Mixup Mitigates Instance-Level Overfitting under Structural Overparameterization

Anonymous authors
Paper under double-blind review

Abstract

Overparameterization is commonly measured by the number of fitted model parameters relative to the training sample size, i.e., the *dimensional overparameterization*. In this paper, we examine mixup augmentation through the complementary notion of *structural overparameterization*: a nonlinear student contains directions that are unnecessary for a linear teacher but can overfit observation noise. For Gaussian covariates, we derive precise high-dimensional characterizations of the prediction risk under full-pair mixup. Our analysis combines a decomposition of pairwise interactions with leave-many-out arguments to handle the complex dependence among n^2 mixup samples generated from n training data. The resulting risk formulas reveal how mixup changes the tradeoff between signal reconstruction error and residual fitting error under *homoscedastic* and *heteroscedastic* noise. In particular, under explicit conditions, we show that mixup achieves *strictly lower risk* than ordinary nonlinear regressors, with the ridge penalty optimized separately for each method. This advantage consequently extends to mixup jointly tuned over its mixing weight and ridge strength. Numerical experiments support the predicted risk comparisons. Motivated by this analysis, we propose a selective mixup rule based on training-time fitting progress in neural networks with learned features on three real-world regression datasets. Under training-label perturbations, selected mixup reduces test error relative to empirical risk minimization and separately tuned weight decay.

1 Introduction

A predictor's complexity depends on more than its number of coefficients. We study *structural overparameterization*, which can arise when a student model has greater expressive capacity than is needed to describe the data-generating regression function, which we call the *teacher*. This additional flexibility can allow the student to fit variation absent from the teacher, including observation noise. This notion complements the familiar *dimensional overparameterization*, measured by the ratio of the number of fitted coefficients to the training sample size. Dimensional overparameterization already gives rise to phenomena such as double descent (Belkin et al., 2019; Hastie et al., 2022) and benign overfitting (Bartlett et al., 2020). The structural perspective additionally distinguishes students with the same parameter count but different functional forms. Modern machine learning models often exhibit both forms of excess complexity.

In this paper, we investigate how mixup affects generalization under these two forms of overparameterization. Mixup is a widely used data augmentation technique that trains a predictor on convex combinations of inputs and their labels (Zhang et al., 2018). To study its interaction with structural overparameterization in an analytically tractable setting, we focus on nonlinear random feature regression with a linear teacher. Here, the nonlinear features introduce flexibility beyond the teacher's linear structure, allowing us to examine how mixup acts on this additional flexibility. For two observations (x_i, y_i) and (x_j, y_j), a fixed mixing weight $\lambda \in (0, 1)$ produces a mixup sample $(x_{ij}^{\text{mix}}, y_{ij}^{\text{mix}})$ with

$$x_{ij}^{\text{mix}} = \lambda x_i + (1-\lambda)x_j, \qquad y_{ij}^{\text{mix}} = \lambda y_i + (1-\lambda)y_j. \tag{1.1}$$

For a linear teacher of the form $y_i = x_i^{\mathrm{T}}\beta^* + \varepsilon_i$, the signal component of y_{ij}^{mix} is exactly the teacher evaluated at x_{ij}^{mix}, and only the observation noise is mixed.

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

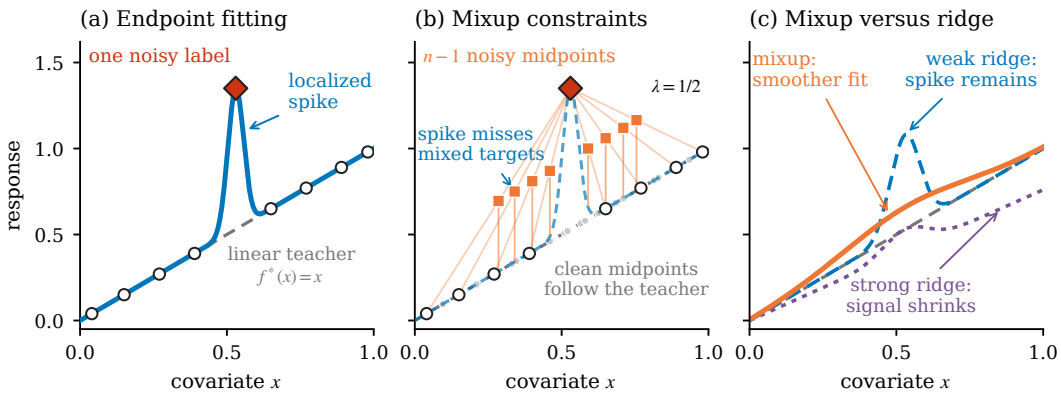

Figure 1: How chord constraints regularize instance-level nonlinear overfitting. Panel (a) shows a flexible endpoint interpolant that absorbs one contaminated label through a sample-specific spike. In panel (b), the contaminated endpoint generates mixed-label targets on many chords, which the narrow spike fails to fit; clean–clean midpoints follow the linear teacher. Panel (c) contrasts weak ridge, which leaves a spike, with strong ridge, which also shrinks the signal, and a smoother mixup fit. Section 4 studies the correspondinn population-risk comparison. The schematic uses the midpoint $\lambda = 1/2$ for clarity; the estimator and risk theory allow λ to vary.

Our theoretical analysis considers *full-pair mixup*, which uses every ordered pair, and applies the same mixing weight to covariates and labels. To establish a baseline without structural overparameterization, we first restrict the student to linear predictors $f_\beta(x) = x^\top\beta$, with $\beta \in \mathbb{R}^d$. This model has d coefficients for n training observations and is dimensionally overparameterized when $d > n$. The following result shows that full-pair mixup leaves the minimum-norm ridgeless estimator *unchanged*, its proof is given in Appendix S.4.1.

Proposition 1.1 (Mixup has no benefit for ridgeless linear students) Consider the noisy linear teacher $y = X\beta^* + \varepsilon$ in Definition 2.1, with $X \in \mathbb{R}^{n\times d}$, $d \geq n$, and $\mathrm{rank}(X) = n$. For any fixed mixing weight $\lambda \in [0, 1]$, let $\widehat{\beta}_{\mathrm{mix}}(\lambda)$ be the minimum-ℓ_2-norm linear least-squares fit to the full-pair data $(X^{\mathrm{mix}}, y^{\mathrm{mix}})$ constructed as in Definition 2.2, with no ridge penalty, and let $\widehat{\beta}_{\mathrm{ridgeless}}$ be the corresponding fit to the ordinary data (X, y). Then

$$\widehat{\beta}_{\mathrm{mix}}(\lambda) = \widehat{\beta}_{\mathrm{ridgeless}} = X^\dagger y,$$

where $(\cdot)^\dagger$ denotes the Moore–Penrose inverse. The two student models therefore give *identical* predictions, training errors, and generalization risks.

Proposition 1.1 ensures that dimensional overparameterization alone does *not* yield a benefit from full-pair mixup in this setting. This invariance provides a baseline for investigating how mixup acts on the additional nonlinear directions present under *structural overparameterization*.

For a nonlinear student model, mixup tests a correction that is useful only at x_i at points along every chord joining x_i to the other observations. Figure 1 illustrates how these additional constraints act on a localized spike and compares their effect with scalar ridge.

The figure alone does not establish that the local change in training space reduces prediction error at a new covariate. It also leaves open whether mixup simply changes the effective amount of global shrinkage. If so, separately tuning scalar ridge should remove its apparent advantage. Resolving these questions requires a risk theory that accounts for the dependence among the n^2 mixed pairs and allows each method to optimize its own ridge parameter.

Our contributions. The main contributions of this paper are summarized as follows.

1. **A structural overparameterization perspective on mixup.** We introduce a novel perspective on the generalization benefits of mixup through the distinction between *dimensional* and *struc-*

tural overparameterization. For a linear teacher and a nonlinear student model, we derive, in Section 3, precise characterizations of the high-dimensional prediction risk for full-pair mixup as the input dimension d, number of random features p, and the sample size n grow proportionally.

2. **Generalization benefits of mixup beyond optimally tuned ridge.** In Section 4, we identify explicit regimes in which mixup *strictly outperforms* ridge regression on the same set of random features, with the ridge parameter optimized separately for each method. The risk decomposition in Corollaries 4.1 and 4.2 explains this gain through an improved tradeoff between signal reconstruction error and residual fitting error. Numerical results in Section 5 and Appendix S.6 support both the asymptotic formulas and the risk comparisons. Section 5.2 then explores instance-selective mixup empirically in neural networks with learned features.

3. **Novel RMT results for learning from dependent data.** From a technical perspective, mixup creates dependence across reused training observations. This dependence is not covered by augmentation analyses based on independent source blocks (Morisset et al., 2026); see Remark 2 in the appendix for a detailed comparison. Building upon random matrix techniques (Liao & Mahoney, 2026; Couillet & Liao, 2022), we combine a decomposition of pairwise interactions with leave-many-out arguments to control this dependence and obtain precise risk limits. The resulting approach may be of independent interest for random matrix analyses of ML methods beyond the standard i.i.d. setting.

1.1 Related work

We briefly review related lines of research and position our contributions relative to prior work.

Mixup as interpolation and regularization. The original mixup formulation connected convex interpolation with reduced memorization of corrupted labels (Zhang et al., 2018). Subsequent analyses interpret it through data-dependent penalties (Zhang et al., 2021), correlated perturbations (Carratino et al., 2022), and directional-derivative regularization (Zou et al., 2023b). Our focus is a precise high-dimensional characterization of mixup prediction risk, which allows us to determine when mixup improves generalization beyond optimally tuned ridge baseline.

Data dependence and feature learning. Mixup exhibits distribution-dependent failure and equivalence regimes (Chidambaram et al., 2022) and can alter feature-learning dynamics (Zou et al., 2023a). Closest to our setting, Liu et al. (2023) study training dynamics with fixed random features and a full synthetic collection, including errors caused by interpolating a nonlinear labeling function. Our theoretical analysis considers a linear teacher, for which signal interpolation is exact, keeps the nonlinear feature map fixed, and retains the dependence among reused samples. This setting allows us to study structural overparameterization without feature learning or signal-interpolation error.

Data augmentation and high-dimensional analyses of ML models. Data augmentation can induce data-dependent spectral reshaping (Lin et al., 2024). Morisset et al. (2026) derive risk characterizations allowing dependence within independent source blocks. Here, our theory considers full-pair mixup (see Definition 2.2 below) that is close to practice yet challenging to analyze. See Remark 2 for further discussions. Building on precise random feature asymptotics (Rahimi & Recht, 2007; Mei & Montanari, 2022; Hastie et al., 2022) and the double descent and benign overfitting literature (Belkin et al., 2019; Bartlett et al., 2020), the proposed analysis connects shared-sample dependence to explicit prediction risk formulas that may be of independent technical interest.

2 Problem setup

In this section, we introduce the model and full-pair mixup procedure under study, as well as the working assumptions and quantities needed for our high-dimensional risk analysis.

Definition 2.1 (Noisy linear teacher) Suppose n training observations $(x_i, y_i) \subset \mathbb{R}^d \times \mathbb{R}$ are drawn from

$$x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d), \quad y_i = x_i^\top \beta^* + \varepsilon_i, \quad \varepsilon_i = \sigma_{n,i}\,\xi_i, \quad i = 1, \dots, n, \quad \overline{\sigma}^2_{\varepsilon,n} := \frac{1}{n}\sum_{i=1}^{n} \sigma^2_{n,i}, \tag{2.1}$$

where the teacher weights $\beta^* \in \mathbb{R}^d$ and noise scales $\sigma_{n,1}, \ldots, \sigma_{n,n}$ are deterministic but could be as large as $O(n^{1/2})$. $\overline{\sigma}^2_{\varepsilon,n}$ is the *average noise variance*. The standardized noises $\xi_1, \ldots, \xi_n$ are mutually independent and independent of all covariates, of zero mean, unit variance, and bounded eighth order moment.

Allowing the noise scales to vary with both i and n accommodates regimes in which a few observations have much larger noise variances than the rest. We can thus examine how noise heterogeneity affects mixup generalization performance under structural overparameterization; see the heteroscedastic teacher in Example 2 versus the homoscedastic teacher in Example 1. We consider the following full-pair mixup procedure.

Definition 2.2 (Full-pair mixup) For a *mixing weight*[1] $\lambda \in [1/2, 1]$, the *full-pair mixup* of a training set $\{(x_i, y_i)\}_{i=1}^n$ uses all n^2 pairs as in (1.1), to obtain the augmented set $\{(x_{ij}^{\text{mix}}, y_{ij}^{\text{mix}})\}_{i,j=1}^n$.

To compare against the linear baseline in Proposition 1.1, we introduce the following nonlinear random feature student to study the interaction between mixup and structural overparameterization beyond simple parameter count.

Definition 2.3 (Nonlinear student) Let $W \in \mathbb{R}^{d\times p}$ be independent of the training covariates and noises, with $W^{\mathsf{T}}W = I_p$. For a Lipschitz activation $\sigma : \mathbb{R} \to \mathbb{R}$, applied coordinatewise, define

$$\phi_W(x) = \sigma(W^{\mathsf{T}}x) \in \mathbb{R}^p, \qquad f_\beta(x) = \phi_W(x)^{\mathsf{T}}\beta, \qquad \beta \in \mathbb{R}^p.$$

Given training observations $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ and a ridge parameter $\gamma > 0$, we consider:

1. ***Vanilla* random feature student**, with $\widehat{\beta}(\gamma) = \operatorname{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n (y_i - \phi_W(x_i)^{\mathsf{T}}\beta)^2 + \gamma\|\beta\|^2$.
2. ***Mixup* random feature student**, for a fixed mixing weight $\lambda \in [1/2, 1]$, with $\widehat{\beta}_{\text{mix}}(\lambda, \gamma) = \operatorname{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{n^2}\sum_{i,j=1}^n (y_{ij}^{\text{mix}} - \phi_W(x_{ij}^{\text{mix}})^{\mathsf{T}}\beta)^2 + \gamma\|\beta\|^2$ from the full-pair mixup in Definition 2.2.

At $\gamma = 0$, each estimator is defined as the least-squares minimizer of smallest Euclidean norm, obtained using the Moore–Penrose inverses (as in Proposition 1.1 for linear student). For $\lambda = 1$, the mixup solution $\widehat{\beta}_{\text{mix}}(\lambda = 1, \gamma)$ reduces to vanilla solution $\widehat{\beta}(\gamma)$ with the same penalty γ.

Note from Definition 2.3 that the nonlinear mixup student $\widehat{\beta}_{\text{mix}}(\lambda, \gamma)$ applies the nonlinear activation *after* mixing the inputs. In general,

$$\sigma(\lambda W^{\mathsf{T}}x_i + (1-\lambda)W^{\mathsf{T}}x_j) \neq \lambda\sigma(W^{\mathsf{T}}x_i) + (1-\lambda)\sigma(W^{\mathsf{T}}x_j).$$

This is *in sharp contrast* to linear students, for which prediction commutes with interpolation, and full-pair mixup leaves the ridgeless solution unchanged in Proposition 1.1. This intuition is made quantitative in the remainder of the paper: Theorem 3.1 provides a precise high-dimensional characterization of the generalization risk of nonlinear students and distinguishes how mixup affects signal reconstruction error and residual fitting error. Section 4 then establishes conditions under which these changes *improve* prediction risk beyond (separately tuned) vanilla ridge solution.

Remark 1 (Structural overparameterization) For a linear teacher $x \mapsto x^{\mathsf{T}}\beta^*$, the nonlinear random feature student $f_\beta(x) = \sigma(W^{\mathsf{T}}x)^{\mathsf{T}}\beta$ in Definition 2.3 allows directions in function space that the teacher does *not* require, but that can fit observation noise. We call this mismatch *structural overparameterization*, distinct from *dimensional overparameterization*, measured by the parameter-to-sample ratio p/n. This terminology describes a mismatch in functional form and does not imply that the student class contains the teacher: even when $p = d$, the nonlinear feature map in Definition 2.3 need *not* fit the linear teacher exactly.

Definition 2.4 (Prediction risk) For a fresh $x_{\text{new}} \sim \mathcal{N}(0, I_d)$, let $y_{\text{new}} = x_{\text{new}}^{\mathsf{T}}\beta^* + \varepsilon_{\text{new}}$ as in Definition 2.1. The test noise ε_{new} is centered and independent of the test covariate, the training

[1] Note that every pair (i, j) appears together with (j, i), the objectives at λ and $1 - \lambda$ coincide. We therefore restrict the study to $\lambda \in [1/2, 1]$.

sample, and W. It has a fixed variance $\sigma^2_{\text{test}} \geq 0$. The (conditional) prediction risk is

$$\text{Risk}(\lambda, \gamma) = \mathbb{E}\Big[\Big(y_{\text{new}} - \phi_W(x_{\text{new}})^{\mathsf{T}}\widehat{\beta}_{\text{mix}}(\lambda, \gamma)\Big)^2 \,\Big|\, X, W\Big]. \tag{2.2}$$

The expectation in (2.2) averages over the training noises and the fresh test pair, and remains random through X, W. In particular, this risk measures the average prediction error of fitting noisy labels, for example, the cost of instance-level nonlinear fitting under structural overparameterization in Example 1 versus Example 2 in Section 4.

3 Main results

In this section, we present our main technical results on the prediction risk (in the sense of Definition 2.4) of mixup nonlinear students. We consider the following high-dimensional regime.

Assumption 1 (High-dimensional asymptotics) As $n \to \infty$, the input dimension $d = d(n)$ and number of features $p = p(n)$ satisfy $1 \leq p \leq d$ and $p/n \longrightarrow c \in (0, \infty)$. The teacher model in Definition 2.1 satisfies $\sup_n \overline{\sigma}^2_{\varepsilon,n} < \infty$ and $\sup_n \|\beta^*\| < \infty$.

Recall from its definition in (2.1), Assumption 1 permits individual noise variances $\sigma^2_{n,i}$ to be as large as order n, despite that the average noise variance satisfies $\overline{\sigma}^2_{\varepsilon,n} = O(1)$. As a consequence, a few noisy observations can have a *nonvanishing* effect on the prediction risk. In Section 4, we use two examples to examine how mixup changes label-noise sensitivity under structural overparameterization, while keeping the level of dimensional overparameterization p/n fixed.

