma 7, for all sufficiently large n, p,

$$\|Xb\|_2^2 \geq \left(\frac{1}{4} - C_\kappa \frac{k_*}{n}\right)\|b\|_2^2 \geq \frac{1}{8}\|b\|_2^2, \qquad \tau_\kappa^2\|b\|_1^2 \lesssim_\kappa \frac{k_*}{n}\|b\|_2^2,$$

where the second estimate follows from (32) and $k_* L/n \to 1$. Since $k_*/n \to 0$, the quadratic penalty term can be absorbed into the Euclidean lower bound. The squared comparison therefore implies $\|b\|_2 \lesssim_\kappa k_*/\sqrt{n}$, and (32) then gives $\|b\|_1 \lesssim_\kappa k_*^2/n$. These bounds also imply

$$\langle q, b\rangle \leq \gamma\|b\|_1 + Z_\gamma\|b\|_2 \lesssim_\kappa \frac{k_*^{3/2}}{n} \lesssim_\kappa R\frac{k_*}{n}.$$

Returning to $\|y - Xb\|_2 \geq R - \langle q, b\rangle$ and using $\|y - Xb\|_2 \leq R$, we obtain, uniformly over $F_\kappa(b) \leq R$,

$$\|b\|_2^2 + \|b\|_1 \lesssim_\kappa \frac{k_*^2}{n}, \qquad 1 - C_\kappa\frac{k_*}{n} \leq \frac{\|y - Xb\|_2}{R} \leq 1. \tag{33}$$

Since $\|\beta_*\|_\infty = 1$, the first bound also implies

$$\left|\mathsf{MSE}_s(b) - 1\right| \leq \frac{\|b\|_2^2 + 2\|b\|_1}{k_*} \lesssim_\kappa \frac{k_*}{n}.$$

Every minimizer satisfies $F_\kappa(\widehat{\beta}_{\kappa,1}) \leq F_\kappa(0_p) = R$. Since

$$\frac{k_*}{n} \approx \frac{1}{L} = \frac{1}{2\log(p/n)} \longrightarrow 0,$$

both normalized errors differ from one by $O_\kappa(L^{-1})$ on the same high-probability event. Consequently,

$$\left(\mathsf{MSE}_s(\widehat{\beta}_{\kappa,1}), \mathcal{E}_s(\widehat{\beta}_{\kappa,1})\right) \xrightarrow{\mathbb{P}} (1, 1),$$

completing the proof.

Remark 1 (Connection to weak recovery)**.** *The recovery part of our proof of Lemma 1 gives a short, quantitative application of the Gaussian-distance approach to sparse recovery and generalized LASSO (Chandrasekaran et al., 2012; Amelunxen et al., 2014; Oymak et al., 2013). A direct calculation of the distance to the dilated ℓ_1 subdifferential establishes recovery at the leading Donoho–Tanner weak-recovery scale $k_* \approx \frac{n}{2\log(p/n)}$ (Donoho and Tanner, 2009), without counting the faces of the random polytope $Q_{n,p}$.*

Remark 2 (Between geometric and CGMT approaches)**.** *Our argument combines the geometric viewpoint on minimum-norm interpolation (Kur et al., 2026) with classical Gaussian comparison. Rather than deriving a full asymptotic characterization of the optimizer through CGMT, we control the objective directly around the ground truth and the zero predictor. This makes explicit the geometric and distributional ingredients responsible for the transition. A natural direction is to determine whether the same zero–one law holds for non-Gaussian designs, such as matrices with independent Rademacher entries. Developing geometric or universality arguments that replace our Gaussian-specific arguments would be of independent interest.*

A.2 Sketch of the proof of Theorem 1

We use the notation of the preceding sketch. The hybrid norm adds a second way to fit the response: a component with small ℓ_∞ norm that completes the prediction without contributing appreciably to its normalized coefficient energy. We first show that this forces exact interpolation, and then recover the same zero–one transition as in Lemma 1.

The cost of completing a prediction. The additional geometric input is that the image XB_∞^p is approximately Euclidean. For the completion cost in (65),

$$f(z) = \sqrt{\frac{2}{\pi}}\frac{p}{\sqrt{n}} \min_{Xv=z} \|v\|_\infty,$$

Lemma 10 gives, with probability at least $1 - 2e^{-n/2}$, simultaneously for all $z \in \mathbb{R}^n$,

$$(1 - \epsilon_{n,p})\|z\|_2 \le f(z) \le (1 + \epsilon_{n,p})\|z\|_2, \qquad \|v\|_2 \lesssim \sqrt{\frac{n}{p}}\,\|z\|_2, \tag{34}$$

where $\epsilon_{n,p} = C\sqrt{n/p} = o(k_*/n)$ and v is any minimum-ℓ_∞ completion of z. Thus, the normalized cost of fitting a response is close to its Euclidean length, while the required coefficient energy is small because $p \gg n$.

Exact interpolation for every κ. The objective in (10), denoted by J_κ in (43), is

$$J_\kappa(b) = \|y - Xb\|_2 + \frac{\tau_\kappa}{2} N_\kappa(b).$$

For any b with nonzero residual $e = y - Xb$, choose a minimum-ℓ_∞ completion v_e satisfying $Xv_e = e$. The scaling in N_κ gives $\tau_\kappa N_\kappa(v_e) \le f(e)$. Since $b + v_e$ interpolates, the triangle inequality yields

$$J_\kappa(b + v_e) - J_\kappa(b) \le -\|e\|_2 + \frac{1}{2} f(e) \le -\frac{1 - \epsilon_{n,p}}{2}\|e\|_2 < 0. \tag{35}$$

Hence every minimizer interpolates exactly and therefore minimizes N_κ subject to $Xb = y$; see (69).

For an optimal decomposition $b = u + v$, fixing u and minimizing over v subject to $Xv = y - Xu$ leaves the objective $H_\kappa(u) = f(y - Xu) + \tau_\kappa\|u\|_1$ from (71). In particular,

$$\min_{Xb=y} N_\kappa(b) = \frac{1}{\tau_\kappa} \min_u H_\kappa(u), \qquad \widetilde{\beta}_\kappa = u + v, \tag{36}$$

where u minimizes H_κ and v is a minimum-ℓ_∞ completion of $y - Xu$. This correspondence holds for every minimizer by (72). It remains to control u; the bound in (34) then controls v.

Exact recovery for $\kappa < 1$. For every perturbation h, the subgradient inequality and (34) give

$$H_\kappa(\beta_* + h) - H_\kappa(\beta_*) \ge (1 - \epsilon_{n,p})\|Xh\|_2 + \tau_\kappa D(h).$$

Using α from the preceding sketch, the right-hand side equals

$$\alpha\left(\|Xh\|_2 + \frac{\nu}{\sqrt{n}} D(h)\right) + (1 - \epsilon_{n,p} - \alpha)\|Xh\|_2 \gtrsim_\kappa \eta\|h\|_2.$$

Here we used Lemma 8 and the fact that $\alpha \to \kappa < 1$ and $\epsilon_{n,p} \to 0$. Thus $u = \beta_*$ is the unique minimizer of H_κ. Its residual is zero, so its minimum-ℓ_∞ completion is $v = 0$. Therefore $\widetilde{\beta}_\kappa = \beta_*$ uniquely, with probability at least $1 - e^{-cn\eta^2} - 2e^{-n/2}$.

The null side: $\kappa > 1$. We work on the intersection of the events in Lemmas 9, 7, and 10, which has probability tending to one. As in the preceding sketch, $R \asymp \sqrt{k_*}$ and $Z_\gamma \lesssim \sqrt{k_*/n}$ on this event. For any w with $H_\kappa(w) \le H_\kappa(0_p) = f(y)$, the cube estimate gives

$$(1 - \epsilon_{n,p})\|y - Xw\|_2 + \tau_\kappa\|w\|_1 \le (1 + \epsilon_{n,p})R.$$

Combining this with $\|y - Xw\|_2 \ge R - \langle q, w\rangle$ and the clipping inequality yields

$$\big(\tau_\kappa - (1 - \epsilon_{n,p})\gamma\big)\|w\|_1 \le (1 - \epsilon_{n,p})Z_\gamma\|w\|_2 + 2\epsilon_{n,p}R.$$

Since the coefficient on the left is at least $\tau_\kappa - \gamma$, it follows that

$$\|w\|_1 \lesssim_\kappa \frac{k_*}{\sqrt{n}}\|w\|_2 + \epsilon_{n,p}k_*. \tag{37}$$

Squaring the residual comparison and applying the global Euclidean lower bound (50), as in (76), then gives

$$\left(\frac{1}{4} - C_\kappa\frac{k_*}{n}\right)\|w\|_2^2 \le C_\kappa\frac{k_*}{\sqrt{n}}\|w\|_2 + C_\kappa\epsilon_{n,p}k_*. \tag{38}$$

For $\|w\|_2 = Mk_*/\sqrt{n}$, dividing by k_*^2/n shows that this is impossible for a sufficiently large fixed $M = M_\kappa$, since $k_*/n \to 0$ and $\epsilon_{n,p} n/k_* \to 0$. Hence $H_\kappa(w) > H_\kappa(0_p)$ on that sphere.

Convexity now confines every minimizer u of H_κ to the ball it bounds. Indeed, otherwise the segment from zero to u would meet the sphere at tu, with $t \in (0, 1]$, and

$$H_\kappa(tu) \le (1-t)H_\kappa(0_p) + tH_\kappa(u) \le H_\kappa(0_p),$$

a contradiction. Thus $\|u\|_2 \lesssim_\kappa k_*/\sqrt{n}$, and (37) implies

$$\|u\|_2^2 + \|u\|_1 \lesssim_\kappa \frac{k_*^2}{n}. \tag{39}$$

Completing the prediction and the risk bound. For every minimizer u, comparison with zero also gives $f(y - Xu) \le H_\kappa(u) \le f(y)$, and therefore $\|y - Xu\|_2 \lesssim R \asymp \sqrt{k_*}$. The coefficient-energy estimate in (34) yields

$$\frac{\|v\|_2^2}{k_*} \lesssim \frac{n}{p}.$$

Using $\|\beta_*\|_\infty = 1$ and $\|\beta_*\|_2 = \sqrt{k_*}$, we conclude that

$$\begin{aligned}\left|\mathsf{MSE}_s(u+v) - 1\right| &\le \frac{\|u+v\|_2^2 + 2\|u\|_1}{k_*} + \frac{2\|v\|_2}{\sqrt{k_*}} \\ &\lesssim_\kappa \frac{k_*}{n} + \sqrt{\frac{n}{p}} = O_\kappa(L^{-1}).\end{aligned}$$

Thus the completed predictor has asymptotically trivial normalized risk above the transition, despite fitting the response exactly. Together with exact recovery below the transition, this proves, for every fixed $\kappa \ne 1$ and every minimizer,

$$\left(\mathsf{MSE}_s(\widetilde{\beta}_\kappa), \mathcal{E}_s(\widetilde{\beta}_\kappa)\right) \xrightarrow{\mathbb{P}} \left(\mathbf{1}_{\{\kappa>1\}}, 0\right).$$

A.3 Preparation for the proofs

We use the notation of Section 4.1 and the preceding sketch. We first prove Lemma 1, bounding the coefficients of all vectors whose objective is no larger than at zero. For Theorem 1, we use the same clipping and Euclidean estimates, together with convexity, to control the ℓ_1 component. The cube-completion estimate controls the ℓ_∞ component.

The dimensional assumptions and (6) imply

$$k_* \approx \frac{n}{L} \longrightarrow \infty, \qquad \frac{k_*}{n} \approx \frac{1}{L} \longrightarrow 0, \qquad n\eta^2 \longrightarrow \infty,$$
$$\sqrt{\frac{n}{p}} = e^{-L/4} = o(k_*/n). \tag{40}$$

Indeed, $L \to \infty$ and $L = o(n)$ imply $n/L \to \infty$, so the floor in k_* has relative size $O(L/n) = o(1)$. Also, $L^{-1} + L/n \ge 2n^{-1/2}$, so the condition $L^{-1} + L/n = o(\eta)$ yields $\eta\sqrt{n} \to \infty$. We will also use $\eta L \to \infty$ and $L/n = o(\eta)$.

Moreover, the Gaussian model implies

$$y \sim N(0, (k_*/n) I_n), \qquad \frac{R^2}{k_*} \overset{\text{law}}{=} \frac{\chi_n^2}{n}, \qquad \frac{R}{\sqrt{k_*}} \xrightarrow{\mathbb{P}} 1. \tag{41}$$

For the hybrid norm in (8), introduce the abbreviations

$$a := \sqrt{\frac{2}{\pi}} \frac{p}{\sqrt{n}}, \qquad m_\kappa := \frac{p}{\kappa}\sqrt{\frac{2}{\pi L}}, \qquad \tau_\kappa m_\kappa = a, \tag{42}$$

and denote the objective in (10) by

$$J_\kappa(b) := \|y - Xb\|_2 + \frac{\tau_\kappa}{2} N_\kappa(b). \tag{43}$$

We prove the limits in (7) and (11) for every minimizer. For the hybrid estimator, we also show that $X\widetilde{\beta}_\kappa = y$ exactly on a high-probability event. All constants in this appendix are positive and independent of n, p; subscripts indicate permitted dependence on fixed parameters. The conclusions are for each fixed $\kappa \ne 1$.

A.4 Preliminaries

We write B_r^d for the unit ℓ_r ball and S^{d-1} for the Euclidean unit sphere. We use dist for Euclidean distance and $\|A\|_{\mathrm{F}} := (\sum_{i,j} A_{ij}^2)^{1/2}$ for the Frobenius norm. For a nonempty compact set $T \subset \mathbb{R}^d$, define its Gaussian width by

$$w(T) := \mathbb{E} \sup_{v \in T} \langle g_d, v \rangle, \qquad g_d \sim N(0, I_d).$$

We use Gaussian concentration: for a standard Gaussian vector Z, a K-Lipschitz function f, and $t > 0$, each of $\mathbb{P}\{f(Z) - \mathbb{E}f(Z) > t\}$ and $\mathbb{P}\{f(Z) - \mathbb{E}f(Z) < -t\}$ is at most $e^{-t^2/(2K^2)}$. Also, $\mathrm{Var}(f(Z)) \le K^2$. These imply

$$\sqrt{n-1} \le \mathbb{E}\|g_n\|_2 \le \sqrt{n}, \tag{44}$$

since $\mathbb{E}\|g_n\|_2^2 = n$ and the norm is one-Lipschitz. See Vershynin (2018) for these Gaussian facts.

Lemma 4 (Gordon's comparison theorem)**.** *Let $A \in \mathbb{R}^{m \times d}$ have independent $N(0,1)$ entries, let $T \subset S^{d-1}$ and $U \subset S^{m-1}$ be deterministic, nonempty compact sets, and let $f : T \to \mathbb{R}$ be deterministic and continuous. Then*

$$\mathbb{E} \inf_{v \in T} \sup_{u \in U} \{\langle u, Av \rangle - f(v)\} \ge w(U) - \mathbb{E} \sup_{v \in T} \{\langle g_d, v \rangle + f(v)\}, \tag{45}$$

$$\mathbb{E} \sup_{v \in T,\, u \in U} \langle u, Av \rangle \le w(T) + w(U). \tag{46}$$

Moreover, for every $t > 0$,

$$\mathbb{P}\left\{ \inf_{v \in T} \sup_{u \in U} \{\langle u, Av \rangle - f(v)\} < w(U) - \mathbb{E} \sup_{v \in T} \{\langle g_d, v \rangle + f(v)\} - t \right\} \le e^{-t^2/2},$$

$$\mathbb{P}\left\{ \sup_{v \in T,\, u \in U} \langle u, Av \rangle > w(T) + w(U) + t \right\} \le e^{-t^2/2}.$$

Lemma 5. *Let $h_0(t) := \mathbb{E}(|G| - t)_+^2$ for $G \sim N(0,1)$, and write $\phi(t) = (2\pi)^{-1/2} e^{-t^2/2}$. For $t > 0$,*

$$h_0(t) \le \frac{4\phi(t)}{t^3}. \tag{47}$$

Under the stated assumptions,

$$\frac{p}{n} h_0(\nu) = O(L^{-1}) = o(\eta), \qquad \frac{p}{k_*} h_0(\mu\sqrt{L}) \le C_\mu \frac{e^{-(\mu^2-1)L/2}}{\sqrt{L}} \longrightarrow 0 \quad (\mu > 1 \text{ fixed}). \tag{48}$$

Lemma 6 (Width of effectively sparse vectors)**.** *For integers $1 \le \ell \le p$, set $T_\ell := \{v \in S^{p-1} : \|v\|_1 \le \sqrt{\ell}\}$. Then*

$$w(T_\ell) \le C\sqrt{\ell \log(ep/\ell)}. \tag{49}$$

Lemma 7. *Under the dimensional assumptions above, suppose that the rows of $\sqrt{n}\, X$ are independent, centered, isotropic sub-Gaussian random vectors with sub-Gaussian norms bounded by a fixed K, namely $\sup_{i \le n} \sup_{v \in S^{p-1}} \|\sqrt{n}\, \langle X_{i,:}, v \rangle\|_{\psi_2} \le K$. Then there exist constants $c_K, C_K > 0$, depending only on K, such that, for all sufficiently large dimensions, with probability at least $1 - e^{-c_K n}$, simultaneously for every $b \in \mathbb{R}^p$,*

$$\|Xb\|_2^2 \ge \frac{1}{4}\|b\|_2^2 - C_K \frac{L}{n}\|b\|_1^2. \tag{50}$$

A.5 Proof of Lemma 1

A.5.1 Recovery below the transition

The subdifferential appearing in (28) is

$$\partial\|\beta_*\|_1 = \{z \in \mathbb{R}^p : z_S = s,\ \|z_{S^c}\|_\infty \le 1\}. \tag{51}$$

The subgradient inequality gives $\|\beta_* + h\|_1 - k_* \ge D(h)$ for every $h \in \mathbb{R}^p$.

Lemma 8 (Weighted recovery margin)**.** *For all sufficiently large dimensions, with probability at least* $1 - e^{-cn\eta^2}$*, simultaneously for every* $h \in \mathbb{R}^p$*,*

$$\|Xh\|_2 + \frac{\nu}{\sqrt{n}} D(h) \geq c\eta \|h\|_2. \tag{52}$$

Proof. The squared distance to the dilated subdifferential has expectation

$$\begin{aligned} \mathbb{E}\,\mathrm{dist}^2(g_p, \nu\partial\|\beta_*\|_1) &= k_*(1+\nu^2) + (p-k_*)h_0(\nu), \\ \frac{1}{n}\mathbb{E}\,\mathrm{dist}^2(g_p, \nu\partial\|\beta_*\|_1) &\leq 1 - \eta + \frac{1}{\nu^2} + \frac{p}{n}h_0(\nu) \leq 1 - \frac{\eta}{2}. \end{aligned} \tag{53}$$

The first inequality uses $k_*\nu^2 \leq (1-\eta)n$; the second follows from Lemma 5 and $\eta L \to \infty$. Write

$$\Delta := \mathbb{E}\|g_n\|_2 - \mathbb{E}\,\mathrm{dist}(g_p, \nu\partial\|\beta_*\|_1).$$

Jensen's inequality and (44) give

$$\Delta \geq \sqrt{n-1} - \sqrt{n(1-\eta/2)} \geq c\eta\sqrt{n},$$

since $\eta\sqrt{n} \to \infty$. For each $g \in \mathbb{R}^p$, choose its Euclidean projection νz_g onto $\nu\partial\|\beta_*\|_1$. For every unit v,

$$\langle g, v\rangle - \nu D(v) \leq \langle g - \nu z_g, v\rangle \leq \mathrm{dist}(g, \nu\partial\|\beta_*\|_1).$$

We now apply the first probability bound in Lemma 4 to $A = \sqrt{n}X$, $T = S^{p-1}$, $U = S^{n-1}$, $f = -\nu D$, and $t = \Delta/2$. With probability at least $1 - e^{-\Delta^2/8}$,

$$\|Xh\|_2 + \frac{\nu}{\sqrt{n}} D(h) \geq \frac{\Delta}{2\sqrt{n}} \|h\|_2$$

for all h, by homogeneity. This proves the claim. □

Fix $0 < \kappa < 1$. On the event of Lemma 8, use the coefficient α from the sketch. Since $\alpha \to \kappa$, eventually $0 < \alpha < 1$ and, for every $h \neq 0$,

$$\begin{aligned} F_\kappa(\beta_* + h) - F_\kappa(\beta_*) &\geq \|Xh\|_2 + \tau_\kappa D(h) \\ &= \alpha\left(\|Xh\|_2 + \frac{\nu}{\sqrt{n}} D(h)\right) + (1-\alpha)\|Xh\|_2 \\ &\geq c_\kappa \eta \|h\|_2 > 0. \end{aligned} \tag{54}$$

Thus $\widehat{\beta}_{\kappa,1} = \beta_*$ uniquely, and both normalized errors are exactly zero on this event. The same event also gives uniqueness of the ℓ_1-MNI: if $Xh = 0$ and $\|\beta_* + h\|_1 \leq k_*$, then $D(h) \leq 0$, contradicting (52) unless $h = 0$.

A.5.2 Clipping above the transition

Fix $\kappa > 1$ and use q, γ, and Z_γ from (30). Set

$$\mu := \frac{1+\kappa}{2}, \qquad \sqrt{n}\,\gamma = \mu\sqrt{L}. \tag{55}$$

Lemma 9 (Initial correlation excess)**.** *There exists* $c_\kappa > 0$*, depending only on* κ*, such that, for all sufficiently large* n, p*, with probability at least* $1 - 2e^{-c_\kappa n} - e^{-k_*/2}$*,*

$$\frac{1}{2}\sqrt{k_*} \leq R \leq 2\sqrt{k_*}, \qquad Z_\gamma \leq 3\sqrt{\frac{k_*}{n}}, \qquad \langle q, b\rangle \leq \gamma\|b\|_1 + Z_\gamma\|b\|_2 \quad (b \in \mathbb{R}^p). \tag{56}$$

Proof. Let $P = I_p - \beta_*\beta_*^\mathsf{T}/k_*$. Since $\|\beta_*\|_2^2 = k_*$, P is the orthogonal projection onto $\beta_*^\perp$. Decompose each row of X into its projection onto $\mathrm{span}(\beta_*)$ and its projection onto $\beta_*^\perp$. Writing $B := XP$ and using $y = X\beta_*$ gives

$$X = X(I_p - P) + XP = \frac{y\beta_*^\mathsf{T}}{k_*} + B.$$

For each row, the two projections are jointly Gaussian and have zero cross-covariance, since $n^{-1}(I_p - P)P = 0$. They are therefore independent. As the rows of X are independent, B is independent of y and has independent $N(0, P/n)$ rows. Consequently, conditional on y,

$$q = \frac{R}{k_*}\beta_* + \frac{g}{\sqrt{n}}, \qquad g := \sqrt{n}\, B^\mathsf{T} \frac{y}{R} \sim N(0, P). \tag{57}$$

Indeed, the conditional covariance of g is

$$n \sum_{i=1}^{n} \frac{y_i^2}{R^2} \frac{P}{n} = P.$$

Let $\mathcal{A}_y$ denote the response event

$$\frac{1}{2}\sqrt{k_*} \le R \le 2\sqrt{k_*}, \qquad \frac{R}{k_*} \le \gamma.$$

Since $k_* L/n \to 1$ and $\mu > 1$, for all sufficiently large dimensions,

$$\mu\sqrt{\frac{k_* L}{n}} \ge \frac{1+\mu}{2} > 1.$$

Also, $R/\sqrt{k_*}$ has the same distribution as $\|g_n\|_2/\sqrt{n}$. Gaussian concentration for the Euclidean norm therefore gives

$$\begin{aligned}\mathbb{P}(\mathcal{A}_y^c) &\le \mathbb{P}\left(\frac{R}{\sqrt{k_*}} < \frac{1}{2}\right) + \mathbb{P}\left(\frac{R}{\sqrt{k_*}} > \min\left\{2, \frac{1+\mu}{2}\right\}\right) \\ &\le 2e^{-c_\kappa n}.\end{aligned}$$

Here the second threshold is a fixed number strictly greater than one, so $c_\kappa > 0$ can be chosen independently of n, p.

On $\mathcal{A}_y$, the conditional mean on S lies in the clipping cube: for $j \in S$, $|(R/k_*)(\beta_*)_j| \le \gamma$. Thus $(|q_j| - \gamma)_+ \le |g_j|/\sqrt{n}$ on S, whereas $q_j = g_j/\sqrt{n}$ off S. It follows that

$$Z_\gamma^2 \le \frac{1}{n}\left(\|g_S\|_2^2 + \sum_{j \notin S}(|g_j| - \mu\sqrt{L})_+^2\right).$$

Conditionally on y, $\mathbb{E}[\|g_S\|_2^2 \mid y] = \operatorname{tr}(P_{S,S}) = k_* - 1$, and the off-support coordinates are independent standard Gaussians. Lemma 5 therefore gives, on $\mathcal{A}_y$,

$$\mathbb{E}[Z_\gamma^2 \mid y] \le \frac{k_* - 1}{n} + \frac{p - k_*}{n} h_0(\mu\sqrt{L}) \le \frac{2k_*}{n}$$

for all sufficiently large n, p.

In conditional distribution, realize $g = Ph$ with $h \sim N(0, I_p)$. The function

$$h \longmapsto \operatorname{dist}\left((R/k_*)\beta_* + n^{-1/2}Ph, \gamma B_\infty^p\right)$$

equals Z_γ in this representation and it is $n^{-1/2}$-Lipschitz as the distance to a fixed set is one-Lipschitz and the operator norm of P is at most one. Its conditional mean is at most $\sqrt{2k_*/n}$ by Jensen's inequality. Gaussian concentration therefore gives, on $\mathcal{A}_y$,

$$\mathbb{P}\left(Z_\gamma > 3\sqrt{k_*/n} \mid y\right) \le \exp\left(-\frac{(3-\sqrt{2})^2}{2}k_*\right) \le e^{-k_*/2}.$$

Integrating over y and combining this with the bound on $\mathbb{P}(\mathcal{A}_y^c)$ gives the stated probability.

Finally, define the clipped vector coordinate-wise by

$$q_j^{\text{clip}} := \operatorname{sign}(q_j)\min\{|q_j|, \gamma\}.$$

Then

$$q_j - q_j^{\mathrm{clip}} = \mathrm{sign}(q_j)(|q_j| - \gamma)_+, \qquad \|q^{\mathrm{clip}}\|_\infty \le \gamma, \qquad \|q - q^{\mathrm{clip}}\|_2 = Z_\gamma.$$

For every $b \in \mathbb{R}^p$, Hölder's and Cauchy–Schwarz inequalities give

$$\begin{aligned}\langle q, b\rangle &= \langle q^{\mathrm{clip}}, b\rangle + \langle q - q^{\mathrm{clip}}, b\rangle \\ &\le \|q^{\mathrm{clip}}\|_\infty \|b\|_1 + \|q - q^{\mathrm{clip}}\|_2 \|b\|_2 \\ &\le \gamma\|b\|_1 + Z_\gamma\|b\|_2.\end{aligned}$$

This proves the last, deterministic inequality in (56). To obtain the full clipping implication in (31), Cauchy–Schwarz also gives

$$\|y - Xb\|_2 \ge \left\langle \frac{y}{R}, y - Xb \right\rangle = R - \langle q, b\rangle.$$

Hence, whenever $F_\kappa(b) \le R$,

$$\tau_\kappa\|b\|_1 \le R - \|y - Xb\|_2 \le \langle q, b\rangle \le \gamma\|b\|_1 + Z_\gamma\|b\|_2,$$

as claimed in the sketch. □

A.5.3 Coefficient bounds above the transition

Proposition 1 (Comparison with the zero predictor). *Fix $\kappa > 1$. For all sufficiently large n, p, with probability at least $1 - 3e^{-c_\kappa n} - e^{-k_*/2}$, simultaneously for every $b \in \mathbb{R}^p$,*

$$F_\kappa(b) \le R \quad \Longrightarrow \quad \frac{\|b\|_2^2 + \|b\|_1}{k_*} \le C_\kappa \frac{k_*}{n}. \tag{58}$$

Proof. Let us intersect the event in Lemma 9 with the event in Lemma 7. Their intersection has the stated probability by a union bound, after decreasing c_κ if necessary. Fix b with $F_\kappa(b) \le R$, and put $B_0 = \|b\|_2$ and $U_0 = \|b\|_1$. By Cauchy–Schwarz and (56),

$$\|y - Xb\|_2 \ge R - \langle q, b\rangle, \qquad \tau_\kappa U_0 \le \langle q, b\rangle \le \gamma U_0 + Z_\gamma B_0.$$

Since $\tau_\kappa - \gamma = (\kappa - 1)\sqrt{L/n}/2$ and $k_* L/n \to 1$, it follows that

$$U_0 \le \frac{Z_\gamma}{\tau_\kappa - \gamma} B_0 \le C_\kappa \frac{k_*}{\sqrt{n}} B_0. \tag{59}$$

The objective comparison gives $\|y - Xb\|_2 \le R - \tau_\kappa U_0$, whose right-hand side is nonnegative. Squaring and using clipping yields

$$\begin{aligned}\|Xb\|_2^2 - \tau_\kappa^2 U_0^2 &\le 2R\big(\langle q, b\rangle - \tau_\kappa U_0\big) \\ &\le 2RZ_\gamma B_0 \le C\frac{k_*}{\sqrt{n}} B_0.\end{aligned} \tag{60}$$

Combining (50) with these bounds and $L/n \asymp 1/k_*$ gives

$$\frac{1}{4}B_0^2 \le \frac{C_\kappa}{k_*} U_0^2 + C\frac{k_*}{\sqrt{n}} B_0 \le C_\kappa \frac{k_*}{n} B_0^2 + C\frac{k_*}{\sqrt{n}} B_0.$$

For sufficiently large n, p, the quadratic term on the right is at most $B_0^2/8$. Thus $B_0 \le C_\kappa k_*/\sqrt{n}$; substitution into (59) gives $U_0 \le C_\kappa k_*^2/n$. All inequalities hold for every b on the same event, proving (58). □

A.5.4 Generalization and training errors

The continuous objective F_κ is coercive, so it attains its minimum. Every minimizer $\widehat{\beta}_{\kappa,1}$ satisfies $F_\kappa(\widehat{\beta}_{\kappa,1}) \le F_\kappa(0_p) = R$. Proposition 1 therefore gives, for every b with $F_\kappa(b) \le R$,

$$\|b\|_2^2 + \|b\|_1 \le C_\kappa \frac{k_*^2}{n}, \qquad |\mathsf{MSE}_s(b) - 1| \le \frac{\|b\|_2^2 + 2\|b\|_1}{k_*} \le C_\kappa \frac{k_*}{n}. \tag{61}$$

The middle inequality uses $|\langle \beta_*, b\rangle| \le \|b\|_1$. For the training error, clipping and (61) give

$$\langle q, b\rangle \le \gamma\|b\|_1 + Z_\gamma\|b\|_2 \le C_\kappa \frac{k_*^{3/2}}{n} \le C_\kappa R\frac{k_*}{n}.$$

The Cauchy–Schwarz inequality used above gives the lower bound

$$\|y - Xb\|_2 \ge R - \langle q, b\rangle \ge R\left(1 - C_\kappa\frac{k_*}{n}\right).$$

For the upper bound, compare the objective at b with the objective at the zero predictor: $F_\kappa(b) \le F_\kappa(0_p) = R$. Dropping the nonnegative penalty gives $\|y - Xb\|_2 \le F_\kappa(b) \le R$. These two comparisons yield the two-sided residual bound

$$1 - C_\kappa\frac{k_*}{n} \le \frac{\|y - Xb\|_2}{R} \le 1 \qquad (F_\kappa(b) \le R). \tag{62}$$

For sufficiently large n, p, the lower bound is positive, so no minimizer interpolates. By the definition of $\mathcal{E}_s$, (62) directly gives $|\mathcal{E}_s(b) - 1| \le C_\kappa k_*/n$ on the same set. Together with (54) and (61), this proves both components of (7) and the coefficient energy bound. The conclusions hold for every minimizer on a common high-probability event for each fixed $\kappa \ne 1$.