Write $y_i = (W^{\mathsf{T}}x_i)^{\mathsf{T}}(W^{\mathsf{T}}\beta^*) + \zeta_i$, and $\zeta_i := x_i^{\mathsf{T}}(I_d - WW^{\mathsf{T}})\beta^* + \varepsilon_i$. We call ζ_i the training residual relative to the teacher signal visible through $W^{\mathsf{T}}x_i$. Conditional on W, these residuals are centered and independent of the projected training covariates, with average variance

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\zeta_i^2 \mid W] = \|(I_d - WW^{\mathsf{T}})\beta^*\|^2 + \overline{\sigma}^2_{\varepsilon,n}.$$

We define the two risk coefficients through two simple fits. On independent inputs $g_i \sim \mathcal{N}(0, I_p)$, use the mixup student in Definition 2.3 with feature map $\sigma(g)$ and the same (λ, γ):

Signal only: fit labels $y_i = g_{i1}$, the first input coordinate; call the fitted predictor $\widehat{f}_{\text{sig}}$.

Residuals only: fit labels $y_i = \xi_i$, the independent, standardized noises in Definition 2.1; call the fitted predictor $\widehat{f}_{\text{res}}$, with test target is zero.

For an independent test input $g_{\text{new}} \sim \mathcal{N}(0, I_p)$, the limiting errors define

$$\mathbb{E}\Big[\{\widehat{f}_{\text{sig}}(g_{\text{new}}) - g_{\text{new},1}\}^2 \mid g_1, \ldots, g_n\Big] \xrightarrow{\mathbb{P}} \mathcal{E}(\lambda, \gamma), \qquad \mathbb{E}\Big[\widehat{f}_{\text{res}}(g_{\text{new}})^2 \mid g_1, \ldots, g_n\Big] \xrightarrow{\mathbb{P}} \chi(\lambda, \gamma), \tag{3.1}$$

where expectations average the test input and any random training labels. Appendix S.2.1 establishes these limits under the assumptions of Theorem 3.1; Appendix S.1.3 gives the fixed point equation that determines the two quantities with explicit dependence on c and σ.

Theorem 3.1 (High-dimensional prediction risk) Under Assumption 1, suppose σ is odd and Lipschitz. For every fixed mixing weight $\lambda \in [1/2, 1]$ and regularization strength $\gamma > 0$, the prediction risk in Definition 2.4 satisfies $\text{Risk}(\lambda, \gamma) - \overline{\text{Risk}}(\lambda, \gamma) \to 0$ in probability as $n \to \infty$, where $\overline{\text{Risk}}(\lambda, \gamma)$ is

$$\underbrace{\mathcal{E}(\lambda, \gamma) \cdot \|W^{\mathsf{T}}\beta^*\|^2}_{\text{signal reconstruction error}} + \underbrace{\chi(\lambda, \gamma) \cdot \Big(\|(I_d - WW^{\mathsf{T}})\beta^*\|^2 + \overline{\sigma}^2_{\varepsilon,n}\Big)}_{\text{residual fitting error}} + \underbrace{\|(I_d - WW^{\mathsf{T}})\beta^*\|^2 + \sigma^2_{\text{test}}}_{\text{irreducible error}}. \tag{3.2}$$

By the construction of $\mathcal{E}$ and χ, the three terms in Theorem 3.1 admit the following interpretation:

1. $\mathcal{E}(\lambda, \gamma)$ is the *signal reconstruction error coefficient*, which measures prediction error per unit of teacher "signal strength" visible to the random feature student, and depends on the nonlinearity $\sigma(\cdot)$, the mixing weight λ, and the regularization strength γ. This term includes the error incurred by the nonlinear student in reconstructing the visible linear signal.
2. $\chi(\lambda, \gamma)$ is the *residual fitting error coefficient*, which measures the prediction error induced by fitting the residuals ζ_i, per unit of their average variance. The second term in (3.2) is the resulting *residual fitting error* on a fresh test observation. For fixed (λ, γ) and test noise variance, $\chi(\lambda, \gamma)$ also gives the increase in asymptotic risk per unit increase in average training noise variance $\overline{\sigma}^2_{\varepsilon,n}$.
3. The final term is an *irreducible error* for *any* predictor based on $W^\top x$.

Comparing $\lambda \in [1/2, 1)$ with the unmixed case $\lambda = 1$, Theorem 3.1 reveals how mixup changes the *tradeoff* between signal reconstruction error and residual fitting error. It also explains why measuring *dimensional overparameterization* using the augmented sample count can be misleading: the (limiting) prediction risk $\overline{\text{Risk}}(\lambda, \gamma)$ depends on the ratio $c = \lim p/n$, rather than p/n^2.

In Section 4, we identify explicit regimes in which mixup achieves *strictly lower* risk than vanilla non-mixup students, even when the ridge strength is optimized separately for each method. In these regimes, the benefit of mixup *cannot* be reproduced by adjusting a scalar ridge penalty alone.

4 The benefits of mixup: an instance-level overfitting perspective

In this section, we examine the benefits of mixup through the prediction error associated with fitting noise in individual training labels. Our main theoretical result in Theorem 3.1 connect this *instance-level overfitting* with out-of-sample prediction risk. We first illustrate this connection in Section 4.1 through two canonical teacher models, then examine in Section 4.2 whether the benefit of mixup persists when ordinary random feature ridge chooses its regularization strength optimally.

4.1 Instance-level overfitting and prediction risk

We begin with the homoscedastic teacher, which provides a baseline with the same level of noise variance across training observations, defined as follow.

Example 1 (*Homoscedastic* teacher) Consider the noisy linear teacher in Definition 2.1 with a *homoscedastic* variance profile, that is $\sigma_{n,i} = \sigma_\varepsilon$ for all i, and test noise variance satisfying $\sigma^2_{\text{test}} = \sigma^2_\varepsilon$. In this case, the *average noise variance* of the teacher is $\overline{\sigma}^2_{\varepsilon,n} = \sum_{i=1}^n \sigma^2_{n,i}/n = \sigma^2_\varepsilon$.

Applying Theorem 3.1 to this homoscedastic variance profile gives the following risk expression.

Corollary 4.1 (High-dimensional prediction risk of full-pair mixup for nonlinear students under *homoscedastic* teacher) Under the assumptions and notations of Theorem 3.1, for the homoscedastic teacher in Example 1 with average noise variance σ^2_ε, the full-pair mixup high-dimensional prediction risk satisfies

$$\overline{\text{Risk}}(\lambda, \gamma) = \overline{\text{Risk}}_{\text{hom}}(\lambda, \gamma) := \mathcal{E}(\lambda, \gamma) \cdot \|W^\top \beta^*\|^2 + (\chi(\lambda, \gamma) + 1) \cdot \left(\|(I_d - WW^\top)\beta^*\|^2 + \sigma^2_\varepsilon \right). \quad (4.1)$$

The factor $\chi + 1$ combines two contributions: $\chi(\lambda, \gamma)(\|(I_d - WW^\top)\beta^*\|^2 + \sigma^2_\varepsilon)$ is the residual fitting error, and $\|(I_d - WW^\top)\beta^*\|^2 + \sigma^2_\varepsilon$ is the irreducible error. Thus, for fixed (λ, γ), increasing the common training and test noise variance increases total risk at rate $\chi(\lambda, \gamma) + 1$. This baseline allows us to isolate the additional cost of fitting an "unusually" noisy training label, as illustrated in Figure 1. To this end, we increase the noise variance at one observation, keeping the remaining training and test noise levels unchanged, see Figure 1 for an illustration.

Example 2 (*Heteroscedastic* teacher) Consider the noisy linear teacher in Definition 2.1 with a *heteroscedastic* variance profile, that is $\sigma^2_{n,i} = \sigma^2_\varepsilon$ for $i \neq i_*$ and $\sigma^2_{n,i_*} = \sigma^2_\varepsilon + n\tau^2$, and test

noise variance satisfying $\sigma^2_{\text{test}} = \sigma^2_\varepsilon$. Here, $\tau^2 \geq 0$ denotes the *noise spike strength*, the index i_* is deterministic, and the *average noise variance* of the teacher is $\overline{\sigma}^2_{\varepsilon,n} = \sum_{i=1}^n \sigma^2_{n,i}/n = \sigma^2_\varepsilon + \tau^2$.

Apply again Theorem 3.1 to this heteroscedastic model gives the corresponding risk expression, with the homoscedastic risk given in Corollary 4.1 serving as the background term.

Corollary 4.2 (High-dimensional prediction risk under *heteroscedastic* teacher) Under the assumptions and notations of Theorem 3.1, for the heteroscedastic teacher in Example 2 with *average noise variance* $\sigma^2_\varepsilon + \tau^2$, the full-pair mixup high-dimensional prediction risk satisfies

$$\overline{\text{Risk}}(\lambda, \gamma) = \overline{\text{Risk}}_{\text{het}}(\lambda, \gamma; \tau^2) := \overline{\text{Risk}}_{\text{hom}}(\lambda, \gamma) + \tau^2 \cdot \chi(\lambda, \gamma),$$

for the residual fitting error coefficient $\chi(\lambda, \gamma)$ given in Theorem 3.1, the homoscedastic teacher risk $\overline{\text{Risk}}_{\text{hom}}(\lambda, \gamma)$ in Corollary 4.1, and noise spike strength $\tau^2 \geq 0$.

The additional term $\tau^2\chi(\lambda, \gamma)$ is the residual fitting error induced by the added noise at a single training observation. Because the test noise variance remains fixed, the irreducible error is unchanged, and total risk increases at rate $\chi(\lambda, \gamma)$ per unit of noise spike strength. This connects the residual fitting error coefficient to instance-level overfitting.[2]

4.2 Mixup versus ordinary random feature regression under optimal regularization

Ridge regularization also controls the influence of noisy training labels. To assess the benefit of mixup after accounting for this effect, we allow each method to optimize its ridge strength separately. By (1.1), ordinary random feature ridge corresponds to $\lambda = 1$, while midpoint mixup corresponds to $\lambda = 1/2$, averaging both the covariates and labels of two training observations. Define, for each mixing weight $\lambda \in [1/2, 1]$, the optimally regularized risk

$$\overline{\text{Risk}}^*_{\text{het}}(\lambda; \tau^2) := \inf_{\gamma > 0} \overline{\text{Risk}}_{\text{het}}(\lambda, \gamma; \tau^2) \tag{4.2}$$

where $\overline{\text{Risk}}_{\text{het}}(\lambda, \gamma; \tau^2)$ is the high-dimensional prediction risk given in Corollary 4.2.

The following result shows that mixup can *strictly* improve on vanilla random feature ridge regression after each method optimizes its ridge strength separately. This improvement already holds for the fixed midpoint mixing weight, without further optimization over λ.

Theorem 4.3 (Midpoint mixup versus optimally regularized non-mixup solution: informal) Under the assumptions and notations of Theorem 3.1, consider the heteroscedastic teacher model in Example 2. There exist a nonaffine activation, a dimension ratio, and teacher parameters with $\|W^\top \beta^*\|^2 > 0$, such that, for every fixed $\tau^2 \geq 0$,

$$\overline{\text{Risk}}^*_{\text{het}}(\lambda = 1/2; \tau^2) < \overline{\text{Risk}}^*_{\text{het}}(\lambda = 1; \tau^2). \tag{4.3}$$

Corollary S.4.14 in the appendix gives *explicit* sufficient conditions for this comparison. Under the stated conditions, mixup improves the tradeoff between signal reconstruction error and residual fitting error, achieving lower prediction risk than ordinary random feature ridge after each method optimizes its ridge strength separately. The comparison includes $\tau^2 = 0$, connecting this conclusion to the homoscedastic baselinein Example 1.

The proof makes this comparison concrete by constructing a midpoint mixup solution whose risk is already below that of the best ordinary ridge solution. Optimizing the midpoint ridge strength can only improve this comparison. The construction is given in Appendix S.4.4; Proposition S.4.10 provides results on the ordinary ridge solution.

Allowing the mixing weight to vary can, of course, only even improve on the midpoint choice. The same comparison therefore extends directly to mixup with jointly optimized mixing weight and ridge strength: for every fixed $\tau^2 \geq 0$, $\inf_{\lambda \in [1/2,1]} \overline{\text{Risk}}^*_{\text{het}}(\lambda; \tau^2) \leq \overline{\text{Risk}}^*_{\text{het}}(1/2; \tau^2) < \overline{\text{Risk}}^*_{\text{het}}(1; \tau^2)$. Section 5.1 illustrates the corresponding risk gap.

[2] These risks average over training noise, as in Definition 2.4. For fixed (λ, γ) and a realized standardized spike ξ_{i_*}, the additional risk is asymptotically $\tau^2\chi(\lambda, \gamma)\xi^2_{i_*}$; see Appendix S.4.3.

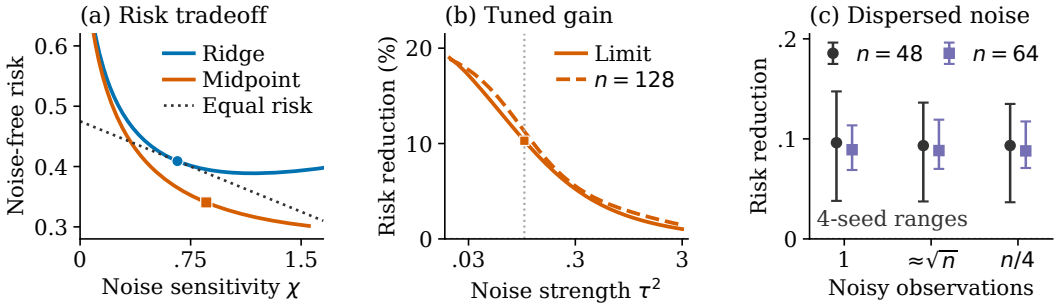

Figure 2: Risk improvement after separate ridge tuning. (a) Limiting noise-free risk versus noise sensitivity at $\tau^2 = 0.1$. (b) Midpoint risk reduction relative to scalar ridge: deterministic limit and one finite design with $(n, p) = (128, 160)$. (c) Absolute risk reduction for jointly tuned mixup when fixed average noise variance 0.05 is spread over one, a few, or many observations. Points show the mean risk reduction across these four samples; bars span the minimum and maximum risk reductions over four repetitions.

5 Numerical studies

We first test the predicted advantage over separately tuned scalar ridge, then examine instance-selective mixup with learned neural features.

5.1 Risk improvement beyond scalar ridge

We use isotropic Gaussian covariates, independent orthonormal tanh features, $p/n = 1.25$, and a unit-norm linear teacher. One predetermined training label has noise variance $n\tau^2$; the other training labels and test labels are noiseless. Corollary 4.2 then gives the limiting risk $\mathcal{E}(\lambda, \gamma) + \tau^2\chi(\lambda, \gamma)$. Figure 2(a) plots the two coefficients along the endpoint (ridge-only) and midpoint ridge paths. At $\tau^2 = 0.1$, the total risk at a point $(\chi, \mathcal{E})$ is $\mathcal{E} + 0.1\chi$. The best scalar-ridge risk, along the endpoint path $\lambda = 1$, is $R^\star_{\text{ridge}} = \inf_{\gamma \geq 0}\{\mathcal{E}(1, \gamma) + 0.1\chi(1, \gamma)\}$, so its equal-risk contour in this plot is $\mathcal{E} = R^\star_{\text{ridge}} - 0.1\chi$. The optimized midpoint point lies below this contour, and therefore has lower total risk than every ridge-only choice.

Panel (b) quantifies this advantage as the noise strength τ^2 varies; the finite-design curve closely tracks the deterministic limit at $n = 128$. Each method tunes its ridge strength separately. A further limiting-risk search over $\lambda \in [1/2, 1]$ selects the midpoint throughout the evaluated $\tau^2 \in [0, 2]$ grid; full tuning details appear in Appendix S.6.

Panel (c) spreads noise over different numbers of observations at each fixed n, keeping the average variance at 0.05. The risk improvement persists at both displayed sample sizes, consistent with the average-noise term in Theorem 3.1. Appendix S.6 also gives weak-ridge scans and anisotropic-covariate stress tests. Under strong anisotropy, gains persist for signals in the bulk eigenspace but diminish as the signal concentrates in the leading covariance direction, where the joint search returns to scalar ridge.

5.2 Instance-selective mixup for neural networks

Our analysis motivates concentrating mixup on observations whose label noise may be overfit. We explore this idea with a learned tanh neural network. A preliminary empirical risk minimization (ERM) fit assigns each observation a fitting-progress score: the decrease in squared training error between early and later checkpoints (model parameter values saved at specified training iterations). We use high scores to identify observations whose label noise the model may be overfitting and select the top fraction q to form a fixed set S_q, while recognizing that they are not perfect scores. In a fresh fit, *selected mixup* draws an observation uniformly. If that observation belongs to S_q, its inputs and label are mixed with a uniform partner from S_q^c, using $\lambda \sim \text{Beta}(0.4, 0.4)$; otherwise it remains unmixed. We compare selected mixup with *ERM*, *tuned weight decay*, *uniform mixup* of

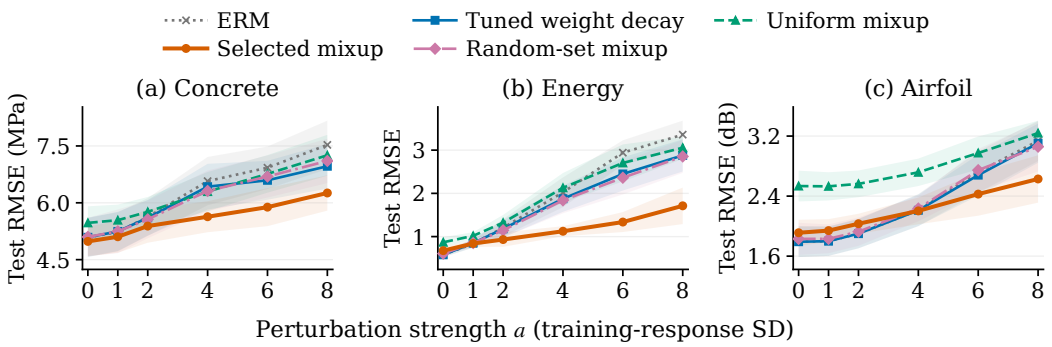

Figure 3: Learned tanh networks under increasing label-perturbation strength at 1% prevalence: (a) Concrete, (b) Energy, (c) Airfoil. Lines show mean test RMSE in original response units; bands show one standard deviation over ten paired splits. Each method uses validation-selected settings and checkpoints.

all samples, and *random-set mixup*, which substitutes a uniform random set for S_q. Each method selects checkpoints with the lowest validation mean squared error; all except zero-decay ERM tune weight decay on a common grid. Selected and random-set mixup also tune q on the same grid.

On Concrete, Energy, and Airfoil (Yeh, 1998; Tsanas & Xifara, 2012; Brooks et al., 1989), we use ten paired 60/20/20 training/validation/test splits. A uniform 1% of training labels receive independent random-sign perturbations $\pm a s_0$, where s_0 is the original training-response standard deviation. Holdout labels remain unchanged, and tuning precedes test scoring. Figure 3 shows the clearest benefits under stronger perturbations. At $a = 4$, Concrete's mean test root mean squared error (RMSE) is 5.63 for selected mixup versus 6.43 for tuned weight decay. Energy shows a larger relative gain; Airfoil's advantage emerges at stronger noise. At stronger perturbations, selected mixup also has lower mean error than uniform and random-set mixup on all three datasets. Appendix S.7 reports prevalence scans, rectified linear unit (ReLU) networks, and fixed-q comparisons: the latter favor selection on Concrete and Energy, while Airfoil is sensitive to set size.