A.6 Proof of Theorem 1

Recall from (8) that

$$N_\kappa(b) = \inf_{u+v=b}\left\{\|u\|_1 + \frac{p}{\kappa}\sqrt{\frac{2}{\pi L}}\,\|v\|_\infty\right\}.$$

A.6.1 The infimal-convolution norm

Proposition 2 (Unit ball and norm comparison)**.** *For every $\kappa > 0$, the infimum in (8) is attained for every $b \in \mathbb{R}^p$, and N_κ is a norm with unit ball*

$$\{b \in \mathbb{R}^p : N_\kappa(b) \le 1\} = \operatorname{conv}\left(B_1^p \cup \frac{\kappa}{p}\sqrt{\frac{\pi L}{2}}\,B_\infty^p\right).$$

For each fixed $\kappa > 0$ and all sufficiently large dimensions,

$$B_1^p \subseteq \{b \in \mathbb{R}^p : N_\kappa(b) \le 1\} \subseteq \kappa\sqrt{\frac{\pi L}{2}}\,B_1^p. \tag{63}$$

Equivalently,

$$\frac{1}{\kappa}\sqrt{\frac{2}{\pi L}}\,\|b\|_1 \le N_\kappa(b) \le \|b\|_1 \qquad (b \in \mathbb{R}^p). \tag{64}$$

Proof. For fixed b, the cost as a function of u, with $v = b - u$, is continuous and tends to infinity as $\|u\|_2 \to \infty$, so it attains its minimum. A decomposition with cost at most one expresses b as a convex combination of a point in B_1^p, a point in $m_\kappa^{-1}B_\infty^p$, and possibly zero. Conversely, grouping the two types of terms in such a convex combination produces a decomposition with cost at most one. Thus the unit sublevel set is the stated convex hull. It is compact, convex, symmetric, and contains a neighborhood of zero. Together with the positive homogeneity of the infimal convolution, this identifies N_κ as its Minkowski functional and proves that N_κ is a norm.

For fixed κ, $m_\kappa/p = \kappa^{-1}\sqrt{2/(\pi L)} < 1$ for all sufficiently large n, p. Every decomposition then satisfies

$$\|u\|_1 + m_\kappa\|v\|_\infty \ge \frac{m_\kappa}{p}\big(\|u\|_1 + \|v\|_1\big) \ge \frac{m_\kappa}{p}\|b\|_1.$$

Taking the infimum proves the lower bound in (64); choosing $u = b$ and $v = 0$ proves the upper bound. Passing to unit balls gives (63). □

For $z \in \mathbb{R}^n$, define the normalized cost of completing a prediction by

$$f(z) := a \min_{Xv=z} \|v\|_\infty. \tag{65}$$

The minimum is attained because its constraint set is nonempty and closed and the objective is coercive. Its unnormalized value is the Minkowski functional of XB_∞^p, which is a norm since X has full row rank. In particular, f is even, convex, and positively homogeneous.

Lemma 10 (Uniform cube completion and its coefficient energy)**.** *There is an absolute $C > 0$ such that, for all sufficiently large dimensions, $\epsilon_{n,p} := C\sqrt{n/p} < 1/2$ and, with probability at least $1 - 2e^{-n/2}$, the following bounds hold simultaneously for every $z \in \mathbb{R}^n$:*

$$(1 - \epsilon_{n,p})\|z\|_2 \le f(z) \le (1 + \epsilon_{n,p})\|z\|_2. \tag{66}$$

On the same event, every minimum-ℓ_∞ completion satisfies

$$v \in \arg\min_{Xw=z} \|w\|_\infty \quad \Longrightarrow \quad \|v\|_2 \le C\sqrt{\frac{n}{p}}\|z\|_2. \tag{67}$$

Moreover, $\epsilon_{n,p} = Ce^{-L/4} = o(k_*/n)$.

Proof. This is the Gaussian projection form of the Dvoretzky–Milman theorem for the cube; see Vershynin (2018, Theorem 11.3.3 and Example 11.3.6). Its high-probability refinement (Exercise 11.3.5 there), obtained by Gaussian concentration, gives in our normalization

$$a(1-\theta)B_2^n \subseteq XB_\infty^p \subseteq a(1+\theta)B_2^n, \qquad \theta = C_0\sqrt{n/p},$$

with probability at least $1 - 2e^{-n/2}$, for an absolute $C_0 > 0$. Here the scale is $a = n^{-1/2}w(B_\infty^p)$. Passing to Minkowski functionals yields

$$\frac{\|z\|_2}{1+\theta} \le f(z) \le \frac{\|z\|_2}{1-\theta} \qquad (z \in \mathbb{R}^n),$$

which implies (66) after choosing $C \ge 2C_0$ and taking sufficiently large n, p. Every minimum-ℓ_∞ completion satisfies

$$\|v\|_2 \le \sqrt{p}\,\|v\|_\infty = \frac{\sqrt{p}}{a} f(z) \le C\sqrt{\frac{n}{p}}\,\|z\|_2,$$

proving (67) on the same event. Finally, $\sqrt{n/p} = e^{-L/4} = o(L^{-1})$ and $k_*/n \approx L^{-1}$ give $\epsilon_{n,p} = o(k_*/n)$. All bounds are uniform in z. □

A.6.2 Exact interpolation of every regularized minimizer

Recall from (43) that

$$J_\kappa(b) = \|y - Xb\|_2 + \frac{\tau_\kappa}{2} N_\kappa(b)$$

is the objective defining $\widetilde{\beta}_\kappa$ in (10). We work on the event of Lemma 10. For any $b \in \mathbb{R}^p$, set $e = y - Xb$ and choose a minimum-ℓ_∞ completion v_e of e. By (8), (42), and (65),

$$\tau_\kappa N_\kappa(v_e) \le \tau_\kappa m_\kappa \|v_e\|_\infty = f(e).$$

Since $b + v_e$ interpolates, the triangle inequality and (66) give

$$\begin{aligned} J_\kappa(b + v_e) - J_\kappa(b) &\le -\|e\|_2 + \frac{\tau_\kappa}{2} N_\kappa(v_e) \\ &\le -\|e\|_2 + \frac{1}{2} f(e) \le -\frac{1 - \epsilon_{n,p}}{2}\|e\|_2. \end{aligned} \tag{68}$$

Thus every noninterpolating vector is strictly suboptimal. The objective J_κ is coercive, and N_κ attains its minimum on the nonempty closed affine interpolation set. Among interpolators, J_κ is a positive multiple of N_κ. Consequently,

$$\arg\min_{b \in \mathbb{R}^p} J_\kappa(b) = \arg\min_{Xb=y} N_\kappa(b). \tag{69}$$

In particular, on this same event, every minimizer satisfies

$$X\widetilde{\beta}_\kappa = y, \qquad \mathcal{E}(\widetilde{\beta}_\kappa) = \mathcal{E}_s(\widetilde{\beta}_\kappa) = 0. \tag{70}$$

This establishes exact training fit before either generalization regime is considered.

A.6.3 ELIMINATING THE ℓ_∞ COMPONENT

We write $b = u + v$ as in the definition of N_κ. For fixed u, interpolation requires $Xv = y - Xu$, and the smallest possible value of $\tau_\kappa m_\kappa \|v\|_\infty$ is $f(y - Xu)$. Thus, minimizing over v leaves the objective

$$H_\kappa(u) := f(y - Xu) + \tau_\kappa \|u\|_1. \tag{71}$$

More precisely,

$$\begin{aligned} \min_{Xb=y} N_\kappa(b) &= \min_{X(u+v)=y} \{\|u\|_1 + m_\kappa \|v\|_\infty\} \\ &= \min_u \left\{ \|u\|_1 + \frac{1}{\tau_\kappa} f(y - Xu) \right\} = \frac{1}{\tau_\kappa} \min_u H_\kappa(u). \end{aligned} \tag{72}$$

The function H_κ is continuous and coercive, so it attains its minimum. If b minimizes N_κ subject to $Xb = y$, every optimal decomposition $b = u + v$ must have u minimizing H_κ and v minimizing $\|v\|_\infty$ subject to $Xv = y - Xu$; otherwise, replacing that component would lower the joint cost in (72). Conversely, every such pair gives a minimum-N_κ interpolator. By (69), every minimizer of J_κ therefore has a decomposition

$$\widetilde{\beta}_\kappa = u + v, \qquad u \in \arg\min_w H_\kappa(w), \qquad v \in \arg\min_{Xw = y - Xu} \|w\|_\infty. \tag{73}$$

We will show that $u = \beta_*$ and $v = 0$ when $\kappa < 1$. When $\kappa > 1$, we bound $\|u\|_2^2 + \|u\|_1$ using H_κ, and $\|v\|_2$ using Lemma 10. These bounds then control the MSE of $u + v$.

A.6.4 GENERALIZATION BELOW THE TRANSITION

Fix $0 < \kappa < 1$ and intersect the events in Lemmas 8 and 10. For every $h \in \mathbb{R}^p$, evenness of f and the subgradient inequality give

$$\begin{aligned} H_\kappa(\beta_* + h) - H_\kappa(\beta_*) &= f(Xh) + \tau_\kappa\big(\|\beta_* + h\|_1 - k_*\big) \\ &\geq (1 - \epsilon_{n,p})\|Xh\|_2 + \tau_\kappa D(h). \end{aligned}$$

Let us use the coefficient α from the sketch. Since $\alpha \to \kappa < 1$ and $\epsilon_{n,p} \to 0$, eventually $0 < \alpha < 1 - \epsilon_{n,p}$. Hence the last expression equals

$$\alpha \left(\|Xh\|_2 + \frac{\nu}{\sqrt{n}} D(h) \right) + (1 - \epsilon_{n,p} - \alpha)\|Xh\|_2 \geq c_\kappa \eta \|h\|_2,$$

by (52). It follows that β_* is the unique minimizer of H_κ. The minimum-ℓ_∞ completion of $y - X\beta_* = 0$ is uniquely $v = 0$. By (69) and (72), every minimizer of J_κ therefore equals β_*. Thus, both MSE_s and $\mathcal{E}_s$ are exactly zero on an event of probability at least $1 - e^{-cn\eta^2} - 2e^{-n/2}$, by a union bound.

A.6.5 GENERALIZATION ABOVE THE TRANSITION

Fix $\kappa > 1$. Let Ω_κ be the intersection of the events in Lemmas 9, 7, and 10. By a union bound, for sufficiently large n, p,

$$\mathbb{P}(\Omega_\kappa) \geq 1 - 3e^{-c_\kappa n} - e^{-k_*/2} - 2e^{-n/2}.$$

We work on this event throughout.

We first show that, for a sufficiently large constant $M > 0$ depending only on κ,

$$H_\kappa(w) > H_\kappa(0_p) \qquad \text{whenever} \qquad \|w\|_2 = M \frac{k_*}{\sqrt{n}}.$$

For any w satisfying $H_\kappa(w) \leq H_\kappa(0_p) = f(y)$, (66) gives

$$(1 - \epsilon_{n,p})\|y - Xw\|_2 + \tau_\kappa \|w\|_1 \leq (1 + \epsilon_{n,p})R. \tag{74}$$

Using $\|y - Xw\|_2 \geq R - \langle q, w \rangle$ and (56) gives

$$\big(\tau_\kappa - (1 - \epsilon_{n,p})\gamma\big)\|w\|_1 \leq (1 - \epsilon_{n,p}) Z_\gamma \|w\|_2 + 2\epsilon_{n,p} R.$$

The coefficient on the left is at least $\tau_\kappa - \gamma$. Thus, as in (59),

$$\|w\|_1 \le C_\kappa \left(\frac{k_*}{\sqrt{n}} \|w\|_2 + \epsilon_{n,p} k_* \right). \tag{75}$$

Next, isolate the residual in (74) and square its nonnegative upper bound. Expanding and applying clipping, with $(1+\epsilon_{n,p})\tau_\kappa/(1-\epsilon_{n,p})^2 > \gamma$, gives

$$\begin{aligned} \|Xw\|_2^2 - \frac{\tau_\kappa^2}{(1-\epsilon_{n,p})^2}\|w\|_1^2 &\le \frac{4\epsilon_{n,p}}{(1-\epsilon_{n,p})^2} R^2 + 2RZ_\gamma \|w\|_2 \\ &\le C\epsilon_{n,p} k_* + C\frac{k_*}{\sqrt{n}}\|w\|_2. \end{aligned}$$

Combining this with (50), (75), and $L/n \asymp 1/k_*$ yields

$$\left(\frac{1}{4} - C_\kappa \frac{k_*}{n} \right) \|w\|_2^2 \le C\frac{k_*}{\sqrt{n}}\|w\|_2 + C_\kappa \epsilon_{n,p} k_*. \tag{76}$$

If $\|w\|_2 = Mk_*/\sqrt{n}$, division by k_*^2/n gives

$$\left(\frac{1}{4} - C_\kappa \frac{k_*}{n} \right) M^2 \le CM + C_\kappa \epsilon_{n,p} \frac{n}{k_*}.$$

Since $k_*/n \to 0$ and $\epsilon_{n,p} n/k_* \asymp Le^{-L/4} \to 0$, this is impossible for a sufficiently large fixed M and all sufficiently large n, p. This proves the claimed boundary inequality.

Now let u be any minimizer of H_κ. If $\|u\|_2 \ge Mk_*/\sqrt{n}$, choose $t \in (0,1]$ so that $\|tu\|_2 = Mk_*/\sqrt{n}$. Convexity gives

$$H_\kappa(tu) \le (1-t)H_\kappa(0_p) + tH_\kappa(u) \le H_\kappa(0_p),$$

contradicting the boundary inequality. Hence $\|u\|_2 < Mk_*/\sqrt{n}$. Applying (75) to u gives

$$\frac{\|u\|_2^2 + \|u\|_1}{k_*} \le C_\kappa \left(\frac{k_*}{n} + \epsilon_{n,p} \right) \le C_\kappa \frac{k_*}{n}. \tag{77}$$

For every associated minimum-ℓ_∞ completion v,

$$f(y - Xu) \le H_\kappa(u) \le f(y), \qquad \|y - Xu\|_2 \le \frac{1+\epsilon_{n,p}}{1-\epsilon_{n,p}} R \le CR.$$

Lemma 10 and $R \le 2\sqrt{k_*}$ imply

$$\|v\|_2 \le C\sqrt{\frac{n}{p}} R, \qquad \frac{\|v\|_2^2}{k_*} \le C\frac{n}{p}. \tag{78}$$

Consequently, for $b = u + v$,

$$\frac{\|b\|_2^2}{k_*} \le \frac{2\|u\|_2^2 + 2\|v\|_2^2}{k_*} \le C_\kappa \left(\frac{k_*}{n} + \frac{n}{p} \right) \le C_\kappa \frac{k_*}{n}. \tag{79}$$

The inequalities $|\langle \beta_*, u \rangle| \le \|u\|_1$ and $|\langle \beta_*, v \rangle| \le \sqrt{k_*}\|v\|_2$ then give

$$\begin{aligned} \left| \mathsf{MSE}_s(b) - 1 \right| &\le \frac{\|b\|_2^2}{k_*} + \frac{2\|u\|_1}{k_*} + \frac{2\|v\|_2}{\sqrt{k_*}} \\ &\le C_\kappa \left(\frac{k_*}{n} + \sqrt{\frac{n}{p}} \right) \le C_\kappa \frac{k_*}{n}. \end{aligned} \tag{80}$$

Here $\sqrt{n/p} = o(k_*/n)$, and therefore also $n/p = o(k_*/n)$. All these estimates hold on Ω_κ for every minimizer u of H_κ and every associated minimizing completion v. By (72), they therefore hold for every minimizer of J_κ. Since $k_*/n \approx L^{-1}$, its normalized MSE differs from one by $O_\kappa(L^{-1})$ on Ω_κ, while its training error is exactly zero by (70). Together with exact recovery for $\kappa < 1$, this proves (11) and completes the proof of Theorem 1.

A.7 PROOFS OF THE PRELIMINARY LEMMAS

We prove the preliminary lemmas used above.

Proof of Lemma 4. First suppose that T and U are finite. Let $z \sim N(0,1)$, g_m, and g_d be independent of each other and of A, and compare

$$P_{v,u} = \langle u, Av \rangle + z, \qquad Q_{v,u} = \langle g_m, u \rangle - \langle g_d, v \rangle.$$

Both processes are centered and have variance two. Their covariance difference is

$$\mathrm{Cov}(P_{v,u}, P_{v',u'}) - \mathrm{Cov}(Q_{v,u}, Q_{v',u'}) = (1 - \langle v, v' \rangle)(1 - \langle u, u' \rangle) \geq 0,$$

with equality when $v = v'$. Apply Gordon's comparison in Thrampoulidis et al. (2015b, Theorem A.1), with its process X equal to Q, its process Y equal to P, and thresholds $c + f(v)$. For every $c \in \mathbb{R}$, it gives

$$\mathbb{P}\left\{ \min_{v \in T} \max_{u \in U} (P_{v,u} - f(v)) \geq c \right\} \geq \mathbb{P}\left\{ \min_{v \in T} \max_{u \in U} (Q_{v,u} - f(v)) \geq c \right\}.$$

Taking expectations, using $\mathbb{E}z = 0$, and separating the two terms in Q yields (45). For the upper bound, apply the single-row case of the same Theorem A.1 (Slepian's comparison), now with $X = P$, $Y = Q$ and index set $T \times U$. The covariance ordering gives

$$\mathbb{E} \max_{v \in T,\, u \in U} P_{v,u} \leq \mathbb{E} \max_{v \in T,\, u \in U} Q_{v,u} = w(T) + w(U).$$

Again $\mathbb{E}z = 0$, so this proves (46).

For compact T, U, choose finite ε-nets contained in the respective sets, with $\varepsilon \downarrow 0$. Let

$$\omega_f(\varepsilon) := \sup\{|f(v) - f(v')| : v, v' \in T,\ \|v - v'\|_2 \leq \varepsilon\}.$$

Uniform continuity gives $\omega_f(\varepsilon) \to 0$. Replacing each index by a nearest net point changes $P_{v,u} - f(v)$ by at most $2\varepsilon\|A\|_{\mathrm{F}} + \omega_f(\varepsilon)$, and $Q_{v,u} - f(v)$ by at most $\varepsilon(\|g_m\|_2 + \|g_d\|_2) + \omega_f(\varepsilon)$. Thus the corresponding extrema converge almost surely. Their absolute values are bounded, respectively, by

$$\|A\|_{\mathrm{F}} + |z| + \|f\|_\infty, \qquad \|g_m\|_2 + \|g_d\|_2 + \|f\|_\infty,$$

where $\|f\|_\infty = \sup_{v \in T} |f(v)| < \infty$. Both bounds are integrable, so dominated convergence passes the expectation comparisons to T, U. The same argument with $f = 0$ applies to the supremum comparison.

Finally, both random extrema in the lemma are 1-Lipschitz in A, since

$$\sup_{u \in U,\, v \in T} |\langle u, (A - B)v \rangle| \leq \|A - B\|_{\mathrm{F}}.$$

Gaussian concentration gives the stated tail bounds. □

Proof of Lemma 5. A change of variables gives

$$h_0(t) = 2\phi(t) \int_0^\infty x^2 e^{-tx - x^2/2}\, \mathrm{d}x \leq 2\phi(t) \int_0^\infty x^2 e^{-tx}\, \mathrm{d}x = \frac{4\phi(t)}{t^3}.$$

Since $p/n = e^{L/2}$, $\nu \asymp \sqrt{L}$, and $e^{-\nu^2/2} \leq Ce^{-L/2}\sqrt{L}$,

$$\frac{p}{n} h_0(\nu) = O(L^{-1}) = o(\eta).$$

For fixed $\mu > 1$, substituting $t = \mu\sqrt{L}$ and using $k_* \asymp n/L$ gives

$$\frac{p}{k_*} h_0(\mu\sqrt{L}) \leq C_\mu \frac{e^{-(\mu^2 - 1)L/2}}{\sqrt{L}} \longrightarrow 0.$$

These are the two bounds in (48). □

Proof of Lemma 6. For $v \in T_\ell$, order the coordinates by decreasing magnitude and partition them into successive blocks $I_1, I_2, \dots$ of size ℓ, except possibly the last. For $j \geq 2$, each entry in I_j is bounded by the average absolute entry in I_{j-1}, so

$$\|v_{I_j}\|_2 \leq \frac{\|v_{I_{j-1}}\|_1}{\sqrt{\ell}}, \qquad \sum_j \|v_{I_j}\|_2 \leq \|v\|_2 + \frac{\|v\|_1}{\sqrt{\ell}} \leq 2.$$

It follows that

$$\sup_{v \in T_\ell} \langle g_p, v \rangle \leq 2 \max_{|I|=\ell} \|(g_p)_I\|_2.$$

Writing $V = \max_{|I|=\ell} \|(g_p)_I\|_2$, we have

$$\mathbb{E}e^{V^2/4} \leq \sum_{|I|=\ell} \mathbb{E}e^{\|(g_p)_I\|_2^2/4} = \binom{p}{\ell} 2^{\ell/2}.$$

Jensen's inequality and $\binom{p}{\ell} \leq (ep/\ell)^\ell$ give

$$\mathbb{E}V \leq (\mathbb{E}V^2)^{1/2} \leq C\sqrt{\ell \log(ep/\ell)},$$

which proves (49). □

Proof of Lemma 7. Set $\ell = \lfloor a_K n/L \rfloor$, where $a_K > 0$ will be chosen sufficiently small. Since $L \to \infty$ and $L = o(n)$, eventually $1 \leq \ell \leq p$ and

$$\log(ep/\ell) = \frac{L}{2} + \log L + O_K(1).$$

Lemma 6 therefore gives

$$w(T_\ell) \leq C\sqrt{\ell \log(ep/\ell)} \leq C\sqrt{a_K n}$$

for all sufficiently large n, p.

By Liaw et al. (2017, Theorem 1.4), applied to $\sqrt{n}\,X$ and $T_\ell \subset S^{p-1}$, for every $t \geq 1$, with probability at least $1 - e^{-t^2}$,

$$\sup_{v \in T_\ell} \big| \|\sqrt{n}\,Xv\|_2 - \sqrt{n} \big| \leq CK^2\big(w(T_\ell) + t\big).$$

Here T_ℓ has Euclidean radius one. Choose $a_K > 0$, $b_K > 0$ sufficiently small and take $t = b_K\sqrt{n}$. For all sufficiently large dimensions, the right-hand side is at most $\sqrt{n}/2$. After decreasing $c_K > 0$, the resulting event has probability at least $1 - e^{-c_K n}$, and on this event

$$\|Xv\|_2 \geq \frac{1}{2} \qquad (v \in T_\ell).$$

By homogeneity,

$$\|Xb\|_2^2 \geq \frac{1}{4}\|b\|_2^2 \quad \text{whenever} \quad \|b\|_1 \leq \sqrt{\ell}\,\|b\|_2.$$

For vectors outside this class,

$$\frac{1}{4}\|b\|_2^2 - \frac{1}{4\ell}\|b\|_1^2 \leq 0,$$

so nonnegativity of $\|Xb\|_2^2$ gives the same lower bound with the correction term. Combining the two cases and using $1/\ell \leq C_K L/n$ proves, simultaneously for all $b \in \mathbb{R}^p$,

$$\|Xb\|_2^2 \geq \frac{1}{4}\|b\|_2^2 - C_K \frac{L}{n}\|b\|_1^2,$$

which is (50). □

B PROOF OF LEMMA 2

Throughout the remaining appendices, for $a, b \in \mathbb{R}^p$ and $1 \le s < \infty$, we use the shorthands

$$\|b\|_{s,p} := p^{-1/s}\|b\|_s, \qquad \langle a, b\rangle_p := p^{-1}\langle a, b\rangle, \qquad \|b\|_{\infty,p} := \max_{1\le j\le p} |b_j|.$$

We note that the proof presented in this appendix gives a result stronger than the convergence in (16), namely the convergence in W_2 of the joint empirical distribution of $(\widehat{\beta}_{\kappa,r}, \beta_*)$, as detailed in the remark below.

Remark 3 (Joint empirical laws)**.** *Under the assumptions of Lemmas 2 and 3, our argument establishes the joint empirical distribution of true and estimated coefficients:*

$$\frac{1}{p}\sum_{j=1}^{p} \delta_{((\beta_*)_j, (\widehat{\beta}_{\kappa,r})_j)} \xrightarrow{W_2} \mathcal{L}(B, f_\kappa), \qquad \frac{1}{p}\sum_{j=1}^{p} \delta_{((\beta_*)_j, (\widehat{\beta}_r)_j)} \xrightarrow{W_2} \mathcal{L}(B, f_{\rm int}), \tag{81}$$

where $f_\kappa := f_{\tau_\kappa,\alpha_\kappa}$ and $f_{\rm int} := f_{\tau_{\rm int},\alpha_{\rm int}}$. This strengthens the results of Lemmas 2-3 on the MSE. In fact, (81) implies the convergence of empirical averages of any continuous function of true and estimated coefficients with at most quadratic growth, including MSE, mean absolute error, normalized correlations between vectors as well as their ℓ_r norms for $r \in [1, 2]$.

B.1 SKETCH OF THE ARGUMENT

We first express the optimization problem (2) in a form to which the convex Gaussian min–max theorem (CGMT) applies. Write $G_X := \sqrt{n}X$ and $w := b - \beta_*$. Dividing the objective by $\sqrt{n}$ and using $\|a\|_2 = \max_{\|u\|_2\le 1} u^\top a$ gives the primary objective (see (134)):

$$P_p(w) = \frac{1}{\sqrt{n}}\left\|\varepsilon - \frac{G_X w}{\sqrt{n}}\right\|_2 + \frac{\kappa}{\delta_p}\|\beta_* + w\|_{r,p} = \max_{\|u\|_2\le 1}\left\{\frac{u^\top\varepsilon}{\sqrt{n}} - \frac{u^\top G_X w}{n} + \frac{\kappa}{\delta_p}\|\beta_* + w\|_{r,p}\right\}.$$

Thus, b solves (2) exactly when $w = b - \beta_*$ minimizes P_p.

Conditionally on ε, the CGMT associates to this minimization an auxiliary problem in which the matrix term $-u^\top G_X w/n$ is replaced by $\big(\|w\|_2 g^\top u - \|u\|_2 h^\top w\big)/n$, where $g \in \mathbb{R}^n$ and $h \in \mathbb{R}^p$ are independent standard Gaussian vectors, independent of (G_X, ε). Maximizing first over the direction of u and then over its length $\|u\|_2 \in [0, 1]$ gives the following auxiliary objective (see (122)):

$$A_p^{\rm raw}(b) = \left[\left\|\frac{\varepsilon}{\sqrt{n}} + \frac{\|b-\beta_*\|_2}{n} g\right\|_2 - \frac{1}{n}h^\top(b - \beta_*)\right]_+ + \frac{\kappa}{\delta_p}\|b\|_{r,p}.$$

Here $[x]_+ = \max\{x, 0\}$; see Appendix B.7 for the derivation.