6 Conclusions and Discussions

The effect of overparameterization depends on more than the parameter count. We study mixup under structural overparameterization, where ERM with a nonlinear student can overfit sample-specific label noise when the teacher is linear. Mixup constrains surplus nonlinear directions through chords between training observations. Accounting for dependence among full-pair mixup samples, our high-dimensional risk formulas quantify the resulting tradeoff between signal reconstruction error and residual fitting error. In the identified regimes, mixup outperforms scalar ridge even after both procedures are tuned separately. Numerical studies support these comparisons and explore their robustness. This perspective motivates selective mixup targeting observations whose label noise the model may be overfitting. Experiments on three real regression datasets show that this strategy can improve neural network prediction under controlled label perturbations.

Our theoretical guarantees rely on isotropic Gaussian covariates, a linear teacher, an independent orthonormal feature frame, and the full-pair squared-loss objective. A central direction is to extend the analysis to learned representations and training dynamics, characterizing when surplus model capacity learns task-relevant structure and when it fits instance-specific noise. This could guide adaptive mixup strategies that select observations, partners, and mixing weights throughout training, with generalization guarantees accounting for data-dependent selection. More broadly, studying nonlinear teachers, classification, and anisotropic or non-Gaussian designs would clarify how the benefit depends on the structural mismatch between the learning task and the fitted model. Efficient sampled-pair algorithms and principled data-driven tuning would further connect these developments to large-scale neural-network training.

AI USE STATEMENT

We used GPT-6 Astral to assist with writing and text editing, literature search and retrieval, and experimental code development. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final content of the paper.

REFERENCES

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. doi: 10.1073/pnas.1907378117.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116.

Thomas Brooks, D. Stuart Pope, and Michael Marcolini. Airfoil Self-Noise. UCI Machine Learning Repository, 1989.

Luigi Carratino, Moustapha Cissé, Rodolphe Jenatton, and Jean-Philippe Vert. On mixup regularization. *Journal of Machine Learning Research*, 23(325):1–31, 2022.

Muthu Chidambaram, Xiang Wang, Yuzheng Hu, Chenwei Wu, and Rong Ge. Towards understanding the data dependency of mixup-style training. In *International Conference on Learning Representations*, 2022.

Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022. ISBN 978-1-009-12323-5. doi: 10.1017/cbo9781009128490.

Victor H. de la Peña and Evarist Giné. *Decoupling: From Dependence to Independence*. Springer, New York, 1999. doi: 10.1007/978-1-4612-0537-1.

Victor H. de la Peña and Stephen J. Montgomery-Smith. Decoupling inequalities for the tail probabilities of multivariate U-statistics. *The Annals of Probability*, 23(2):806–816, 1995. doi: 10.1214/aop/1176988291.

Sjoerd Dirksen, Johannes Maly, and Holger Rauhut. Covariance estimation under one-bit quantization. arXiv preprint arXiv:2104.01280, 2022. Version 2; theorem numbering follows this version.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022. doi: 10.1214/21-AOS2133.

Seong-Hyeon Hwang, Minsu Kim, and Steven Euijong Whang. RC-Mixup: A data augmentation strategy against noisy data for regression tasks. *arXiv preprint arXiv:2405.17938*, 2024. doi: 10.48550/arXiv.2405.17938.

Marius Junge and Quanhua Xu. Noncommutative burkholder/rosenthal inequalities. *The Annals of Probability*, 31(2):948–995, 2003. doi: 10.1214/aop/1048516542.

Marius Junge and Quanhua Xu. Noncommutative Burkholder/Rosenthal inequalities II: Applications. *Israel Journal of Mathematics*, 167(1):227–282, 2008. doi: 10.1007/s11856-008-1048-4.

Zhenyu Liao and Michael W. Mahoney. Random matrix theory for deep learning: Beyond eigenvalues of linear models. *IEEE Signal Processing Magazine*, 43(2):93–106, 2026. doi: 10.1109/MSP.2025.3618012.

Chi-Heng Lin, Chiraag Kaushik, Eva L. Dyer, and Vidya Muthukumar. The good, the bad and the ugly sides of data augmentation: An implicit spectral regularization perspective. *Journal of Machine Learning Research*, 25(91):1–85, 2024.

Zixuan Liu, Ziqiao Wang, Hongyu Guo, and Yongyi Mao. Over-training with mixup may hurt generalization. In *International Conference on Learning Representations*, 2023.

Sho Matsumoto. General moments of the inverse real Wishart distribution and orthogonal Weingarten functions. *Journal of Theoretical Probability*, 25(3):798–822, 2012. doi: 10.1007/s10959-011-0340-0. Theorem 2 and Appendix A are cited from arXiv:1004.4717v3.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022. doi: 10.1002/cpa.22008.

Stanislav Minsker and Xiaohan Wei. Moment inequalities for matrix-valued U-statistics of order 2. arXiv preprint arXiv:1801.05921, 2019. Version 4; lemma numbering follows this version.

Lucas Morisset, Alain Durmus, and Adrien Hardy. Characterizing the generalization error of random feature regression with arbitrary data-augmentation. *arXiv preprint arXiv:2605.10290*, 2026.

Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994. doi: 10.1214/aop/1176988477. Revised author draft available as arXiv:1208.2200v2.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pp. 1177–1184, 2007.

Mark Rudelson and Roman Vershynin. The smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739, 2009. doi: 10.1002/cpa.20294.

Jack W Silverstein and Zhi Dong Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate analysis*, 54(2):175–192, 1995.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z.

Athanasios Tsanas and Angeliki Xifara. Energy Efficiency. UCI Machine Learning Repository, 2012.

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

Huaxiu Yao, Yiping Wang, Linjun Zhang, James Y. Zou, and Chelsea Finn. C-Mixup: Improving generalization in regression. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0243.

I-Cheng Yeh. Concrete Compressive Strength. UCI Machine Learning Repository, 1998.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Zou. How does mixup help with robustness and generalization? In *International Conference on Learning Representations*, 2021.

Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. The benefits of mixup for feature learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 43423–43479. PMLR, 2023a.

Yingtian Zou, Vikas Verma, Sarthak Mittal, Wai Hoh Tang, Hieu Pham, Juho Kannala, Yoshua Bengio, Arno Solin, and Kenji Kawaguchi. MixupE: Understanding and improving mixup from directional derivative perspective. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2597–2607. PMLR, 2023b.

An overview

The appendix presents the proofs of the theoretical results, additional risk formulas, and further experimental details. Section S.1 introduces the spectral notation, fixed-point equations, and resolvents used in the proofs. Section S.2 first proves the main random feature risk theorem. Section S.3 then gives the ridgeless, noncentered positive-ridge, and Haar-frame extensions alongside their proofs, followed by the technical estimates supporting these results. Section S.4 proves the linear benchmark, the instance-level influence identities, and the risk comparisons with optimally tuned ridge. Section S.5 collects the standard matrix identities, concentration inequalities, and random matrix and analytic tools used in these arguments. Section S.6 reports additional random feature numerical studies, including tuning protocols and stress tests for covariate anisotropy and heterogeneous noise. Section S.7 details the neural-network experiments, including data preparation, observation selection, training, validation, and additional comparisons.

Guide to the proofs of the main results. The following list maps each main result to its proof in the order used in the main text.

(i) Proposition 1.1 is proved in Section S.4.1. The original and full-pair linear least-squares problems have the same set of interpolating solutions, so their minimum-norm estimators coincide.

(ii) Theorem 3.1 is proved in Section S.2.1. Resolvent estimates give the signal contribution to prediction risk, while a uniform bound on each observation's noise sensitivity reduces the noise contribution to the average training variance. Substitution into the exact risk decomposition yields the stated deterministic equivalent. The centered resolvent estimate used in this proof is established in Section S.3.5.

(iii) Corollary 4.1 is proved at the end of Section S.2.1 by specializing Theorem 3.1 to equal training and test noise variances.

(iv) Corollary 4.2 is proved in Section S.4.3. Increasing one training noise variance by $n\tau^2$ raises the average variance by τ^2; substituting this profile into Theorem 3.1 gives the additional risk $\tau^2\chi(\lambda,\gamma)$.

(v) Theorem 4.3 is proved in Section S.4.4. The argument first characterizes the optimally regularized non-mixup risk, then constructs a midpoint ridge choice with strictly smaller risk. Corollary S.4.14 provides the sufficient conditions for this construction; optimizing the midpoint ridge strength preserves the strict improvement.

Appendix contents

S.1 Random Matrix Preliminaries, Fixed Points, and Resolvents **3**

S.1.1 Spectral measures, Stieltjes transforms, and resolvents 3

S.1.2 Full-pair decomposition and effective channels 3

S.1.3 Limiting-risk calculation dictionary 3

S.1.4 Reduced fixed point and finite-dimensional coefficients 5

S.1.5 The two-scalar fixed-point system 6

S.2 Proof of the Main Random Feature Risk Theorem **6**

S.2.1 Centered odd-activation theorem 6

S.3 Supplementary Random Feature Results and Technical Estimates **8**

S.3.1 Ridgeless results 8

S.3.2 Positive-ridge generalization 12

S.3.3 Noncentered ridgeless continuation 16

S.3.4 Haar-frame specialization 17
S.3.5 Centered score–resolvent estimates 17
S.3.6 Supporting lemmas for the fixed-ridge proofs 28
S.3.7 Supporting lemmas for zero ridge and Haar averaging 81
S.3.8 Technical lemmas used in the supporting arguments 85

S.4 Proofs of the Mechanism Results **92**
S.4.1 Linear-model baseline 92
S.4.2 Finite-sample transfer 92
S.4.3 Critical-coordinate identity and finite-sample oracle transfer 93
S.4.4 Ridgeless and oracle comparisons 101
S.4.5 Additional theorem-matched diagnostics 111

S.5 Auxiliary Lemmas and Standard Results **111**
S.5.1 Deterministic matrix and resolvent identities 111
S.5.2 Moment and concentration inequalities 112
S.5.3 Matrix concentration and dependent sums 112
S.5.4 Standard random-matrix limits 113
S.5.5 Haar projection facts 113
S.5.6 Analytic tools 113

S.6 Additional random feature numerical studies **114**
S.6.1 Experimental setup and finite-sample comparison 114
S.6.2 Joint tuning of mixing weight and ridge strength 114
S.6.3 Heterogeneous training-noise profiles 115
S.6.4 Anisotropic covariates 115
S.6.5 Weak regularization near interpolation 116

S.7 Learned-feature neural-network experiments **117**
S.7.1 Data and controlled perturbations 117
S.7.2 Selection rule and comparison methods 118
S.7.3 Training and validation 119
S.7.4 Perturbation strength and prevalence on Concrete 119
S.7.5 Other datasets and fixed-q comparisons 119
S.7.6 Selection and noise-fitting diagnostics 120
S.7.7 ReLU extension 120

S.1 Random Matrix Preliminaries, Fixed Points, and Resolvents

S.1.1 Spectral measures, Stieltjes transforms, and resolvents

Write $c_n := p/n$, so that $c_n \to c$. For a symmetric matrix $M \in \mathbb{R}^{p\times p}$ with empirical spectral distribution $\mu_M = p^{-1}\sum_{k=1}^{p}\delta_{\lambda_k(M)}$, we use the Stieltjes-transform convention

$$m_M(z) = \int \frac{1}{x-z}\, d\mu_M(x) = \frac{1}{p}\operatorname{tr}(M - zI_p)^{-1}, \qquad z \in \mathbb{C}\setminus \operatorname{supp}(\mu_M).$$

For the shifted resolvent

$$Q_M(\omega) = \{M + (\gamma+\omega)I_p\}^{-1},$$

the normalization used in the proofs is $n^{-1}\operatorname{tr} Q_M(\omega) = c_n m_M(-\gamma-\omega)$. In particular,

$$\frac{d}{d\omega}\frac{1}{n}\operatorname{tr} Q_M(\omega) = -\frac{1}{n}\operatorname{tr} Q_M(\omega)^2.$$

We use the same convention for the Marchenko–Pastur transform in the endpoint calculations. Basic finite-rank resolvent identities appear in Section S.5.1.

S.1.2 Full-pair decomposition and effective channels

We first compare full-pair mixup with augmentation based on independent source blocks. The following remark highlights the dependence created by reusing training observations and motivates our pairwise decomposition and leave-many-out analysis.

Remark 2 (Comparison to Morisset et al. (2026)) Morisset et al. (2026) consider augmentation driven by randomness *independent* of the training sample, yielding independent blocks across source observations. This *independence* allows their leave-two-out analysis to statistically "decouple" blocks from the retained data and derive limiting prediction risks. Full-pair mixup in Definition 2.2 instead constructs mixup samples from the *same* training sample. This better reflects the empirical pairing *commonly used in practice*, but is more challenging to analyze, since each observation now appears in many overlapping pairs, and deleting individual mixed *rows* does not restore independence from their source observations. To address this challenge, we develop a *leave-many-out* reduction in Lemma S.3.22: removing two original observations deletes all $4n-4$ associated mixed rows. Despite this growing deletion size, we show that the resulting difference is approximately of rank four, plus a remainder that is shown to vanish uniformly.

For independent standard Gaussians G, G', define

$$\begin{aligned} m_\lambda(u) &= \begin{pmatrix}\mathbb{E}_G\sigma(\lambda u + (1-\lambda)G)\\ \mathbb{E}_G\sigma(\lambda G + (1-\lambda)u)\end{pmatrix}, \qquad \bar m_\lambda = \mathbb{E}\{m_\lambda(G')\},\\ V_\lambda &= \operatorname{Cov}\{m_\lambda(G')\}, \qquad \kappa(\lambda) = \operatorname{Var}[\sigma(\lambda G + (1-\lambda)G')] - \operatorname{tr}(V_\lambda). \end{aligned} \tag{S.1.1}$$

The entries of m_λ average the mixed feature while fixing either observation. The *interaction factor* $\kappa(\lambda)$ is the variance left after removing both conditional averages. The two coordinates of m_λ in (S.1.1) are the conditional endpoint means denoted by f_1, f_2. Suppress the subscript λ and write

$$V = V_\lambda, \qquad w = (\lambda, 1-\lambda)^{\mathrm{T}}.$$

The canonical Hoeffding interaction h_0 is the part of the mixed feature left after removing its grand mean and the two conditional endpoint means. Its covariance $\kappa(\lambda)I_p$ supplies the isotropic bulk channel, V describes the two reused endpoint channels, and a nonzero feature mean contributes a rank-one channel. These three components reduce the nominal n^2 rows to a fixed-point problem with aspect ratio p/n.

S.1.3 Limiting-risk calculation dictionary

We give explicit formulas for the coefficients defined in (3.1), together with the notation needed for the proofs and noncentered activations. Put

$$\alpha^2 = \lambda^2 + (1-\lambda)^2, \qquad G_\alpha \sim \mathcal{N}(0, \alpha^2),$$

$$\beta_\sigma = \mathbb{E}\sigma(G_\alpha), \qquad \nu_\alpha = \mathbb{E}\sigma(G_\alpha)^2, \qquad d_\alpha = \mathbb{E}\sigma'(G_\alpha).$$

For a Lipschitz activation, σ' is understood almost everywhere. Then $\bar{m}_\lambda = \beta_\sigma \mathbf{1}_2$ and $\kappa(\lambda) = \nu_\alpha - \beta_\sigma^2 - \operatorname{tr} V$. For an independent $G \sim \mathcal{N}(0,1)$, also define

$$\mu_0 = \mathbb{E}\sigma(G), \qquad \mu_1 = \mathbb{E}\{G\sigma(G)\}, \qquad \nu = \mathbb{E}\sigma(G)^2, \qquad \kappa_\sigma = \nu - \mu_0^2.$$

For the sample-side calculations, let

$$g_i := W^\top x_i, \qquad a := W^\top \beta^* \in \mathbb{R}^p,$$
$$\beta_\perp^* := (I_d - WW^\top)\beta^*, \qquad z_{ij} := \phi_W(x_{ij}^{\mathrm{mix}}) = \sigma\Big(\lambda g_i + (1-\lambda) g_j\Big).$$

Let $Z \in \mathbb{R}^{n^2 \times p}$ have row (i,j) equal to $z_{ij}^\top$, and let $\widetilde{y} \in \mathbb{R}^{n^2}$ have entry $\widetilde{y}_{ij} := y_{ij}^{\mathrm{mix}}$, with the ordered pairs indexed in the same way. Set

$$S_\lambda := \frac{1}{n^2} Z^\top Z, \qquad r_\lambda := \frac{1}{n^2} Z^\top \widetilde{y}, \qquad Q_{\lambda,\gamma} := (S_\lambda + \gamma I_p)^{-1} \quad (\gamma > 0), \qquad \widehat{z} := \frac{1}{n^2} \sum_{i,j=1}^{n} z_{ij}.$$

In centered Gram calculations, $H_n := I_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top$. Write $S = S_\lambda$, $Q = Q_{\lambda,\gamma}$, $r = r_\lambda$, and $H_\sigma = \kappa_\sigma I_p + \mu_0^2 \mathbf{1}\mathbf{1}^\top$. Define the endpoint score vectors and the conditional score moments by

$$\begin{aligned} \phi_i &= \frac{1}{n} \sum_{j=1}^{n} \{\lambda z_{ij} + (1-\lambda) z_{ji}\}, \qquad r = \frac{1}{n} \sum_i \phi_i y_i, \\ r_\mu := \mathbb{E}_\xi[r \mid X, W] &= \frac{1}{n} \sum_i \phi_i x_i^\top \beta^*, \qquad C_\varepsilon := \operatorname{Cov}_\xi(r \mid X, W) = \frac{1}{n^2} \sum_i \sigma_{n,i}^2 \phi_i \phi_i^\top. \end{aligned} \tag{S.1.2}$$

This expectation averages only the observation noise: the orthogonal teacher component $x_i^\top \beta_\perp^*$ is fixed conditional on X, W. The risk $\mathrm{Risk}_{\lambda,\gamma,n}$ below is $\mathrm{Risk}(\lambda,\gamma)$ from Definition 2.4; for fixed λ, γ, we also write Risk_n, or $\mathrm{Risk}_n(W)$ when displaying the frame dependence. Likewise, $\mathrm{Risk}_n(U)$ denotes this conditional risk with the frame fixed at U. Applying Gaussian integration by parts to the test covariate and then averaging over the training noise gives the exact identity

$$\mathrm{Risk}_{\lambda,\gamma,n} = r_\mu^\top Q H_\sigma Q r_\mu + \operatorname{tr}(Q H_\sigma Q C_\varepsilon) - 2\mu_1 a^\top Q r_\mu + \|\beta^*\|^2 + \sigma_{\mathrm{test}}^2. \tag{S.1.3}$$

Let $\mathrm{Risk}_{\lambda,\gamma,n}^{\mathrm{real}}$ instead denote the test risk conditional on both the realized training labels and X, W. Then

$$\mathrm{Risk}_{\lambda,\gamma,n}^{\mathrm{real}} = r^\top Q H_\sigma Q r - 2\mu_1 a^\top Q r + \|\beta^*\|^2 + \sigma_{\mathrm{test}}^2, \qquad \mathrm{Risk}_{\lambda,\gamma,n} = \mathbb{E}_\xi[\mathrm{Risk}_{\lambda,\gamma,n}^{\mathrm{real}} \mid X, W]. \tag{S.1.4}$$

The same identities hold at zero ridge with $Q = S_\lambda^\dagger$.

For the positive-ridge estimates in Sections S.1–S.3 that are uniform on a compact ridge interval, γ_- and γ_+ denote fixed constants with $0 < \gamma_- < \gamma_+ < \infty$ and $\gamma \in [\gamma_-, \gamma_+]$. Such an interval can always be chosen around a fixed $\gamma > 0$.

For positive ridge, let $t_* > 0$ be the unique solution of

$$c = t_* \Big[\gamma + \kappa(\lambda) + \operatorname{tr}\{V(I_2 + t_* V)^{-1}\}\Big]. \tag{S.1.5}$$

This solution is unique because the right-hand side is strictly increasing from zero to infinity.

Set

$$\begin{aligned} C_* &= (I_2 + t_* V)^{-1}, \qquad q = t_*/c, \\ \theta = d_\alpha w^\top C_* w, \qquad \tau_{\mathrm{fp}} &= \frac{t_*}{\gamma + \kappa(\lambda) + \operatorname{tr}(V C_*^2)}, \qquad \ell = \tau_{\mathrm{fp}} w^\top V C_*^2 w, \\ K &= \theta^2 \frac{\tau_{\mathrm{fp}}}{c} + (1 - 2 d_\alpha \theta q)\ell, \\ \mathfrak{K} = \kappa_\sigma K - 2\mu_1 \theta q + 1, \qquad \mathfrak{L} &= \kappa_\sigma \ell + 1, \\ \mathcal{E}(\lambda,\gamma) := \mathfrak{K}, \qquad \chi(\lambda,\gamma) &:= \mathfrak{L} - 1 = \kappa_\sigma \ell. \end{aligned} \tag{S.1.6}$$

Thus θq is the limiting teacher–prediction alignment, ℓ is residual-label susceptibility before multiplying by κ_σ, and $\mathcal{E}, \chi$ are the two coefficients used in the statement of Theorem 3.1 in the main text. For odd σ, $\mu_0 = 0$ and $\kappa_\sigma = \nu$.

Write $V = (v_{jk})_{j,k=1}^2$. For a noncentered activation, the extra feature-mean direction is stable under the explicit margin

$$\beta_\sigma \neq 0, \qquad 1 - t_* \chi_+ > 0, \qquad \chi_+ = \sqrt{v_{11} v_{22}} + v_{12}. \tag{S.1.7}$$

The inequality prevents the resolvent in the mean direction from diverging.

For an interior nonaffine weight at zero ridge, let $t_0 > 0$ solve

$$c = t_0 \left[\kappa(\lambda) + \operatorname{tr}\{V(I_2 + t_0 V)^{-1}\} \right], \tag{S.1.8}$$

and define

$$\begin{aligned} C_0 &= (I_2 + t_0 V)^{-1}, \quad q_0 = t_0/c, \quad \theta_0 = d_\alpha w^{\mathrm{T}} C_0 w, \\ \tau_0 &= \frac{t_0}{\kappa(\lambda) + \operatorname{tr}(V C_0^2)}, \quad \ell_0 = \tau_0 w^{\mathrm{T}} V C_0^2 w, \\ K_0^{\mathrm{rid}} &= \theta_0^2 \frac{\tau_0}{c} + (1 - 2 d_\alpha \theta_0 q_0) \ell_0, \\ \mathfrak{K}_0^{\mathrm{rid}} &= \kappa_\sigma K_0^{\mathrm{rid}} - 2\mu_1 \theta_0 q_0 + 1, \qquad \mathfrak{L}_0^{\mathrm{rid}} = \kappa_\sigma \ell_0 + 1, \\ \mathcal{E}(\lambda, 0) &:= \mathfrak{K}_0^{\mathrm{rid}}, \qquad \chi(\lambda, 0) := \mathfrak{L}_0^{\mathrm{rid}} - 1 = \kappa_\sigma \ell_0. \end{aligned} \tag{S.1.9}$$

S.1.4 Reduced fixed point and finite-dimensional coefficients

The proof uses finite-n versions of the limiting coefficients in (S.1.5) and (S.1.6). Throughout the supplementary noncentered branch, write

$$e = \frac{\mathbf{1}}{\sqrt{p}}, \qquad P_\perp = I_p - e e^{\mathrm{T}}, \qquad \kappa_\sigma = \nu - \mu_0^2,$$

and let $e^{\mathrm{T}} a \to \eta$ describe teacher alignment with the feature-mean direction. Under the bounded variance profile (S.3.10), we retain the training-label notation $\rho_n^2 = \|\beta^*\|^2 + \overline{\sigma}_{\varepsilon,n}^2$.

Let $t_{*,n} > 0$ be the unique solution of

$$c_n = t_{*,n} \left[\gamma + \kappa(\lambda) + \operatorname{tr}\{V(I_2 + t_{*,n} V)^{-1}\} \right], \tag{S.1.10}$$

and define

$$\begin{aligned} C_{*,n} &= (I_2 + t_{*,n} V)^{-1}, \qquad q_n = t_{*,n}/c_n, \\ \theta_n = d_\alpha w^{\mathrm{T}} C_{*,n} w, \qquad \tau_n &= \frac{t_{*,n}}{\gamma + \kappa(\lambda) + \operatorname{tr}(V C_{*,n}^2)}, \qquad \ell_n = \tau_n w^{\mathrm{T}} V C_{*,n}^2 w, \\ K_{\perp,n} &= K_{0,n} = \theta_n^2 \frac{\tau_n}{c_n} + (1 - 2 d_\alpha \theta_n q_n) \ell_n. \end{aligned} \tag{S.1.11}$$

These quantities converge to the coefficients in (S.1.6). Indeed, for

$$\mathfrak{f}(t) := t \left[\gamma + \kappa(\lambda) + \operatorname{tr}\{V(I_2 + tV)^{-1}\} \right],$$

one has $\mathfrak{f}(t_{*,n}) = c_n$, $\mathfrak{f}(t_*) = c$, and

$$\mathfrak{f}'(t) = \gamma + \kappa(\lambda) + \operatorname{tr}\{V(I_2 + tV)^{-2}\} \geq \gamma_-.$$

The mean value theorem therefore gives

$$|t_{*,n} - t_*| \leq \gamma_-^{-1} |c_n - c| \longrightarrow 0. \tag{S.1.12}$$

Consequently, by continuity,

$$\begin{aligned} C_{*,n} &\longrightarrow C_*, \qquad q_n \longrightarrow q, \qquad \theta_n \longrightarrow \theta, \\ \tau_n &\longrightarrow \tau_{\mathrm{fp}}, \qquad \ell_n \longrightarrow \ell, \qquad K_{\perp,n} = K_{0,n} \longrightarrow K. \end{aligned} \tag{S.1.13}$$

When $\beta_\sigma \neq 0$, we also use $r_{\perp,n}^2 = \|P_\perp a\|^2 = \|a\|^2 - (e^{\mathrm{T}} a)^2$. For the nonzero-mean proof, also set

$$\Delta_{*,n} = \mathbf{1}_2^{\mathrm{T}} C_{*,n} \mathbf{1}_2 - 1 = \frac{(1 - t_{*,n}\chi_+)(1 + t_{*,n}\chi_-)}{\det(I_2 + t_{*,n} V)}, \qquad \chi_- = \sqrt{v_{11} v_{22}} - v_{12}.$$

Condition (S.1.7) and (S.1.12) imply that $1 - t_{*,n}\chi_+$ and $\Delta_{*,n}$ are uniformly positive for all sufficiently large n.

S.1.5 The two-scalar fixed-point system

The proof uses the following four-dimensional representation. Put

$$J_k = \mathbf{1}_k \mathbf{1}_k^\top, \qquad v = (v_{11}, v_{12}, v_{12}, v_{22})^\top.$$

$$\bar{Q} = \left(\beta_\sigma^2(1 - \mathbf{1}_4^\top \bar{v})\mathbf{1}\mathbf{1}^\top + (v_\alpha - \beta_\sigma^2 - v^\top \bar{v})I_p + \gamma I_p\right)^{-1},$$

with

$$\bar{v} = \left[I_4 + \frac{1}{n}\operatorname{tr}(\bar{Q})(I_2 \otimes V) + \frac{\beta_\sigma^2}{n}\mathbf{1}^\top \bar{Q}\mathbf{1}J_4\right]^{-1}\left[\frac{1}{n}\operatorname{tr}(\bar{Q})v + \frac{\beta_\sigma^2}{n}\mathbf{1}^\top \bar{Q}\mathbf{1}\mathbf{1}_4\right].$$

Define

$$z_B = \begin{pmatrix} \lambda^2 \\ \lambda(1-\lambda) \\ \lambda(1-\lambda) \\ (1-\lambda)^2 \end{pmatrix}, \qquad c_B = z_B^\top \bar{v}.$$

Set

$$t = \frac{1}{n}\operatorname{tr}(\bar{Q}), \qquad s = \frac{1}{n}\mathbf{1}^\top \bar{Q}\mathbf{1},$$

and reserve these two letters for values at the deterministic fixed point throughout the calculations. We distinguish reduced, perturbed, random, and expectation-level counterparts by their subscripts or arguments. For free scalar arguments (τ, ς), define

$$m(\tau, \varsigma) = \left[I_4 + \tau(I_2 \otimes V) + \beta_\sigma^2 \varsigma J_4\right]^{-1}\left(\tau v + \beta_\sigma^2 \varsigma \mathbf{1}_4\right).$$

Thus $\bar{v} = m(t, s)$. Let $m_\tau = \partial_\tau m$ and $m_\varsigma = \partial_\varsigma m$, and define the scalar functions

$$a_\tau = v^\top m_\tau, \quad a_\varsigma = v^\top m_\varsigma, \qquad b_\tau = \mathbf{1}_4^\top m_\tau, \quad b_\varsigma = \mathbf{1}_4^\top m_\varsigma.$$

When no arguments are displayed, $a_\tau, a_\varsigma, b_\tau, b_\varsigma$ are evaluated at the fixed point (t, s). Define

$$K(u) = \left(v_\alpha - \beta_\sigma^2 - v^\top u\right) I_p + \beta_\sigma^2 (1 - \mathbf{1}_4^\top u)\, \mathbf{1}\mathbf{1}^\top, \qquad u \in \mathbb{C}^4.$$

Let

$$R_0 = \frac{1}{n}\operatorname{tr}(\bar{Q}^2), \qquad R_1 = \frac{1}{n}\mathbf{1}^\top \bar{Q}^2 \mathbf{1}, \qquad R_2 = \frac{1}{n}(\mathbf{1}^\top \bar{Q}\mathbf{1})^2,$$

and define

$$L = \begin{pmatrix} 1 - a_\tau R_0 - \beta_\sigma^2 b_\tau R_1 & -a_\varsigma R_0 - \beta_\sigma^2 b_\varsigma R_1 \\ -a_\tau R_1 - \beta_\sigma^2 b_\tau R_2 & 1 - a_\varsigma R_1 - \beta_\sigma^2 b_\varsigma R_2 \end{pmatrix}. \tag{S.1.14}$$

S.2 Proof of the Main Random Feature Risk Theorem

This section proves Theorem 3.1 and its homoscedastic specialization. The centered score–resolvent estimates used here are stated first and proved with the other technical estimates in Section S.3.5. The supplementary ridgeless, noncentered, and Haar results appear in Section S.3.

For realized training labels, the two risk components are $A = r^\top Q H_\sigma Q r$ and $B = -2\mu_1 a^\top Q r$. Hence (S.1.4) gives $\mathrm{Risk}_n = \mathbb{E}_\xi[A + B \mid X, W] + \|\beta^*\|^2 + \sigma_{\mathrm{test}}^2$. For the odd theorem, we use the clean-label equivalents below and handle the training noise contribution through its conditional covariance.

S.2.1 Centered odd-activation theorem

The next lemma gives the clean-label estimates used in the centered theorem. Its proof follows the supplementary results.

Lemma S.2.1 (Centered score–resolvent equivalents) Assume the standing high-dimensional conditions, let σ be odd and Lipschitz, and fix $\lambda \in [1/2, 1]$ and $\gamma > 0$. Take the clean labels

$y_i = x_i^{\mathrm{T}}\beta^*$, whose score is r_μ in (S.1.2). Let Γ be a sufficiently small fixed circle about zero with $\mathrm{Re}(\gamma + \omega) > 0$ on Γ. Define $t_n(\omega)$ by (S.3.30), set

$$C_{*,n}(\omega) = (I_2 + t_n(\omega)V)^{-1}, \quad q_n(\omega) = t_n(\omega)/c_n, \quad \theta_n(\omega) = d_\alpha w^{\mathrm{T}} C_{*,n}(\omega) w,$$

$$c_{B,n}(\omega) = w^{\mathrm{T}}\{I_2 - C_{*,n}(\omega)\}w, \qquad Q_I(\omega) = \{S + (\gamma + \omega)I_p\}^{-1}.$$

Then

$$\sup_{\omega\in\Gamma} \left\| a^{\mathrm{T}} Q_I(\omega) r_\mu - \theta_n(\omega) q_n(\omega) \|a\|^2 \right\|_{L^2} \longrightarrow 0,$$

$$\sup_{\omega\in\Gamma} \left\| r_\mu^{\mathrm{T}} Q_I(\omega) r_\mu - \{\theta_n(\omega)^2 q_n(\omega)\|a\|^2 + \|\beta^*\|^2 c_{B,n}(\omega)\} \right\|_{L^2} \longrightarrow 0.$$

These are the estimates recorded, with their original label, in (S.3.43) below.

S.2.1.1 Proof of the centered fixed-ridge theorem

Proof of Theorem 3.1. We apply the two uniform L^2 estimates in (S.3.43) to r_μ. The orthogonal teacher residual is Gaussian, with uniformly bounded moments, and its energy is $\|\beta_\perp^*\|^2$. We retain the fixed Cauchy circle Γ and write $Q = Q_I(0)$.

Cauchy's formula at zero applied to the first estimate gives

$$a^{\mathrm{T}} Q r_\mu - \theta_n q_n \|a\|^2 \xrightarrow{L^2} 0.$$

Differentiating (S.3.30) and its coefficient functions with respect to ω at zero, where $t_n(0) = t_{*,n}$, gives

$$t_n'(0) = -\tau_n, \qquad q_n'(0) = -\frac{\tau_n}{c_n}, \qquad \theta_n'(0) = d_\alpha \ell_n, \qquad c_{B,n}'(0) = -\ell_n.$$

Since $\|\beta^*\|^2 = \|a\|^2 + \|\beta_\perp^*\|^2$, the definition of $K_{0,n}$ in (S.1.11), Cauchy's formula, and Minkowski's inequality yield

$$\begin{aligned} &\left\| r_\mu^{\mathrm{T}} Q^2 r_\mu - \{K_{0,n}\|a\|^2 + \ell_n \|\beta_\perp^*\|^2\} \right\|_{L^2} \\ &\quad \le C_\Gamma \sup_{\omega\in\Gamma} \left\| r_\mu^{\mathrm{T}} Q_I(\omega) r_\mu - \{\theta_n(\omega)^2 q_n(\omega)\|a\|^2 + \|\beta^*\|^2 c_{B,n}(\omega)\} \right\|_{L^2} \longrightarrow 0, \end{aligned}$$

where $Q_I'(0) = -Q^2$ and C_Γ depends only on the fixed circle.

For the observation-noise contribution, oddness gives $H_\sigma = \nu I_p$. Define

$$L_{i,n} := \frac{\nu}{n} \phi_i^{\mathrm{T}} Q^2 \phi_i.$$

Lemma S.4.3 depends only on the design. The bounds are uniform over deterministic indices, so

$$\max_{1\le i\le n} \|L_{i,n} - \nu\ell_n\|_{L^2} \longrightarrow 0.$$

By (S.1.2) and Minkowski's inequality,

$$\begin{aligned} &\left\| \nu \,\mathrm{tr}(Q^2 C_\varepsilon) - \nu \ell_n \overline{\sigma}_{\varepsilon,n}^2 \right\|_{L^2} \\ &\quad = \left\| \frac{1}{n} \sum_i \sigma_{n,i}^2 (L_{i,n} - \nu\ell_n) \right\|_{L^2} \le \overline{\sigma}_{\varepsilon,n}^2 \max_i \|L_{i,n} - \nu\ell_n\|_{L^2} \longrightarrow 0. \end{aligned} \tag{S.2.1}$$

Substituting the clean-signal estimates and (S.2.1) into (S.1.3) gives the intermediate centering with $K_{0,n}$, ℓ_n, θ_n, and q_n from (S.1.11). For the deterministic frame sequence (U_n) covered by the theorem, abbreviate at dimension n $U = U_n \in \mathrm{St}(d_n, p_n)$. Set

$$\overline{\mathrm{Risk}}_n^{\mathrm{fin,odd}}(U) = \{\nu K_{0,n} - 2\mu_1\theta_n q_n + 1\}\|U^{\mathrm{T}}\beta^*\|^2 + (\nu\ell_n + 1)\{\|\beta^*\|^2 - \|U^{\mathrm{T}}\beta^*\|^2 + \overline{\sigma}_{\varepsilon,n}^2\} + \sigma_{\mathrm{test}}^2 - \overline{\sigma}_{\varepsilon,n}^2.$$

The preceding estimates give, along the chosen deterministic frame sequence,

$$\mathrm{Risk}_n(U) - \overline{\mathrm{Risk}}_n^{\mathrm{fin,odd}}(U) \xrightarrow{L^2} 0.$$

For the limiting coefficients, define the conditional centering

$$\overline{\mathrm{Risk}}_n^{\mathrm{odd}}(U) := \mathfrak{K}\|U^{\mathsf{T}}\beta^*\|^2 + \mathfrak{L}\{\|\beta^*\|^2 - \|U^{\mathsf{T}}\beta^*\|^2 + \overline{\sigma}_{\varepsilon,n}^2\} + \sigma_{\mathrm{test}}^2 - \overline{\sigma}_{\varepsilon,n}^2, \tag{S.2.2}$$

where $\kappa_\sigma = \nu$ in (S.1.6). By (S.1.13), the teacher norm bound, and bounded average noise variance,

$$\sup_{U\in\mathrm{St}(d_n,p_n)} \left|\overline{\mathrm{Risk}}_n^{\mathrm{fin,odd}}(U) - \overline{\mathrm{Risk}}_n^{\mathrm{odd}}(U)\right| \longrightarrow 0.$$

Hence

$$\mathrm{Risk}_n(U) - \overline{\mathrm{Risk}}_n^{\mathrm{odd}}(U) \xrightarrow{\mathbb{P}} 0.$$

Since $\mathcal{E}(\lambda,\gamma) = \mathfrak{K}$ and $\chi(\lambda,\gamma) = \nu\ell$, this is exactly (3.2). The two limits in (3.1) follow by taking $d = p$ and $W = I_p$ in the preceding argument. For the signal fit, set $\beta^* = e_1$, the first coordinate vector, and $\overline{\sigma}_{\varepsilon,n}^2 = \sigma_{\mathrm{test}}^2 = 0$; the limiting risk is $\mathcal{E}(\lambda,\gamma)$. For the residual fit, set $\beta^* = 0$, $\overline{\sigma}_{\varepsilon,n}^2 = 1$, and $\sigma_{\mathrm{test}}^2 = 0$; the limiting risk is $\chi(\lambda,\gamma)$. □

Proof of Corollary 4.1. When $\sigma_{n,i} \equiv \sigma_\varepsilon$, $\overline{\sigma}_{\varepsilon,n}^2 = \sigma_\varepsilon^2$ for every n. The test variance is also σ_ε^2, so the correction in (3.2) vanishes and (4.1) follows. □

S.3 Supplementary Random Feature Results and Technical Estimates

We first give the zero-ridge, noncentered positive-ridge, and Haar-frame extensions.