The norm in this expression converges to $\sqrt{\sigma^2 + \delta_p^{-1}\|b - \beta_*\|_{2,p}^2}$, uniformly when $\|b - \beta_*\|_{2,p}$ is bounded. Thus, $A_p^{\rm raw}(b)$ is approximated by

$$\overline{A}_p(b) = [C_p(b)]_+ + \frac{\kappa}{\delta_p}\|b\|_{r,p},$$

$$C_p(b) = \sqrt{\sigma^2 + \frac{1}{\delta_p}\|b - \beta_*\|_{2,p}^2} - \frac{1}{\delta_p}\langle h, b - \beta_*\rangle_p,$$

see (125). Appendix B.5 bounds $\|b - \beta_*\|_{2,p}$ and $\|b\|_{r,p}$ uniformly over fixed sublevel sets of both auxiliary objectives (see (126)), so the approximation applies at their minimizers. Dropping the positive part gives the empirical functional in (107):

$$Q_p(b) = \sqrt{\sigma^2 + \frac{1}{\delta_p}\|b - \beta_*\|_{2,p}^2} - \frac{1}{\delta_p}\langle h, b - \beta_*\rangle_p + \frac{\kappa}{\delta_p}\|b\|_{r,p}.$$

Thus, $Q_p(b) \le \overline{A}_p(b)$, with equality whenever $C_p(b) \ge 0$. We will show that this lower bound is asymptotically attained.

The population counterpart of Q_p replaces coordinate averages by expectations and δ_p by δ. The true coefficient, fitted coefficient, and auxiliary Gaussian coordinate are represented by B, f, and G, respectively, where f ranges over $L_2 \cap L_r$ on the probability space supporting (B, G). This gives the population objective in (85):

$$\mathcal{Q}(f) = \sqrt{\sigma^2 + \frac{1}{\delta}\mathbb{E}(f-B)^2} - \frac{1}{\delta}\mathbb{E}[G(f-B)] + \frac{\kappa}{\delta}\|f\|_r, \qquad \|f\|_r = (\mathbb{E}|f|^r)^{1/r}.$$

The fixed point equations in (14) ensure that the random variable in (86),

$$f_\kappa = \tau_\kappa \eta_{\alpha_\kappa, r}\left(\frac{B}{\tau_\kappa} + G\right),$$

satisfies the optimality conditions of $\mathcal{Q}$. Appendix B.2 proves that it is the unique minimizer, including when $r = 1$.

The minimizing vectors depend on the Gaussian data, so passing to the population problem requires norm bounds and stability. Appendix B.3 establishes convergence of the joint empirical law of $((\beta_*)_j, h_j)$ to $\mathcal{L}(B, G)$ and bounds the Gaussian inner products. These bounds imply $Q_p(b) \geq c\|b\|_{2,p} - C$ as in (110), so near-minimizers have bounded normalized ℓ_2 norms. A compactness argument then gives $\min_b Q_p(b) \to \mathcal{Q}(f_\kappa)$ in (111) and shows that every sequence of near-minimizers has limiting squared error $\mathbb{E}(f_\kappa - B)^2$, coefficient moment $\mathbb{E}|f_\kappa|^r$, and joint empirical law $\mathcal{L}(B, f_\kappa)$ in W_2; see Appendix B.4.

To see why dropping the positive part gives the correct auxiliary minimum, evaluate the objectives at $b^\circ_{p,j} = \tau_\kappa \eta_{\alpha_\kappa, r}((\beta_*)_j/\tau_\kappa + h_j)$. The empirical convergence results give $Q_p(b^\circ_p) \to \mathcal{Q}(f_\kappa)$ and $C_p(b^\circ_p) \to R_\kappa > 0$. Hence $\overline{A}_p(b^\circ_p) = Q_p(b^\circ_p)$ eventually. Combining this upper bound with $\overline{A}_p \geq Q_p$ and the uniform norm approximation shows that all three minimum values converge almost surely to the value in (128) (see (111) and (127)):

$$\ell_\kappa = \mathcal{Q}(f_\kappa) = R_\kappa + \frac{\kappa}{\delta}\|f_\kappa\|_r.$$

The norm bounds and approximation also carry the stability of Q_p to A^{raw}_p; see Appendix B.5.

Finally, vectors whose error, coefficient moment, or empirical law stays away from these limits have an auxiliary objective value separated from its minimum. Appendix B.6 first confines the primary and auxiliary minimizers to deterministic sets with bounded normalized norms. An application of CGMT (as in Thrampoulidis et al. (2015b)) exclude such deviations for every primary minimizer almost surely, giving that $\min_w P_p(w) \to \ell_\kappa$ almost surely (see as in (137)). Subtracting the regularization gives

$$\frac{1}{\sqrt{n}}\|y - X\widehat{\beta}_{\kappa,r}\|_2 \to \ell_\kappa - \frac{\kappa}{\delta}\|f_\kappa\|_r = R_\kappa,$$

as detailed in Appendix B.7.

We note that several intermediate results are established under the weaker condition $D_r(\tau_\kappa, \alpha_\kappa) \leq \delta$, allowing their reuse in the interpolation regime considered in Appendix C.

B.2 Uniqueness of the minimizer of the population objective

For $1 < r < \infty$ and $\alpha \geq 0$, recall

$$\eta_{\alpha,r}(z) := \arg\min_{u \in \mathbb{R}} \left\{\frac{1}{2}(u-z)^2 + \frac{\alpha}{r}|u|^r\right\}.$$

and let

$$\psi_r(u) := \mathrm{sign}(u)|u|^{r-1}.$$

For $r = 1$, set $\eta_{\alpha,1}(z) := \mathrm{ST}(z;\alpha) := \mathrm{sign}(z)(|z| - \alpha)_+$. Setting the derivative to 0 in the optimization problem above gives the following equation for the proximal operator $\eta_{\alpha,r}$:

$$z = \eta_{\alpha,r}(z) + \alpha\psi_r(\eta_{\alpha,r}(z)). \tag{82}$$

We start with two auxiliary lemmas that will be used throughout the argument.

Lemma 11 (Results on proximal operator). *For every $1 < r < \infty$ and $\alpha > 0$, $\eta_{\alpha,r}$ is odd, nondecreasing, and 1-Lipschitz. Moreover,*

$$|\eta_{\alpha,r}(z)| \le |z|, \qquad \eta'_{\alpha,r}(z) = \frac{1}{1 + \alpha(r-1)|\eta_{\alpha,r}(z)|^{r-2}} \in [0,1], \tag{83}$$

with the continuous value 0 at $\eta_{\alpha,r}(z) = 0$ when $1 < r < 2$. The same oddness, monotonicity, Lipschitz, and growth conclusions hold for $\mathrm{ST}(\,\cdot\,;\alpha)$.

Proof. The objective defining $\eta_{\alpha,r}(z)$ has a unique minimizer. In fact, the objective is continuous and coercive, because its quadratic term tends to $+\infty$ as $|u| \to +\infty$. It is also strictly convex, since the quadratic term is strictly convex and $|u|^r$ is convex.

The minimizer satisfies (82), due to the first-order optimality condition. Furthermore, the map $T_\alpha(u) := u + \alpha\psi_r(u)$ is an odd, continuous, strictly increasing bijection of $\mathbb{R}$, as $T'_\alpha(u) = 1 + \alpha(r-1)|u|^{r-2} > 0$ for $u \neq 0$. This readily implies that the inverse $\eta_{\alpha,r} = T_\alpha^{-1}$ is odd and nondecreasing.

To see that $|\eta_{\alpha,r}(z)| \le |z|$, note that, if $z \ge 0$ and $u = \eta_{\alpha,r}(z)$, then $u \ge 0$ and $z = u + \alpha u^{r-1} \ge u$; furthermore, for $z < 0$, the claim follows from the fact that $\eta_{\alpha,r}$ is odd.

Formula (83) for $\eta'_{\alpha,r}$ holds wherever the inverse-function theorem applies, and its RHS extends continuously at the origin. To see this, differentiate $z = u + \alpha\psi_r(u)$. For $1 < r < 2$, the denominator tends to $+\infty$ as $u \to 0$; for $r \ge 2$, (83) is immediate. Its value lies in $[0,1]$, so $\eta_{\alpha,r}$ is globally 1-Lipschitz.

The assertions for $\mathrm{ST}(\,\cdot\,;\alpha)$ follow directly from the fact that the function has three affine pieces. □

Lemma 12 (Zero-prior gap). *If $B = 0$ almost surely, then, for every $\tau, \alpha > 0$,*

$$D_r(\tau,\alpha) = H_r(\tau,\alpha) + \alpha U_r(\tau,\alpha) > H_r(\tau,\alpha). \tag{84}$$

Proof. Set $u = \eta_{\alpha,r}(G)$. Gaussian integration by parts and (82) give

$$D_r(\tau,\alpha) = \mathbb{E}[Gu] = \mathbb{E}u^2 + \alpha\mathbb{E}|u|^r = H_r(\tau,\alpha) + \alpha U_r(\tau,\alpha).$$

For $r = 1$, write $G - u = \alpha\xi$, where $\xi \in \partial|u|$, and use $\xi u = |u|$. The inequality is strict because $\mathbb{P}(u \neq 0) > 0$. □

For a random variable f on the probability space supporting (B, G), write $\|f\|_s = (\mathbb{E}|f|^s)^{1/s}$, and define on $L_2 \cap L_r$ the following population objective:

$$\mathcal{Q}(f) := \sqrt{\sigma^2 + \frac{1}{\delta}\mathbb{E}(f-B)^2} - \frac{1}{\delta}\mathbb{E}[G(f-B)] + \frac{\kappa}{\delta}\|f\|_r. \tag{85}$$

Set

$$A_\kappa := \frac{B}{\tau_\kappa}, \qquad u_\kappa := \eta_{\alpha_\kappa,r}(A_\kappa + G), \qquad f_\kappa := \tau_\kappa u_\kappa. \tag{86}$$

The following lemma characterizes the unique minimizer of the population objective.

Lemma 13 (Population optimizer). *Assume that (14) holds and that $D_r(\tau_\kappa,\alpha_\kappa) \le \delta$. Then f_κ is the unique minimizer of $\mathcal{Q}$.*

Proof. Note that

$$\sqrt{\sigma^2 + \delta^{-1}\mathbb{E}(f_\kappa - B)^2} = \tau_\kappa. \tag{87}$$

In fact, by (86), $f_\kappa - B = \tau_\kappa(u_\kappa - A_\kappa)$, so the square of the LHS is $\sigma^2 + \delta^{-1}\tau_\kappa^2 H_r(\tau_\kappa,\alpha_\kappa)$, which equals τ_κ^2 by (14).

We now show that, if $1 < r < \infty$, then $0 \in \partial\mathcal{Q}(f_\kappa)$. First, note that the L_r norm is differentiable at $f_\kappa \neq 0$ and

$$\nabla\|f_\kappa\|_r = \|f_\kappa\|_r^{1-r}\psi_r(f_\kappa).$$

In fact, for $1 < r < \infty$, this is the standard derivative of the L_r norm, and u_κ cannot vanish almost surely, because $A_\kappa + G$ has a nondegenerate Gaussian conditional law and $\eta_{\alpha,r}(z) = 0$ only when

$z = 0$. Next, as $\|f_\kappa\|_r = \tau_\kappa U_r(\tau_\kappa, \alpha_\kappa)^{1/r}$ and $\psi_r(f_\kappa) = \tau_\kappa^{r-1}\psi_r(u_\kappa)$, the second equation in (14) implies that

$$\kappa \|f_\kappa\|_r^{1-r}\psi_r(f_\kappa) = \alpha_\kappa \psi_r(u_\kappa). \tag{88}$$

Combining (88) with the proximal equation (82) and using that $f_\kappa - B = \tau_\kappa(u_\kappa - A_\kappa)$ gives

$$\frac{f_\kappa - B}{\tau_\kappa} - G + \kappa\|f_\kappa\|_r^{1-r}\psi_r(f_\kappa) = 0, \tag{89}$$

which is equivalent to $0 = \nabla\mathcal{Q}(f_\kappa)$.

If $r = 1$, then $0 \in \partial\mathcal{Q}(f_\kappa)$. To see this, note that the soft-thresholding relation is equivalent to the existence of $\xi_\kappa \in \partial|u_\kappa|$ such that

$$A_\kappa + G - u_\kappa = \kappa\xi_\kappa.$$

In fact, $\alpha_\kappa = \kappa$ by the second equation in (14), and this is the proximal optimality condition for $\kappa|u|$. Since $\partial|\tau u| = \partial|u|$ for $\tau > 0$, the preceding identity becomes

$$\frac{f_\kappa - B}{\tau_\kappa} - G + \kappa\xi_\kappa = 0, \qquad \xi_\kappa \in \partial|f_\kappa|.$$

By (87), this is exactly $0 \in \partial\mathcal{Q}(f_\kappa)$.

The subgradient calculations for $r > 1$ and $r = 1$, together with convexity of $\mathcal{Q}$, imply that f_κ is a global minimizer. In fact, the square root in (85) is the Hilbert norm of $(\sigma, (f-B)/\sqrt{\delta})$; the second term is affine and the last is a norm. Hence $\mathcal{Q}$ is convex, and a zero subgradient is sufficient for global optimality.

If $\sigma = 0$, we first show that $\mathbb{P}(B \neq 0) > 0$. Indeed, if $B = 0$ almost surely, then Lemma 12, evaluated at $(\tau_\kappa, \alpha_\kappa)$, gives $D_r(\tau_\kappa, \alpha_\kappa) > H_r(\tau_\kappa, \alpha_\kappa) = \delta$, where the equality follows from (14). This contradicts $D_r(\tau_\kappa, \alpha_\kappa) \leq \delta$.

It remains to show the uniqueness of the minimizer, and we split this argument into three cases: *(i)* $1 < r < \infty$ and $\sigma > 0$, *(ii)* $1 < r < \infty$ and $\sigma = 0$, and *(iii)* $r = 1$.

As for first case ($1 < r < \infty$ and $\sigma > 0$), note that the map $f \mapsto \sqrt{\sigma^2 + \delta^{-1}\|f - B\|_2^2}$ is strictly convex when $\sigma > 0$.

As for the second case ($1 < r < \infty$ and $\sigma = 0$), let $f_0 = f_\kappa$ and f_1 be any other minimizer. Then, for $0 < t < 1$, equality must hold separately in the convexity inequalities for $\|f - B\|_2$ and $\|f\|_r$ along $(1-t)f_0 + tf_1$. This implies that there exist $a, b \geq 0$ such that

$$f_1 - B = a(f_0 - B), \qquad f_1 = bf_0. \tag{90}$$

In fact, equality in the L_2 triangle inequality gives the first relation; since L_r is strictly convex for $1 < r < \infty$, equality in its triangle inequality gives the second. If $a = b$, subtracting the two identities in (90) gives $(1-a)B = 0$. As B is not almost surely zero, we have $a = 1$, which implies $f_1 = f_0$. If $a \neq b$, $f_0 = cB$ for a deterministic constant c. However, (89) would express G as a measurable function of B, which is impossible.

As for the third case ($r = 1$), let us define the Moreau envelope and scalar profile

$$M_1(z;\kappa) := \min_u\{\tfrac{1}{2}(u-z)^2 + \kappa|u|\},$$
$$\Phi_1(\tau) := \frac{\tau}{2} + \frac{\sigma^2}{2\tau} + \frac{\tau}{\delta}\left(\mathbb{E}M_1(B/\tau + G;\kappa) - \frac{1}{2}\right).$$

Denote $H_\kappa(\tau) := H_1(\tau, \kappa)$. We show that

$$H_\kappa'(\tau) = -\frac{2}{\tau^3}\mathbb{E}\left[B^2\mathbf{1}_{\{|B/\tau+G|\leq\kappa\}}\right] \leq 0. \tag{91}$$

Set $A = B/\tau$, $z = A + G$, $u = \mathrm{ST}(z;\kappa)$, and $q = u - A$. Away from $|z| = \kappa$, q is independent of τ on the active set $|z| > \kappa$, whereas $q = -B/\tau$ on the inactive set. Therefore, $\partial_\tau q^2 = -2B^2\tau^{-3}\mathbf{1}_{\{|z|\leq\kappa\}}$ almost surely. The threshold event has conditional Gaussian probability zero, and the derivative is dominated on compact τ intervals by CB^2. Hence, dominated convergence proves (91).

Next, we show that the derivative of the scalar profile is

$$\Phi_1'(\tau) = \frac{1}{2\tau^2}\left(\tau^2\left(1 - \frac{H_\kappa(\tau)}{\delta}\right) - \sigma^2\right). \tag{92}$$

Note that the envelope theorem gives $\partial_z M_1(z;\kappa) = z - u$. With $e = z - u = G - q$, the proximal condition gives $eu = \kappa|u|$. Hence,

$$M_1(z;\kappa) - Ae = \tfrac{1}{2}e^2 + eu - Ae = \tfrac{1}{2}e^2 + eq = \tfrac{1}{2}G^2 - \tfrac{1}{2}q^2.$$

Differentiating $\tau\{\mathbb{E}M_1(B/\tau + G;\kappa) - 1/2\}$, we obtain the following derivative:

$$\mathbb{E}[M_1 - 1/2 - A\partial_z M_1] = -\tfrac{1}{2}H_\kappa(\tau),$$

because $\mathbb{E}G^2 = 1$. Adding the derivatives of $\tau/2 + \sigma^2/(2\tau)$ proves (92).

At this point, we show that the function Φ_1 has the unique minimizer τ_κ. Let $K(\tau) := \tau^2(1 - H_\kappa(\tau)/\delta)$. By (14), $K(\tau_\kappa) = \sigma^2$. If $\sigma > 0$, then $H_\kappa(\tau_\kappa) < \delta$. If $0 < s < t$ and $H_\kappa(s) < \delta$, then

$$K(t) - K(s) = (t^2 - s^2)\left(1 - \frac{H_\kappa(s)}{\delta}\right) + \frac{t^2}{\delta}\{H_\kappa(s) - H_\kappa(t)\} > 0.$$

We now compare $K(\tau)$ with $K(\tau_\kappa) = \sigma^2$. If $\tau < \tau_\kappa$ and $H_\kappa(\tau) \geq \delta$, then $K(\tau) \leq 0 < \sigma^2$. If instead $H_\kappa(\tau) < \delta$, applying the preceding identity with $s = \tau$ and $t = \tau_\kappa$ gives $K(\tau) < K(\tau_\kappa) = \sigma^2$. Finally, if $\tau > \tau_\kappa$, then $H_\kappa(\tau_\kappa) < \delta$, and applying the same identity with $s = \tau_\kappa$ and $t = \tau$ gives $K(\tau) > K(\tau_\kappa) = \sigma^2$. Therefore, (92) shows that Φ_1 is strictly decreasing on $(0, \tau_\kappa)$ and strictly increasing on (τ_κ, ∞).

If $\sigma = 0$, then $H_\kappa(\tau_\kappa) = \delta$. Since $\mathbb{P}(B \neq 0) > 0$, the expectation in (91) is strictly positive for every $\tau > 0$, so H_κ is strictly decreasing. Hence $K(\tau) < 0$ before τ_κ and $K(\tau) > 0$ after it. As a result, (92) gives that Φ_1 decreases before τ_κ and increases after it, which implies the desired claim on the uniqueness of the minimizer of Φ_1.

As $\sqrt{x} = \inf_{\tau>0}\{\tau/2 + x/(2\tau)\}$, we have

$$\mathcal{Q}(f) = \inf_{\tau>0}\left[\frac{\tau}{2} + \frac{\sigma^2}{2\tau} + \frac{1}{\delta}\mathbb{E}\left[\frac{(f-B)^2}{2\tau} - G(f-B) + \kappa|f|\right]\right] \tag{93}$$

For fixed τ, the term inside the expectation is strictly convex in f and its unique pointwise minimizer is $f = \tau\,\mathrm{ST}(B/\tau + G;\kappa)$. Minimizing first in f produces $\Phi_1(\tau)$. If $\sigma > 0$, or if $\|f - B\|_2 > 0$, the infimum over τ in (93) is attained at a positive scale. The strict minimality of τ_κ for Φ_1, together with pointwise strict convexity, then shows that $\mathcal{Q}(f) = \mathcal{Q}(f_\kappa)$ forces $f = f_\kappa$.

It remains, when $\sigma = 0$, to exclude $f = B$, whose optimal scale is the boundary $\tau \downarrow 0$. Fix $0 < \tau_0 < \tau_\kappa$. At $f = B$, the expression in brackets in (93) is $\frac{\tau}{2} + \frac{\kappa}{\delta}\mathbb{E}|B|$, which is bounded below by $\Phi_1(\tau)$. Hence, for $0 < \tau \leq \tau_0$, it is at least $\Phi_1(\tau_0) > \Phi_1(\tau_\kappa)$. Letting $\tau \downarrow 0$ gives

$$\mathcal{Q}(B) \geq \Phi_1(\tau_0) > \Phi_1(\tau_\kappa) = \mathcal{Q}(f_\kappa).$$

Thus, f_κ is the unique minimizer of $\mathcal{Q}$ and the proof is complete. □

B.3 Joint empirical laws and bounds on Gaussian inner products

Lemma 14 (Convergence of joint empirical distribution)**.** *Let $h = (h_1, \ldots, h_p)$ have i.i.d. $N(0,1)$ entries and be independent of β_*. Almost surely,*

$$\frac{1}{p}\sum_{j=1}^{p}\delta_{((\beta_*)_j, h_j)} \xrightarrow[W_q]{} \mathcal{L}(B, G). \tag{94}$$

Consequently, for fixed $\tau, \alpha > 0$, if $b_j(\tau,\alpha) := \tau\eta_{\alpha,r}((\beta_*)_j/\tau + h_j)$, then

$$\|b(\tau,\alpha) - \beta_*\|_{2,p}^2 \to \tau^2 H_r(\tau,\alpha), \tag{95}$$

$$\|b(\tau,\alpha)\|_{r,p}^r \to \tau^r U_r(\tau,\alpha), \tag{96}$$

$$\langle h, b(\tau,\alpha) - \beta_*\rangle_p \to \tau D_r(\tau,\alpha). \tag{97}$$

For $r = 1$, (96) corresponds to convergence of the first absolute moment, and the empirical law of $((\beta_*)_j, b_j(\tau,\alpha))$ converges in W_2 to the corresponding population law.

Proof. We start by showing that, for every bounded Lipschitz $f : \mathbb{R}^2 \to \mathbb{R}$,

$$\frac{1}{p}\sum_{j=1}^{p} f((\beta_*)_j, h_j) \to \mathbb{E}f(B, G) \qquad \text{almost surely.} \tag{98}$$

Let $\bar{f}(x) := \mathbb{E}f(x, G)$. Then, $p^{-1}\sum_j \bar{f}((\beta_*)_j) \to \mathbb{E}\bar{f}(B)$, as $\bar{f}$ is bounded and continuous, and W_q convergence of the empirical law of β_* implies weak convergence. Consider the centered average

$$Z_p := \frac{1}{p}\sum_{j=1}^{p}\{f((\beta_*)_j, h_j) - \bar{f}((\beta_*)_j)\}.$$

Conditional on β_*, the summands are independent, centered, and uniformly bounded. Let $M > 0$ satisfy $\|f\|_\infty \le M$. Hoeffding's inequality (Vershynin, 2018, Theorem 2.2.6), applied to both tails, gives, for every $t > 0$,

$$\mathbb{P}(|Z_p| > t) \le 2\exp\left(-\frac{pt^2}{2M^2}\right).$$

Thus, an application of Borel–Cantelli gives that Z_p converges almost surely to zero, which gives (98).

The deterministic measures $p^{-1}\sum_j \delta_{(\beta_*)_j}$ are uniformly q-integrable, due to their assumed W_q convergence. Applying first the strong law of large numbers to the truncated functions $|h|^q \wedge K$ and then letting $K \to \infty$ gives that, almost surely,

$$\lim_{L\to\infty}\limsup_{p\to\infty}\frac{1}{p}\sum_{j=1}^{p}|h_j|^q \mathbf{1}_{\{|h_j|>L\}} = 0.$$

Since $\|(x, g)\|^q \le C_q(|x|^q + |g|^q)$, we obtain that the q-th moments of the empirical measures are uniformly integrable and converge to the q-th moment of (B, G). Thus, a union bound over a countable class of bounded Lipschitz function estalishes the weak convergence in (94).

Consider the continuous map

$$(x, g) \mapsto \big(x, \tau\eta_{\alpha,r}(x/\tau + g)\big).$$

Lemma 11 gives linear growth of this map. Furthermore, the squared-error function has quadratic growth, and the rth-power function has growth at most $C(1+|x|^q+|g|^q)$ because $q = 2 \vee r$. Hence, (94) implies (95) and (96), and the asserted W_2 convergence for $r = 1$.

It remains to show (97). To that aim, consider the function

$$(x, g) \mapsto g(\eta_{\alpha,r}(x/\tau + g) - x/\tau).$$

This function is continuous except on a null set when $r = 1$, and has at most quadratic growth due to Lemma 11. Hence, using (94), we have that its empirical average converges to $\mathbb{E}[G\{\eta_{\alpha,r}(B/\tau + G) - B/\tau\}]$. Conditional on B, the map of G is Lipschitz and absolutely continuous. Thus, Gaussian integration by parts gives

$$\mathbb{E}_G[G\eta_{\alpha,r}(B/\tau + G)] = \mathbb{E}_G[\eta'_{\alpha,r}(B/\tau + G)],$$

while $\mathbb{E}_G[GB/\tau] = 0$. Thus, integrating over B and using dominated convergence gives (97) and concludes the proof. □

Define

$$\gamma_r(\kappa) := \sup_{v\in L_2\cap L_r,\ \|v\|_2\le 1}\{\mathbb{E}[Gv] - \kappa\|v\|_r\}. \tag{99}$$

Here, $L_s = L_s(P_B \otimes N(0, 1))$, we identify with B and G the coordinate random variables on this product space and for $v \in L_2 \cap L_r$, $\mathbb{E}[Gv]$ denotes the L_2 inner product between G and $v = v(B, G)$.

Lemma 15 (Upper bound on $\gamma_r(\kappa)$)**.** *Assume that* (14) *holds and that* $D_r(\tau_\kappa, \alpha_\kappa) \le \delta$*. Then*

$$\gamma_r(\kappa) < \sqrt{\delta}. \tag{100}$$

Proof. We start by showing that the quantity in (99) has the dual representation

$$\gamma_r(\kappa) = \mathrm{dist}_{L_2}\big(G, K_\kappa\big), \qquad K_\kappa := \{s \in L_2 : \|s\|_{r^*} \le \kappa\}, \tag{101}$$

where r^* denotes the conjugate exponent of r, i.e., $1/r + 1/r^* = 1$, with $r^* = \infty$ when $r = 1$. Note that the set K_κ is convex in L_2. We now prove that K_κ is closed with respect to the L_2 topology. This requires showing that, if $s_n \in K_\kappa$ and $\|s_n - s\|_2 \to 0$, then $\|s\|_{r^*} \le \kappa$. If $r^* \le 2$, then

$$|\|s_n\|_{r^*} - \|s\|_{r^*}| \le \|s_n - s\|_{r^*} \le \|s_n - s\|_2 \to 0.$$

If $2 < r^* < \infty$, then L_2 convergence does not imply L_{r^*} convergence. However, L_2 convergence implies convergence in probability and, hence, there is a subsequence (s_{n_k}) such that $s_{n_k} \to s$ almost surely. Fatou's lemma then gives

$$\mathbb{E}|s|^{r^*} \le \liminf_{k\to\infty} \mathbb{E}|s_{n_k}|^{r^*} \le \kappa^{r^*}.$$

Finally, if $r^* = \infty$, then $\|s_n\|_\infty \le \kappa$ implies that $|s_n| \le \kappa$ almost surely. Thus, taking the almost-surely convergent subsequence above, we obtain pointwise $|s| = \lim_{k\to\infty} |s_{n_k}| \le \kappa$ almost surely, which gives that $\|s\|_\infty \le \kappa$. Since K_κ is closed and convex, G has a unique projection onto it, which we denote by s_0. Moreover, s_0 is measurable with respect to G. Indeed, if $s \in K_\kappa$, then $\bar{s} := \mathbb{E}[s \mid G]$ is feasible by conditional Jensen, and

$$\mathbb{E}(G - \bar{s})^2 = \mathbb{E}(G - s)^2 - \mathbb{E}(s - \bar{s})^2.$$

Uniqueness of the projection therefore gives $s_0 = \mathbb{E}[s_0 \mid G]$ almost surely. By symmetry and uniqueness, s_0 is an odd function of G.

For every $s \in K_\kappa$ and every $v \in L_2 \cap L_r$,

$$\mathbb{E}[Gv] - \kappa\|v\|_r \le \mathbb{E}[(G - s)v] \le \|G - s\|_2 \|v\|_2,$$

where the first inequality follows from Hölder's inequality and the fact that $s \in K_\kappa$, and the second one follows from Cauchy–Schwarz. Taking the supremum in v and infimum in s proves one side of the equivalence in (101).