S.3.1 Ridgeless results

We treat interior mixup and endpoint training separately. For interior mixup, the canonical interaction supplies a spectral floor. At the endpoint, the limit is governed by the Moore–Penrose inverse and the Marchenko–Pastur hard edge.

Remark 3 (Affine activation) For fixed $\lambda \in [1/2, 1)$ and a Lipschitz activation, the interaction factor in (S.1.1) satisfies $\kappa(\lambda) > 0$ *if and only if* σ is nonaffine; see Lemma S.3.1 below. An affine activation creates *no* such interaction, while the unmixed case $\lambda = 1$ has $\kappa(1) = 0$ for all activations. As a consequence, we focus on nonaffine activations when studying structural overparameterization in Section 4.

Remark 4 (Ridgeless setting) Theorem 3.1 assumes positive ridge regularization, $\gamma > 0$. At $\gamma = 0$, we select the least-squares solution of minimum Euclidean norm, a setting widely studied in connection with double descent and benign overfitting (Bartlett et al., 2020; Mei & Montanari, 2022; Hastie et al., 2022). As discussed in Remark 3, a nonaffine activation yields positive interaction factor $\kappa(\lambda) > 0$. This keeps the limiting risk *finite* at $c = 1$, the interpolation threshold of the unmixed student. In contrast, the ridgeless risk of ordinary random feature regression (i.e., $\lambda = 1$) *diverges* as $c \to 1$. Theorems S.3.2 and S.3.3 give the precise statements. This is another manifestation of how mixup can improve the tradeoff between signal reconstruction error and residual fitting error, even without an explicit ridge penalty. Section 4 examines whether this advantage persists after separate and optimal ridge tuning.

In addition to being bounded, the teacher and noise energies must converge separately for the limiting risk formulas in this subsection:

$$\begin{aligned} \|a\|^2 &\longrightarrow r_a^2, \qquad \overline{\sigma}_{\varepsilon,n}^2 \longrightarrow \overline{\sigma}_\varepsilon^2, \\ \rho_{\perp,n}^2 &:= \|\beta_\perp^*\|^2 + \overline{\sigma}_{\varepsilon,n}^2 \longrightarrow \rho_{\perp,\infty}^2. \end{aligned} \tag{S.3.1}$$

For $\gamma = 0$, define

$$\widehat{\beta}_{\mathrm{mix}}(\lambda, 0) = S_\lambda^\dagger r_\lambda.$$

At an interior weight and a nonaffine activation, $\kappa(\lambda) > 0$ supplies an interaction-induced spectral floor. Let $\mathcal{E}(\lambda,0)$ and $\chi(\lambda,0)$ denote the explicit zero-ridge coefficients in Appendix S.1.3. They are the continuous $\gamma \downarrow 0$ limits of the positive-ridge coefficients for an interior weight.

S.3.1.1 The canonical interaction is positive for nonlinear activations

Lemma S.3.1 (Nonlinearity and the pair-interaction floor) Fix $\lambda \in (0,1)$ and let σ be Lipschitz. Then $\kappa(\lambda) = 0$ if and only if σ is affine. Consequently every nonaffine activation has $\kappa(\lambda) > 0$ at a fixed interior mixing weight.

Proof of Lemma S.3.1. The affine implication follows by direct substitution. Conversely, suppose that $\kappa(\lambda) = 0$. Since $\kappa(\lambda) = \mathbb{E}h_0(G_1, G_2)^2$, the canonical Hoeffding component vanishes almost surely. Define

$$H(x,y) = \sigma(\lambda x + (1-\lambda)y) - f_1(x) - f_2(y) + \beta_\sigma.$$

The Lipschitz property of σ implies that f_1, f_2, and H are continuous. The product Gaussian density is strictly positive on $\mathbb{R}^2$, so the almost-sure identity $H(G_1, G_2) = 0$ extends by continuity to

$$H(x,y) = 0, \qquad (x,y) \in \mathbb{R}^2.$$

Taking the rectangular difference at $(x,y), (x',y), (x,y')$, and (x',y') gives

$$\begin{aligned} 0 = {} & \sigma(\lambda x' + (1-\lambda)y') - \sigma(\lambda x' + (1-\lambda)y) \\ & - \sigma(\lambda x + (1-\lambda)y') + \sigma(\lambda x + (1-\lambda)y). \end{aligned}$$

Because both λ and $1-\lambda$ are nonzero, this is equivalent to

$$\sigma(u+s+t) - \sigma(u+s) - \sigma(u+t) + \sigma(u) = 0 \tag{S.3.2}$$

for every $u, s, t \in \mathbb{R}$. For fixed s, (S.3.2) says that $\sigma(u+s) - \sigma(u)$ is independent of u. Hence, with $A(s) = \sigma(s) - \sigma(0)$,

$$A(s+t) = A(s) + A(t).$$

The function A is continuous, so the continuous Cauchy equation gives $A(s) = \mathfrak{a}s$. Thus $\sigma(u) = \mathfrak{a}u + \mathfrak{b}$ with $\mathfrak{b} = \sigma(0)$. This proves the equivalence. Direct evaluation in the affine case also gives

$$\beta_\sigma = \mathfrak{b}, \qquad f_1(u) = \mathfrak{a}\lambda u + \mathfrak{b}, \qquad f_2(u) = \mathfrak{a}(1-\lambda)u + \mathfrak{b}, \qquad V = \mathfrak{a}^2 w w^\top.$$

In particular, an odd affine activation is linear, completing the proof. □

Theorem S.3.2 (Interior ridgeless risk) Under the assumptions of Theorem 3.1 and (S.3.1), let σ be nonaffine, fix $\lambda \in (0,1)$, and set $\gamma = 0$. Then S_λ is invertible with probability tending to one and

$$\mathrm{Risk}_{\lambda,0,n} \xrightarrow{\mathbb{P}} \mathcal{E}(\lambda,0) r_a^2 + \{\chi(\lambda,0) + 1\}\rho_{\perp,\infty}^2 + \sigma_{\mathrm{test}}^2 - \overline{\sigma}_\varepsilon^2. \tag{S.3.3}$$

S.3.1.2 Interior mixup at zero ridge

Proof of the interior-mixup part of Theorem S.3.2. Under oddness, put

$$u_i = f_1(g_i), \qquad v_i = f_2(g_i), \qquad \mathsf{U} = [u_1, \dots, u_n], \qquad \mathsf{V} = [v_1, \dots, v_n].$$

The centered Gram estimate (S.3.41) remains valid at $\gamma = 0$:

$$S = \kappa(\lambda) I_p + \frac{1}{n}\mathsf{U} H_n \mathsf{U}^\top + \frac{1}{n}\mathsf{V} H_n \mathsf{V}^\top + \widetilde{z}\widetilde{z}^\top + E_{S,n}, \qquad \|E_{S,n}\|_{\mathrm{op}} = o_{\mathbb{P}}(1). \tag{S.3.4}$$

Lemma S.3.1 gives $\kappa(\lambda) > 0$, and all four displayed terms before $E_{S,n}$ are positive semidefinite. Set

$$\mathcal{G}_n = \{\lambda_{\min}(S) \geq \kappa(\lambda)/2\}.$$

Then $\mathbb{P}(\mathcal{G}_n) \to 1$. Thus S is invertible with high probability, uniformly when c_n stays in a compact subset of $(0,\infty)$. On $\mathcal{G}_n$, $S^\dagger = S^{-1}$. With $Q_\gamma = (S + \gamma I_p)^{-1}$ and $Q_0 = S^{-1}$ on this event,

$$\|Q_\gamma - Q_0\|_{\mathrm{op}} \leq \gamma \|Q_\gamma\|_{\mathrm{op}} \|Q_0\|_{\mathrm{op}} \leq \frac{4\gamma}{\kappa(\lambda)^2}.$$

To compare the noise-averaged risks, define

$$\mathcal{M}_n := \mathbb{E}_\xi[\|r\|^2 \mid X, W] = \|r_\mu\|^2 + \mathrm{tr}\, C_\varepsilon. \tag{S.3.5}$$

For the clean labels, (S.3.40) and (S.3.41) give $\mathbb{E}\|r_\mu\|^2 \le C$. Jensen's inequality in the definition of ϕ_i, the Lipschitz Gaussian feature moments, and $p = O(n)$ give $\max_i \mathbb{E}\|\phi_i\|^2 \le Cn$. Hence

$$\mathbb{E}\,\mathrm{tr}\, C_\varepsilon = \frac{1}{n^2}\sum_i \sigma_{n,i}^2 \mathbb{E}\|\phi_i\|^2 \le C\overline{\sigma}_{\varepsilon,n}^2 \le C.$$

Thus $\mathbb{E}\mathcal{M}_n \le C$ and $\mathcal{M}_n = O_{\mathbb{P}}(1)$. The same estimates hold at the endpoint. Put $b_{\gamma,n} = Q_\gamma r$ and $b_{0,n} = S^\dagger r$. On $\mathcal{G}_n$,

$$\|b_{\gamma,n} - b_{0,n}\| \le \frac{4\gamma}{\kappa(\lambda)^2}\|r\|, \qquad \|b_{\gamma,n}\| + \|b_{0,n}\| \le \frac{4}{\kappa(\lambda)}\|r\|.$$

For an odd activation, the risk of a deterministic coefficient b is exactly

$$\mathcal{E}_n(b) = \nu\|b\|^2 - 2\mu_1 a^\top b + \|\beta^*\|^2 + \sigma_{\text{test}}^2.$$

Compare the two realized test risks and then take $\mathbb{E}_\xi[\,\cdot \mid X, W]$. Because $\mathcal{G}_n$ depends only on the design, its indicator can be taken outside the conditional expectation. Conditional Cauchy–Schwarz and the teacher norm bound give

$$\mathbf{1}_{\mathcal{G}_n}|\,\mathrm{Risk}_{\gamma,n} - \mathrm{Risk}_{0,n}\,| \le C\gamma\{\mathcal{M}_n + \|a\|\sqrt{\mathcal{M}_n}\}.$$

Since the quantity in braces is uniformly tight and $\mathbb{P}(\mathcal{G}_n^c) \to 0$, we obtain the two-stage continuity statement

$$\lim_{\gamma\downarrow 0}\limsup_{n\to\infty}\mathbb{P}\{|\,\mathrm{Risk}_{\gamma,n} - \mathrm{Risk}_{0,n}\,| > \varepsilon\} = 0, \qquad \varepsilon > 0. \tag{S.3.6}$$

At $\gamma = 0$, the right-hand side of the reduced scalar equation is

$$F_0(t) = t\kappa(\lambda) + \sum_{k=1}^{2}\frac{t\lambda_k(V)}{1 + t\lambda_k(V)},$$

where $\lambda_1(V), \lambda_2(V)$ are the eigenvalues of V. It is strictly increasing, $F_0(0) = 0$, and $F_0(t) \to \infty$ because $\kappa(\lambda) > 0$. Consequently (S.1.8) has a unique finite root t_0. If t_γ is the positive root at ridge $\gamma > 0$, then

$$F_0(t_0) - F_0(t_\gamma) = \gamma t_\gamma, \qquad F_0'(t) \ge \kappa(\lambda),$$

and hence

$$0 \le t_0 - t_\gamma \le \frac{\gamma t_\gamma}{\kappa(\lambda)} \le \frac{c\gamma}{\kappa(\lambda)^2}.$$

Thus $t_\gamma \to t_0$, and every coefficient in (S.1.6) converges to its counterpart in (S.1.9). Let R_γ and R_0 denote the corresponding deterministic odd risks. Then $R_\gamma \to R_0$.

For every fixed $\gamma > 0$, Theorem 3.1 gives $\mathrm{Risk}_{\gamma,n} \to R_\gamma$ in probability. For $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}\{|\,\mathrm{Risk}_{0,n} - R_0| > 3\varepsilon\} \le{}& \mathbb{P}\{|\,\mathrm{Risk}_{0,n} - \mathrm{Risk}_{\gamma,n}\,| > \varepsilon\} \\ &+ \mathbb{P}\{|\,\mathrm{Risk}_{\gamma,n} - R_\gamma| > \varepsilon\} + \mathbf{1}_{\{|R_\gamma - R_0| > \varepsilon\}}.\end{aligned}$$

Taking first $n \to \infty$ and then $\gamma \downarrow 0$, and using (S.3.6), proves (S.3.3).

□

The endpoint behaves differently because its pair-interaction floor vanishes. The endpoint formulas below divide by the feature variance and therefore assume

$$\nu = \mathbb{E}\sigma(G)^2 > 0.$$

Define

$$r_{\text{eff}}^2 = \rho_{\perp,\infty}^2 + \left(1 - \frac{\mu_1^2}{\nu}\right) r_a^2.$$

Theorem S.3.3 (Endpoint ridgeless risk) Under the assumptions of Theorem 3.1 and (S.3.1), suppose $\nu > 0$, set $\lambda = 1$, $\gamma = 0$, and assume $c \neq 1$. Then

$$\mathrm{Risk}_{1,0,n} \xrightarrow{\mathbb{P}} \begin{cases} \dfrac{r_{\mathrm{eff}}^2}{1-c}, & 0 < c < 1, \\ \dfrac{c\, r_{\mathrm{eff}}^2}{c-1} + \dfrac{\mu_1^2}{\nu} r_a^2 \left(1 - \dfrac{1}{c}\right), & c > 1, \end{cases} + \sigma_{\mathrm{test}}^2 - \overline{\sigma}_\varepsilon^2. \tag{S.3.7}$$

S.3.1.3 Endpoint Moore–Penrose risk

Proof of Theorem S.3.3. At $\lambda = 1$, the assumption $\nu > 0$ ensures a nondegenerate feature variance. Let

$$\Phi = \begin{pmatrix} \sigma(g_1)^\top \\ \vdots \\ \sigma(g_n)^\top \end{pmatrix} \in \mathbb{R}^{n \times p}, \qquad S_{\mathrm{end}} = \frac{1}{n}\Phi^\top \Phi, \qquad r_{\mathrm{end}} = \frac{1}{n}\Phi^\top y.$$

Oddness and Gaussian coordinate independence imply that the entries of Φ are iid, centered, sub-Gaussian, and have variance $\nu > 0$. The endpoint ridge and minimum-norm coefficients are

$$\widehat{\beta}_{\mathrm{end},\gamma} = (S_{\mathrm{end}} + \gamma I_p)^{-1} r_{\mathrm{end}}, \qquad \widehat{\beta}_{\mathrm{end},0} = S_{\mathrm{end}}^\dagger r_{\mathrm{end}} = \Phi^\dagger y.$$

Fix $c \neq 1$. Put $\widetilde{\Phi} = \nu^{-1/2}\Phi$, and transpose $\widetilde{\Phi}$ when $p > n$. The resulting rectangular matrix has independent centered unit-variance sub-Gaussian entries. Since its smaller dimension divided by its larger dimension stays below one by a fixed amount, the lower-tail estimate (Rudelson & Vershynin, 2009, Theorem 1.1) gives a deterministic $b_c > 0$ such that, with probability tending to one, its smallest singular value is at least b_c times the square root of the larger dimension. Rescaling back and using $p/n \to c$ gives a deterministic $a_c > 0$ for which

$$\mathbb{P}\{\lambda_{\min}^+(S_{\mathrm{end}}) \geq a_c\} \to 1,$$

where $\lambda_{\min}^+$ is the smallest nonzero eigenvalue. Furthermore,

$$\|r_{\mathrm{end}}\| \leq n^{-1}\|\Phi\|_{\mathrm{op}}\|y\| = O_{\mathbb{P}}(1), \qquad r_{\mathrm{end}} \in \mathrm{ran}(S_{\mathrm{end}}).$$

Here $n^{-1}\|y\|^2 = O_{\mathbb{P}}(1)$ follows from Markov's inequality and $n^{-1}\sum_i \mathbb{E}(y_i^2 \mid W) = \rho_n^2 = O(1)$. Moreover, r_μ and every ϕ_i lie in $\mathrm{ran}(S_{\mathrm{end}})$. We therefore compare the resolvent and pseudoinverse only on this range. On the preceding spectral-gap event,

$$\|\widehat{\beta}_{\mathrm{end},\gamma} - \widehat{\beta}_{\mathrm{end},0}\| \leq \frac{\gamma}{a_c(a_c+\gamma)}\|r_{\mathrm{end}}\| = O_{\mathbb{P}}(\gamma), \qquad \|\widehat{\beta}_{\mathrm{end},\gamma}\| + \|\widehat{\beta}_{\mathrm{end},0}\| \leq \frac{2}{a_c}\|r_{\mathrm{end}}\| = O_{\mathbb{P}}(1). \tag{S.3.8}$$