Set $q_0 := G - s_0$. As s_0 is the projection of G, $\mathbb{E}[(G - s_0)(s - s_0)] \le 0$ for all $s \in K_\kappa$. Consequently,

$$\sup_{s\in K_\kappa} \mathbb{E}[q_0 s] = \mathbb{E}[q_0 s_0] < \infty. \tag{102}$$

The support function of K_κ is

$$\sup_{s\in K_\kappa} \mathbb{E}[qs] = \begin{cases} \kappa\|q\|_r, & q \in L_r, \\ +\infty, & q \notin L_r. \end{cases} \tag{103}$$

Indeed, the upper bound in the first case follows from Hölder's inequality. The reverse bound follows by truncating the formal optimizer $s^* = \kappa \frac{\mathrm{sign}(q)|q|^{r-1}}{\|q\|_r^{r-1}}$ when $1 < r < \infty$. For $r = 1$, the optimizer is $s^* = \kappa\,\mathrm{sign}(q)$. The second case in (103) follows by applying the same truncation argument.

Since the LHS of (102) is finite, (103) implies that $q_0 \in L_r$. Moreover, $\mathbb{E}[q_0 s_0] = \kappa\|q_0\|_r$. Therefore,

$$\mathbb{E}[Gq_0] - \kappa\|q_0\|_r = \mathbb{E}[(G - s_0)q_0] = \|q_0\|_2^2.$$

If $q_0 \neq 0$, taking $v = q_0/\|q_0\|_2$ in (99) gives

$$\gamma_r(\kappa) \ge \|q_0\|_2 = \mathrm{dist}_{L_2}(G, K_\kappa).$$

Together with the opposite inequality proved above, this establishes (101). If $q_0 = 0$, the conclusion is immediate.

Recall the definitions in (86) and consider the random variable

$$s_\kappa := A_\kappa + G - u_\kappa = A_\kappa + G - \eta_{\alpha_\kappa, r}(A_\kappa + G). \tag{104}$$

If $r = 1$, then $s_\kappa = A_\kappa + G - \mathrm{ST}(A_\kappa + G; \kappa)$, which implies that $\|s_\kappa\|_\infty \le \kappa$. If $1 < r < \infty$, then $s_\kappa = \alpha_\kappa \psi_r(u_\kappa)$ by (82), and $\|s_\kappa\|_{r^*} = \alpha_\kappa(\mathbb{E}|u_\kappa|^r)^{(r-1)/r} = \kappa$ by the second equation in (14). Therefore, for all $r \ge 1$, $s_\kappa \in K_\kappa$ and we have

$$\gamma_r(\kappa)^2 \le \|G - s_\kappa\|_2^2 = \|u_\kappa - A_\kappa\|_2^2 = H_r(\tau_\kappa, \alpha_\kappa) \le \delta, \tag{105}$$

where the first inequality follows from (101), the next two equalities respectively use (104) and (13), and the last inequality follows from (14). If $\sigma > 0$, then the last inequality in (105) is strict and (100) follows immediately.

It remains to show that the inequality in (105) is strict for $\sigma = 0$. We will do that by contradiction and suppose that equality in (105) holds. Equality in (105) would imply that s_κ is a metric projection of G onto K_κ. By uniqueness of the projection, $s_\kappa = s_0$. By the observation following the definition of s_0, it follows that s_κ is an odd measurable function of G.

Consider $1 < r < \infty$. From $s_\kappa = \alpha_\kappa \psi_r(u_\kappa)$, the invertibility of ψ_r makes u_κ an odd measurable function of G. Furthermore, $u_\kappa - A_\kappa = G - s_\kappa$ is odd and measurable in G. Hence A_κ is both measurable in G and independent of G, so it is constant; being odd, we have that $A_\kappa = 0$ almost surely. For $r = 1$, we also have that $A_\kappa = 0$ almost surely. In fact, the metric projection onto the L_∞ ball is uniquely $\mathrm{clip}(G; [-\kappa, \kappa])$, where $\mathrm{clip}(z; [a, b]) := \min\{b, \max\{a, z\}\}$. Hence, $\mathrm{ST}(A_\kappa + G; \kappa) - A_\kappa = \mathrm{ST}(G; \kappa)$. Conditional on $A_\kappa = a$, this equality holds for almost every G. On the interval $|G| < \kappa$ it can hold only for $a = 0$. Independence therefore gives $A_\kappa = 0$ almost surely.

As $A_\kappa = B/\tau_\kappa$, equality in (105) would therefore imply $B = 0$ almost surely. Since equality also forces $\sigma = 0$, (14) gives $H_r(\tau_\kappa, \alpha_\kappa) = \delta$. On the other hand, Lemma 12 would give $D_r(\tau_\kappa, \alpha_\kappa) > H_r(\tau_\kappa, \alpha_\kappa) = \delta$, contradicting the hypothesis. Hence (105) is strict, and the proof is complete. □

Lemma 16 (Convergence to $\gamma_r(\kappa)$)**.** *Let*

$$\gamma_{r,p}(\kappa) := \sup_{\|v\|_{2,p} \le 1} \{\langle h, v\rangle_p - \kappa \|v\|_{r,p}\}. \tag{106}$$

Then, for every fixed $\kappa > 0$,

$$\gamma_{r,p}(\kappa) \to \gamma_r(\kappa)$$

almost surely and in L_1.

Proof. Repeating the same argument of Lemma 15 and replacing expectations with empirical averages, we have

$$\gamma_{r,p}(\kappa) = \min_{\|s\|_{r^*,p} \le \kappa} \|h - s\|_{2,p},$$

with r^* the conjugate exponent of r. Let $s_0(G)$ be the population metric projection in (101). Evaluating it at $h_1, \ldots, h_p$ and, if necessary, rescaling by $\frac{\kappa}{\|s_0(h)\|_{r^*,p} \vee \kappa}$ gives

$$\limsup_{p\to\infty} \gamma_{r,p}(\kappa) \le \gamma_r(\kappa),$$

from the strong law of large numbers applied to $|h - s_0(h)|^2$ and, when $r^* < \infty$, to $|s_0(h)|^{r^*}$.

Conversely, let s_p be an empirical projection. Then $\|h - s_p\|_{2,p} \le \|h\|_{2,p}$, and hence $\|s_p\|_{2,p} \le 2\|h\|_{2,p}$. Along every subsequence, the empirical laws of $(h_j, s_{p,j})$ have a further weakly convergent subsequence, say to the law of (G, S), with

$$\|S\|_{r^*} \le \kappa, \qquad \|G - S\|_2 \le \liminf_{p\to\infty} \|h - s_p\|_{2,p}.$$

Here tightness follows from the L_2 bound; lower semicontinuity gives the constraint and the distance inequality, with truncation when $r^* > 2$. The limit variable S need not be measurable with respect to G, so put $\overline{S} := \mathbb{E}[S \mid G]$. Conditional Jensen gives

$$\|\overline{S}\|_{r^*} \le \|S\|_{r^*} \le \kappa, \qquad \|G - \overline{S}\|_2 \le \|G - S\|_2.$$

Thus $\overline{S}$, viewed as a function of the G-coordinate on the product space supporting (B, G), is feasible in (101). Consequently,

$$\gamma_r(\kappa) \le \|G - \overline{S}\|_2 \le \|G - S\|_2 \le \liminf_{p\to\infty} \|h - s_p\|_{2,p}.$$

This proves almost-sure convergence. Finally,

$$0 \le \gamma_{r,p}(\kappa) \le \|h\|_{2,p},$$

and the variables on the right have uniformly bounded second moments. Thus the convergence also holds in L_1. □

B.4 Stability of the empirical objective

Define the empirical active functional

$$Q_p(b) := \sqrt{\sigma^2 + \frac{1}{\delta_p}\|b-\beta_*\|_{2,p}^2} - \frac{1}{\delta_p}\langle h, b-\beta_*\rangle_p + \frac{\kappa}{\delta_p}\|b\|_{r,p}. \tag{107}$$

Set

$$M_\kappa^{\mathrm{err}} := \mathbb{E}(f_\kappa - B)^2 = \tau_\kappa^2 H_r(\tau_\kappa,\alpha_\kappa), \qquad M_\kappa^{\mathrm{coef}} := \mathbb{E}|f_\kappa|^r = \tau_\kappa^r U_r(\tau_\kappa,\alpha_\kappa), \qquad \mu_\kappa := \mathcal{L}(B, f_\kappa).$$

For $b \in \mathbb{R}^p$, define

$$\mu_{p,b} := \frac{1}{p}\sum_{j=1}^p \delta_{((\beta_*)_j, b_j)}$$

and

$$d_p(b) := \max\left\{\left|\|b-\beta_*\|_{2,p}^2 - M_\kappa^{\mathrm{err}}\right|, \left|\|b\|_{r,p}^r - M_\kappa^{\mathrm{coef}}\right|, W_2(\mu_{p,b},\mu_\kappa)\right\}. \tag{108}$$

For $\epsilon > 0$, define

$$\mathcal{B}_{p,\epsilon} := \{b \in \mathbb{R}^p : d_p(b) \geq \epsilon\}. \tag{109}$$

Lemma 17 (Stability of empirical objective)**.** *Assume that (14) holds and that $D_r(\tau_\kappa,\alpha_\kappa) \leq \delta$. Then, almost surely, the following statements hold.*

(i) There are deterministic constants $c, C > 0$ such that, for every sufficiently large p and every $b \in \mathbb{R}^p$,

$$Q_p(b) \geq c\|b\|_{2,p} - C. \tag{110}$$

(ii) We have

$$\inf_{b\in\mathbb{R}^p} Q_p(b) \to \mathcal{Q}(f_\kappa). \tag{111}$$

(iii) If $Q_p(b_p) \leq \min_b Q_p(b) + o(1)$, then

$$\|b_p - \beta_*\|_{2,p}^2 \to \tau_\kappa^2 H_r(\tau_\kappa,\alpha_\kappa), \tag{112}$$

$$\|b_p\|_{r,p}^r \to \tau_\kappa^r U_r(\tau_\kappa,\alpha_\kappa), \tag{113}$$

$$\frac{1}{p}\sum_{j=1}^p \delta_{((\beta_*)_j, b_{p,j})} \xrightarrow[W_2]{} \mu_\kappa. \tag{114}$$

(iv) For every $\epsilon > 0$,

$$\liminf_{p\to\infty}\left(\inf_{b\in\mathcal{B}_{p,\epsilon}} Q_p(b) - \inf_{b\in\mathbb{R}^p} Q_p(b)\right) > 0. \tag{115}$$

Proof. By Lemmas 15 and 16, and since $\delta_p \to \delta$, there is a deterministic $c_1 > 0$ such that, eventually,

$$\sqrt{\delta_p}\|v\|_{2,p} - \langle h, v\rangle_p + \kappa\|v\|_{r,p} \geq c_1\|v\|_{2,p}, \tag{116}$$

for $v \in \mathbb{R}^p$. An application of the triangle inequality gives

$$\sqrt{\sigma^2 + \delta_p^{-1}\|b-\beta_*\|_{2,p}^2} \geq \delta_p^{-1/2}\|b\|_{2,p} - \delta_p^{-1/2}\|\beta_*\|_{2,p}.$$

As $\|\beta_*\|_{2,p} = O(1)$ and $|\langle h, \beta_*\rangle_p| = O(1)$ almost surely, applying (116) with $v = b$ gives that (110) holds.

The bound in (110) readily implies that every sequence with $Q_p(b_p) = O(1)$ has bounded $\|b_p\|_{2,p}$. Furthermore, rearranging (107) gives

$$\frac{\kappa}{\delta_p}\|b_p\|_{r,p} \leq Q_p(b_p) + \frac{1}{\delta_p}|\langle h, b_p - \beta_*\rangle_p|,$$

which implies that $\|b_p\|_{r,p}$ is also bounded whenever $Q_p(b_p)$ is bounded.

Let $b^\circ_{p,j} := \tau_\kappa \eta_{\alpha_\kappa,r}((\beta_*)_j/\tau_\kappa + h_j)$. Then, by Lemma 14, we have $Q_p(b^\circ_p) \to \mathcal{Q}(f_\kappa)$. This implies that

$$\limsup_p \min_b Q_p(b) \le \mathcal{Q}(f_\kappa). \tag{117}$$

Let b_p be any sequence satisfying

$$Q_p(b_p) \le \inf_b Q_p(b) + o(1).$$

Then, the sequences $\|b_p - \beta_*\|_{2,p}$ and $\|b_p\|_{r,p}$ are bounded. Fix an arbitrary subsequence. Passing to a further subsequence, we may assume that

$$a_p := \|b_p - \beta_*\|^2_{2,p} \to a, \qquad c_p := \|b_p\|^r_{r,p} \to c,$$

and that the empirical laws

$$\nu_p := \frac{1}{p}\sum_{j=1}^p \delta_{((\beta_*)_j, h_j, b_{p,j})}$$

converge weakly to the law of a triple (B, G, F). Such a subsequence exists by the previous moment bounds and Prokhorov's theorem.

We first verify convergence of the Gaussian cross term. Set $e_{p,j} := b_{p,j} - (\beta_*)_j$. For $K, T > 0$,

$$\frac{1}{p}\sum_{j=1}^p |h_j e_{p,j}| \mathbf{1}_{\{|h_j e_{p,j}| > T\}} \le \left(\frac{1}{p}\sum_{j=1}^p h_j^2 \mathbf{1}_{\{|h_j|>K\}}\right)^{1/2} \|e_p\|_{2,p} + \frac{K^2}{T}\|e_p\|^2_{2,p}.$$

Indeed, the first term controls the contribution from $\{|h_j| > K\}$; on its complement, $|h_j e_{p,j}| > T$ implies $|e_{p,j}| > T/K$. Letting first $T \to \infty$ and then $K \to \infty$ proves uniform integrability of the products $h_j e_{p,j}$. Consequently,

$$\langle h, b_p - \beta_* \rangle_p \to \mathbb{E}[G(F - B)]. \tag{118}$$

It follows that

$$\lim_{p\to\infty} Q_p(b_p) = \sqrt{\sigma^2 + \frac{a}{\delta}} - \frac{1}{\delta}\mathbb{E}[G(F-B)] + \frac{\kappa}{\delta} c^{1/r}. \tag{119}$$

Let

$$\overline{F} := \mathbb{E}[F \mid B, G].$$

Lower semicontinuity and conditional Jensen's inequality give

$$a \ge \mathbb{E}(F - B)^2 \ge \mathbb{E}(\overline{F} - B)^2, \qquad c^{1/r} \ge \|F\|_r \ge \|\overline{F}\|_r,$$

whereas

$$\mathbb{E}[G(F - B)] = \mathbb{E}[G(\overline{F} - B)].$$

Hence the RHS of (119) is at least $\mathcal{Q}(\overline{F})$, which in turn is at least $\mathcal{Q}(f_\kappa)$ by Lemma 13. On the other hand, (117) and the near-minimality of b_p give

$$\limsup_{p\to\infty} Q_p(b_p) \le \mathcal{Q}(f_\kappa).$$

Therefore equality holds throughout:

$$\lim_{p\to\infty} Q_p(b_p) = \mathcal{Q}(\overline{F}) = \mathcal{Q}(f_\kappa).$$

Taking b_p to be an exact minimizer proves (111).

The uniqueness of the population minimizer gives that $\overline{F} = f_\kappa$ almost surely. Moreover, $\mathbb{E}(F - B)^2 = \mathbb{E}(f_\kappa - B)^2 + \mathbb{E}(F - f_\kappa)^2$. If $\mathbb{E}(F - f_\kappa)^2 > 0$, the square-root term in (119) would be strictly larger than the corresponding term in $\mathcal{Q}(f_\kappa)$, while the norm term cannot be smaller. This contradicts equality above. Thus, $F = f_\kappa$ almost surely.

It remains to exclude escape of moment mass. Lower semicontinuity now gives $a \ge M^{\mathrm{err}}_\kappa$, $c \ge M^{\mathrm{coef}}_\kappa$. Using $F = f_\kappa$ in (119) and comparing with $\mathcal{Q}(f_\kappa)$, we obtain

$$0 = \sqrt{\sigma^2 + \frac{a}{\delta}} - \sqrt{\sigma^2 + \frac{M^{\mathrm{err}}_\kappa}{\delta}} + \frac{\kappa}{\delta}\left(c^{1/r} - (M^{\mathrm{coef}}_\kappa)^{1/r}\right).$$

Both terms on the right are nonnegative. Since $\kappa > 0$, both must vanish, and hence $a = M_\kappa^{\mathrm{err}}, c = M_\kappa^{\mathrm{coef}}$. Because the original subsequence was arbitrary, this proves (112) and (113) for the full sequence.

Finally, the empirical laws of $((\beta_*)_j, b_{p,j} - (\beta_*)_j)$ converge weakly to $\mathcal{L}(B, f_\kappa - B)$. Their second moments also converge: the first-coordinate second moments converge by the W_q assumption, and the second-coordinate second moments converge by (112). Thus these empirical laws converge in W_2. Applying the continuous linear map $(x, e) \longmapsto (x, x + e)$ gives (114).

It remains to show statement *(iv)*. By contradiction, suppose it fails for some $\epsilon > 0$. Then, there is a subsequence p_k and $b_{p_k} \in \mathcal{B}_{p_k,\epsilon}$ such that $Q_{p_k}(b_{p_k}) - \inf_b Q_{p_k}(b) \to 0$. For indices outside the subsequence, choose an exact minimizer of Q_p. Then, $Q_p(b_p) \le \min_b Q_p(b) + o(1)$ and by statement *(iii)* we have that $d_p(b_p) \to 0$, which is a contradiction. □

B.5 Analysis of the auxiliary optimization

We next consider the auxiliary optimization problem and relate it to the original one via CGMT. Let $X = G_X/\sqrt{n}$ and $w = b - \beta_*$. Let $g \in \mathbb{R}^n$ and $h \in \mathbb{R}^p$ be independent standard Gaussian vectors, independent of (G_X, ε). Define

$$C_p(b) := \sqrt{\sigma^2 + \delta_p^{-1}\|b - \beta_*\|_{2,p}^2} - \delta_p^{-1}\langle h, b - \beta_*\rangle_p, \tag{120}$$

$$C_p^{\mathrm{raw}}(b) := \left\| \frac{\varepsilon}{\sqrt{n}} + \frac{\|b - \beta_*\|_2}{n} g \right\|_2 - \frac{1}{n} h^\top (b - \beta_*). \tag{121}$$

The CGMT auxiliary objective is

$$A_p^{\mathrm{raw}}(b) := [C_p^{\mathrm{raw}}(b)]_+ + \frac{\kappa}{\delta_p}\|b\|_{r,p}. \tag{122}$$

Its deterministic-norm counterpart is

$$\overline{A}_p(b) := [C_p(b)]_+ + \frac{\kappa}{\delta_p}\|b\|_{r,p}. \tag{123}$$

Define the minima

$$\overline{a}_p^* := \inf_b \overline{A}_p(b), \qquad a_{p,\mathrm{raw}}^* := \inf_b A_p^{\mathrm{raw}}(b), \tag{124}$$

and the set

$$C_{p,M} := \{b \in \mathbb{R}^p : \|b - \beta_*\|_{2,p} \le M, \|b\|_{r,p} \le M\}.$$

Lemma 18 (Characterization of the auxiliary optimization)**.** *Assume that (14) holds and that $D_r(\tau_\kappa, \alpha_\kappa) \le \delta$. Then, almost surely, the following statements hold.*

(i) For any $M < \infty$,

$$\sup_{\|b - \beta_*\|_{2,p} \le M} |A_p^{\mathrm{raw}}(b) - \overline{A}_p(b)| \longrightarrow 0. \tag{125}$$

(ii) For any $L < \infty$, there exists $M_L < \infty$ such that, for all sufficiently large p,

$$\{b : \overline{A}_p(b) \le L\} \cup \{b : A_p^{\mathrm{raw}}(b) \le L\} \subseteq C_{p,M_L}. \tag{126}$$

(iii) The two auxiliary minimum values satisfy

$$\overline{a}_p^* \to \ell_\kappa, \qquad a_{p,\mathrm{raw}}^* \to \ell_\kappa, \tag{127}$$

where

$$\ell_\kappa := R_\kappa + \frac{\kappa}{\delta}\|f_\kappa\|_r. \tag{128}$$

(iv) For every $\epsilon > 0$,

$$\liminf_{p\to\infty} \left(\inf_{b \in \mathcal{B}_{p,\epsilon}} \overline{A}_p(b) - \overline{a}_p^* \right) > 0, \tag{129}$$

$$\liminf_{p\to\infty} \left(\inf_{b \in \mathcal{B}_{p,\epsilon}} A_p^{\mathrm{raw}}(b) - a_{p,\mathrm{raw}}^* \right) > 0. \tag{130}$$

Proof. Let $\varepsilon = \sigma z$ with $z \sim N(0, I_n)$ independent of g, and set $\rho := \|b - \beta_*\|_2/\sqrt{n}$. Then,

$$\left\| \frac{\varepsilon}{\sqrt{n}} + \frac{\|b - \beta_*\|_2}{n} g \right\|_2 = T_n(\rho) := \frac{1}{\sqrt{n}} \|\sigma z + \rho g\|_2.$$

For every fixed $\rho \geq 0$, $T_n(\rho)^2 \to \sigma^2 + \rho^2$ almost surely. As $\sqrt{\sigma^2 + \rho^2}$ is 1-Lipschitz, for all $\rho, \rho' \geq 0$,

$$|T_n(\rho) - T_n(\rho')| \leq \frac{\|g\|_2}{\sqrt{n}} |\rho - \rho'|.$$

Since $\|g\|_2/\sqrt{n} \to 1$, a finite-net argument gives uniform convergence on every compact ρ interval and (125) readily follows.

Consider the sublevel sets of $\overline{A}_p(b)$, and set $w = b - \beta_*$. Using $[x]_+ \geq x$ and $\|b\|_{r,p} \geq \|w\|_{r,p} - \|\beta_*\|_{r,p}$, we have

$$\overline{A}_p(b) \geq \frac{1}{\delta_p} \left(\sqrt{\delta_p} \|w\|_{2,p} - \langle h, w \rangle_p + \kappa \|w\|_{r,p} \right) - C.$$

By (116), $\overline{A}_p(b) \geq c\|w\|_{2,p} - C$ eventually. Hence, a bounded objective sublevel has bounded $\|w\|_{2,p}$. Since $\overline{A}_p(b) \geq (\kappa/\delta_p)\|b\|_{r,p}$, the same sublevel has bounded $\|b\|_{r,p}$.

Moving to the sublevel sets of $A_p^{\mathrm{raw}}(b)$, the reverse triangle inequality yields

$$\frac{1}{\sqrt{n}} \|\sigma z + \rho g\|_2 \geq \frac{\|g\|_2}{\sqrt{n}} \rho - \sigma \frac{\|z\|_2}{\sqrt{n}}.$$

Let $a_n := \|g\|_2/\sqrt{n} \to 1$. Then, the same argument as above gives

$$A_p^{\mathrm{raw}}(b) \geq \frac{1}{\delta_p} \left(a_n \sqrt{\delta_p} \|w\|_{2,p} - \langle h, w \rangle_p + \kappa \|w\|_{r,p} \right) - C.$$

Since $a_n \sqrt{\delta_p} - \gamma_{r,p}(\kappa) \to \sqrt{\delta} - \gamma_r(\kappa) > 0$, the RHS is at least $c\|w\|_{2,p} - C$ eventually. Again, $A_p^{\mathrm{raw}}(b) \geq (\kappa/\delta_p)\|b\|_{r,p}$, which concludes the proof of (126).

We next prove statement *(iii)*. For every b, $\overline{A}_p(b) \geq Q_p(b)$. Lemma 17 and (111) therefore give

$$\liminf_{p \to \infty} \overline{a}_p^* \geq \mathcal{Q}(f_\kappa).$$

For the sequence $b_{p,j}^\circ := \tau_\kappa \eta_{\alpha_\kappa, r} \left(\frac{(\beta_*)_j}{\tau_\kappa} + h_j \right)$, Lemma 14, (14), and Gaussian integration by parts give

$$C_p(b_p^\circ) \to R_\kappa, \qquad \|b_p^\circ\|_{r,p} \to \|f_\kappa\|_r.$$

The hypothesis $D_r(\tau_\kappa, \alpha_\kappa) \leq \delta$ gives $R_\kappa \geq 0$. Consequently,

$$\limsup_{p \to \infty} \overline{a}_p^* \leq R_\kappa + \frac{\kappa}{\delta} \|f_\kappa\|_r = \ell_\kappa.$$

A similar calculation gives

$$\mathcal{Q}(f_\kappa) = R_\kappa + \frac{\kappa}{\delta} \|f_\kappa\|_r = \ell_\kappa.$$

Thus, $\overline{a}_p^* \to \ell_\kappa$.

Notice also that every near-minimizer of $\overline{A}_p$ is a near-minimizer of Q_p. Indeed, if $\overline{A}_p(b_p) \leq \overline{a}_p^* + o(1)$, then

$$0 \leq Q_p(b_p) - \inf_b Q_p(b) \leq \overline{A}_p(b_p) - \inf_b Q_p(b) \longrightarrow 0.$$

We now compare the minima of the two objectives (122) and (123). Evaluating both objectives at $b = \beta_*$ gives

$$\overline{A}_p(\beta_*) = \sigma + \frac{\kappa}{\delta_p} \|\beta_*\|_{r,p} = O(1)$$

and

$$A_p^{\mathrm{raw}}(\beta_*) = \frac{\|\varepsilon\|_2}{\sqrt{n}} + \frac{\kappa}{\delta_p} \|\beta_*\|_{r,p} = O(1)$$

almost surely. Hence there is a deterministic $L < \infty$ such that both minimum values are at most L for all sufficiently large p. By statement *(ii)*, the minimizers of both objectives are then contained in C_{p,M_L} eventually. Set

$$\Delta_p := \sup_{\|b-\beta_*\|_{2,p} \le M_L} \left| A_p^{\mathrm{raw}}(b) - \overline{A}_p(b) \right|. \tag{131}$$

Statement *(i)* gives $\Delta_p \to 0$ almost surely. Evaluating each objective at a minimizer of the other one yields

$$a^*_{p,\mathrm{raw}} \le \overline{a}^*_p + \Delta_p, \qquad \overline{a}^*_p \le a^*_{p,\mathrm{raw}} + \Delta_p.$$

Therefore,

$$\left| a^*_{p,\mathrm{raw}} - \overline{a}^*_p \right| \le \Delta_p \longrightarrow 0.$$

Since $\overline{a}^*_p \to \ell_\kappa$, it follows that $a^*_{p,\mathrm{raw}} \to \ell_\kappa$.

We finally prove statement *(iv)*. Fix $\epsilon > 0$ and work on the probability-one event on which Lemma 17*(iv)* and statements *(i)–(iii)* above hold. Since $\overline{A}_p(b) \ge Q_p(b)$,

$$\begin{aligned}\inf_{b \in \mathcal{B}_{p,\epsilon}} \overline{A}_p(b) - \overline{a}^*_p \ge \inf_{b \in \mathcal{B}_{p,\epsilon}} Q_p(b) - \inf_b Q_p(b) \\ + \inf_b Q_p(b) - \overline{a}^*_p.\end{aligned}$$

The last line tends to zero by (111) and statement *(iii)*. Lemma 17*(iv)* therefore implies (129).

We prove (130) by contradiction. Suppose that

$$\liminf_{p\to\infty} \left(\inf_{b \in \mathcal{B}_{p,\epsilon}} A_p^{\mathrm{raw}}(b) - a^*_{p,\mathrm{raw}} \right) = 0.$$

The expression inside braces is nonnegative. Hence, there exists a subsequence, still denoted by p, and points $b_p \in \mathcal{B}_{p,\epsilon}$ such that $A_p^{\mathrm{raw}}(b_p) - a^*_{p,\mathrm{raw}} \longrightarrow 0$. Here, if the infimum is not attained, choose b_p within $1/p$ of it. By statement *(iii)*, $a^*_{p,\mathrm{raw}} \to \ell_\kappa$. It follows that, by taking $L > \ell_\kappa + 1$, $A_p^{\mathrm{raw}}(b_p) \le L$ for all sufficiently large p in the subsequence. By statement *(ii)*, there exists $M_L < \infty$ such that $b_p \in C_{p,M_L}$ eventually. Setting Δ_p as in (131) and recalling that $\Delta_p \to 0$, we have

$$\begin{aligned}0 &\le \overline{A}_p(b_p) - \overline{a}^*_p \\ &\le \left| \overline{A}_p(b_p) - A_p^{\mathrm{raw}}(b_p) \right| + A_p^{\mathrm{raw}}(b_p) - a^*_{p,\mathrm{raw}} + \left| a^*_{p,\mathrm{raw}} - \overline{a}^*_p \right| \\ &\le \Delta_p + A_p^{\mathrm{raw}}(b_p) - a^*_{p,\mathrm{raw}} + \left| a^*_{p,\mathrm{raw}} - \overline{a}^*_p \right| \longrightarrow 0.\end{aligned}$$

On the other hand, $b_p \in \mathcal{B}_{p,\epsilon}$, so (129) implies that $\overline{A}_p(b_p) - \overline{a}^*_p$ is bounded away from zero along every sufficiently large subsequence, which gives a contradiction and concludes the argument. □

B.6 Application of CGMT

We use the following nonasymptotic form of CGMT; see Theorem II.1*(i)–(ii)* (and in particular equations (16) and (21)) of Thrampoulidis et al. (2015b). Let $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{U} \subseteq \mathbb{R}^n$ be nonempty compact sets, let $G \in \mathbb{R}^{n\times p}$, $g \in \mathbb{R}^n$, and $h \in \mathbb{R}^p$ have independent standard Gaussian entries, and let $\psi : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ be continuous. Define

$$\begin{aligned}\Phi(G) &:= \min_{x\in\mathcal{X}} \max_{u\in\mathcal{U}} \left\{ u^\top G x + \psi(x,u) \right\}, \\ \phi(g,h) &:= \min_{x\in\mathcal{X}} \max_{u\in\mathcal{U}} \left\{ \|x\|_2 g^\top u + \|u\|_2 h^\top x + \psi(x,u) \right\}.\end{aligned}$$

Then, for every $t \in \mathbb{R}$,

$$\mathbb{P}(\Phi(G) \le t) \le 2\mathbb{P}(\phi(g,h) \le t). \tag{132}$$

No convexity of $\mathcal{X}$ or $\mathcal{U}$ is required for (132) to hold. If, in addition, $\mathcal{X}$ and $\mathcal{U}$ are convex and ψ is convex in x and concave in u, then

$$\mathbb{P}(\Phi(G) \ge t) \le 2\mathbb{P}(\phi(g,h) \ge t). \tag{133}$$

We note that Thrampoulidis et al. (2015b) use strict inequalities in the events on the LHS. To include equality at the threshold, it suffices to replace t by $t \pm \eta$ and then let $\eta \downarrow 0$, using continuity of probability for decreasing events.

In our case, ψ also depends on the Gaussian noise ε, which is independent of G_X, g, and h. Thus, we apply these results conditionally on ε and then average over ε.

Lemma 19 (Localization of minimizers and application of CGMT). *For $w \in \mathbb{R}^p$, define*

$$P_p(w) := \max_{\|u\|_2 \le 1} \left\{ \frac{1}{\sqrt{n}} u^\top \varepsilon - \frac{1}{n} u^\top G_X w + \frac{\kappa}{\delta_p} \|\beta_* + w\|_{r,p} \right\}. \tag{134}$$

For all p, let $S_p \subseteq \mathbb{R}^p$ be a deterministic closed set, and suppose that

$$\liminf_{p\to\infty} \left(\inf_{b \in S_p} A_p^{\mathrm{raw}}(b) - a_{p,\mathrm{raw}}^* \right) > 0 \qquad \textit{almost surely}. \tag{135}$$

Then, almost surely, there exists $p_0 < \infty$ such that, for every $p \ge p_0$,

$$\arg\min_{w \in \mathbb{R}^p} P_p(w) \cap \{w \in \mathbb{R}^p : \beta_* + w \in S_p\} = \varnothing. \tag{136}$$

Moreover,

$$\inf_{w \in \mathbb{R}^p} P_p(w) \to \ell_\kappa \qquad \textit{almost surely}. \tag{137}$$

Proof. We first prove that all minimizers of P_p eventually belong to a deterministic bounded set. A direct calculation gives that

$$P_p^\infty(v) := \lim_{t\to\infty} \frac{P_p(w+tv) - P_p(w)}{t} = \frac{1}{n}\|G_X v\|_2 + \frac{\kappa}{\delta_p}\|v\|_{r,p}. \tag{138}$$

We claim that there exists a deterministic $c_0 > 0$ such that, almost surely, for all sufficiently large p,

$$P_p^\infty(v) \ge c_0 \|v\|_{2,p} \qquad \text{for every } v \in \mathbb{R}^p. \tag{139}$$

To show this, first note that, by positive homogeneity, it is enough to minimize $P_p^\infty(v)$ over the sphere $\|v\|_{2,p} = 1$. Define

$$R_p^{\mathrm{PO}} := \inf_{\|v\|_{2,p}=1} P_p^\infty(v) = \inf_{\|v\|_{2,p}=1} \left\{ \frac{1}{n}\|G_X v\|_2 + \frac{\kappa}{\delta_p}\|v\|_{r,p} \right\}.$$

This quantity measures the smallest linear growth rate of P_p at infinity. Using the dual representation of the Euclidean norm,

$$\frac{1}{n}\|G_X v\|_2 = \sup_{\|u\|_2 \le 1} \left\{ -\frac{1}{n} u^\top G_X v \right\},$$

the Gaussian min–max theorem associates with R_p^{PO} the auxiliary optimization

$$R_p^{\mathrm{AO}} := \inf_{\|v\|_{2,p}=1} \sup_{\|u\|_2 \le 1} \left\{ \frac{\|v\|_2}{n} g^\top u - \frac{\|u\|_2}{n} h^\top v + \frac{\kappa}{\delta_p}\|v\|_{r,p} \right\},$$

where $g \in \mathbb{R}^n$ and $h \in \mathbb{R}^p$ are independent standard Gaussian vectors. Since $\|v\|_{2,p} = 1$, we have $\|v\|_2 = \sqrt{p}$. For fixed v, write $\rho = \|u\|_2 \in [0,1]$. Maximizing first over the direction of u and then over its norm gives

$$\sup_{\|u\|_2 \le 1} \left\{ \frac{\|v\|_2}{n} g^\top u - \frac{\|u\|_2}{n} h^\top v \right\} = \sup_{0 \le \rho \le 1} \rho \left\{ \frac{\|g\|_2}{\sqrt{n}\sqrt{\delta_p}} - \frac{1}{\delta_p}\langle h, v\rangle_p \right\} = \left[\frac{\|g\|_2}{\sqrt{n}\sqrt{\delta_p}} - \frac{1}{\delta_p}\langle h, v\rangle_p \right]_+ .$$

Indeed, the maximizing radius is $\rho = 1$ when the expression in braces is positive and $\rho = 0$ when it is negative. Consequently,

$$R_p^{\mathrm{AO}} = \inf_{\|v\|_{2,p}=1} \left\{ \left[\frac{\|g\|_2}{\sqrt{n}\sqrt{\delta_p}} - \frac{1}{\delta_p}\langle h, v\rangle_p \right]_+ + \frac{\kappa}{\delta_p}\|v\|_{r,p} \right\}.$$

Using $[x]_+ \ge x$ and the definition of $\gamma_{r,p}(\kappa)$, we have

$$R_p^{\mathrm{AO}} \ge \frac{1}{\delta_p} \inf_{\|v\|_{2,p}=1} \left\{ \frac{\|g\|_2}{\sqrt{n}}\sqrt{\delta_p} - \langle h, v\rangle_p + \kappa\|v\|_{r,p} \right\} \ge \frac{1}{\delta_p} \left(\frac{\|g\|_2}{\sqrt{n}}\sqrt{\delta_p} - \gamma_{r,p}(\kappa) \right). \tag{140}$$

Since

$$\frac{\|g\|_2}{\sqrt{n}} \longrightarrow 1, \qquad \delta_p \to \delta, \qquad \gamma_{r,p}(\kappa) \to \gamma_r(\kappa)$$

almost surely, the RHS of (140) converges to

$$c_* := \frac{\sqrt{\delta} - \gamma_r(\kappa)}{\delta} > 0,$$

where the inequality follows from Lemma 15. The RHS of (140) also converges in L_1 as $0 \le \gamma_{r,p}(\kappa) \le \|h\|_{2,p}$, hence

$$\liminf_{p\to\infty} \mathbb{E} R_p^{\mathrm{AO}} \ge c_*.$$

Furthermore, R_p^{AO} is a $C/\sqrt{p}$-Lipschitz function of (g, h), for a deterministic C independent of p. Gaussian concentration therefore gives deterministic constants $a_0, c > 0$ such that

$$\mathbb{P}(R_p^{\mathrm{AO}} \le a_0) \le e^{-cp} \tag{141}$$

for all sufficiently large p. Applying (132) (cf. (Thrampoulidis et al., 2015b, Theorem II.1*(i)*)), to $\inf_{\|v\|_{2,p}=1} P_p^{\infty}(v)$, we conclude that

$$\mathbb{P}\left(\inf_{\|v\|_{2,p}=1} P_p^{\infty}(v) \le a_0\right) \le 2\mathbb{P}(R_p^{\mathrm{AO}} \le a_0) \le 2e^{-cp}.$$

The probabilities are summable, so Borel–Cantelli proves (139), with $c_0 = a_0$.

The triangle inequality and (139) imply

$$P_p(w) \ge P_p^{\infty}(w) - \frac{\|\varepsilon\|_2}{\sqrt{n}} - \frac{\kappa}{\delta_p}\|\beta_*\|_{r,p} \ge c_0\|w\|_{2,p} - C \tag{142}$$

eventually almost surely, where $C < \infty$ is deterministic. Since $P_p(0) = O(1)$ almost surely, every minimizer of P_p has bounded $\|w\|_{2,p}$. Moreover, $P_p(w) \ge \frac{\kappa}{\delta_p}\|\beta_* + w\|_{r,p}$, so $\|\beta_* + w\|_{r,p}$ is bounded as well. Consequently, there exists a deterministic $M < \infty$ such that, almost surely, all minimizers of P_p eventually belong to

$$\mathcal{C}_{p,M}^{w} := \{w \in \mathbb{R}^p : \|w\|_{2,p} \le M,\ \|\beta_* + w\|_{r,p} \le M\}.$$

Increasing M if necessary, Lemma 18*(ii)-(iii)* ensures that all minimizers of A_p^{raw} eventually satisfy $b - \beta_* \in \mathcal{C}_{p,M}^{w}$. Define

$$V_p := \inf_{\substack{b\in\mathbb{R}^p \\ b-\beta_*\in\mathcal{C}_{p,M}^{w}}} A_p^{\mathrm{raw}}(b), \qquad W_p := \inf_{\substack{b\in S_p \\ b-\beta_*\in\mathcal{C}_{p,M}^{w}}} A_p^{\mathrm{raw}}(b).$$

If the feasible set defining W_p is empty, then (136) trivially holds. We will therefore assume that the feasible set is not empty. By the localization of the minimizers of A_p^{raw} proved in Lemma 18, we have that $V_p = a_{p,\mathrm{raw}}^*$ eventually almost surely. Hence, (135) implies

$$\Gamma := \liminf_{p\to\infty}(W_p - V_p) > 0 \qquad \text{almost surely,} \tag{143}$$

which also implies that

$$c_S := \frac{1}{4}\mathbb{E}[\min\{\Gamma, 1\}] > 0.$$

Fatou's lemma gives $\liminf_{p\to\infty} \mathbb{E}\big[\min\{W_p - V_p, 1\}\big] \ge \mathbb{E}[\min\{\Gamma, 1\}] = 4c_S$. Since $W_p \ge V_p$, it follows that, for all sufficiently large p,

$$\mathbb{E}W_p - \mathbb{E}V_p \ge 2c_S. \tag{144}$$

We now apply the CGMT results. Define

$$\Phi_p := \inf_{w\in\mathcal{C}_{p,M}^{w}} P_p(w), \qquad \Phi_p^S := \inf_{\substack{w\in\mathcal{C}_{p,M}^{w} \\ \beta_*+w\in S_p}} P_p(w).$$

For any $t \in \mathbb{R}$, (132) (cf. (Thrampoulidis et al., 2015b, Theorem II.1*(i)*)) gives

$$\mathbb{P}(\Phi_p^S \le t) \le 2\mathbb{P}(W_p \le t). \tag{145}$$

Furthermore, $\mathcal{C}^w_{p,M}$ is convex, $\{u : \|u\|_2 \le 1\}$ is convex, and $\psi_p(w,u) = \frac{1}{\sqrt{n}} u^\top \varepsilon + \frac{\kappa}{\delta_p}\|\beta_* + w\|_{r,p}$ is convex in w and affine, hence concave, in u. Thus, (133) (cf. (Thrampoulidis et al., 2015b, Theorem II.1*(ii)*)) gives

$$\mathbb{P}(\Phi_p \ge t) \le 2\mathbb{P}(V_p \ge t). \tag{146}$$

Let $t_p := (\mathbb{E}V_p + \mathbb{E}W_p)/2$. By (144), $t_p - \mathbb{E}V_p \ge c_S$ and $\mathbb{E}W_p - t_p \ge c_S$. Note that both V_p and W_p are $C_M/\sqrt{n}$-Lipschitz functions of their standard Gaussian data. Gaussian concentration therefore yields

$$\mathbb{P}(W_p \le t_p) + \mathbb{P}(V_p \ge t_p) \le 2e^{-c'_S n} \tag{147}$$

for a deterministic $c'_S > 0$ and all sufficiently large p. Combining (145), (146), and (147) gives

$$\mathbb{P}(\Phi^S_p \le \Phi_p) \le 4e^{-c'_S n}.$$

As these probabilities are summable, Borel-Cantelli allows us to conclude that $\Phi^S_p > \Phi_p$ eventually almost surely. Thus, no minimizer of P_p in $\mathcal{C}^w_{p,M}$ satisfies $\beta_* + w \in S_p$ and since all primary minimizers are eventually contained in $\mathcal{C}^w_{p,M}$, (136) follows.

It remains to prove (137). By Lemma 18*(iii)*,

$$V_p = a^*_{p,\mathrm{raw}} \to \ell_\kappa \qquad \text{almost surely}$$

Moreover, $0 \le V_p \le A^{\mathrm{raw}}_p(\beta_*)$ and the latter variables have uniformly bounded second moments. Hence, $(V_p)_p$ is uniformly integrable and $\mathbb{E}V_p \to \ell_\kappa$. Gaussian concentration then implies that, for any $t > 0$,

$$\mathbb{P}(|V_p - \ell_\kappa| > t) \le 2e^{-c_t n},$$

for all sufficiently large p. Hence, using (132)-(133) (cf. (Thrampoulidis et al., 2015b, Theorem II.1*(i)-(ii)*)), we conclude that

$$\mathbb{P}(|\Phi_p - \ell_\kappa| > t) \le 4e^{-c_t n}.$$

Another application of Borel-Cantelli then implies that

$$\Phi_p \to \ell_\kappa \qquad \text{almost surely.}$$

Finally, all primary minimizers are eventually contained in $\mathcal{C}^w_{p,M}$, so

$$\Phi_p = \inf_{w \in \mathbb{R}^p} P_p(w)$$

eventually almost surely, which concludes the argument. □

B.7 Completing the proof

Proof. Set $w = b - \beta_*$, so that $y - Xb = \varepsilon - Xw$. Since $X = G_X/\sqrt{n}$, dividing the objective in (2) by $\sqrt{n}$ gives

$$\frac{1}{\sqrt{n}}\left\|\varepsilon - \frac{1}{\sqrt{n}} G_X w\right\|_2 + \frac{\kappa}{\delta_p}\|\beta_* + w\|_{r,p}.$$

As $\|a\|_2 = \max_{\|u\|_2 \le 1} u^\top a$, the preceding objective equals

$$P_p(w) := \max_{\|u\|_2 \le 1} \left\{ \frac{1}{\sqrt{n}} u^\top \varepsilon - \frac{1}{n} u^\top G_X w + \frac{\kappa}{\delta_p}\|\beta_* + w\|_{r,p} \right\}.$$

Thus, b minimizes the objective in (2) if and only if $w = b - \beta_*$ minimizes P_p.

We next identify the auxiliary optimization appearing in Lemmas 18-19. Conditionally on ε, the CGMT auxiliary problem associated with the bilinear term in P_p is

$$\min_w \max_{\|u\|_2 \le 1} \left\{ \frac{\|w\|_2}{n} g^\top u - \frac{\|u\|_2}{n} h^\top w + \frac{1}{\sqrt{n}} u^\top \varepsilon + \frac{\kappa}{\delta_p}\|\beta_* + w\|_{r,p} \right\}.$$

For fixed w, the same radial maximization used in the proof of Lemma 19 gives

$$\max_{\|u\|_2 \le 1} \left\{ u^\top \left(\frac{\varepsilon}{\sqrt{n}} + \frac{\|w\|_2}{n} g \right) - \frac{h^\top w}{n}\|u\|_2 \right\} = \left[\left\| \frac{\varepsilon}{\sqrt{n}} + \frac{\|w\|_2}{n} g \right\|_2 - \frac{h^\top w}{n} \right]_+ .$$

Consequently, the scalar auxiliary objective is

$$A_p^{\mathrm{raw}}(b) = \left[\left\| \frac{\varepsilon}{\sqrt{n}} + \frac{\|b-\beta_*\|_2}{n} g \right\|_2 - \frac{1}{n} h^\top (b-\beta_*)\right]_+ + \frac{\kappa}{\delta_p}\|b\|_{r,p},$$

which is studied in Lemma 18.

We now transfer the stability conclusions from the auxiliary optimization to the original primary problem. Note that $\mathcal{B}_{p,\varepsilon}$ is a closed set. Indeed, the two moment maps appearing in d_p are continuous, and the coupling that pairs the j-th atom of $\mu_{p,b}$ with the j-th atom of $\mu_{p,b'}$ gives $W_2(\mu_{p,b},\mu_{p,b'}) \le \|b-b'\|_{2,p}$. Thus, $b \mapsto W_2(\mu_{p,b},\mu_\kappa)$ is continuous.

For any $\varepsilon > 0$, Lemma 18*(iv)* gives

$$\liminf_{p\to\infty} \left(\inf_{b\in\mathcal{B}_{p,\varepsilon}} A_p^{\mathrm{raw}}(b) - \inf_{b\in\mathbb{R}^p} A_p^{\mathrm{raw}}(b) \right) > 0 \qquad \text{almost surely.}$$

Thus, by Lemma 19, almost surely, for all sufficiently large p, no primary minimizer w satisfies $\beta_* + w \in \mathcal{B}_{p,\varepsilon}$, which gives

$$\sup_{\widehat{\beta}\in\mathcal{M}_p} d_p(\widehat{\beta}) \longrightarrow 0 \qquad \text{almost surely,} \tag{148}$$

where $\mathcal{M}_p$ denotes the set of minimizers of (2). In particular, uniformly over all choices of minimizer,

$$\|\widehat{\beta}-\beta_*\|_{2,p}^2 \to M_\kappa^{\mathrm{err}} = \tau_\kappa^2 H_r(\tau_\kappa,\alpha_\kappa).$$

Hence, the convergence of the MSE in (16) follows from (14). Furthermore, (148) implies that

$$W_2(\mu_{p,\widehat{\beta}},\mu_\kappa) \longrightarrow 0,$$

which proves the first convergence in (81) of Remark 3. Note that (148) also implies that

$$\|\widehat{\beta}\|_{r,p} \to \|f_\kappa\|_r.$$

Lemma 19 gives

$$\min_w P_p(w) \to \ell_\kappa \qquad \text{almost surely,}$$

where, by (128), $\ell_\kappa = R_\kappa + \frac{\kappa}{\delta}\|f_\kappa\|_r$. For any $\widehat{\beta} \in \mathcal{M}_p$, setting $\widehat{w} = \widehat{\beta} - \beta_*$ gives

$$\min_w P_p(w) = P_p(\widehat{w}) = \frac{1}{\sqrt{n}}\|y - X\widehat{\beta}\|_2 + \frac{\kappa}{\delta_p}\|\widehat{\beta}\|_{r,p}.$$

Since $\delta_p \to \delta$, we conclude that

$$\frac{1}{\sqrt{n}}\|y - X\widehat{\beta}\|_2 \to \ell_\kappa - \frac{\kappa}{\delta}\|f_\kappa\|_r = R_\kappa,$$

which gives the convergence of the training error in (16) and concludes the proof. □

C Proof of Lemma 3

We start by isolating the ingredient used to pass from penalized optimization to exact interpolation. First, we show that every residual can be corrected with a small ℓ_r cost. This implies that, if the penalty on the residual is large enough, the solution of the regularized problem is an exact interpolator.

Lemma 20. *Fix $1 \le r < \infty$ and $\delta \in (0,1)$. There is a deterministic constant $C_{r,\delta} < \infty$ such that, almost surely for all sufficiently large p, for every $z \in \mathbb{R}^n$ there exists $u \in \mathbb{R}^p$ satisfying*

$$Xu = z, \qquad \|u\|_{r,p} \le C_{r,\delta}\frac{\|z\|_2}{\sqrt{n}}. \tag{149}$$

Proof. Let r^* be the conjugate exponent of r and $N_p(u) = \|u\|_{r,p}$. Its dual norm is $N_p^*(a) = p^{1/r}\|a\|_{r^*}$. We claim that there is a deterministic $c_{r,\delta} > 0$ such that, with probability at least $1 - e^{-cp}$,

$$\inf_{\|u\|_2=1} N_p^*(X^\top u) \geq c_{r,\delta}\sqrt{n}. \tag{150}$$

Suppose first that $r^* \geq 2$. Note that $p^{1/r}\|a\|_{r^*} \geq \sqrt{p}\,\|a\|_2$. Thus, Corollary 5.35 of Vershynin (2012) gives

$$\mathbb{P}\big(s_{\min}(\sqrt{n}X^\top) \leq \sqrt{p} - \sqrt{n} - t\big) \leq 2e^{-t^2/2},$$

where $s_{\min}(\cdot)$ denotes the smallest singular value. This proves (150), as $n/p \to \delta < 1$.

It remains to consider $1 < r^* < 2$. For every fixed $u \in \mathbb{S}^{n-1}$ and $0 < a \leq 1$, taking the exponential and applying Markov's inequality gives

$$\mathbb{P}\left(p^{-1/r^*}\|\sqrt{n}X^\top u\|_{r^*} \leq a\right) \leq e^p\left(\mathbb{E}e^{-|G|^{r^*}/a^{r^*}}\right)^p \leq (Ca)^p,$$

where $G \sim N(0,1)$ and $C > 0$ is a deterministic constant depending only on r. Moreover,

$$\sup_{\|u\|_2=1} p^{-1/r^*}\|\sqrt{n}X^\top u\|_{r^*} \leq p^{-1/2}\|\sqrt{n}X^\top\|_{\mathrm{op}},$$

and hence this supremum is at most a deterministic constant K except on an event of probability e^{-cp}. On the complementary event, take an $a/(2K)$-net of $\mathbb{S}^{n-1}$. Its cardinality is at most $(1 + 4K/a)^n$ by Corollary 4.2.13 of Vershynin (2018), so the union bound yields

$$\mathbb{P}\left(\inf_{\|u\|_2=1} p^{-1/r^*}\|\sqrt{n}X^\top u\|_{r^*} < \frac{a}{2}\right) \leq e^{-cp} + (1 + 4K/a)^n(Ca)^p.$$

Since $n/p \to \delta < 1$, the RHS is at most $2e^{-c'p}$ when $a > 0$ is chosen sufficiently small. Thus,

$$N_p^*(X^\top u) = \frac{p^{1/r}}{\sqrt{n}}\|\sqrt{n}X^\top u\|_{r^*} \geq c\frac{p}{\sqrt{n}}\|u\|_2 \geq c_{r,\delta}\sqrt{n}\,\|u\|_2.$$

The failure probabilities are summable, so (150) holds eventually almost surely. Eventually X has full row rank. Convex duality then gives, for every $z \in \mathbb{R}^n$,

$$\inf_{Xu=z} N_p(u) = \sup\left\{u^\top z : N_p^*(X^\top u) \leq 1\right\} \leq \frac{\|z\|_2}{c_{r,\delta}\sqrt{n}},$$

which is the desired result. □

Lemma 21 (Exact penalty for interpolation)**.** *Fix $L > C_{r,\delta}$, where $C_{r,\delta}$ is the deterministic constant of Lemma 20. Almost surely, for all sufficiently large p and every $y \in \mathbb{R}^n$,*

$$\arg\min_{b\in\mathbb{R}^p}\left\{\|b\|_{r,p} + \frac{L}{\sqrt{n}}\|y - Xb\|_2\right\} = \arg\min_{Xb=y}\|b\|_{r,p}. \tag{151}$$

Proof. Let $V_p(y) := \inf_{X\widetilde{b}=y}\|\widetilde{b}\|_{r,p}$. Lemma 20 shows that the feasible set is nonempty; the infimum is attained because it is a finite-dimensional norm minimization over a closed affine set. For arbitrary b, applying Lemma 20 to $z = y - Xb$ gives that there exists u such that $X(b+u) = y$ and

$$\|u\|_{r,p} \leq C_{r,\delta}\frac{\|y - Xb\|_2}{\sqrt{n}}.$$

Therefore

$$\|b\|_{r,p} + \frac{L}{\sqrt{n}}\|y - Xb\|_2 \geq V_p(y) + \frac{L - C_{r,\delta}}{\sqrt{n}}\|y - Xb\|_2.$$

Since $L > C_{r,\delta}$, equality is possible precisely when $Xb = y$ and $\|b\|_{r,p} = V_p(y)$. □

We next extend the results of Lemma 19 by allowing the loss and regularizer to have independent coefficients.

Lemma 22 (Extension of Lemma 19)**.** *Let $\rho_p, \lambda_p > 0$ be deterministic sequences such that $\rho_p \to \rho \in (0,\infty)$ and $\lambda_p \to \lambda \in (0,\infty)$. Fix $\kappa_0 > 0$ such that $\gamma_r(\kappa_0) < \sqrt{\delta}$, and suppose that, eventually,*

$$\lambda_p \geq \frac{\delta_p \rho_p}{\kappa_0}. \tag{152}$$

Define

$$P_p^{\rho,\lambda}(w) := \rho_p \|\beta_* + w\|_{r,p} + \frac{\lambda_p}{\sqrt{n}} \left\| \varepsilon - \frac{1}{\sqrt{n}} G_X w \right\|_2, \tag{153}$$

$$A_p^{\rho,\lambda,\mathrm{raw}}(b) := \rho_p \|b\|_{r,p} + \lambda_p [C_p^{\mathrm{raw}}(b)]_+. \tag{154}$$

For every p, let $S_p \subseteq \mathbb{R}^p$ be deterministic and closed. Suppose that, for some deterministic $\ell < \infty$,

$$\inf_b A_p^{\rho,\lambda,\mathrm{raw}}(b) \to \ell \qquad \textit{almost surely}, \tag{155}$$

and

$$\liminf_{p\to\infty} \left(\inf_{b \in S_p} A_p^{\rho,\lambda,\mathrm{raw}}(b) - \inf_b A_p^{\rho,\lambda,\mathrm{raw}}(b) \right) > 0 \qquad \textit{almost surely}. \tag{156}$$

Then, almost surely, for all sufficiently large p,

$$\arg\min_{w \in \mathbb{R}^p} P_p^{\rho,\lambda}(w) \cap \{w : \beta_* + w \in S_p\} = \varnothing. \tag{157}$$

Moreover,

$$\inf_w P_p^{\rho,\lambda}(w) \to \ell \qquad \textit{almost surely}. \tag{158}$$

Proof. By the dual representation of the Euclidean norm,

$$P_p^{\rho,\lambda}(w) = \max_{\|u\|_2 \leq 1} \left\{ \frac{\lambda_p}{\sqrt{n}} u^\top \varepsilon - \frac{\lambda_p}{n} u^\top G_X w + \rho_p \|\beta_* + w\|_{r,p} \right\}.$$

The associated CGMT auxiliary problem is obtained exactly as in the proof of Lemma 19. Maximizing first over the direction and then over the norm of u shows that its scalar objective is

$$A_p^{\rho,\lambda,\mathrm{raw}}(\beta_* + w) = \rho_p \|\beta_* + w\|_{r,p} + \lambda_p [C_p^{\mathrm{raw}}(\beta_* + w)]_+.$$

We first verify the localization needed to apply the CGMT on deterministic compact sets. Define

$$R_p^{\mathrm{AO},\rho,\lambda} := \inf_{\|v\|_{2,p}=1} \left\{ \rho_p \|v\|_{r,p} + \lambda_p \left[\frac{\|g\|_2}{\sqrt{n}\sqrt{\delta_p}} - \frac{1}{\delta_p} \langle h, v \rangle_p \right]_+ \right\}.$$

Set $\nu_p := \frac{\delta_p \rho_p}{\kappa_0}$. Since $\lambda_p \geq \nu_p$ eventually and $\lambda_p [x]_+ \geq \nu_p x$ for every $x \in \mathbb{R}$, we obtain

$$R_p^{\mathrm{AO},\rho,\lambda} \geq \frac{\rho_p}{\kappa_0} \inf_{\|v\|_{2,p}=1} \left\{ \frac{\|g\|_2}{\sqrt{n}} \sqrt{\delta_p} - \langle h, v \rangle_p + \kappa_0 \|v\|_{r,p} \right\} \geq \frac{\rho_p}{\kappa_0} \left(\frac{\|g\|_2}{\sqrt{n}} \sqrt{\delta_p} - \gamma_{r,p}(\kappa_0) \right).$$

By Lemma 16, the last expression converges almost surely and in L_1 to $c_* := \frac{\rho}{\kappa_0} \{ \sqrt{\delta} - \gamma_r(\kappa_0) \} > 0$. Moreover, since λ_p is bounded, $R_p^{\mathrm{AO},\rho,\lambda}$ is a $C/\sqrt{p}$-Lipschitz function of (g,h). Gaussian concentration and an application of (Thrampoulidis et al., 2015b, Theorem II.1*(i)*) as in (139)–(141) therefore gives that, almost surely eventually,

$$(P_p^{\rho,\lambda})^\infty(v) \geq c \|v\|_{2,p}, \tag{159}$$

for $v \in \mathbb{R}^p$ and a deterministic $c > 0$. The triangle inequality then gives

$$P_p^{\rho,\lambda}(w) \geq c \|w\|_{2,p} - \rho_p \|\beta_*\|_{r,p} - \frac{\lambda_p}{\sqrt{n}} \|\varepsilon\|_2.$$

Since the last two terms are eventually bounded and $P_p^{\rho,\lambda}(0) = O(1)$, all primary minimizers have bounded $\|w\|_{2,p}$. Furthermore,

$$P_p^{\rho,\lambda}(w) \geq \rho_p \|\beta_* + w\|_{r,p},$$

and ρ_p is bounded away from zero, so their $\|\beta_* + w\|_{r,p}$ norms are bounded as well.