For an odd activation, the conditional test risk of a coefficient b is

$$\mathcal{E}_n(b) = \nu\|b\|^2 - 2\mu_1 a^\top b + \|\beta^*\|^2 + \sigma_{\mathrm{test}}^2.$$

Thus $\mathrm{Risk}_{\mathrm{end},\gamma,n} = \mathbb{E}_\xi[\mathcal{E}_n(\widehat{\beta}_{\mathrm{end},\gamma}) \mid X, W]$. Let $\mathcal{G}_n$ denote the preceding spectral-gap event, which depends only on the design. Apply (S.3.8) to the realized-risk difference and take the conditional expectation. Using $\mathcal{M}_n$ from (S.3.5), we obtain

$$\mathbf{1}_{\mathcal{G}_n}|\mathrm{Risk}_{\mathrm{end},\gamma,n} - \mathrm{Risk}_{\mathrm{end},0,n}| \leq C\gamma\{\mathcal{M}_n + \|a\|\sqrt{\mathcal{M}_n}\}.$$

The preceding bound shows that the multiplier is tight under bounded average variance. Since the spectral-gap event has probability tending to one, for every $\varepsilon > 0$ this implies

$$\lim_{\gamma \downarrow 0} \limsup_{n \to \infty} \mathbb{P}\{|\mathrm{Risk}_{\mathrm{end},\gamma,n} - \mathrm{Risk}_{\mathrm{end},0,n}| > \varepsilon\} = 0. \tag{S.3.9}$$

To complete the argument, take $\gamma \downarrow 0$ in the fixed-ridge equivalent. At the endpoint,

$$V = \mathrm{diag}(\nu, 0), \qquad \kappa(1) = 0, \qquad d_\alpha = \mu_1, \qquad w = (1,0)^\top.$$

Put $x_\gamma = \nu t_\gamma$ and $\delta = \gamma/\nu$. The scalar equation becomes

$$c = x_\gamma \left\{\delta + \frac{1}{1+x_\gamma}\right\}.$$

If $0 < c < 1$, then $x_\gamma \to c/(1-c)$. If $c > 1$, then

$$x_\gamma \to \infty, \qquad \delta x_\gamma \to c - 1.$$

Writing $s_\sigma = \mu_1^2/\nu$, substitution in the endpoint specialization of (S.1.6) gives

$$(\mathfrak{R}, \mathfrak{Q}) \longrightarrow \begin{cases} ((1-s_\sigma)/(1-c),\ 1/(1-c)), & 0 < c < 1, \\ (c(1-s_\sigma)/(c-1) + s_\sigma(1-1/c),\ c/(c-1)), & c > 1. \end{cases}$$

Substituting these limits in (3.2) and retaining the correction $\sigma_{\text{test}}^2 - \overline{\sigma}_\varepsilon^2$ gives the two branches in (S.3.7). Denote the fixed-ridge deterministic limit by $R_{\text{end},\gamma}$ and the appropriate branch in (S.3.7) by $R_{\text{end},0}$. For every $\varepsilon > 0$ and fixed $\gamma > 0$,

$$\begin{aligned} \mathbb{P}\{|\operatorname{Risk}_{\text{end},0,n} - R_{\text{end},0}| > 3\varepsilon\} \le{}& \mathbb{P}\{|\operatorname{Risk}_{\text{end},0,n} - \operatorname{Risk}_{\text{end},\gamma,n}| > \varepsilon\} \\ &+ \mathbb{P}\{|\operatorname{Risk}_{\text{end},\gamma,n} - R_{\text{end},\gamma}| > \varepsilon\} + \mathbf{1}_{\{|R_{\text{end},\gamma} - R_{\text{end},0}| > \varepsilon\}}. \end{aligned}$$

First let $n \to \infty$ and use Theorem 3.1 together with Proposition S.3.7; then let $\gamma \downarrow 0$ and use (S.3.9) together with the displayed coefficient limits. This proves the endpoint assertion. □

The endpoint risk diverges as $c \to 1$ whenever $r_{\text{eff}}^2 > 0$, whereas the interior risk remains finite because $\kappa(\lambda) > 0$. Section 4 establishes a strict risk comparison and then addresses the harder question of whether the advantage survives separate ridge tuning. The endpoint hard-edge behavior comes from the lower spectral edge of a high-dimensional sample covariance, the Marchenko–Pastur mechanism that drives ridgeless double descent (Hastie et al., 2022).

S.3.2 Positive-ridge generalization

The noncentered generalization theorem follows from the two component limits below. Sections S.3.6.4 and S.3.6.5 provide the corresponding resolvent arguments.

The proofs below for the realized risk components, and the noncentered positive-ridge and zero-ridge results derived from them, use Assumption 1 and the noise conditions in Definition 2.1. For these results, we impose the stronger profile condition

$$\varepsilon_i = \sigma_{n,i}\xi_i, \qquad \sup_n \max_i \sigma_{n,i} < \infty, \qquad \overline{\sigma}_{\varepsilon,n}^2 := \frac{1}{n}\sum_{i=1}^n \sigma_{n,i}^2 \longrightarrow \overline{\sigma}_\varepsilon^2. \tag{S.3.10}$$

The scales are deterministic, and the standardized noises are independent of the design. The observation noises and residuals therefore have uniformly bounded eighth moments. Using the notation of Section 2, set

$$D_{\zeta,n} := \operatorname{diag}(\|\beta_\perp^*\|^2 + \sigma_{n,1}^2, \dots, \|\beta_\perp^*\|^2 + \sigma_{n,n}^2), \quad \rho_{\perp,n}^2 := \frac{1}{n}\operatorname{tr} D_{\zeta,n}, \quad \rho_n^2 := \|\beta^*\|^2 + \overline{\sigma}_{\varepsilon,n}^2.$$

The odd positive-ridge theorem and its odd zero-ridge and Haar consequences do not require this stronger condition.

S.3.2.1 Limit of the Cross Term B

Theorem S.3.4 Assume (S.1.7) and suppose that

$$\|a\|^2 \longrightarrow r_a^2, \qquad e^{\mathsf{T}} a \longrightarrow \eta.$$

Then

$$B \xrightarrow{\mathbb{P}} B_\infty, \qquad B_\infty := -2\mu_1\theta q(r_a^2 - \eta^2). \tag{S.3.11}$$

Proof of Theorem S.3.4. Condition on W, and write $W = [w_1, \dots, w_p]$. Since $\|w_k\| = 1$, Gaussian integration by parts gives, for every $k \le p$,

$$\mathbb{E}_{x_{\text{new}}}\left[x_{\text{new}}\sigma(w_k^{\mathsf{T}} x_{\text{new}}) \,\middle|\, W\right] = w_k \mathbb{E}[G\sigma(G)] = \mu_1 w_k.$$

Applying this identity columnwise yields

$$\mathbb{E}_{x_{\text{new}}}\left[x_{\text{new}}\sigma(W^{\mathrm{T}}x_{\text{new}})^{\mathrm{T}}\,\middle|\,W\right]=\mu_1 W.$$

Since $a=W^{\mathrm{T}}\beta^*$ and $\widehat{\beta}=Qr$,

$$B=-2\mu_1 a^{\mathrm{T}}Qr. \tag{S.3.12}$$

Write

$$g_i=W^{\mathrm{T}}x_i,\qquad \beta_\perp^*=(I_d-WW^{\mathrm{T}})\beta^*,\qquad \zeta_i=x_i^{\mathrm{T}}\beta_\perp^*+\varepsilon_i.$$

Since $a=W^{\mathrm{T}}\beta^*$ and $W^{\mathrm{T}}W=I_p$,

$$\beta^*=Wa+\beta_\perp^*,\qquad W^{\mathrm{T}}\beta_\perp^*=0.$$

Hence, conditionally on W,

$$\mathrm{Cov}\left(g_i,x_i^{\mathrm{T}}\beta_\perp^*\,\middle|\,W\right)=W^{\mathrm{T}}\beta_\perp^*=0.$$

The vector g_i and the scalar $x_i^{\mathrm{T}}\beta_\perp^*$ are jointly Gaussian, so their zero covariance implies independence. Since the noise is also independent, $\zeta=(\zeta_1,\dots,\zeta_n)^{\mathrm{T}}$ is centered and conditionally independent of $G=(g_1,\dots,g_n)^{\mathrm{T}}$. To include independent noises that need not be identically distributed, write

$$\sigma_{n,i}^2=\mathbb{E}\varepsilon_i^2,\qquad D_{\varepsilon,n}=\mathrm{diag}(\sigma_{n,1}^2,\dots,\sigma_{n,n}^2),$$

and set

$$D_{\zeta,n}:=\|\beta_\perp^*\|^2 I_n+D_{\varepsilon,n}.$$

Then

$$\mathbb{E}[\zeta\mid G,W]=0,\qquad \mathbb{E}[\zeta\zeta^{\mathrm{T}}\mid G,W]=D_{\zeta,n}.$$

In the homoscedastic specialization, $D_{\zeta,n}=\rho_{\perp,n}^2 I_n$. Under the bounded-profile condition (S.3.10), the uniform eighth-moment bound implies

$$\|D_{\zeta,n}\|_{\mathrm{op}}\le\|\beta_\perp^*\|^2+\max_i(\mathbb{E}|\varepsilon_i|^8)^{1/4}\le C. \tag{S.3.13}$$

Let M be the full-pair label-mixing matrix in (S.5.1), and set

$$T=\frac{1}{n^2}Z^{\mathrm{T}}\mathsf{M}.$$

Define the projected mixed-label scalar

$$\ell_{ij}^{\mathrm{proj}}:=\{\lambda g_i+(1-\lambda)g_j\}^{\mathrm{T}}a.$$

Since $y=Ga+\zeta$, the score has the exact decomposition

$$r=r_\parallel+r_{\perp,\varepsilon},\qquad r_\parallel=\frac{1}{n^2}\sum_{i,j}z_{ij}\ell_{ij}^{\mathrm{proj}},\qquad r_{\perp,\varepsilon}=T\zeta. \tag{S.3.14}$$

The matrix bound in (S.5.1) gives

$$TT^{\mathrm{T}}=\frac{1}{n^4}Z^{\mathrm{T}}\mathsf{M}\mathsf{M}^{\mathrm{T}}Z\le\frac{1}{n}S. \tag{S.3.15}$$

Conditioning on (G,W), the scalar $a^{\mathrm{T}}Qr_{\perp,\varepsilon}$ is centered. Using (S.3.15) yields

$$\begin{aligned}\mathbb{E}\left[|a^{\mathrm{T}}Qr_{\perp,\varepsilon}|^2\,\middle|\,G,W\right]&=a^{\mathrm{T}}QTD_{\zeta,n}T^{\mathrm{T}}Qa\\&\le\|D_{\zeta,n}\|_{\mathrm{op}}\,a^{\mathrm{T}}QTT^{\mathrm{T}}Qa\\&\le\frac{\|D_{\zeta,n}\|_{\mathrm{op}}}{n}a^{\mathrm{T}}QSQa.\end{aligned}$$

Since $Q=(S+\gamma I_p)^{-1}$,

$$QSQ=Q(S+\gamma I_p)Q-\gamma Q^2=Q-\gamma Q^2\preceq Q\preceq\gamma_-^{-1}I_p.$$

Together with (S.3.13) and $\|a\|=O(1)$, this gives

$$\mathbb{E}\left[|a^{\mathrm{T}}Qr_{\perp,\varepsilon}|^2\,\middle|\,G,W\right]\le\frac{C}{n}.$$

Conditional Chebyshev's inequality therefore gives $a^{\mathrm{T}}Qr_{\perp,\varepsilon} = O_{\mathbb{P}}(n^{-1/2}) = o_{\mathbb{P}}(1)$. Applying Lemma S.3.13 to the first term in (S.3.14), and using

$$a^{\mathrm{T}}Qr_{\|} = \frac{1}{n^2}\sum_{i,j=1}^{n} \ell_{ij}^{\mathrm{proj}}\, a^{\mathrm{T}}Qz_{ij},$$

gives

$$a^{\mathrm{T}}Qr = d_\alpha(\alpha^2 - c_B)a^{\mathrm{T}}\bar{Q}a + o_{\mathbb{P}}(1). \tag{S.3.16}$$

The coefficient and projector estimates (S.3.197) and (S.3.198), proved in Lemma S.3.13, apply here with $q_0 = q_n$ and $\theta_0 = \theta_n$. They give

$$d_\alpha(\alpha^2 - c_B) = \theta_n + O(n^{-1}), \qquad \|\bar{Q} - q_n P_\perp\| = O(n^{-1}), \quad P_\perp = I_p - ee^{\mathrm{T}}.$$

Since $\|a\| = O(1)$,

$$a^{\mathrm{T}}\bar{Q}a = q_n r_{\perp,n}^2 + O(n^{-1}).$$

Substituting these estimates into (S.3.16) and (S.3.12) gives the finite-dimensional intermediate equivalent

$$B - \overline{B}_n^{\mathrm{fin}} \xrightarrow{\mathbb{P}} 0, \qquad \overline{B}_n^{\mathrm{fin}} := -2\mu_1\theta_n q_n r_{\perp,n}^2. \tag{S.3.17}$$

The alignment assumptions in Theorem S.3.4 give $r_{\perp,n}^2 \to r_a^2 - \eta^2$, while (S.1.13) gives $\theta_n q_n \to \theta q$. Hence $\overline{B}_n^{\mathrm{fin}} \to B_\infty$, and (S.3.17) proves (S.3.11). □

S.3.2.2 Limit of the Quadratic Term A

Theorem S.3.5 Assume (S.1.7) and suppose that

$$\|a\|^2 \longrightarrow r_a^2, \qquad e^{\mathrm{T}}a \longrightarrow \eta, \qquad \rho_{\perp,n}^2 \longrightarrow \rho_{\perp,\infty}^2.$$

Then

$$A \xrightarrow{\mathbb{P}} A_\infty, \qquad A_\infty := \kappa_\sigma\left[K(r_a^2 - \eta^2) + \ell\{\eta^2 + \rho_{\perp,\infty}^2\}\right]. \tag{S.3.18}$$

Proof of Theorem S.3.5. At $\omega = 0$, the definitions in Lemma S.3.17 agree with the finite-dimensional notation in (S.1.10)– (S.1.11):

$$t_0 = t_{*,n}, \qquad C_2 = C_{*,n}, \qquad \theta = \theta_n, \qquad \tau_0 = \tau_n, \qquad \ell = \ell_n.$$

Consequently, the coefficient denoted by $K_{\perp,n}$ in (S.3.245) is exactly the coefficient defined in (S.1.11). Set

$$\overline{A}_n^{\mathrm{fin}} := \kappa_\sigma\left[K_{\perp,n}r_{\perp,n}^2 + \ell_n\{(e^{\mathrm{T}}a)^2 + \rho_{\perp,n}^2\}\right].$$

By Lemma S.3.17 and (S.3.245),

$$\{\nu - \mu_0(\sigma)^2\}r^{\mathrm{T}}Q^2r = \overline{A}_n^{\mathrm{fin}} + o_{\mathbb{P}}(1).$$

Lemma S.3.18, in particular (S.3.246) and (S.3.247), therefore gives the finite-dimensional intermediate equivalent

$$A - \overline{A}_n^{\mathrm{fin}} \xrightarrow{\mathbb{P}} 0. \tag{S.3.19}$$

The alignment assumptions in Theorem S.3.5 and (S.1.13) imply $\overline{A}_n^{\mathrm{fin}} \to A_\infty$. Equation (S.3.19) now proves (S.3.18). □

S.3.2.3 Noncentered positive-ridge theorem

The centered theorem in the main text suppresses the feature-mean direction. The following result gives the stable branch needed for noncentered activations such as ReLU.

Theorem S.3.6 (Noncentered positive-ridge risk) Assume (S.3.10) and (S.1.7), and suppose

$$\|a\|^2 \to r_a^2, \qquad e^{\mathrm{T}}a \to \eta, \qquad \rho_{\perp,n}^2 \to \rho_{\perp,\infty}^2.$$

For every fixed $\lambda \in (0, 1]$ and $\gamma > 0$,

$$\mathrm{Risk}_{\lambda,\gamma,n} \xrightarrow{\mathbb{P}} \{\kappa_\sigma K - 2\mu_1\theta q + 1\}(r_a^2 - \eta^2) + (\kappa_\sigma \ell + 1)(\eta^2 + \rho_{\perp,\infty}^2) + \sigma_{\mathrm{test}}^2 - \overline{\sigma}_\varepsilon^2. \tag{S.3.20}$$

At $\lambda = 1$, this is ordinary noncentered random feature ridge.

The stability inequality has a direct spectral meaning: the finite-rank feature-mean resolvent stays on the branch continuously connected to positive ridge. Its numerical gap is imposed only in the nonzero-mean branch to control the mean direction. At the noncentered endpoint, $\chi_+ = 0$, so the numerical inequality is automatic.

S.3.2.4 Proof of the noncentered generalization theorem

The proof combines the two finite-dimensional component equivalents and then passes to the limiting coefficients.

Proof of Theorem S.3.6. The proofs of the finite-dimensional intermediate equivalents (S.3.17) and (S.3.19) give L^2 bounds for the errors in the cross term and the bulk quadratic term. The remaining nonnegative mean term satisfies $\mathbb{E}A_{11,n} = O(n^{-1}\log^C n)$, by Lemma S.3.18. The difference between the realized risk and the centering below therefore tends to zero in L^1. Conditional Jensen's inequality and (S.1.4) give the same L^1 convergence after averaging over the training noise. In particular,

$$\mathrm{Risk}_n - \left[\kappa_\sigma\{K_{\perp,n} r_{\perp,n}^2 + \ell_n((e^{\mathsf{T}}a)^2 + \rho_{\perp,n}^2)\} - 2\mu_1\theta_n q_n r_{\perp,n}^2 + \|\beta^*\|^2 + \sigma_{\mathrm{test}}^2\right] \xrightarrow{\mathbb{P}} 0. \tag{S.3.21}$$

For the deterministic frame sequence (U_n) covered by the theorem, abbreviate at dimension n $U = U_n \in \mathrm{St}(d_n, p_n)$. Put

$$a(U) = U^{\mathsf{T}}\beta^*, \qquad \rho_{\perp,n}^2(U) = \|\beta^*\|^2 - \|a(U)\|^2 + \overline{\sigma}_{\varepsilon,n}^2,$$

and define the limit-coefficient conditional centering

$$\overline{\mathrm{Risk}}_n^{\mathrm{nz}}(U) := \Re\|P_\perp a(U)\|^2 + \mathfrak{L}\{(e^{\mathsf{T}}a(U))^2 + \rho_{\perp,n}^2(U)\} + \sigma_{\mathrm{test}}^2 - \overline{\sigma}_{\varepsilon,n}^2. \tag{S.3.22}$$

The teacher norm bound and (S.1.13) show that the bracket in (S.3.21) differs from $\overline{\mathrm{Risk}}_n^{\mathrm{nz}}(U)$ by $o(1)$, uniformly over $U \in \mathrm{St}(d_n, p_n)$. Therefore, along the chosen deterministic frame sequence,

$$\mathrm{Risk}_n(U) - \overline{\mathrm{Risk}}_n^{\mathrm{nz}}(U) \xrightarrow{\mathbb{P}} 0.$$

By the alignment assumptions, the conditional centering converges to

$$A_\infty + B_\infty + r_a^2 + \rho_{\perp,\infty}^2 + \sigma_{\mathrm{test}}^2 - \overline{\sigma}_\varepsilon^2.$$

Substituting the definitions of A_∞ and B_∞ gives exactly the right-hand side of (S.3.20). □

S.3.2.5 Endpoint specialization

Proposition S.3.7 (Endpoint specialization) The centered conclusion of Theorem 3.1 and the noncentered conclusion of Theorem S.3.6 remain valid at $\lambda = 1$ after evaluating the displayed moments at that endpoint. The resulting estimator is ordinary random feature ridge because each original observation is repeated exactly n times.

Proof of Proposition S.3.7. At $\lambda = 1$, for every ordered pair (i, j),

$$x_{ij}^{\mathrm{mix}} = x_i, \qquad \widetilde{y}_{ij} = y_i, \qquad z_{ij} = \sigma(W^{\mathsf{T}}x_i).$$

Consequently, summing over the unused second index repeats each endpoint row exactly n times and gives

$$\frac{1}{n^2}Z^{\mathsf{T}}Z = \frac{1}{n}\sum_{i=1}^n \phi_W(x_i)\phi_W(x_i)^{\mathsf{T}}, \qquad \frac{1}{n^2}Z^{\mathsf{T}}\widetilde{y} = \frac{1}{n}\sum_{i=1}^n \phi_W(x_i)y_i.$$

Thus the full-pair estimator is exactly ordinary endpoint ridge.

We also check that specializing the deterministic-equivalent proof for a nonzero feature mean introduces no singular step. At $\lambda = 1$ and $\beta_\sigma \neq 0$,

$$f_1(u) = \sigma(u), \qquad f_2(u) = \beta_\sigma, \qquad h_0 \equiv 0, \qquad \kappa(1) = 0,$$

and $V = \mathrm{diag}(\kappa_\sigma, 0)$. Hence every identity in the two-way Hoeffding decomposition remains valid with the second conditional channel and the canonical interaction identically zero. The concentration, leave-two-out, and resolvent bounds in Sections S.1, S.3.2, and S.3.6–S.3.8 use only the Lipschitz constants λL_σ and $(1-\lambda)L_\sigma$, the positive-semidefiniteness of V, and the resolvent bound $\|Q\| \le \gamma_-^{-1}$. The endpoint values therefore delete terms from those arguments. Moreover $\chi_+ = 0$, so the mean-gap factor $1 - t_*\chi_+$ equals one. Applying the same component equivalents with these specialized moments proves the auxiliary finite-dimensional relation (S.3.21); coefficient convergence in (S.1.13), together with the inherited alignment assumptions, then proves (S.3.20) for endpoint ridge.

For the centered branch, oddness gives $\beta_\sigma = 0$, $V = \mathrm{diag}(\nu, 0)$, and $\kappa(1) = 0$ at the endpoint. The proof of Theorem 3.1 remains valid with these values. □

S.3.3 Noncentered ridgeless continuation

Theorem S.3.8 (Noncentered interior ridgeless risk) Assume the standing conditions, including (S.3.10), without a positive lower bound on ridge, fix $\lambda \in (0,1)$, and suppose σ is nonaffine and $\beta_\sigma \neq 0$. Let t_0 and the zero-ridge coefficients be defined by (S.1.8)–(S.1.9), and assume

$$1 - t_0\chi_+ > 0. \tag{S.3.23}$$

Then S_λ is invertible with probability tending to one. If $\|a\|^2 \to r_a^2$, $e^\top a \to \eta$, and $\rho_{\perp,n}^2 \to \rho_{\perp,\infty}^2$, then

$$\mathrm{Risk}_{\lambda,0,n} \xrightarrow{\mathbb{P}} \mathfrak{R}_0^{\mathrm{rid}}(r_a^2 - \eta^2) + \mathfrak{Q}_0^{\mathrm{rid}}(\eta^2 + \rho_{\perp,\infty}^2) + \sigma_{\mathrm{test}}^2 - \overline{\sigma}_\varepsilon^2, \tag{S.3.24}$$

where $\kappa_\sigma = \nu - \mu_0^2$ in the displayed coefficients $\mathfrak{R}_0^{\mathrm{rid}}$ and $\mathfrak{Q}_0^{\mathrm{rid}}$. If $\chi_+ > 0$, the stability condition is equivalent to

$$c < c_{\mathrm{mean}} := \frac{\kappa(\lambda)}{\chi_+} + \mathrm{tr}\{V(\chi_+ I_2 + V)^{-1}\}; \tag{S.3.25}$$

when $\chi_+ = 0$, set $c_{\mathrm{mean}} = +\infty$.

Proof of Theorem S.3.8. Let

$$F_0(t) = t\left[\kappa(\lambda) + \mathrm{tr}\{V(I_2 + tV)^{-1}\}\right].$$

If $\lambda_1(V), \lambda_2(V) \ge 0$ are the eigenvalues of V, then

$$F_0(t) = \kappa(\lambda)t + \sum_{j=1}^{2} \frac{t\lambda_j(V)}{1 + t\lambda_j(V)}, \qquad F_0'(t) = \kappa(\lambda) + \mathrm{tr}\{V(I_2 + tV)^{-2}\} \ge \kappa(\lambda).$$

Thus F_0 is strictly increasing from zero to infinity, proving existence and uniqueness of t_0. If $\chi_+ > 0$, direct substitution gives

$$F_0(\chi_+^{-1}) = \frac{\kappa(\lambda)}{\chi_+} + \mathrm{tr}\{V(\chi_+ I_2 + V)^{-1}\} = c_{\mathrm{mean}}.$$

Strict monotonicity proves the equivalence between (S.3.23) and $c < c_{\mathrm{mean}}$. When $\chi_+ = 0$, the gap holds automatically.

For fixed $\gamma > 0$, let t_γ solve the limiting positive-ridge equation. Equivalently,

$$c = F_0(t_\gamma) + \gamma t_\gamma.$$

Let

$$(C_\gamma, q_\gamma, \theta_\gamma, \tau_\gamma, \ell_\gamma, K_\gamma)$$

denote the quantities in (S.1.6) evaluated at ridge γ and root t_γ. Since $V \ge 0$ implies $\chi_+ \ge 0$, the inequality $t_\gamma < t_0$ shows that the zero-ridge gap implies $1 - t_\gamma\chi_+ > 0$. Theorem S.3.6 therefore applies at every fixed $\gamma > 0$. Write its deterministic risk limit as R_γ.

The mean value theorem gives

$$0 \le t_0 - t_\gamma \le \frac{\gamma t_\gamma}{\kappa(\lambda)} \le \frac{c\gamma}{\kappa(\lambda)^2}.$$

The scalar denominator in the ridge-dependent coefficients is bounded below by $\kappa(\lambda)$; moreover, $c > 0$ is fixed and $I_2 + t_\gamma V \succeq I_2$. Consequently, as $\gamma \downarrow 0$,

$$C_\gamma \to C_0, \quad q_\gamma \to q_0, \quad \theta_\gamma \to \theta_0, \quad \tau_\gamma \to \tau_0, \quad \ell_\gamma \to \ell_0, \quad K_\gamma \to K_0^{\mathrm{rid}},$$

and R_γ converges to the right-hand side R_0 of (S.3.24).

For $\varepsilon > 0$ and fixed $\gamma > 0$,

$$\begin{aligned}\mathbb{P}\{|\operatorname{Risk}_{0,n} - R_0| > 3\varepsilon\} \le \mathbb{P}\{|\operatorname{Risk}_{0,n} - \operatorname{Risk}_{\gamma,n}| > \varepsilon\} \\ + \mathbb{P}\{|\operatorname{Risk}_{\gamma,n} - R_\gamma| > \varepsilon\} + \mathbf{1}_{\{|R_\gamma - R_0| > \varepsilon\}}.\end{aligned}$$

First take $n \to \infty$ and then $\gamma \downarrow 0$. Lemma S.3.19, the fixed-ridge theorem, and $R_\gamma \to R_0$ show that all three terms vanish. This proves (S.3.24). □

S.3.4 Haar-frame specialization

Theorem 3.1 conditions on W. A Haar frame is an orthonormal p-frame drawn uniformly over orientations, independently of the sample. It turns the frame-specific teacher projection into a typical visible-energy fraction.

Corollary S.3.9 (Haar-frame risk) Suppose

$$W \sim \mathrm{Haar}\{\mathrm{St}(d,p)\}, \qquad \frac{p}{d} \to \varpi \in [0,1], \qquad \|\beta^*\|^2 \to r_\beta^2, \qquad \overline{\sigma}_{\varepsilon,n}^2 \to \overline{\sigma}_\varepsilon^2,$$

and retain the assumptions of Theorem 3.1. Jointly over the frame and the sample,

$$\operatorname{Risk}_{\lambda,\gamma,n} \xrightarrow{\mathbb{P}} \mathcal{E}(\lambda,\gamma)\,\varpi r_\beta^2 + \{\chi(\lambda,\gamma) + 1\}\{(1-\varpi)r_\beta^2 + \overline{\sigma}_\varepsilon^2\} + \sigma_{\mathrm{test}}^2 - \overline{\sigma}_\varepsilon^2. \tag{S.3.26}$$

The fraction ϖ is the asymptotic teacher energy seen by a typical random frame. This corollary gives the unconditional risk curves used in Section 5.

We derive the unconditional specialization for a Haar-random orthonormal feature frame by first fixing the frame and applying the conditional deterministic equivalents above. Averaging over the Haar orientation then requires only the concentration of a deterministic teacher's energy in a Haar-random subspace.

Proof of Corollary S.3.9. By Lemma S.5.1,

$$\frac{k_n}{d} \longrightarrow \varpi, \qquad T_n - \frac{k_n}{d}\|\beta^*\|^2 \xrightarrow{L^2(W)} 0.$$

For the noncentered branch, (S.5.4) also gives $e^{\mathsf{T}} W^{\mathsf{T}} \beta^* \to 0$ in $L^2(W)$; thus no nonzero mean-direction teacher alignment survives Haar averaging. Consequently, (S.3.259) converges to the right-hand side of (S.3.26). Proposition S.3.21 and Slutsky's theorem complete the proof. □

S.3.5 Centered score–resolvent estimates

Proof of Lemma S.2.1. Centered covariance estimates. For the intermediate Gram and score estimates, write $y_i = g_i^{\mathsf{T}} a + \zeta_i$, where the ζ_i are centered, independent across observations, independent of the projected Gaussian design, and have uniformly bounded eighth moments. This includes both the clean labels and the homoscedastic baseline labels. We specialize to the clean labels for the final score–resolvent equivalents. With f_1, f_2 denoting the two scalar endpoint conditional means, put

$$u_i = f_1(g_i), \qquad v_i = f_2(g_i), \qquad F_i = [u_i, v_i], \qquad w = (\lambda, 1-\lambda)^{\mathsf{T}}.$$

Oddness and Gaussian symmetry imply that f_1, f_2 are odd and centered. Consequently, the rows of F_i are independent centered two-dimensional vectors with covariance V. Let

$$H_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}},$$

$$M_n = \frac{1}{n}\sum_i F_i, \qquad \check{C}_n = \kappa(\lambda) I_p + \frac{1}{n}\sum_i F_i F_i^{\mathrm{T}}, \qquad \check{s}_n = \frac{1}{n}\sum_i F_i w\, y_i,$$

and define

$$C_n = \check{C}_n - M_n M_n^{\mathrm{T}}, \qquad s_n = \check{s}_n - M_n w \bar{y}, \qquad \bar{y} = \frac{1}{n}\sum_i y_i. \tag{S.3.27}$$

For a fixed circle Γ in a sufficiently small complex neighborhood of zero, chosen so that $\inf_{\omega\in\Gamma} \mathrm{Re}(\gamma+\omega) > 0$, write

$$\check{R}_\omega = \{\check{C}_n + (\gamma+\omega) I_p\}^{-1}, \qquad \mathcal{R}_\omega = \{C_n + (\gamma+\omega) I_p\}^{-1}.$$

All these resolvents have uniformly bounded operator norm on Γ.

We first establish the two-dimensional cavity estimate. Delete observation i and set

$$\check{R}_{i,\omega} = \left\{ \kappa(\lambda) I_p + \frac{1}{n}\sum_{j\neq i} F_j F_j^{\mathrm{T}} + (\gamma+\omega) I_p \right\}^{-1},$$

$$T_{i,\omega} = \frac{1}{n} F_i^{\mathrm{T}} \check{R}_{i,\omega} F_i, \qquad C_{i,\omega} = (I_2 + T_{i,\omega})^{-1}.$$

Conditionally on $\check{R}_{i,\omega}$, the rows of F_i are independent centered two-dimensional vectors. We use the following estimate only for deterministic complex symmetric matrices A, so $A^{\mathrm{T}} = A$. For a diagonal channel, apply the Hanson–Wright inequality (Vershynin, 2018, Theorem 6.2.1) separately to the real and imaginary symmetric parts of A. For a mixed channel, polarization gives

$$x^{\mathrm{T}} A y = \frac{1}{4}\{(x+y)^{\mathrm{T}} A (x+y) - (x-y)^{\mathrm{T}} A (x-y)\},$$

where x and y are the two centered endpoint-channel vectors. The coordinates of $x+y$ and $x-y$ are independent, centered, and uniformly sub-Gaussian. Applying the same Hanson–Wright bound to these two quadratic forms and integrating its tail at the fixed eighth moment gives

$$\left\| F_i^{\mathrm{T}} A F_i - \mathrm{tr}(A) V \right\|_{L^8} \le C \|A\|_{\mathrm{F}} = C\{\mathrm{tr}(AA^*)\}^{1/2}.$$

The deleted resolvent $\check{R}_{i,\omega}$ is complex symmetric, so this estimate applies with $A = \check{R}_{i,\omega}$. Thus

$$\left\| T_{i,\omega} - \frac{1}{n}\mathrm{tr}(\check{R}_{i,\omega}) V \right\|_{L^8} \le C n^{-1/2}. \tag{S.3.28}$$

Let

$$\check{t}_n(\omega) = \mathbb{E}\frac{1}{n}\mathrm{tr}\,\check{R}_\omega, \qquad \check{t}_{i,n}(\omega) = \frac{1}{n}\mathrm{tr}\,\check{R}_{i,\omega}.$$

Returning one rank-two block changes the normalized trace by $O(n^{-1})$. The bounded-difference martingale estimate for the independent blocks gives

$$\max_i \|\check{t}_{i,n}(\omega) - \mathbb{E}\check{t}_{i,n}(\omega)\|_{L^8} = O(n^{-1/2}), \qquad \max_i |\mathbb{E}\check{t}_{i,n}(\omega) - \check{t}_n(\omega)| = O(n^{-1}).$$

Taking the normalized trace in the resolvent equation and using

$$\frac{1}{n} F_i^{\mathrm{T}} \check{R}_\omega F_i = T_{i,\omega}(I_2 + T_{i,\omega})^{-1}$$

gives

$$c_n = (\gamma + \omega + \kappa(\lambda))\frac{1}{n}\mathrm{tr}\,\check{R}_\omega + \frac{1}{n}\sum_i \mathrm{tr}\{T_{i,\omega}(I_2 + T_{i,\omega})^{-1}\}.$$

Taking expectations and using (S.3.28) gives

$$\left|(\gamma+\omega+\kappa(\lambda))\check{t}_n(\omega) + \mathrm{tr}\{\check{t}_n(\omega) V (I_2 + \check{t}_n(\omega) V)^{-1}\} - c_n\right| \le C n^{-1/2}. \tag{S.3.29}$$

Let $t_n(\omega)$ be the analytic root of

$$c_n = t_n(\omega)\left[\gamma + \omega + \kappa(\lambda) + \mathrm{tr}\{V(I_2 + t_n(\omega) V)^{-1}\}\right]. \tag{S.3.30}$$

At $\omega = 0$, the derivative with respect to t is

$$\gamma + \kappa(\lambda) + \mathrm{tr}\{V(I_2 + tV)^{-2}\} \ge \gamma.$$

The roots $t_n(0)$ remain in a fixed compact subset $K \Subset (0,\infty)$. The resolvent identity gives

$$|\check{t}_n(\omega) - \check{t}_n(0)| \le C|\omega|,$$

and the analytic implicit-function estimate for (S.3.30) gives

$$|t_n(\omega) - t_n(0)| \le C|\omega|.$$

At $\omega = 0$, the real mean value theorem applied to (S.3.29) gives

$$|\check{t}_n(0) - t_n(0)| = O(n^{-1/2}).$$

Choose a fixed thin complex tube about K on which

$$\mathrm{Re}\left[\gamma + \omega + \kappa(\lambda) + \mathrm{tr}\{V(I_2 + tV)^{-2}\}\right] \ge \frac{\gamma}{2}.$$

After shrinking the ω-disk, the approximate trace, the analytic root, and the line segment joining them remain in this tube. Hence the segment integral of the derivative has real part at least $\gamma/2$, and the line-segment identity applied to (S.3.29) yields

$$\sup_{\omega\in\Gamma} |\check{t}_n(\omega) - t_n(\omega)| = O(n^{-1/2}).$$

Consequently, with $C_{*,n}(\omega) = \{I_2 + t_n(\omega)V\}^{-1}$,

$$\sup_{\omega\in\Gamma} \max_i \{\|T_{i,\omega} - t_n(\omega)V\|_{L^8} + \|C_{i,\omega} - C_{*,n}(\omega)\|_{L^8}\} = O(n^{-1/2}). \tag{S.3.31}$$

We next derive the bilinear and energy equivalents. Put

$$q_n(\omega) = \frac{t_n(\omega)}{c_n}, \qquad \theta_n(\omega) = d_\alpha w^{\mathrm{T}} C_{*,n}(\omega) w,$$

$$c_{B,n}(\omega) = w^{\mathrm{T}}\{I_2 - C_{*,n}(\omega)\}w.$$

When ω is fixed in this argument, we abbreviate $C_{*,n}(\omega)$ and $\theta_n(\omega)$ by $C_{*,n}$ and θ_n, respectively. Woodbury gives the exact score identity

$$u^{\mathrm{T}}\check{R}_\omega \check{s}_n = \frac{1}{n}\sum_i y_i w^{\mathrm{T}} C_{i,\omega} F_i^{\mathrm{T}} \check{R}_{i,\omega} u$$

for every deterministic bounded vector u. Replacing $C_{i,\omega}$ by $C_{*,n}(\omega)$ costs $O(n^{-1/2})$ in L^2 by (S.3.31). Gaussian integration by parts gives

$$\mathbb{E}(y_i F_i) = d_\alpha a w^{\mathrm{T}}.$$

Define

$$X_i = y_i w^{\mathrm{T}} C_{*,n} F_i^{\mathrm{T}} \check{R}_{i,\omega} u - \theta_n a^{\mathrm{T}} \check{R}_{i,\omega} u.$$