The minimizers of $A_p^{\rho,\lambda,\mathrm{raw}}$ satisfy the same localization. Indeed, with $a_n := \|g\|_2/\sqrt{n}$, the reverse triangle inequality gives, almost surely eventually,

$$A_p^{\rho,\lambda,\mathrm{raw}}(\beta_* + w) \geq \frac{\rho_p}{\kappa_0}\{a_n\sqrt{\delta_p}\|w\|_{2,p} - \langle h, w\rangle_p + \kappa_0\|w\|_{r,p}\} - C,$$

for a deterministic $C < \infty$. By homogeneity and the definition of $\gamma_{r,p}(\kappa_0)$, the expression in braces is at least $\{a_n\sqrt{\delta_p} - \gamma_{r,p}(\kappa_0)\}\|w\|_{2,p}$. As $a_n\sqrt{\delta_p} - \gamma_{r,p}(\kappa_0)$ converges to $\sqrt{\delta} - \gamma_r(\kappa_0) > 0$, we conclude that

$$A_p^{\rho,\lambda,\mathrm{raw}}(\beta_* + w) \geq c\|w\|_{2,p} - C$$

eventually. Since also $A_p^{\rho,\lambda,\mathrm{raw}}(b) \geq \rho_p\|b\|_{r,p}$, the minimizers of $A_p^{\rho,\lambda,\mathrm{raw}}$ are localized as claimed.

Consequently, there is a deterministic $M < \infty$ such that all minimizers of $P_p^{\rho,\lambda}$ and $A_p^{\rho,\lambda,\mathrm{raw}}$ eventually belong to

$$\mathcal{C}_{p,M}^w := \{w : \|w\|_{2,p} \leq M,\ \|\beta_* + w\|_{r,p} \leq M\}.$$

By localization and (156),

$$\liminf_{p\to\infty}\Big(\inf_{\substack{w\in\mathcal{C}_{p,M}^w\\ \beta_*+w\in S_p}} A_p^{\rho,\lambda,\mathrm{raw}}(\beta_* + w) - \inf_{w\in\mathcal{C}_{p,M}^w} A_p^{\rho,\lambda,\mathrm{raw}}(\beta_* + w)\Big) > 0 \qquad \text{almost surely.}$$

The argument in the proof of Lemma 19 starting from (143) and involving the application of (Thrampoulidis et al., 2015b, Theorem II.1*(ii)*)) now applies without change, giving (157).

It remains to transfer the minimum value. By localization and (155),

$$V_p \to \ell \qquad \text{almost surely.}$$

Moreover,

$$0 \leq V_p \leq \rho_p\|\beta_*\|_{r,p} + \frac{\lambda_p}{\sqrt{n}}\|\varepsilon\|_2,$$

so $(V_p)_p$ is uniformly integrable. Thus, the argument at the end of the proof of Lemma 19 therefore gives

$$\Phi_p \to \ell \qquad \text{almost surely.}$$

Since all minimizers of $P_p^{\rho,\lambda}$ are eventually contained in $\mathcal{C}_{p,M}^w$, $\Phi_p = \inf_w P_p^{\rho,\lambda}(w)$ eventually, which proves (158). □

At this point, we are ready to give the proof of Lemma 3.

Proof of Lemma 3. Let us define

$$\kappa_{\mathrm{int}} := \begin{cases} \alpha_{\mathrm{int}}, & r = 1,\\ \alpha_{\mathrm{int}} U_r(\tau_{\mathrm{int}}, \alpha_{\mathrm{int}})^{(r-1)/r}, & 1 < r < \infty, \end{cases}$$

where U_r is defined in (13). We first verify that Lemma 18 applies at the interpolation parameters. Equation (17) and the definition of $\kappa_{\mathrm{int}} > 0$ above imply (14) with $(\tau_\kappa, \alpha_\kappa, \kappa, f_\kappa) = (\tau_{\mathrm{int}}, \alpha_{\mathrm{int}}, \kappa_{\mathrm{int}}, f_{\mathrm{int}})$. Moreover, the second equation in (17) gives $D_r(\tau_{\mathrm{int}}, \alpha_{\mathrm{int}}) = \delta$, which satisfies the remaining hypothesis of the lemma and implies $R_\kappa = 0$. Moreover, $\kappa_{\mathrm{int}} > 0$.

Let $L_p > 0$ be any deterministic sequence such that

$$L_p \to L \in (0, \infty), \qquad L_p \geq \frac{\delta_p}{\kappa_{\mathrm{int}}} \quad \text{eventually.} \tag{160}$$

Let $c_p := \frac{\delta_p}{\kappa_{\mathrm{int}}}$, and define

$$I_{p,L_p}^{\mathrm{raw}}(b) := \|b\|_{r,p} + L_p[C_p^{\mathrm{raw}}(b)]_+.$$

At interpolation, $R_\kappa = 0$. Therefore, Lemma 18*(iii)–(iv)* gives

$$a_p^* := \inf_b A_p^{\mathrm{raw}}(b) \to \frac{\kappa_{\mathrm{int}}}{\delta}\|f_{\mathrm{int}}\|_r \qquad \text{almost surely,} \tag{161}$$

and, for every $\eta > 0$,

$$\liminf_{p\to\infty}\left(\inf_{b\in\mathcal{B}_{p,\eta}} A_p^{\mathrm{raw}}(b) - a_p^*\right) > 0 \qquad \text{almost surely.} \tag{162}$$

Here and below, d_p and $\mathcal{B}_{p,\eta}$ (defined respectively in (108) and (109)) are evaluated at interpolation, i.e., under the substitution $(\tau_\kappa, \alpha_\kappa, \kappa, f_\kappa) = (\tau_{\mathrm{int}}, \alpha_{\mathrm{int}}, \kappa_{\mathrm{int}}, f_{\mathrm{int}})$.

Since $L_p \geq c_p$, we have

$$I_{p,L_p}^{\mathrm{raw}}(b) \geq c_p A_p^{\mathrm{raw}}(b), \tag{163}$$

for $b \in \mathbb{R}^p$. Let $b_{p,j}^\circ := \tau_{\mathrm{int}}\eta_{\alpha_{\mathrm{int}},r}\left(\frac{(\beta_*)_j}{\tau_{\mathrm{int}}} + h_j\right)$. Lemma 14, (17) and Gaussian integration by parts give

$$C_p(b_p^\circ) \longrightarrow 0, \qquad \|b_p^\circ - \beta_*\|_{2,p} = O(1), \qquad \|b_p^\circ\|_{r,p} \to \|f_{\mathrm{int}}\|_r.$$

Equation (125) therefore gives

$$[C_p^{\mathrm{raw}}(b_p^\circ)]_+ \longrightarrow 0.$$

Since L_p is bounded,

$$I_{p,L_p}^{\mathrm{raw}}(b_p^\circ) \to \|f_{\mathrm{int}}\|_r.$$

On the other hand, (163) and (161) give

$$\liminf_{p\to\infty}\inf_b I_{p,L_p}^{\mathrm{raw}}(b) \geq \lim_{p\to\infty} c_p a_p^* = \|f_{\mathrm{int}}\|_r.$$

Consequently,

$$i_p^* := \inf_b I_{p,L_p}^{\mathrm{raw}}(b) \to \|f_{\mathrm{int}}\|_r \qquad \text{almost surely,} \tag{164}$$

and

$$\inf_{b\in\mathcal{B}_{p,\eta}} I_{p,L_p}^{\mathrm{raw}}(b) - i_p^* \geq c_p\left(\inf_{b\in\mathcal{B}_{p,\eta}} A_p^{\mathrm{raw}}(b) - a_p^*\right) + c_p a_p^* - i_p^*.$$

The last two terms differ by $o(1)$ by (161) and (164). Since $c_p \to \delta/\kappa_{\mathrm{int}} > 0$, (162) gives

$$\liminf_{p\to\infty}\left(\inf_{b\in\mathcal{B}_{p,\eta}} I_{p,L_p}^{\mathrm{raw}}(b) - i_p^*\right) > 0 \qquad \text{almost surely.} \tag{165}$$

Letting

$$P_{p,L_p}(w) := \|\beta_* + w\|_{r,p} + \frac{L_p}{\sqrt{n}}\left\|\varepsilon - \frac{1}{\sqrt{n}}G_X w\right\|_2, \tag{166}$$

an application of Lemma 22 gives

$$\inf_w P_{p,L_p}(w) \to \|f_{\mathrm{int}}\|_r \qquad \text{almost surely.} \tag{167}$$

and

$$\sup_{\widehat{w}\in\arg\min_w P_{p,L_p}(w)} d_p(\beta_* + \widehat{w}) \longrightarrow 0 \qquad \text{almost surely.} \tag{168}$$

To conclude, choose $L > \max\left\{\frac{\delta}{\kappa_{\mathrm{int}}}, C_{r,\delta}\right\}$. Since $\delta_p/\kappa_{\mathrm{int}} \to \delta/\kappa_{\mathrm{int}}$, the constant sequence $L_p = L$ satisfies (160). Hence (168) applies to the minimizers of

$$b \longmapsto \|b\|_{r,p} + \frac{L}{\sqrt{n}}\|y - Xb\|_2.$$

By Lemma 21, these minimizers are exactly the minimum-ℓ_r interpolators. Therefore, uniformly over all such interpolators,

$$d_p(\widehat{\beta}_r) \longrightarrow 0 \qquad \text{almost surely.}$$

In particular, the second convergence in (81) of Remark 3 holds, and

$$\|\widehat{\beta}_r - \beta_*\|_{2,p}^2 \to \tau_{\mathrm{int}}^2 H_r(\tau_{\mathrm{int}}, \alpha_{\mathrm{int}}),$$

Finally, (17) gives

$$\tau_{\mathrm{int}}^2 H_r(\tau_{\mathrm{int}}, \alpha_{\mathrm{int}}) = \delta(\tau_{\mathrm{int}}^2 - \sigma^2),$$

which proves (18). □

D PROOFS FOR SECTION 4.3

D.1 AUXILIARY RESULTS

Lemma 23 (Monotonicity of the residual)**.** *Let $1 \le r < \infty$. For each finite (n, p) and $\kappa \ge 0$, let $\widehat{\beta}_{\kappa,r}$ be any minimizer in (2). If $0 \le \kappa_1 < \kappa_2$, then, for any choices of minimizers,*

$$\|y - X\widehat{\beta}_{\kappa_1,r}\|_2 \le \|y - X\widehat{\beta}_{\kappa_2,r}\|_2.$$

Proof. For $i = 1, 2$, write

$$L_i := \|y - X\widehat{\beta}_{\kappa_i,r}\|_2, \qquad P_i := \frac{p}{\sqrt{n}}\|\widehat{\beta}_{\kappa_i,r}\|_{r,p}.$$

Optimality at κ_1 and κ_2 gives

$$L_1 + \kappa_1 P_1 \le L_2 + \kappa_1 P_2, \qquad L_2 + \kappa_2 P_2 \le L_1 + \kappa_2 P_1.$$

Equivalently,

$$\kappa_2(P_2 - P_1) \le L_1 - L_2 \le \kappa_1(P_2 - P_1).$$

Since $\kappa_2 > \kappa_1$, these inequalities imply $P_2 \le P_1$, and hence

$$L_1 - L_2 \le \kappa_1(P_2 - P_1) \le 0.$$

Thus $L_1 \le L_2$. Dividing by $\sqrt{n}$ and passing to the limit proves the final assertion. □

Lemma 24 (Limit towards interpolation)**.** *Suppose that $0 < \delta < 1$, $1 < r < 2$, $\sigma \ge 0$, and $\mathbb{E}B^2 < \infty$. If $\sigma = 0$, assume that the fixed point equations in (17) admit a solution in $(0, \infty)^2$. Then, these equations admit a unique solution $(\tau_{\mathrm{int}}, \alpha_{\mathrm{int}})$, and*

$$\kappa_r^\star = \kappa_{\mathrm{int}}, \qquad (\tau_\kappa, \alpha_\kappa) \to (\tau_{\mathrm{int}}, \alpha_{\mathrm{int}}) \quad \text{as } \kappa \downarrow \kappa_r^\star$$

along any non-interpolating set of solutions. Moreover, $R_\kappa \to 0$ as $\kappa \downarrow \kappa_r^\star$.

Proof. We first prove existence when $\sigma > 0$. For every $\tau \ge \sigma$, the map $\alpha \mapsto D_r(\tau, \alpha)$ is continuous and strictly decreasing from 1 to 0. Indeed, if $z \ne 0$, $u = \eta_{\alpha,r}(z)$, and $t = \alpha|u|^{r-2}$, then

$$\eta'_{\alpha,r}(z) = \frac{1}{1 + (r-1)t}, \qquad \frac{d\log t}{d\log \alpha} = \frac{1+t}{1+(r-1)t} > 0.$$

Thus, for every $z \ne 0$, the map $\alpha \mapsto \eta'_{\alpha,r}(z)$ is continuous and strictly decreasing from 1 to 0. Since, conditionally on B, the variable $B/\tau + G$ has a continuous law, dominated convergence gives that $D_r(\tau, \alpha)$ is jointly continuous in (τ, α) and satisfies

$$\lim_{\alpha \downarrow 0} D_r(\tau, \alpha) = 1, \qquad \lim_{\alpha \to \infty} D_r(\tau, \alpha) = 0.$$

Consequently, there is a unique continuous function $\alpha(\tau) > 0$ such that

$$D_r(\tau, \alpha(\tau)) = \delta.$$

Set

$$g(\tau) := H_r(\tau, \alpha(\tau)) - \delta\left(1 - \frac{\sigma^2}{\tau^2}\right), \qquad \tau \ge \sigma.$$

We have

$$g(\sigma) = H_r(\sigma, \alpha(\sigma)) > 0.$$

Indeed, equality would imply $\eta_{\alpha(\sigma),r}(B/\sigma + G) = B/\sigma$ almost surely. By (82), this would give $G = \alpha(\sigma)\psi_r(B/\sigma)$ almost surely, which gives a contradiction. We next determine the limit of g at infinity. If $\tau_j \to \infty$, the identity $D_r(\tau_j, \alpha(\tau_j)) = \delta$ prevents $\alpha(\tau_j)$ from having a subsequence tending to either 0 or ∞. Every subsequential limit $\bar{\alpha} \in (0, \infty)$ satisfies

$$\mathbb{E}\eta'_{\bar{\alpha},r}(G) = \delta.$$

This equation has a unique solution, denoted by α_0, and therefore $\alpha(\tau) \to \alpha_0$. Lemma 11 and dominated convergence then give

$$H_r(\tau, \alpha(\tau)) \to \mathbb{E}\eta_{\alpha_0,r}(G)^2.$$

Writing $u_0 = \eta_{\alpha_0,r}(G)$, Gaussian integration by parts and (82) yield

$$\delta = \mathbb{E}[Gu_0] = \mathbb{E}u_0^2 + \alpha_0 \mathbb{E}|u_0|^r.$$

Consequently,

$$\lim_{\tau\to\infty} g(\tau) = -\alpha_0 \mathbb{E}|u_0|^r < 0.$$

The intermediate-value theorem then gives the existence of a solution with $\tau_{\text{int}} > \sigma$.

We now move to proving uniqueness. On the probability space supporting (B, G), define

$$\mathcal{C}(f) := \left(\sigma^2 + \frac{1}{\delta}\mathbb{E}(f - B)^2\right)^{1/2} - \frac{1}{\delta}\mathbb{E}[G(f - B)], \tag{169}$$

$$\mathcal{A}_\kappa(f) := [\mathcal{C}(f)]_+ + \frac{\kappa}{\delta}\|f\|_r. \tag{170}$$

Since $1 < r < 2$, these functionals are well defined on L_2. Let $(\tau, \alpha) \in (0, \infty)^2$ satisfy (17) and set

$$f := \tau\eta_{\alpha,r}(B/\tau + G), \qquad \kappa := \alpha U_r(\tau, \alpha)^{(r-1)/r}.$$

Then,

$$\mathcal{C}(f) = 0, \qquad \|f - B\|_2^2 = \delta(\tau^2 - \sigma^2).$$

Moreover, (17) and the definition of κ, are precisely the fixed point equations in (14) with $D_r(\tau, \alpha) = \delta$. Hence Lemma 13 shows that f is the unique minimizer of

$$h \longmapsto \mathcal{C}(h) + \frac{\kappa}{\delta}\|h\|_r. \tag{171}$$

The optimality calculation in the proof of that lemma also gives

$$0 = \nabla\mathcal{C}(f) + \frac{\kappa}{\delta}\nabla\|f\|_r. \tag{172}$$

If $\mathcal{C}(h) \le 0$, optimality in (171) yields

$$\frac{\kappa}{\delta}\|h\|_r \ge \mathcal{C}(h) + \frac{\kappa}{\delta}\|h\|_r \ge \frac{\kappa}{\delta}\|f\|_r,$$

with equality only for $h = f$, thus proving uniqueness of (τ, α), as well as of κ_{int}.

Let $(\tau_{\text{int}}, \alpha_{\text{int}})$ denote this unique solution, and write $f_{\text{int}} := \tau_{\text{int}}\eta_{\alpha_{\text{int}},r}(B/\tau_{\text{int}} + G)$. We now study the minimizers of $\mathcal{A}_\kappa$, beginning with their existence for every $\kappa > 0$. Lemma 15, applied at $(\tau_{\text{int}}, \alpha_{\text{int}}, \kappa_{\text{int}})$, gives

$$\gamma_r(\kappa_{\text{int}}) < \sqrt{\delta}. \tag{173}$$

Since $[\mathcal{C}]_+ \ge \mathcal{C}$, the definition of γ_r, its monotonicity, and $\mathbb{E}[GB] = 0$ imply, for every $\kappa \ge \kappa_{\text{int}}$,

$$\mathcal{A}_\kappa(f) \ge \frac{\sqrt{\delta} - \gamma_r(\kappa_{\text{int}})}{\delta}\|f\|_2 - \frac{\|B\|_2}{\sqrt{\delta}}. \tag{174}$$

Moreover, since both terms defining $\mathcal{A}_\kappa$ are nonnegative, for every $\kappa > 0$,

$$\mathcal{A}_\kappa(f) \ge \min\left(1, \frac{\kappa}{\kappa_{\text{int}}}\right)\mathcal{A}_{\kappa_{\text{int}}}(f).$$

Thus, for each fixed $\kappa > 0$, $\mathcal{A}_\kappa(f) \to \infty$ as $\|f\|_2 \to \infty$. In particular, every sequence with objective values bounded above is bounded in L_2. Since $\mathcal{A}_\kappa$ is also proper, convex, and weakly lower semicontinuous, it attains its minimum.

We next identify these minimizers. The previous argument shows that f_{int} is the unique minimizer of $\|f\|_r$ over $\{f \in L_2 : \mathcal{C}(f) \le 0\}$. For $0 < \kappa \le \kappa_{\text{int}}$, we claim that f_{int} also uniquely minimizes $\mathcal{A}_\kappa$. Indeed, if $\mathcal{C}(f) \ge 0$, then

$$\frac{\delta}{\kappa}\mathcal{A}_\kappa(f) = \|f\|_r + \frac{\delta}{\kappa}\mathcal{C}(f) \ge \|f\|_r + \frac{\delta}{\kappa_{\text{int}}}\mathcal{C}(f) \ge \|f_{\text{int}}\|_r,$$

where the last inequality follows from the optimality of f_{int} in (171) with $\kappa = \kappa_{\text{int}}$. If $\mathcal{C}(f) \le 0$, optimality of f_{int} in (171) at $\kappa = \kappa_{\text{int}}$ gives

$$\frac{\kappa_{\text{int}}}{\delta}\|f\|_r \ge \mathcal{C}(f) + \frac{\kappa_{\text{int}}}{\delta}\|f\|_r \ge \frac{\kappa_{\text{int}}}{\delta}\|f_{\text{int}}\|_r.$$

Since $\mathcal{A}_\kappa(f) = (\kappa/\delta)\|f\|_r$ in this case, the same bound follows. Moreover, $\mathcal{C}(f_{\text{int}}) = 0$, so the bound is attained at f_{int}. In either case, equality forces $f = f_{\text{int}}$ by uniqueness of the minimizer in (171) at $\kappa = \kappa_{\text{int}}$.

Now, let $\kappa > \kappa_{\text{int}}$. We first show that f_{int} cannot minimize $\mathcal{A}_\kappa$. Since $\mathcal{C}(f_{\text{int}}) = 0$,

$$\partial[\mathcal{C}]_+(f_{\text{int}}) = \{\theta \nabla \mathcal{C}(f_{\text{int}}) : 0 \le \theta \le 1\}.$$

Together with (172), the first-order optimality condition would therefore require $\kappa = \theta\kappa_{\text{int}}$ for some $\theta \in [0,1]$, which is impossible. Furthermore, if a minimizer f_κ of $\mathcal{A}_\kappa$ satisfied $\mathcal{C}(f_\kappa) \le 0$, comparison with f_{int} would give $\|f_\kappa\|_r \le \|f_{\text{int}}\|_r$ and the minimality of f_{int} would then imply $f_\kappa = f_{\text{int}}$, a contradiction. Consequently, every minimizer satisfies

$$\mathcal{C}(f_\kappa) > 0, \qquad \kappa > \kappa_{\text{int}}. \tag{175}$$

We now prove that these minimizers converge to f_{int} as $\kappa \downarrow \kappa_{\text{int}}$. For any minimizer f_κ,

$$\mathcal{A}_\kappa(f_\kappa) \le \mathcal{A}_\kappa(f_{\text{int}}) = \frac{\kappa}{\delta}\|f_{\text{int}}\|_r.$$

Together with (174), this shows that all minimizers remain uniformly bounded in L_2 as $\kappa \downarrow \kappa_{\text{int}}$. Let $\kappa_j \downarrow \kappa_{\text{int}}$ and choose $f_j \in \arg\min \mathcal{A}_{\kappa_j}$. On bounded subsets of L_2,

$$\left|\mathcal{A}_{\kappa_j}(h) - \mathcal{A}_{\kappa_{\text{int}}}(h)\right| \le \frac{|\kappa_j - \kappa_{\text{int}}|}{\delta}\|h\|_2 \longrightarrow 0.$$

Weak compactness, lower semicontinuity, and uniqueness of the minimizer of $\mathcal{A}_{\kappa_{\text{int}}}$ therefore show that f_j converges weakly to f_{int} in L_2. Since f_j minimizes $\mathcal{A}_{\kappa_j}$, comparison with f_{int} and (175) give

$$0 < \mathcal{C}(f_j) \le \frac{\kappa_j}{\delta}\left(\|f_{\text{int}}\|_r - \|f_j\|_r\right).$$

In particular, $\|f_j\|_r \le \|f_{\text{int}}\|_r$. On the other hand, weak lower semicontinuity gives $\|f_{\text{int}}\|_r \le \liminf_j \|f_j\|_r$. Therefore,

$$\|f_j\|_r \to \|f_{\text{int}}\|_r, \qquad \mathcal{C}(f_j) \to 0.$$

Furthermore, weak convergence implies

$$\mathbb{E}[G(f_j - B)] \to \mathbb{E}[G(f_{\text{int}} - B)].$$

Using (169), $\mathcal{C}(f_{\text{int}}) = 0$ and (17), we obtain

$$\left(\sigma^2 + \frac{1}{\delta}\|f_j - B\|_2^2\right)^{1/2} = \mathcal{C}(f_j) + \frac{1}{\delta}\mathbb{E}[G(f_j - B)] \to \tau_{\text{int}}.$$

It follows that

$$\|f_j - B\|_2 \to \|f_{\text{int}} - B\|_2,$$

and hence

$$\|f_j - f_{\text{int}}\|_2 \to 0. \tag{176}$$

We next show that these minimizers determine positive solutions of (14) with $R_\kappa > 0$. For all sufficiently large j, define

$$\tau_j := \left(\sigma^2 + \frac{1}{\delta}\mathbb{E}(f_j - B)^2\right)^{1/2}, \qquad \alpha_j := \kappa_j\left(\frac{\tau_j}{\|f_j\|_r}\right)^{r-1}. \tag{177}$$

These quantities are positive because $\tau_j \to \tau_{\text{int}} > 0$ and $\|f_j\|_r \to \|f_{\text{int}}\|_r > 0$. Since $\mathcal{C}(f_j) > 0$, the first-order optimality condition for $\mathcal{A}_{\kappa_j}$ is

$$\frac{f_j - B}{\tau_j} - G + \kappa_j\|f_j\|_r^{1-r}\psi_r(f_j) = 0.$$

Equivalently,

$$\frac{B}{\tau_j} + G = \frac{f_j}{\tau_j} + \alpha_j \psi_r\left(\frac{f_j}{\tau_j}\right).$$

By (82),

$$f_j = \tau_j \eta_{\alpha_j, r}(B/\tau_j + G).$$

Substituting this expression into (177) gives

$$\tau_j^2 = \sigma^2 + \frac{\tau_j^2}{\delta} H_r(\tau_j, \alpha_j), \qquad \kappa_j = \alpha_j U_r(\tau_j, \alpha_j)^{(r-1)/r}.$$

Thus (τ_j, α_j) satisfies (14). Gaussian integration by parts also gives $\mathbb{E}[G(f_j - B)] = \tau_j D_r(\tau_j, \alpha_j)$, and hence

$$R_{\kappa_j} = \tau_j \left(1 - \frac{D_r(\tau_j, \alpha_j)}{\delta}\right) = \mathcal{C}(f_j) > 0.$$

Finally, (177) and the definition of κ_{int} imply

$$\tau_j \to \tau_{\text{int}}, \qquad \alpha_j \to \kappa_{\text{int}} \left(\frac{\tau_{\text{int}}}{\|f_{\text{int}}\|_r}\right)^{r-1} = \alpha_{\text{int}}, \qquad R_{\kappa_j} \to 0.$$

This argument applies to every sequence $\kappa_j \downarrow \kappa_{\text{int}}$ and every choice of minimizers f_j. In particular, the fixed point equations in (14) admit a positive solution with $R_\kappa > 0$ for every $\kappa > \kappa_{\text{int}}$ sufficiently close to κ_{int}.

We now show that the same convergence holds for every positive solution $(\tau_\kappa, \alpha_\kappa)$ of (14) with $R_\kappa > 0$. Set

$$f_\kappa := \tau_\kappa \eta_{\alpha_\kappa, r}\left(B/\tau_\kappa + G\right).$$

Gaussian integration by parts and (14) give $\mathcal{C}(f_\kappa) = R_\kappa > 0$. In particular, $D_r(\tau_\kappa, \alpha_\kappa) < \delta$, so Lemma 13 applies and yields, for every $f \in L_2$,

$$\mathcal{A}_\kappa(f) \geq \mathcal{C}(f) + \frac{\kappa}{\delta}\|f\|_r \geq \mathcal{C}(f_\kappa) + \frac{\kappa}{\delta}\|f_\kappa\|_r = \mathcal{A}_\kappa(f_\kappa).$$

Thus f_κ minimizes $\mathcal{A}_\kappa$. The convergence proved for arbitrary sequences of minimizers therefore applies to these random variables as well. Moreover, (14) expresses τ_κ and α_κ in terms of f_κ by the same formulas as (177). Consequently, for every such choice of solutions,

$$(\tau_\kappa, \alpha_\kappa) \to (\tau_{\text{int}}, \alpha_{\text{int}}), \qquad R_\kappa \to 0, \qquad \kappa \downarrow \kappa_{\text{int}}.$$

It remains to prove that κ_{int} equals $\kappa_r^\star$ as defined in (19). Set $L_p := \delta_p/\kappa_{\text{int}}$. Since $L_p \to \delta/\kappa_{\text{int}} > 0$, this choice satisfies (160). Recall the definition of $P_{p,L_p}(w)$ in (166). Then, (167)-(168) give, almost surely,

$$\inf_w P_{p,L_p}(w) \to \|f_{\text{int}}\|_r, \qquad \sup_{\widehat{w} \in \arg\min_w P_{p,L_p}(w)} \left|\|\beta_* + \widehat{w}\|_{r,p} - \|f_{\text{int}}\|_r\right| \to 0.$$

Note that

$$P_{p,L_p}(b - \beta_*) = \frac{\delta_p}{\kappa_{\text{int}}\sqrt{n}}\left(\|y - Xb\|_2 + \frac{p\kappa_{\text{int}}}{\sqrt{n}}\|b\|_{r,p}\right).$$

Thus $\widehat{w}$ minimizes P_{p,L_p} if and only if $\widehat{\beta} = \beta_* + \widehat{w}$ minimizes (2) at $\kappa = \kappa_{\text{int}}$. For every such $\widehat{\beta}$,

$$\frac{1}{\sqrt{n}}\|y - X\widehat{\beta}\|_2 = \frac{1}{L_p}\left(\inf_w P_{p,L_p}(w) - \|\widehat{\beta}\|_{r,p}\right) \to 0 \qquad \text{almost surely.}$$

Lemma 23 then gives, for every fixed $0 \leq \kappa \leq \kappa_{\text{int}}$ and every choice of minimizer in (2),

$$\frac{1}{\sqrt{n}}\|y - X\widehat{\beta}_{\kappa,r}\|_2 \to 0 \qquad \text{almost surely.}$$

Conversely, fix any $\kappa > \kappa_{\text{int}}$. Choose $\widetilde{\kappa} \in (\kappa_{\text{int}}, \kappa)$ close enough to κ_{int} that (14) admits a positive solution with $R_{\widetilde{\kappa}} > 0$, as established above. Lemma 2, specifically (16), then gives

$$\frac{1}{\sqrt{n}}\|y - X\widehat{\beta}_{\widetilde{\kappa},r}\|_2 \to R_{\widetilde{\kappa}} > 0 \qquad \text{almost surely.}$$

By Lemma 23,

$$\liminf_{p\to\infty} \frac{1}{\sqrt{n}}\|y - X\widehat{\beta}_{\kappa,r}\|_2 \geq R_{\tilde{\kappa}} > 0 \qquad \text{almost surely.}$$

Therefore $\kappa_r^\star = \kappa_{\text{int}}$. Together with the convergence of $(\tau_\kappa, \alpha_\kappa)$ and R_κ proved above, this completes the proof. □

Lemma 25 (Derivative at the interpolation threshold)**.** *Suppose that $\mathbb{E}B^2 < \infty$, $\delta \in (0,1)$, $\sigma \geq 0$, and $1 < r < 2$. If $\sigma = 0$, assume in addition that $\mathbb{P}(B \neq 0) > 0$ and that (17) admits a solution $(\tau_{\text{int}}, \alpha_{\text{int}}) \in (0,\infty)^2$. Then the limit in (20) exists, and it is given by*

$$S_r = \frac{2\delta^2 \partial_\alpha H_r}{\left(\partial_\tau H_r - \dfrac{2\delta\sigma^2}{\tau_{\text{int}}^3}\right)\partial_\alpha D_r - (\partial_\alpha H_r)(\partial_\tau D_r)}. \tag{178}$$

Here H_r, D_r, and all their partial derivatives are evaluated at $(\tau_{\text{int}}, \alpha_{\text{int}})$.

Proof. By Lemma 24,

$$\kappa_r^\star = \kappa_{\text{int}}, \qquad (\tau_\kappa, \alpha_\kappa) \to (\tau_{\text{int}}, \alpha_{\text{int}}) \quad \text{as } \kappa \downarrow \kappa_r^\star.$$

We can therefore compute S_r by differentiating the system of fixed point equations (14) at $(\tau_{\text{int}}, \alpha_{\text{int}})$. For (τ, α) in the neighborhood of $(\tau_{\text{int}}, \alpha_{\text{int}})$, define

$$v := \eta_{\alpha,r}\left(\frac{B}{\tau} + G\right), \qquad \omega := |v|^{r-2}\eta'_{\alpha,r}\left(\frac{B}{\tau} + G\right). \tag{179}$$

At $v = 0$, the second expression is understood by continuity as $\{\alpha(r-1)\}^{-1}$. Differentiating the proximal equation (82) gives

$$\partial_\alpha v = -\operatorname{sign}(v)|v|^{r-1}\eta'_{\alpha,r}\left(\frac{B}{\tau} + G\right), \qquad \partial_\tau v = -\frac{B}{\tau^2}\eta'_{\alpha,r}\left(\frac{B}{\tau} + G\right), \tag{180}$$

and

$$1 - \eta'_{\alpha,r}\left(\frac{B}{\tau} + G\right) = \alpha(r-1)\omega. \tag{181}$$

Since $0 < \omega \leq \{\alpha(r-1)\}^{-1}$ and $|v| \leq |B/\tau + G|$, differentiation under the expectations defining H_r and U_r is justified by dominated convergence and we have

$$\partial_\alpha H_r(\tau,\alpha) = -2\left(\mathbb{E}[v^2\omega] - \mathbb{E}\left[\frac{B}{\tau}v\omega\right]\right), \tag{182}$$

$$\partial_\tau H_r(\tau,\alpha) = \frac{2\alpha(r-1)}{\tau}\left(\mathbb{E}\left[\frac{B}{\tau}v\omega\right] - \mathbb{E}\left[\left(\frac{B}{\tau}\right)^2\omega\right]\right), \tag{183}$$

$$\partial_\tau U_r(\tau,\alpha) = -\frac{r}{\tau}\mathbb{E}\left[\frac{B}{\tau}v\omega\right], \tag{184}$$

$$\partial_\alpha U_r(\tau,\alpha) = -r\mathbb{E}[|v|^r\omega]. \tag{185}$$

Equation (181) also gives

$$\alpha(r-1)\mathbb{E}[|v|^r\omega] = U_r(\tau,\alpha) - \mathbb{E}[v^2\omega]. \tag{186}$$

If (τ, α) solves (14), then

$$1 - \frac{1}{\delta}H_r(\tau,\alpha) = \frac{\sigma^2}{\tau^2}. \tag{187}$$

Substituting (182)-(186) into the Jacobian of (14) gives

$$\begin{aligned}\det \frac{\partial\left(\tau^2 - \sigma^2 - \dfrac{\tau^2}{\delta}H_r(\tau,\alpha),\ \alpha U_r(\tau,\alpha)^{(r-1)/r}\right)}{\partial(\tau,\alpha)} \\ = \frac{2\tau U_r(\tau,\alpha)^{-1/r}}{\delta}\left[\frac{\delta\sigma^2}{\tau^2}\mathbb{E}[v^2\omega] + \alpha(r-1)\left(\mathbb{E}[v^2\omega]\mathbb{E}\left[\left(\frac{B}{\tau}\right)^2\omega\right] - \mathbb{E}\left[\frac{B}{\tau}v\omega\right]^2\right)\right].\end{aligned} \tag{188}$$

Note that the RHS is strictly positive by Cauchy-Schwartz when $\sigma > 0$. If $\sigma = 0$, equality would imply that v is a deterministic multiple of B/τ almost surely. This is impossible when $\mathbb{P}(B \neq 0) > 0$, because, conditionally on $B = b$, the random variable $\eta_{\alpha,r}(b/\tau + G)$ is nondegenerate for P_B-almost every $b \neq 0$. This proves that the Jacobian in (188) is strictly positive.

The implicit-function theorem therefore gives a unique C^1 local solution curve through $(\tau_{\text{int}}, \alpha_{\text{int}})$. By Lemma 24, for $\kappa > \kappa_r^\star$ sufficiently close to $\kappa_r^\star$, this local solution curve agrees with the parameters $(\tau_\kappa, \alpha_\kappa)$ corresponding to positive training residual R_κ. Thus, we have

$$\frac{d\tau}{d\kappa} = \frac{\tau_{\text{int}}^2 \partial_\alpha H_r(\tau_{\text{int}}, \alpha_{\text{int}})}{\delta} \left[\det \frac{\partial\left(\tau^2 - \sigma^2 - \frac{\tau^2}{\delta} H_r(\tau,\alpha),\ \alpha U_r(\tau,\alpha)^{(r-1)/r}\right)}{\partial(\tau,\alpha)} \Bigg|_{(\tau,\alpha)=(\tau_{\text{int}},\alpha_{\text{int}})} \right]^{-1}. \tag{189}$$

For the remainder of the proof, all quantities are evaluated at $\tau = \tau_{\text{int}}$ and $\alpha = \alpha_{\text{int}}$. Gaussian integration by parts gives $D_r = \mathbb{E}[Gv]$. Differentiating this identity using (180) and the proximal equation yields

$$\begin{aligned}
\partial_\alpha D_r &= -\mathbb{E}[Gv\omega] = \frac{1}{2}\partial_\alpha H_r - \alpha \mathbb{E}[|v|^r \omega], \\
\partial_\tau D_r &= -\frac{1}{\tau}\mathbb{E}\left[G\frac{B}{\tau}\eta'_{\alpha,r}\left(\frac{B}{\tau} + G\right)\right] = \frac{1}{2}\partial_\tau H_r - \frac{\delta\sigma^2}{\tau^3} + \frac{\alpha}{\tau}\left(\mathbb{E}|v|^r - \mathbb{E}\left[\frac{B}{\tau}v\omega\right]\right).
\end{aligned}$$

For the second identity, we also used (181) and

$$\mathbb{E}\left[\frac{B}{\tau}v\right] - \mathbb{E}\left[\left(\frac{B}{\tau}\right)^2\right] + \alpha\mathbb{E}|v|^r = D_r - H_r = \frac{\delta\sigma^2}{\tau^2},$$

where the first equality follows from (82), the definition of H_r in (13), and $D_r = \mathbb{E}[Gv]$, while the second follows from (17). The bounds $|\eta'_{\alpha,r}| \le 1$, $\omega \le (\alpha(r-1))^{-1}$, and $|v| \le |B/\tau + G|$ justify differentiation under the expectations and continuity of these derivatives near the interpolation point.

Substituting these identities and (182), (183), and (186) give

$$\begin{aligned}
&\left(\partial_\tau H_r - \frac{2\delta\sigma^2}{\tau^3}\right)\partial_\alpha D_r - (\partial_\alpha H_r)(\partial_\tau D_r) \\
&\quad = \frac{2\alpha}{\tau}\left[\left(\mathbb{E}\left[\left(\frac{B}{\tau} - v\right)v\omega\right]\right)^2 + \mathbb{E}[|v|^r\omega]\left(\frac{\delta\sigma^2}{\tau^2} + \alpha(r-1)\mathbb{E}\left[\left(\frac{B}{\tau} - v\right)^2\omega\right]\right)\right] > 0.
\end{aligned} \tag{190}$$

Finally, we differentiate

$$\overline{\mathsf{MSE}}_r(\kappa) = \delta(\tau_\kappa^2 - \sigma^2), \qquad R_\kappa = \tau_\kappa\left(1 - \frac{D_r(\tau_\kappa, \alpha_\kappa)}{\delta}\right).$$

Since $D_r = \delta$ at interpolation, $\lim_{\kappa \downarrow \kappa_r^\star} dR_\kappa/d\kappa$ equals the LHS of (190) multiplied by τ^3/δ^2 and by the reciprocal of the Jacobian in (188). This limit is therefore finite and strictly positive. Combining this expression with (189) and $d\overline{\mathsf{MSE}}_r/d\kappa = 2\delta\tau_\kappa\, d\tau_\kappa/d\kappa$ proves (178). □

D.2 Proof of Theorem 2

Proof. Let $s = 1/(r-1)$ and rescale the interpolation parameters as

$$t = \sqrt{\delta}\,\tau, \qquad a = \delta^{r-1}\alpha.$$

We first let $\delta \downarrow 0$ with $\lambda > 0$ and σ fixed. All convergence statements below are uniform over $r \in [r_-, r_+]$ and locally uniform in positive t, a. The notation $o_{C^k}(1)$ means that the remainder and its partial derivatives of total order at most k tend uniformly to zero; before evaluating at interpolation these are derivatives in (t, a, r), and afterwards they are derivatives in r.

Define the rescaled proximal function

$$v_\delta(z) := \delta^{-1}\eta_{a\delta^{1-r},r}(z),$$

suppressing its dependence on a, r. By (82), it satisfies

$$z = \delta v_\delta(z) + a\,\mathrm{sign}(v_\delta(z))|v_\delta(z)|^{r-1}.$$

Consequently, $|v_\delta(z)| \le a^{-s}|z|^s$, $\delta|v_\delta(z)| \le |z|$, and

$$\left|v_\delta(z) - a^{-s}z|z|^{s-1}\right| \le C|z|^s \min\{1, \delta|z|^{s-1}\}, \tag{191}$$

$$\delta v_\delta(z)^2 \le C|z|^{s+1}\min\{1, \delta|z|^{s-1}\}. \tag{192}$$

Implicit differentiation gives

$$\partial_z v_\delta = \frac{|v_\delta|^{2-r}}{a(r-1)+\delta|v_\delta|^{2-r}},\ \partial_a v_\delta = -\frac{v_\delta}{a(r-1)+\delta|v_\delta|^{2-r}},\ \partial_r v_\delta = -\frac{av_\delta\log|v_\delta|}{a(r-1)+\delta|v_\delta|^{2-r}}. \tag{193}$$

We now explain why these approximations may be differentiated twice under the expectations used below. Differentiating (193) once more gives, for $z \neq 0$,

$$|\partial_z v_\delta| \le C|z|^{s-1}, \qquad |\partial_{zz} v_\delta| \le C|z|^{s-2}.$$

Derivatives in a preserve the polynomial growth, while each derivative in r adds at most one factor $1 + |\log|z||$. To identify the expectations being differentiated, expand the definition of H_r in (13):

$$H_r(\tau,\alpha) = \mathbb{E}\left[(\delta v_\delta(B/\tau+G) - B/\tau)^2\right] = \frac{\delta}{t^2} - 2\delta\,\mathbb{E}\left[\frac{B}{\tau}v_\delta(B/\tau+G)\right] + \delta^2\,\mathbb{E}\left[v_\delta(B/\tau+G)^2\right],$$

where we used $\mathbb{E}B^2 = 1$ and $\tau = t/\sqrt{\delta}$. Similarly,

$$D_r(\tau,\alpha) = \delta\,\mathbb{E}\left[\partial_z v_\delta(B/\tau+G)\right] = \delta\,\mathbb{E}\left[Gv_\delta(B/\tau+G)\right].$$

On $\{\xi_\rho = 1\}$, we have $B/\tau = W/(t\sqrt{\lambda})$. Consider first the two integrands

$$\frac{W}{t\sqrt{\lambda}}\,v_\delta\left(\frac{W}{t\sqrt{\lambda}}+G\right), \qquad Gv_\delta\left(\frac{W}{t\sqrt{\lambda}}+G\right),$$

which appear in the second term of the expansion of H_r and in the expression for D_r, respectively. Using the derivative bounds above and integrating over G, the conditional expectations given W of their absolute values, and of the absolute values of all their first and second partial derivatives in (t,a,r), are bounded by

$$C\left(1 + |W|^{s+1}(1+\log^2(2+|W|))\right).$$

The preceding bounds, the moment assumption, and (192) justify differentiation under the expectations and the uniform $C^2(t,a,r)$ expansions below, by dominated convergence with truncation near $z = 0$.

Using (13), conditioning on ξ_ρ, and applying Gaussian integration by parts to D_r, we obtain

$$\frac{H_r(t/\sqrt{\delta}, a\delta^{1-r})}{\delta} = \frac{1}{t^2} + \delta a^{-2s}\mathbb{E}|G|^{2s} - \frac{2\delta\sqrt{\lambda}}{ta^s}\mathbb{E}\left[W\left(\frac{W}{t\sqrt{\lambda}}+G\right)\left|\frac{W}{t\sqrt{\lambda}}+G\right|^{s-1}\right] + o_{C^2}(\delta),$$

$$\frac{D_r(t/\sqrt{\delta}, a\delta^{1-r})}{\delta} = a^{-s}\mathbb{E}|G|^{s+1} + o_{C^2}(1), \tag{194}$$

where $o_{C^2}(\delta)$ denotes δ times an $o_{C^2}(1)$ remainder. In the variables (t,a), (17) corresponds to

$$t^2\left(1 - \frac{H_r(t/\sqrt{\delta}, a\delta^{1-r})}{\delta}\right) - \delta\sigma^2 = 0, \qquad \frac{D_r(t/\sqrt{\delta}, a\delta^{1-r})}{\delta} - 1 = 0. \tag{195}$$

By (194), as $\delta \downarrow 0$ with λ, σ fixed, the two left-hand sides converge to $t^2 - 1$ and $a^{-s}\mathbb{E}|G|^{s+1} - 1$, respectively. Then, the equations obtained by setting these limits equal to zero are

$$t^2 - 1 = 0, \qquad a^{-s}\mathbb{E}|G|^{s+1} - 1 = 0,$$

with the corresponding Jacobian in (t, a) given by

$$\begin{pmatrix} 2 & 0 \\ 0 & -s/a \end{pmatrix}, \tag{196}$$

which is nonsingular for $r \in [r_-, r_+]$. The implicit-function theorem thus gives a positive solution, continuously differentiable in r, with

$$t = 1 + o_{C^1}(1), \qquad a^s = \mathbb{E}|G|^{s+1} + o_{C^1}(1). \tag{197}$$

By Lemma 24, this solution corresponds to the interpolation endpoint, also when $\sigma = 0$. Thus, we can apply Lemma 25 and obtain

$$S_r = \frac{2\delta^2 \partial_\alpha H_r}{\left(\partial_\tau H_r - \dfrac{2\delta\sigma^2}{\tau^3}\right)\partial_\alpha D_r - (\partial_\alpha H_r)(\partial_\tau D_r)}, \tag{198}$$

where all quantities are evaluated at interpolation.

To evaluate the ratio, we differentiate (194) using $\partial_\tau = \sqrt{\delta}\,\partial_t$ and $\partial_\alpha = \delta^{r-1}\partial_a$. Together with (197), this gives

$$\delta^{-3/2}\partial_\tau H_r = -2 + o_{C^1}(1), \qquad \delta^{-r}\partial_\alpha D_r = -s/a + o_{C^1}(1), \qquad \delta^{-3/2}\partial_\tau D_r = o_{C^1}(1),$$

and

$$\delta^{-(r+1)}\partial_\alpha H_r = 2sa^{-s-1}\left(\sqrt{\lambda}\,\mathbb{E}\left[W\left(\frac{W}{\sqrt{\lambda}} + G\right)\left|\frac{W}{\sqrt{\lambda}} + G\right|^{s-1}\right] - a^{-s}\mathbb{E}|G|^{2s}\right) + o_{C^1}(1).$$

Since $2\delta\sigma^2/\tau^3 = O(\delta^{5/2})$, substitution into (198) yields

$$\delta^{-3/2}S_r = \frac{2\sqrt{\lambda}}{\mathbb{E}|G|^{s+1}}\mathbb{E}\left[W\left(\frac{W}{\sqrt{\lambda}} + G\right)\left|\frac{W}{\sqrt{\lambda}} + G\right|^{s-1}\right] - \frac{2\mathbb{E}|G|^{2s}}{(\mathbb{E}|G|^{s+1})^2} + o_{C^1}(1). \tag{199}$$

The C^2 estimates above justify the differentiated expansions and show that $r \mapsto S_r$ is continuously differentiable.

We finally choose λ sufficiently small. As $\lambda \to 0$, we have

$$\lambda^{(s-1)/2}\sqrt{\lambda}\,\mathbb{E}\left[W\left(\frac{W}{\sqrt{\lambda}} + G\right)\left|\frac{W}{\sqrt{\lambda}} + G\right|^{s-1}\right] = \mathbb{E}|W|^{s+1} + o_{C^1}(1).$$

Since $\mathbb{E}|W|^{s+1} \geq 1$ and $s > 1$, combining the preceding estimate with (199) gives

$$\lim_{\delta \downarrow 0} \delta^{-3/2}S_r = \frac{2\mathbb{E}|W|^{s+1}}{\mathbb{E}|G|^{s+1}}\,\lambda^{-(s-1)/2}(1 + o_{C^1}(1)), \tag{200}$$

for all sufficiently small $\lambda > 0$. Here the limit in δ is taken at fixed λ, and $o_{C^1}(1)$ denotes a remainder whose value and first derivative in r converge uniformly to zero on $[r_-, r_+]$ as $\lambda \downarrow 0$.The coefficient in (200) is positive, bounded away from zero, and has bounded logarithmic derivative. As $ds/dr = -s^2$, it follows that, for small λ,

$$\frac{d}{dr}\log\left(\lim_{\delta \downarrow 0}\delta^{-3/2}S_r\right) = -\frac{s^2}{2}\log(1/\lambda) + O(1),$$

uniformly in r. We may therefore choose $\lambda_0 := \lambda_0(r_-, r_+, W) > 0$ so that the limit in (199) is strictly positive and has strictly negative derivative throughout $[r_-, r_+]$ for every $0 < \lambda < \lambda_0$. The C^1 convergence in (199) then gives $\delta_0 = \delta_0(r_-, r_+, W, \lambda, \sigma) > 0$ such that $S_r > 0$ and $dS_r/dr < 0$ whenever $0 < \delta < \delta_0$ and $r \in [r_-, r_+]$, which concludes the argument. □

D.3 Proof of Theorem 3

Proof. Define

$$h(\alpha) := \rho(1+\alpha^2) + (1-\rho)\mathbb{E}(|G|-\alpha)_+^2, \qquad \alpha \geq 0, \tag{201}$$

so that $\delta_c = \min_{\alpha \geq 0} h(\alpha)$ by (23). Computing derivatives gives

$$h''(\alpha) = 2(\rho + (1-\rho)\mathbb{P}(|G| > \alpha)) > 0, \qquad h(\alpha) - \frac{\alpha}{2}h'(\alpha) = \rho + (1-\rho)\mathbb{P}(|G| > \alpha).$$

Since $h'(0) < 0$ and $h(\alpha) \to \infty$, its minimizer α_c is unique and positive. In particular,

$$h(\alpha_c) = \delta_c, \qquad h'(\alpha_c) = 0, \qquad h''(\alpha_c) = 2\delta_c, \qquad \rho + (1-\rho)\mathbb{P}(|G| > \alpha_c) = \delta_c.$$

Define

$$e(x,\alpha) := 2\int_x^\infty z\,\mathbb{P}(|z+G| \leq \alpha)\,dz, \qquad x \geq 0.$$

Conditioning on the support of B gives

$$H_1(\tau,\alpha) = h(\alpha) - \rho\mathbb{E}e\left(\frac{|W|}{\tau\sqrt{\rho}}, \alpha\right),$$

$$D_1(\tau,\alpha) = \rho + (1-\rho)\mathbb{P}(|G| > \alpha) - \rho\mathbb{P}\left(\left|\frac{W}{\tau\sqrt{\rho}} + G\right| \leq \alpha\right),$$

which implies that $\partial_\tau H_1 < 0$, $\partial_\tau D_1 < 0$ and $\partial_\alpha D_1 < 0$.

To identify the limit as $\tau \downarrow 0$, note that (22) implies $\mathbb{P}(W = 0) = 0$, and hence $|W|/(\tau\sqrt{\rho}) \to \infty$ almost surely. Since $0 \leq e(x,\alpha) \leq e(0,\alpha) < \infty$ and $e(x,\alpha) \to 0$ as $x \to \infty$, dominated convergence gives $H_1(\tau,\alpha) \to h(\alpha)$. For each $\alpha > \alpha_c$, the uniqueness of the minimizer of h gives

$$\lim_{\tau\downarrow 0} H_1(\tau,\alpha) = h(\alpha) > h(\alpha_c) = \delta_c.$$

Furthermore, the limit as $\tau \to \infty$ is $\mathbb{E}(|G|-\alpha)_+^2 < \delta_c$. Thus, there exists a unique smooth solution of $H_1(\tau,\alpha) = \delta_c$, which satisfies $D_1(\tau,\alpha) < \delta_c$. For $r = 1$ and $\sigma = 0$, the fixed-point equations (14) reduce, for $\tau_\kappa > 0$, to

$$H_1(\tau_\kappa, \alpha_\kappa) = \delta_c, \qquad \alpha_\kappa = \kappa.$$

Thus, for $\kappa = \alpha > \alpha_c$, the pair $(\tau_\kappa, \alpha_\kappa) = (\tau(\alpha), \alpha)$ solves these equations. Moreover, $D_1(\tau(\alpha),\alpha) < \delta_c$ ensures that $R_\kappa > 0$, as required by Lemma 2. Applying that lemma gives the limiting training error and MSE:

$$R_\alpha = \tau(\alpha)\left(1 - \frac{D_1(\tau(\alpha),\alpha)}{\delta_c}\right) > 0, \qquad \overline{\mathsf{MSE}}_1(\alpha) = \delta_c\tau(\alpha)^2.$$

For every fixed $\tau > 0$, $H_1(\tau,\alpha_c) < \delta_c$. By continuity, this inequality holds for α close to α_c, so the solution $\tau(\alpha)$ must lie below that fixed τ. Thus $\tau(\alpha) \to 0$ as $\alpha \downarrow \alpha_c$, and $0 < R_\alpha \leq \tau(\alpha)$ implies $R_\alpha \to 0$. For any fixed $\kappa \leq \alpha_c$, the training error is bounded by Lemma 23 and hence an application of Lemma 2 gives that the training error at κ converges almost surely to zero, which in turn implies $\kappa_1^\star = \alpha_c$.

We next estimate the partial derivatives of $H_1(\tau,\alpha)$ and $D_1(\tau,\alpha)$ as $\tau \downarrow 0$, evaluating these estimates at $\tau = \tau(\alpha)$, where $H_1(\tau(\alpha),\alpha) = \delta_c$. In the calculations below, $\tau \downarrow 0$ and α ranges over a fixed compact neighborhood of α_c contained in $(0,\infty)$. Gaussian tail estimates give, for $x \geq 0$,

$$e(x,\alpha) + \frac{e_\alpha(x,\alpha)^2 + x^2 e_x(x,\alpha)^2}{e(x,\alpha)} \leq Ce^{-cx^2}, \qquad \mathbb{P}(|x+G| \leq \alpha) + x\,|\partial_x\mathbb{P}(|x+G| \leq \alpha)| \leq Ce^{-cx^2}.$$

Furthermore, (22) implies that, for each $c > 0$,

$$\frac{1}{\tau^2}\mathbb{E}\exp\left(-\frac{cW^2}{\rho\tau^2}\right) = \int_0^\infty 2cxe^{-cx^2}\frac{\mathbb{P}(|W| \leq \tau\sqrt{\rho}\,x)}{\tau^2}\,dx \longrightarrow 0.$$

As a result, we have

$$\begin{aligned} 0 < h(\alpha) - H_1(\tau,\alpha) &\leq C\mathbb{E}\exp\left(-\frac{cW^2}{\rho\tau^2}\right) = o(\tau^2), \\ \frac{|h'(\alpha) - \partial_\alpha H_1| + \tau|\partial_\tau H_1|}{\sqrt{h(\alpha) - H_1(\tau,\alpha)}} &\leq C\left(\mathbb{E}\exp\left(-\frac{cW^2}{\rho\tau^2}\right)\right)^{1/2} = o(\tau). \end{aligned} \tag{202}$$

Since h is three times continuously differentiable near α_c, with $h(\alpha_c) = \delta_c$, $h'(\alpha_c) = 0$ and $h''(\alpha_c) = 2\delta_c$, a Taylor expansion gives

$$h(\alpha) - \delta_c = \delta_c(\alpha - \alpha_c)^2 + O\big((\alpha - \alpha_c)^3\big), \qquad h'(\alpha) = 2\delta_c(\alpha - \alpha_c) + O\big((\alpha - \alpha_c)^2\big),$$

which combined with (202) yields

$$\alpha - \alpha_c = o(\tau), \qquad \partial_\alpha H_1 \sim 2\delta_c(\alpha - \alpha_c) > 0, \qquad \partial_\tau H_1 = o(\alpha - \alpha_c).$$

Since $\partial_\tau H_1 < 0$, $\tau(\alpha)$ is strictly increasing near α_c and it therefore has a local inverse $\alpha(\tau)$, satisfying

$$0 < \alpha'(\tau) = -\frac{\partial_\tau H_1}{\partial_\alpha H_1} = o(1),$$

as $\tau \downarrow 0$.

Note that

$$|D_1 - \rho - (1-\rho)\mathbb{P}(|G| > \alpha)| + \tau|\partial_\tau D_1| \le C\mathbb{E}\exp\left(-\frac{cW^2}{\rho\tau^2}\right) = o(\tau^2),$$

which implies that

$$D_1 = \rho + (1-\rho)\mathbb{P}(|G| > \alpha) + o(\tau^2), \qquad \partial_\tau D_1 = o(\tau).$$

Furthermore, $|\partial_\alpha D_1| \le 2/\sqrt{2\pi}$, since $-\partial_\alpha D_1$ is an average of sums of two Gaussian densities.

We now evaluate D_1 and its partial derivatives at $(\tau, \alpha(\tau))$. As $\rho + (1-\rho)\mathbb{P}(|G| > \alpha_c) = \delta_c$ and $\alpha(\tau) - \alpha_c = o(\tau)$, we have

$$1 - D_1/\delta_c = o(\tau).$$

Differentiating $R_{\alpha(\tau)} = \tau(1 - D_1/\delta_c)$ therefore gives

$$0 < \frac{d}{d\tau}R_{\alpha(\tau)} = 1 - \frac{D_1}{\delta_c} - \frac{\tau}{\delta_c}\left(\partial_\tau D_1 + \partial_\alpha D_1\, \alpha'(\tau)\right) = o(\tau).$$

Finally, as $\kappa = \alpha(\tau)$, $\kappa \downarrow \kappa_1^\star = \alpha_c$ implies $\tau \downarrow 0$. Using $\overline{\mathsf{MSE}}_1(\alpha(\tau)) = \delta_c\tau^2$, we therefore obtain

$$S_1 = \lim_{\kappa\downarrow\kappa_1^\star} \frac{d\overline{\mathsf{MSE}}_1/d\kappa}{dR_\kappa/d\kappa} = \lim_{\tau\downarrow 0} \frac{2\delta_c\tau}{\dfrac{d}{d\tau}R_{\alpha(\tau)}} = +\infty,$$

which gives the desired result. □

D.4 Exact interpolation hurts generalization for noisy data and ℓ_1 regularization

Proposition 3 ($S_1 < 0$ for ℓ_1 regularization and $\sigma > 0$)**.** *Let $r = 1$, $\delta \in (0,1)$, $\sigma > 0$, and $0 < \mathbb{E}B^2 < \infty$. Then the limit S_1 in (20) exists and is strictly negative.*

Proof. For every $\tau \ge \sigma$, the equation $D_1(\tau, \alpha) = \mathbb{P}(|B/\tau + G| > \alpha) = \delta$ yields a unique positive α, which is continuous in τ. Furthermore, $H_1(\tau, \alpha) - \delta(1 - \sigma^2/\tau^2)$ is positive at $\tau = \sigma$. As $\tau \to \infty$, α converges to the positive solution of $\mathbb{P}(|G| > \alpha) = \delta$, and the same expression has a strictly negative limit by Lemma 12. Hence, by the intermediate-value theorem, there exists a solution $(\tau_{\mathrm{int}}, \alpha_{\mathrm{int}})$ of (17), with $\tau_{\mathrm{int}} > \sigma$.

In the calculations below, we write $(\tau, \alpha) = (\tau_{\mathrm{int}}, \alpha_{\mathrm{int}})$ and recall $Z = B/\tau + G$ from (12). Using (91) and Gaussian integration by parts, we have

$$\partial_\tau H_1 - \frac{2\delta\sigma^2}{\tau^3} = -\frac{2}{\tau^3}\left(\delta\sigma^2 + \mathbb{E}\left[B^2\mathbf{1}_{\{|Z|\le\alpha\}}\right]\right) < 0,$$
$$\partial_\alpha H_1 = -\frac{2}{\alpha}\left(\frac{\delta\sigma^2}{\tau^2} + \mathbb{E}\left[\frac{B}{\tau}Z\mathbf{1}_{\{|Z|\le\alpha\}}\right]\right) < 0.$$

By continuity of the partial derivatives, we can apply the implicit-function theorem to the first equation in (14) and obtain a continuously differentiable $\tau(\alpha)$ satisfying

$$\tau'(\alpha_{\text{int}}) = -\frac{\partial_\alpha H_1}{\partial_\tau H_1 - 2\delta\sigma^2/\tau^3} < 0.$$

Differentiating D_1 and using (17) gives

$$\partial_\alpha D_1 = \frac{1}{2}\partial_\alpha H_1 - \alpha\delta, \qquad \partial_\tau D_1 = \frac{1}{2}\left(\partial_\tau H_1 - \frac{2\delta\sigma^2}{\tau^3}\right) - \frac{\alpha}{2\tau}\partial_\alpha H_1.$$

Consequently,

$$\frac{d}{d\alpha} D_1(\tau(\alpha), \alpha)\Big|_{\alpha=\alpha_{\text{int}}} = -\alpha\delta - \frac{\alpha}{2\tau}\partial_\alpha H_1\, \tau'(\alpha_{\text{int}}) < 0.$$

For $\alpha > \alpha_{\text{int}}$ sufficiently close to it, we therefore have $D_1 < \delta$. Since $\kappa = \alpha$ for $r = 1$, Lemma 2 gives

$$R_\alpha = \tau(\alpha)\left(1 - \frac{D_1(\tau(\alpha), \alpha)}{\delta}\right) > 0, \qquad \overline{\mathsf{MSE}}_1(\alpha) = \delta\big(\tau(\alpha)^2 - \sigma^2\big).$$

As $\alpha \downarrow \alpha_{\text{int}}$, $R_\alpha \to 0$ and, following the same argument as for Theorem 3, we obtain that $\kappa_1^\star = \alpha_{\text{int}}$. Moreover,

$$\frac{d}{d\alpha} R_\alpha\Big|_{\alpha=\alpha_{\text{int}}} = \alpha\tau + \frac{\alpha}{2\delta}\partial_\alpha H_1\, \tau'(\alpha_{\text{int}}) > 0.$$

This allows us to conclude that

$$S_1 = \lim_{\alpha\downarrow\alpha_{\text{int}}} \frac{2\delta\tau(\alpha)\tau'(\alpha)}{\frac{d}{d\alpha}R_\alpha} = \frac{2\delta\tau_{\text{int}}\tau'(\alpha_{\text{int}})}{\frac{d}{d\alpha}R_\alpha\Big|_{\alpha=\alpha_{\text{int}}}} < 0,$$

which proves the claim. □

D.5 Exact interpolation hurts generalization for ℓ_2 regularization

Proposition 4 ($S_2 \le 0$ for ℓ_2 regularization)**.** *Let $r = 2$, $\delta \in (0, 1)$, and $\sigma \ge 0$. Let $B \sim P_B$ have second moment $m := \mathbb{E}B^2$, where $0 < m + \sigma^2 < \infty$. Then the limit S_2 in (20) exists and*

$$S_2 = -\frac{2\delta^2\sigma^2}{(1-\delta)^2\tau_{\text{int}}} \le 0, \qquad \tau_{\text{int}}^2 = \frac{\sigma^2}{1-\delta} + \frac{1-\delta}{\delta}m. \tag{203}$$

Moreover, the limit

$$\lim_{\kappa\downarrow\kappa_2^\star} \frac{d}{d\kappa}\overline{\mathsf{MSE}}_2(\kappa)$$

exists and is strictly negative when $\sigma > 0$ and zero when $\sigma = 0$.

Proof. Since $\eta_{\alpha,2}(z) = z/(1+\alpha)$, we have $D_2(\tau, \alpha) = 1/(1+\alpha)$, so the interpolation condition $D_2 = \delta$ gives $\alpha_{\text{int}} = (1-\delta)/\delta$. Substituting $\eta_{\alpha,2}(z) = z/(1+\alpha)$ into the definitions of H_2 and U_2 in (13), and then into the fixed-point equations (14), gives

$$\tau(\alpha)^2 = \frac{\delta\sigma^2(1+\alpha)^2 + m\alpha^2}{\delta(1+\alpha)^2 - 1}, \qquad \kappa(\alpha) = \frac{\alpha}{1+\alpha}\sqrt{1 + \frac{m}{\tau(\alpha)^2}}. \tag{204}$$

At $\alpha = \alpha_{\text{int}}$, the formula for τ gives

$$\tau(\alpha_{\text{int}})^2 = \tau_{\text{int}}^2 = \frac{\sigma^2}{1-\delta} + \frac{1-\delta}{\delta}m > 0,$$

ensuring that $\tau(\alpha)$ and $\kappa(\alpha)$ are smooth near α_{int}. For $\alpha > \alpha_{\text{int}}$ and sufficiently close to it, $(\tau(\alpha), \alpha)$ solves (14) at $\kappa = \kappa(\alpha)$ and satisfies $D_2 < \delta$. By Lemma 2, we thus have

$$R_{\kappa(\alpha)} = \tau(\alpha)\left(1 - \frac{1}{\delta(1+\alpha)}\right) > 0, \qquad \overline{\mathsf{MSE}}_2(\kappa(\alpha)) = \delta\big(\tau(\alpha)^2 - \sigma^2\big).$$

Differentiating at α_{int} gives

$$\tau'(\alpha_{\text{int}}) = -\frac{\delta^2\sigma^2}{(1-\delta)^2\tau_{\text{int}}}, \qquad \frac{d}{d\alpha}R_{\kappa(\alpha)}\Big|_{\alpha=\alpha_{\text{int}}} = \delta\tau_{\text{int}} > 0.$$

Since $\tau'(\alpha_{\text{int}}) \le 0$, differentiating the formula for κ also gives $\kappa'(\alpha_{\text{int}}) \ge \delta^2\sqrt{1+m/\tau_{\text{int}}^2} > 0$. Thus, $\kappa(\alpha)$ is locally invertible and increasing. As $\alpha \downarrow \alpha_{\text{int}}$, $R_{\kappa(\alpha)} \to 0$ and, following the same argument as for Theorem 3, we obtain that $\kappa_2^\star = \kappa(\alpha_{\text{int}})$. Consequently, $\kappa \downarrow \kappa_2^\star$ corresponds to $\alpha \downarrow \alpha_{\text{int}}$ and we have

$$S_2 = \lim_{\alpha\downarrow\alpha_{\text{int}}} \frac{2\delta\tau(\alpha)\tau'(\alpha)}{\frac{d}{d\alpha}R_{\kappa(\alpha)}} = -\frac{2\delta^2\sigma^2}{(1-\delta)^2\tau_{\text{int}}}.$$

As $\kappa \downarrow \kappa_2^\star$, the MSE derivative tends to $2\delta\tau_{\text{int}}\tau'(\alpha_{\text{int}})/\kappa'(\alpha_{\text{int}})$, which is strictly negative when $\sigma > 0$ and zero when $\sigma = 0$, thus concluding the argument. □

D.6 Proof of Theorem 4

Proof. Let $s = 1/(r-1) > 1$, define

$$F(x) := \mathbb{E}\left[(x+G)|x+G|^{s-1}\right],$$

and set

$$\mathcal{S}(\lambda) := \frac{2}{\mathbb{E}|G|^{s+1}}\left(\sqrt{\lambda}\,\mathbb{E}\left[WF\left(\frac{W}{\sqrt{\lambda}}\right)\right] - \frac{\mathbb{E}|G|^{2s}}{\mathbb{E}|G|^{s+1}}\right). \tag{205}$$

As in the proof of Theorem 2, we rescale the interpolation parameters and proximal function as

$$t = \sqrt{\delta}\,\tau, \qquad a = \delta^{r-1}\alpha, \qquad v_\delta(z) = \delta^{-1}\eta_{a\delta^{1-r},r}(z).$$

Here r is fixed, and the partial derivatives in the expansions below are taken with respect to (t, a, λ). Differentiation in λ has the same polynomial growth as differentiation in t, and it introduces no logarithmic factors. Thus, using (191), (193), and (192), the same argument gives (194) in $C^2(t, a, \lambda)$, uniformly on compact positive parameter sets. Indeed, after integration over G, the bounds for the linear terms and their derivatives are $C(1+|W|^{s+1})$. The cutoff (192) similarly controls δv_δ^2 and its derivatives on $\{\xi_\rho = 1\}$, whose expectations over W, G tend to zero by dominated convergence.

The argument based on the implicit function theorem which leads to (197) gives a positive solution $(\tau_{\text{int}}, \alpha_{\text{int}})$ of the fixed point equations in (17), namely

$$\tau_{\text{int}}^2 = \sigma^2 + \frac{\tau_{\text{int}}^2}{\delta}H_r(\tau_{\text{int}}, \alpha_{\text{int}}), \qquad D_r(\tau_{\text{int}}, \alpha_{\text{int}}) = \delta,$$

also when $\sigma = 0$. By Lemma 24, this solution is unique and $(\tau_\kappa, \alpha_\kappa) \to (\tau_{\text{int}}, \alpha_{\text{int}})$ as $\kappa \downarrow \kappa_r^\star$. Substituting into (178), as in the derivation of (199), gives

$$\delta^{-3/2}S_r = \mathcal{S}(\lambda) + o_{C^1}(1), \tag{206}$$

as $\delta \downarrow 0$ with $\rho = \lambda\delta$ and σ fixed. Here the remainder and its first derivative in λ converge uniformly to zero on $[\lambda_-, \lambda_+]$.

We next show that $\mathcal{S}'(\lambda) < 0$ for every $\lambda > 0$. By symmetry, F is odd, so

$$\sqrt{\lambda}\,\mathbb{E}\left[WF\left(\frac{W}{\sqrt{\lambda}}\right)\right] = \mathbb{E}\left[W^2\frac{F(|W|/\sqrt{\lambda})}{|W|/\sqrt{\lambda}}\right], \tag{207}$$

where the ratio at zero is defined as $F'(0)$. Moreover, $F'(x) = s\mathbb{E}|x+G|^{s-1}$ is strictly increasing for $x > 0$: for $x, u > 0$,

$$\frac{\partial}{\partial x}\mathbb{P}(|x+G| > u) = \frac{e^{-(u-x)^2/2} - e^{-(u+x)^2/2}}{\sqrt{2\pi}} > 0.$$

Since $F(0) = 0$, it follows that

$$F'(x) - \frac{F(x)}{x} = F'(x) - \frac{1}{x}\int_0^x F'(u)\,du > 0, \qquad x > 0. \tag{208}$$

The bound $|F'(x)| + |F(x)|/x \le C(1 + x^{s-1})$ and $\mathbb{E}|W|^{s+1} < \infty$ justify differentiation under the expectation in (207), yielding

$$\mathcal{S}'(\lambda) = -\frac{1}{\lambda \mathbb{E}|G|^{s+1}} \mathbb{E}\left[W^2 \left(F'\left(\frac{|W|}{\sqrt{\lambda}}\right) - \frac{F(|W|/\sqrt{\lambda})}{|W|/\sqrt{\lambda}} \right) \right] < 0, \tag{209}$$

where the term in brackets is defined as zero when $W = 0$. The strict inequality follows from (208) and $\mathbb{E}W^2 = 1$, and the continuity of $\mathcal{S}'$ follows by dominated convergence.

For fixed δ, (206) and $\partial_\rho = \delta^{-1}\partial_\lambda$ give

$$\frac{\partial S_r}{\partial \rho} = \delta^{1/2} \left\{ \mathcal{S}'(\rho/\delta) + o(1) \right\}, \qquad \rho \in [\lambda_- \delta, \lambda_+ \delta].$$

Since $\mathcal{S}'$ has a strictly negative maximum on $[\lambda_-, \lambda_+]$, this derivative is negative throughout the required interval for all sufficiently small δ. Finally, the same argument used for (200) gives

$$\lim_{\lambda \downarrow 0} \lambda^{(s-1)/2} \mathcal{S}(\lambda) = \frac{2\mathbb{E}|W|^{s+1}}{\mathbb{E}|G|^{s+1}} > 0.$$

Let us choose $\lambda_c = \lambda_c(r, W) > 0$ so that $\mathcal{S}(\lambda) > 0$ for $0 < \lambda < \lambda_c$. If $\lambda_+ < \lambda_c$, then $\mathcal{S}(\lambda) \ge \mathcal{S}(\lambda_+) > 0$ on $[\lambda_-, \lambda_+]$. The uniform convergence in (206) therefore gives $S_r > 0$ throughout the required interval, concluding the argument. □

D.7 Exact interpolation hurts generalization without structure

Proposition 5 ($S_r < 0$ for Gaussian B)**.** *Let $\delta \in (0, 1)$ and $\sigma \ge 0$. Suppose that $B \sim N(0, m)$ for some $m > 0$. Then, for every $1 \le r < 2$, the limits below exist, are finite, and satisfy*

$$\lim_{\kappa \downarrow \kappa_r^\star} \frac{d}{d\kappa} \overline{\mathsf{MSE}}_r(\kappa) < 0, \qquad S_r < 0,$$

where S_r is defined in (20).

Proof. Fix $1 \le r < 2$ and write $v_a := \eta_{a,r}(G)$ for $a > 0$, with $G \sim N(0, 1)$. From the proximal characterization, we have that $a \mapsto \mathbb{E}[G v_a]$ is continuously differentiable with strictly negative derivative, decreasing from 1 to 0. Thus, there is a unique $a_* > 0$ such that

$$\mathbb{E}[G v_{a_*}] = \mathbb{E}\eta'_{a_*,r}(G) = \delta, \qquad \delta - \mathbb{E}v_{a_*}^2 = a_* \mathbb{E}|v_{a_*}|^r > 0,$$

where the first equality follows from Gaussian integration by parts and the second identity from the proximal equation. For a near a_*, define

$$\tau(a)^2 = \frac{m\mathbb{E}(G - v_a)^2 + \delta\sigma^2}{\delta - \mathbb{E}v_a^2}, \qquad \alpha(a) = a\left(1 + \frac{m}{\tau(a)^2}\right)^{(2-r)/2},$$
$$\kappa(a) = a\sqrt{1 + \frac{m}{\tau(a)^2}} \left(\mathbb{E}|v_a|^r\right)^{(r-1)/r}. \tag{210}$$

These functions are positive and continuously differentiable. From the definition of $\alpha(a)$, we readily have

$$\eta_{\alpha(a),r}\left(\sqrt{1 + \frac{m}{\tau(a)^2}}\, G\right) = \sqrt{1 + \frac{m}{\tau(a)^2}}\, v_a.$$

Since $B/\tau + G$ is Gaussian with variance $1 + m/\tau^2$ and $\mathbb{E}[B/\tau \mid B/\tau + G] = m(B/\tau + G)/(m + \tau^2)$, we have

$$D_r(\tau(a), \alpha(a)) = \mathbb{E}[G v_a],$$
$$H_r(\tau(a), \alpha(a)) = \mathbb{E}v_a^2 + \frac{m}{\tau(a)^2}\mathbb{E}(G - v_a)^2,$$
$$U_r(\tau(a), \alpha(a)) = \left(1 + \frac{m}{\tau(a)^2}\right)^{r/2} \mathbb{E}|v_a|^r.$$

Consequently, $(\tau(a), \alpha(a))$ solves the fixed-point equations (14) at $\kappa = \kappa(a)$. For $a > a_*$ and sufficiently close to it, $D_r < \delta$, so Lemma 2 gives

$$R_{\kappa(a)} = \tau(a)\left(1 - \frac{\mathbb{E}[Gv_a]}{\delta}\right) > 0, \qquad \overline{\mathsf{MSE}}_r(\kappa(a)) = \delta\big(\tau(a)^2 - \sigma^2\big).$$

We next prove that $\tau'(a_*) < 0$. Suppose first that $1 < r < 2$. At $a = a_*$, (82) and $\mathbb{E}[Gv_{a_*}] = \delta$ imply

$$\delta - \mathbb{E}v_{a_*}^2 = a_*\mathbb{E}|v_{a_*}|^r, \qquad \mathbb{E}(G - v_{a_*})^2 = a_*^2\mathbb{E}|v_{a_*}|^{2r-2}.$$

Differentiating the formula for $\tau(a)$ in (210), then using these identities and $\sigma^2 \geq 0$, gives

$$\frac{\tau'(a_*)}{\tau(a_*)} \leq \frac{1}{a_*}\left(\frac{\mathbb{E}[|v_{a_*}|^{2r-2}\eta'_{a_*,r}(G)]}{\mathbb{E}|v_{a_*}|^{2r-2}} - \frac{\mathbb{E}[|v_{a_*}|^{r}\eta'_{a_*,r}(G)]}{\mathbb{E}|v_{a_*}|^{r}}\right) < 0. \tag{211}$$

To justify the strict inequality, observe that

$$\eta'_{a,r}(G) = \frac{1}{1 + a(r-1)|v_a|^{r-2}}.$$

Under the probability distribution obtained by weighting the law of $|v_a|$ by $|v_a|^{2r-2}$, the functions x^{2-r} and $1/(1+a(r-1)x^{r-2})$ have strictly positive covariance, since both are strictly increasing and the distribution is nondegenerate. This is precisely the strict inequality between the two weighted averages in (211).

For $r = 1$, we have $v_a = \operatorname{sign}(G)(|G| - a)_+$ and $\mathbb{E}[Gv_a] = \mathbb{P}(|G| > a)$. Differentiating the formula for $\tau(a)$ at a_* gives

$$\tau(a_*)\tau'(a_*) = -\frac{m\mathbb{E}[G^2\mathbf{1}_{\{|G|\leq a_*\}}] + \delta\sigma^2}{a_*^2\mathbb{E}(|G| - a_*)_+} < 0,$$

where the inequality is strict because $m > 0$ and $a_* > 0$.

For every $1 \leq r < 2$, by computing the derivative, we have

$$\frac{d}{da}\left[a\big(\mathbb{E}|v_a|^r\big)^{(r-1)/r}\right] = \big(\mathbb{E}|v_a|^r\big)^{-1/r}\mathbb{E}[|v_a|^r\eta'_{a,r}(G)] > 0.$$

Together with $\tau'(a_*) < 0$, this implies $\kappa'(a_*) > 0$ in (210). Thus, $\kappa(a)$ is locally invertible and increasing. As $a \downarrow a_*$, $R_{\kappa(a)} \to 0$ and, following the same argument as for Theorem 3, we obtain $\kappa_r^\star = \kappa(a_*)$. At a_*, the fixed point equations in (17) also hold, so the limiting MSE of the MNI equals $\delta(\tau(a_*)^2 - \sigma^2)$ by Lemma 3. Finally,

$$\frac{d}{da}R_{\kappa(a)}\bigg|_{a=a_*} = -\frac{\tau(a_*)}{\delta}\,\frac{d}{da}\mathbb{E}[Gv_a]\bigg|_{a=a_*} > 0.$$

Since $\kappa \downarrow \kappa_r^\star$ corresponds to $a \downarrow a_*$, the chain rule gives

$$\lim_{\kappa\downarrow\kappa_r^\star}\frac{d}{d\kappa}\overline{\mathsf{MSE}}_r(\kappa) = \frac{2\delta\tau(a_*)\tau'(a_*)}{\kappa'(a_*)} < 0, \qquad S_r = \frac{2\delta\tau(a_*)\tau'(a_*)}{\dfrac{d}{da}R_{\kappa(a)}\Big|_{a=a_*}} < 0,$$

which proves the desired result. □

E Additional numerical experiments and implementation details

E.1 Early-stopped gradient descent as Q_k-regularized regression

Azulay et al. (2021) show that, when gradient flow on a diagonal linear network converges to an interpolating solution $\hat{\beta}(\infty)$, this solution satisfies $\hat{\beta}(\infty) = \arg\min_{X\beta=y} Q_k(\beta)$, where Q_k is the initialization-dependent regularizer defined in Section 5. Here, we show experimentally that this correspondence extends beyond interpolation: the early-stopped iterates $\hat{\beta}(t)$ closely track the solutions of

$$\min_{\beta\in\mathbb{R}^p} \|y - X\beta\|_2 + \frac{\kappa}{\sqrt{n}}\,Q_k(\beta)$$

along the regularization path, with larger stopping times corresponding to smaller κ.

Experimental Setup. We draw $\beta \in \mathbb{R}^p$ from the three-point prior of Section 5 with $\rho = 0.1$, i.e., B follows (21) with W a Rademacher random variable. The design has i.i.d. entries $X_{ij} \sim \mathcal{N}(0, 1/n)$, and $y = X\beta + \sigma z$ with $z \sim \mathcal{N}(0, I_n)$, where $n = 2000$, $p = 4000$, and $\sigma = 0.03$. We run paired simulations of the two methods described below using the same sampled prior and design matrix, and repeat this procedure 15 times.

Gradient descent. We train the two-layer diagonal linear networks $\beta = u_+ \odot v_+ - u_- \odot v_-$ with the square-root loss $\|y - X\beta\|_2/\sqrt{n}$. All four weight vectors are initialized at the same constant, $u_+(0) = u_-(0) = v_+(0) = v_-(0) = \frac{k^{1/4}}{2}\mathbf{1}$. The induced regularizer is then Q_k, and we sweep $k \in \{0.01, 0.1, 1, 10, 100\}$. We use a fixed step size $\eta = 0.01/\|X\|_2^2$. Time is measured in gradient-flow units, $t = \eta \times$ (number of iterations).

Explicit regularization. For each k, we solve the Q_k-regularized program above with CVXPY at ten values of κ. These are fixed multiples $\{0.5, 0.95, 1, 1.02, 1.05, 1.1, 1.2, 1.4, 1.7, 2.2\}$ of the interpolation threshold $\kappa_*(k)$ predicted by state evolution; for $\kappa < \kappa_*$, the estimator interpolates the data.

Results. In Figure 3, we overlay the gradient-descent trajectories and the solutions to the Q_k-regularized regression problems, plotting parameter MSE as a function of training error. Each solid line is a gradient-descent trajectory for one value of k, averaged over seeds, with a standard-deviation band. Training starts at $\hat{\beta}(0) = 0$ on the right and ends at the interpolating solution on the left, so time flows from right to left. Each square represents one explicitly regularized solution. The star at the left edge, labeled "Azulay-norm SE plateau", marks the interpolating branch prediction from Lemma 3.

For every k, the gradient-descent curve passes through the squares: stopping gradient descent at a given training error yields a parameter MSE close to that obtained by solving the explicit program at the value of κ producing the same training error. To make the correspondence between stopping time and regularization strength explicit, we annotate each square on the $k = 100$ curve with its κ as a multiple of κ_* and with the time t at which gradient descent reaches the same training error.

E.2 Transformer on a modular arithmetic task

Architecture. We use a causal decoder-only transformer following the setup of Power et al. (2022), with one transformer block in place of two. Each block applies pre-layer normalization followed by causal multi-head self-attention with four heads and a residual connection, then pre-layer normalization followed by an MLP with another residual connection. The attention mask excludes future positions (the strictly upper-triangular entries of the attention matrix). The MLP has the form $\mathrm{Linear}(128 \to 512) \to \mathrm{GELU} \to \mathrm{Linear}(512 \to 128)$. Token embeddings and learned positional embeddings each have dimension 128. A final layer normalization precedes a bias-free linear output head over a vocabulary of 99 tokens: 97 residues, <op>, and <=>. The model has 224,512 trainable parameters, of which 13,312 (5.9%) belong to the embeddings.

Task. We consider modular arithmetic over $\mathbb{Z}/p\mathbb{Z}$ with $p = 97$, and compute $c = xy \bmod p$. Each input is the token sequence x <op> y <=>, the model predicts the target c at the final position, and we evaluate cross-entropy loss and accuracy only at this position. We split the $p(p-1) = 9{,}312$ pairs equally between training and validation, with 4,656 pairs in each split.

Optimizer and training. We use AdamW with cross-entropy loss, base learning rate 10^{-3}, weight decay 1, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and batch size 512. The learning rate is warmed up linearly over the first 10 optimizer steps. Training runs for at most 3×10^5 optimizer steps, with early stopping when validation accuracy exceeds 0.99. Each configuration uses five seeds. Within each repetition, the same seed is used for the data split and model initialization.

Output-scale parametrization. Following the rich/lazy output-scale axis of Kunin et al. (2024), we multiply the logits by α immediately before computing the loss: $\tilde{\mathbf{z}} = \alpha\mathbf{z}$. The learning rate is adjusted according to $\eta(\alpha) = \frac{10^{-3}}{\alpha^2}$. We refer to this as the original recipe and we run it for $\alpha > 1$. This recipe compensates the learning rate but leaves the weight-decay coefficient unchanged.

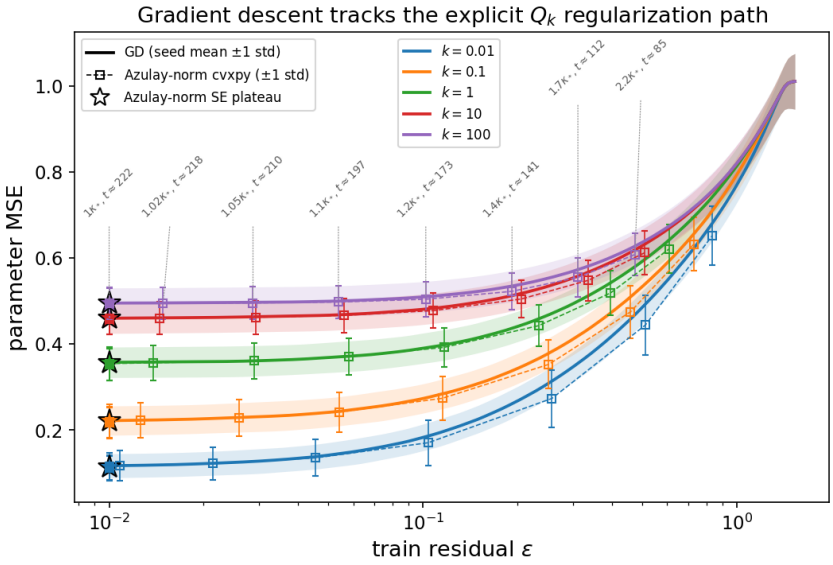

Figure 3: Early-stopped gradient descent on diagonal linear networks (solid lines) versus Q_k-regularized regression solved with CVXPY (squares), for several initialization scales k. Both are plotted as parameter MSE against training error; time along the gradient-descent trajectories flows from right to left. Stars mark the state-evolution prediction for the minimum-Q_k interpolator, the common endpoint of both methods. For $k = 100$, each square is annotated with its κ (as a multiple of the interpolation threshold κ_*) and the gradient-descent time t at which it reaches the same training error.

Consequently, smaller α is associated with both richer training dynamics and stronger effective weight decay.

To control for this confounder, for $\alpha \leq 1$, we run with a compensated weight-decay coefficient $\lambda(\alpha) = \alpha^2$, while keeping the learning rate at $\eta(\alpha) = 10^{-3}/\alpha^2$ as before. This keeps $\eta(\alpha)\lambda(\alpha) = 10^{-3}$ equal to its value at $\alpha = 1$, allowing the output-scale effect to be examined at fixed effective decay.

Training and generalization error curves. We use the validation cross-entropy loss at its final recorded epoch T as an empirical proxy for $\mathcal{L}(\hat{\beta}(\infty))$, $\mathcal{L}$ being the cross-entropy loss. The residual loss is therefore zero at epoch T. Let a_t^{train} denote the training accuracy at epoch t. We define the training error as $\epsilon_t := 1 - \max_{t' \leq t} a_{t'}^{\text{train}}$. When ϵ_t reaches the tolerance value of 0.01 for some $t = t_1$, we set it to 0 for all remaining epochs $t > t_1$.