It is centered after observation i is deleted and has uniformly bounded fourth moment. If $\check{R}_{ij,\omega}$ deletes i, j, the version $X_i^{(ij)}$ obtained by replacing $\check{R}_{i,\omega}$ with $\check{R}_{ij,\omega}$ is conditionally independent of $X_j^{(ij)}$, and

$$X_i - X_i^{(ij)} = -\frac{1}{n}\{y_i w^{\mathrm{T}} C_{*,n} F_i^{\mathrm{T}} \check{R}_{i,\omega} F_j - \theta_n a^{\mathrm{T}} \check{R}_{i,\omega} F_j\} F_j^{\mathrm{T}} \check{R}_{ij,\omega} u.$$

Successive conditioning and the fixed moments of the two-dimensional rows give $\max_{i\ne j} \|X_i - X_i^{(ij)}\|_{L^2} = O(n^{-1/2})$. Hence

$$\mathbb{E}\left|\frac{1}{n}\sum_i X_i\right|^2 \le \frac{C}{n} + \frac{C}{\sqrt{n}} = o(1).$$

Restoring the same rank-two block gives $a^{\mathrm{T}}\check{R}_{i,\omega}u - a^{\mathrm{T}}\check{R}_\omega u = O_{L^2}(n^{-1})$. For deterministic bounded vectors b, d, the resolvent equation and Woodbury's identity give

$$b^{\mathrm{T}} d = (\gamma + \omega + \kappa(\lambda)) b^{\mathrm{T}} \check{R}_\omega d + \frac{1}{n}\sum_i b^{\mathrm{T}} \check{R}_{i,\omega} F_i C_{i,\omega} F_i^{\mathrm{T}} d.$$

With this abbreviation, put

$$Y_i^{\mathrm{bil}} = b^{\mathrm{T}} \check{R}_{i,\omega} F_i C_{*,n} F_i^{\mathrm{T}} d - \mathrm{tr}(V C_{*,n}) b^{\mathrm{T}} \check{R}_{i,\omega} d.$$

The scalar Y_i^{bil} is centered after block i is deleted. If $\check{R}_{ij,\omega}$ deletes i, j, and $Y_i^{\mathrm{bil},(ij)}$ is defined by replacing $\check{R}_{i,\omega}$ with $\check{R}_{ij,\omega}$, then

$$\begin{aligned} Y_i^{\mathrm{bil}} - Y_i^{\mathrm{bil},(ij)} = & -\frac{1}{n}(b^{\mathrm{T}}\check{R}_{i,\omega}F_j)(F_j^{\mathrm{T}}\check{R}_{ij,\omega}F_i)C_{*,n}(F_i^{\mathrm{T}}d) \\ & + \frac{\mathrm{tr}(VC_{*,n})}{n}(b^{\mathrm{T}}\check{R}_{i,\omega}F_j)(F_j^{\mathrm{T}}\check{R}_{ij,\omega}d). \end{aligned}$$

Successive conditioning and the fixed moments of the two-dimensional rows give

$$\max_{i\neq j}\|Y_i^{\mathrm{bil}} - Y_i^{\mathrm{bil},(ij)}\|_{L^2} = O(n^{-1/2}), \qquad \sup_i \|Y_i^{\mathrm{bil}}\|_{L^4} = O(1).$$

The two leave-two versions are conditionally independent and centered. Expanding the covariance of their average therefore gives

$$\left\|\frac{1}{n}\sum_i Y_i^{\mathrm{bil}}\right\|_{L^2} = o(1).$$

Moreover, returning the rank-two block gives

$$\max_i \|b^{\mathrm{T}}(\check{R}_{i,\omega} - \check{R}_\omega)d\|_{L^2} = O(n^{-1}).$$

By (S.3.31), replacing $C_{i,\omega}$ by $C_{*,n}(\omega)$ costs an $o_{L^2}(1)$ error. Consequently,

$$b^{\mathrm{T}}d = \{\gamma + \omega + \kappa(\lambda) + \mathrm{tr}(VC_{*,n}(\omega))\}b^{\mathrm{T}}\check{R}_\omega d + o_{L^2}(1).$$

The reduced-root equation identifies the coefficient in braces with $c_n/t_n(\omega) = 1/q_n(\omega)$. Thus, uniformly on Γ,

$$b^{\mathrm{T}}\check{R}_\omega d - q_n(\omega)b^{\mathrm{T}}d = o_{L^2}(1)$$

for every deterministic bounded pair b, d. Therefore

$$\sup_{\omega\in\Gamma}\left\|u^{\mathrm{T}}\check{R}_\omega\check{s}_n - \theta_n(\omega)q_n(\omega)u^{\mathrm{T}}a\right\|_{L^2} = o(1). \tag{S.3.32}$$

For the quadratic form, put $\check{s}^{(i)} = n^{-1}\sum_{j\neq i}F_j w y_j$. Expanding one score and using Woodbury gives

$$\check{s}_n^{\mathrm{T}}\check{R}_\omega\check{s}_n = A_n + B_n,$$

where

$$A_n = \frac{1}{n}\sum_i y_i w^{\mathrm{T}}C_{i,\omega}F_i^{\mathrm{T}}\check{R}_{i,\omega}\check{s}^{(i)}, \qquad B_n = \frac{1}{n}\sum_i y_i^2 w^{\mathrm{T}}(I_2 - C_{i,\omega})w.$$

The deleted scores have uniformly bounded eighth moments. Indeed, $\mathbb{E}\|F_i w y_i\|^8 = O(p^4)$ follows by splitting $y_i = g_i^{\mathrm{T}}a + \zeta_i$, using Gaussian moments for the first part and independence and the residual eighth-moment bound for the second. Also, $\mathbb{E}(F_i w y_i) = d_\alpha\|w\|^2 a$ has bounded norm. Since $p/n = O(1)$, centering these independent summands and applying Hilbert Rosenthal (Pinelis, 1994, Theorem 5.3, equation (5.8)) gives

$$\max_i \|\check{s}^{(i)}\|_{L^8} + \max_{i\neq j}\|\check{s}^{(ij)}\|_{L^8} = O(1), \qquad \check{s}^{(ij)} = \frac{1}{n}\sum_{k\notin\{i,j\}}F_k w y_k. \tag{S.3.33}$$

By (S.3.31), we can now replace $C_{i,\omega}$ by $C_{*,n}(\omega)$ with an $O(n^{-1/2})$ error in L^2. Define the centered summands by

$$\begin{aligned} Z_i &= \{y_i w^{\mathrm{T}}C_{*,n}F_i^{\mathrm{T}} - \theta_n a^{\mathrm{T}}\}\check{R}_{i,\omega}\check{s}^{(i)}, \\ Z_i^{(ij)} &= \{y_i w^{\mathrm{T}}C_{*,n}F_i^{\mathrm{T}} - \theta_n a^{\mathrm{T}}\}\check{R}_{ij,\omega}\check{s}^{(ij)}. \end{aligned}$$

Conditionally on the observations other than i, j, the variables $Z_i^{(ij)}$ and $Z_j^{(ij)}$ are independent and centered, by $\mathbb{E}(y_i F_i) = d_\alpha a w^{\mathrm{T}}$. Using $\check{s}^{(i)} = \check{s}^{(ij)} + n^{-1}F_j w y_j$, the exact return is

$$\begin{aligned} Z_i - Z_i^{(ij)} = & -\frac{1}{n}\{y_i w^{\mathrm{T}}C_{*,n}F_i^{\mathrm{T}}\check{R}_{i,\omega}F_j - \theta_n a^{\mathrm{T}}\check{R}_{i,\omega}F_j\}F_j^{\mathrm{T}}\check{R}_{ij,\omega}\check{s}^{(ij)} \\ & + \frac{1}{n}\{y_i w^{\mathrm{T}}C_{*,n}F_i^{\mathrm{T}}\check{R}_{i,\omega}F_j - \theta_n a^{\mathrm{T}}\check{R}_{i,\omega}F_j\}w y_j. \end{aligned}$$

Successive conditioning, using (S.3.33), gives $\sup_i \|Z_i\|_{L^4} = O(1)$ and $\max_{i\neq j} \|Z_i - Z_i^{(ij)}\|_{L^2} = O(n^{-1/2})$. The covariance expansion therefore yields

$$\mathbb{E}\left|\frac{1}{n}\sum_i Z_i\right|^2 \le \frac{C}{n} + \frac{C}{\sqrt{n}} = o(1).$$

To express the remaining average in terms of the full random score, use the exact Woodbury identity

$$\check{R}_\omega \check{s}_n - \check{R}_{i,\omega}\check{s}^{(i)} = \frac{1}{n}\check{R}_{i,\omega}F_i C_{i,\omega}\{wy_i - F_i^{\mathrm{T}}\check{R}_{i,\omega}\check{s}^{(i)}\}. \tag{S.3.34}$$

The Hermitian part of $T_{i,\omega}$ is positive semidefinite, so $\|C_{i,\omega}\|_{\mathrm{op}} \le 1$. Conditional Gaussian moment bounds and (S.3.33) imply, uniformly in i and $\omega \in \Gamma$,

$$\|a^{\mathrm{T}}\check{R}_{i,\omega}F_i\|_{L^8} = O(1), \qquad \|wy_i - F_i^{\mathrm{T}}\check{R}_{i,\omega}\check{s}^{(i)}\|_{L^4} = O(1).$$

Consequently, Hölder's inequality and (S.3.34) give

$$\left\|\frac{1}{n}\sum_i a^{\mathrm{T}}\check{R}_{i,\omega}\check{s}^{(i)} - a^{\mathrm{T}}\check{R}_\omega\check{s}_n\right\|_{L^2} \le \frac{1}{n^2}\sum_i \|a^{\mathrm{T}}\check{R}_{i,\omega}F_i\|_{L^8}\|C_{i,\omega}\|_{L^8,\mathrm{op}}\|wy_i - F_i^{\mathrm{T}}\check{R}_{i,\omega}\check{s}^{(i)}\|_{L^4} \le \frac{C}{n}.$$

Applying (S.3.32) with $u = a$ gives

$$\begin{aligned} A_n &= \frac{1}{n}\sum_i Z_i + \frac{\theta_n}{n}\sum_i a^{\mathrm{T}}\check{R}_{i,\omega}\check{s}^{(i)} + o_{L^2}(1) \\ &= \theta_n a^{\mathrm{T}}\check{R}_\omega\check{s}_n + o_{L^2}(1) = \theta_n(\omega)^2 q_n(\omega)\|a\|^2 + o_{L^2}(1). \end{aligned}$$

For the diagonal term, (S.3.31) and the uniform eighth moments of y_i give

$$B_n - c_{B,n}(\omega)\frac{1}{n}\sum_i y_i^2 = o_{L^2}(1).$$

Independence and the fourth-moment bound yield

$$\left\|\frac{1}{n}\sum_i y_i^2 - \left(\|a\|^2 + \frac{1}{n}\sum_i \mathbb{E}\zeta_i^2\right)\right\|_{L^2} = O(n^{-1/2}).$$

Writing $\rho_n^2 = \|a\|^2 + n^{-1}\sum_i \mathbb{E}\zeta_i^2$, we obtain

$$\sup_{\omega\in\Gamma}\left\|\check{s}_n^{\mathrm{T}}\check{R}_\omega\check{s}_n - \{\theta_n(\omega)^2 q_n(\omega)\|a\|^2 + \rho_n^2 c_{B,n}(\omega)\}\right\|_{L^2} = o(1).$$

We restore empirical centering using the reverse Woodbury identities

$$\mathcal{R}_\omega = \check{R}_\omega + \check{R}_\omega M_n J_n M_n^{\mathrm{T}}\check{R}_\omega, \qquad J_n = (I_2 - M_n^{\mathrm{T}}\check{R}_\omega M_n)^{-1} = I_2 + M_n^{\mathrm{T}}\mathcal{R}_\omega M_n. \tag{S.3.35}$$

Scalar Rosenthal applied coordinatewise gives $\|M_n\|_{L^q,\mathrm{op}} = O(1)$ for every fixed q, and hence $\|J_n\|_{L^q,\mathrm{op}} = O(1)$. For a deterministic bounded vector b, the exact leave-one identity is

$$M_n^{\mathrm{T}}\check{R}_\omega b = \frac{1}{n}\sum_i C_{i,\omega}F_i^{\mathrm{T}}\check{R}_{i,\omega}b.$$

After replacing $C_{i,\omega}$ by $C_{*,n}(\omega)$, each summand is centered conditional on the sample with observation i deleted. A common leave-two resolvent makes the versions indexed by i and j conditionally independent and centered. The rank-two Woodbury return and the fixed Gaussian moments give an $O(n^{-1/2})$ L^2 change per returned block. Expanding the covariances of the average, and then interpolating with its uniform higher moments, proves

$$M_n^{\mathrm{T}}\check{R}_\omega b = o_{L^8}(1)$$

uniformly on Γ.

For $h_n = M_n^{\mathrm{T}} \check{R}_\omega \check{s}_n$, returning observation i gives the exact identity

$$h_n = \frac{1}{n}\sum_i C_{i,\omega} F_i^{\mathrm{T}} \check{R}_{i,\omega} \check{s}^{(i)} + \frac{1}{n}\sum_i (I_2 - C_{i,\omega}) w\, y_i.$$

In the first sum replace $C_{i,\omega}$ by $C_{*,n}(\omega)$, and write the resulting summand as

$$Y_i^{\mathrm{score}} = C_{*,n}(\omega) F_i^{\mathrm{T}} \check{R}_{i,\omega} \check{s}^{(i)} \in \mathbb{C}^2.$$

This replacement is $O_{L^5}(n^{-3/10})$, uniformly on Γ. Indeed, $\|C_{i,\omega}\|_{\mathrm{op}} \le 1$, and the circle can be chosen so that $\|C_{*,n}(\omega)\|_{\mathrm{op}} \le 1$. Interpolating the cavity L^8 estimate (S.3.31) with this deterministic bound gives

$$\sup_{\omega\in\Gamma} \max_i \|C_{i,\omega} - C_{*,n}(\omega)\|_{L^{40/3},\mathrm{op}} = O(n^{-3/10}).$$

The deleted scalar $F_i^{\mathrm{T}} \check{R}_{i,\omega} \check{s}^{(i)}$ has a uniform L^8 bound. Hölder's inequality with $1/5 = 3/40 + 1/8$, followed by Minkowski's inequality, gives the replacement bound. The summand Y_i^{score} is centered after deleting i. Let $Y_i^{\mathrm{score},(ij)}$ use the common leave-two resolvent and score. Conditional endpoint linear-form moments, (S.3.33), and the rank-two return give

$$\max_{i\ne j} \|Y_i^{\mathrm{score}} - Y_i^{\mathrm{score},(ij)}\|_{L^2} = O(n^{-1/2}), \qquad \sup_i \|Y_i^{\mathrm{score}}\|_{L^8} = O(1).$$

By conditional independence and centering of $Y_i^{\mathrm{score},(ij)}, Y_j^{\mathrm{score},(ij)}$, the covariance expansion gives

$$\left\| \frac{1}{n}\sum_i Y_i^{\mathrm{score}} \right\|_{L^2}^2 \le \frac{C}{n} + \frac{C}{\sqrt{n}}.$$

Minkowski's inequality and the displayed uniform L^8 bound give an $O(1)$ L^8 norm for the same average. Interpolation yields $\|n^{-1}\sum_i Y_i^{\mathrm{score}}\|_{L^5} = O(n^{-1/20})$. The second sum has the exact decomposition

$$(I_2 - C_{*,n}(\omega)) w\bar{y} + \frac{1}{n}\sum_i \{C_{*,n}(\omega) - C_{i,\omega}\} w\, y_i.$$

The first term is $O_{L^5}(n^{-1/2})$. The same $L^{40/3}$-L^8 Hölder bound, now using $\sup_i \|y_i\|_{L^8} < \infty$, makes the second $O_{L^5}(n^{-3/10})$. Therefore, uniformly on Γ, for every deterministic bounded vector u,

$$M_n^{\mathrm{T}} \check{R}_\omega u = o_{L^8}(1), \qquad M_n^{\mathrm{T}} \check{R}_\omega \check{s}_n = o_{L^5}(1). \tag{S.3.36}$$

Substitution of (S.3.27) into (S.3.35) gives

$$a^{\mathrm{T}} \mathcal{R}_\omega s_n - a^{\mathrm{T}} \check{R}_\omega \check{s}_n = o_{L^2}(1)$$

and the exact quadratic correction

$$\begin{aligned} s_n^{\mathrm{T}} \mathcal{R}_\omega s_n - \check{s}_n^{\mathrm{T}} \check{R}_\omega \check{s}_n = & -2\bar{y}\, w^{\mathrm{T}} M_n^{\mathrm{T}} \check{R}_\omega \check{s}_n + \bar{y}^2 w^{\mathrm{T}} M_n^{\mathrm{T}} \check{R}_\omega M_n w \\ & + \{M_n^{\mathrm{T}} \check{R}_\omega \check{s}_n - M_n^{\mathrm{T}} \check{R}_\omega M_n w\bar{y}\}^{\mathrm{T}} J_n \\ & \cdot \{M_n^{\mathrm{T}} \check{R}_\omega \check{s}_n - M_n^{\mathrm{T}} \check{R}_\omega M_n w\bar{y}\}. \end{aligned}$$

Since $\|\bar{y}\|_{L^8} = O(n^{-1/2})$, (S.3.36) shows that this correction is $o_{L^2}(1)$.

For the empirical pair mean, write

$$\eta_{ij} = z_{ij} - f_1(g_i) - f_2(g_j).$$

The off-diagonal kernel is canonical in both endpoints. In the sixteenth moment of its n^{-2} ordered-pair average, every surviving sixteen-edge index graph has no vertex of degree one and therefore at most sixteen vertices. There are $O(n^{16})$ surviving tuples, with normalization n^{32}. The diagonal sum is smaller. Thus

$$\widehat{z} = \frac{1}{n^2}\sum_{i,j} z_{ij} = M_n \mathbf{1}_2 + \epsilon_{m,n}, \qquad \|\epsilon_{m,n}\|_{L^{16}} = O(n^{-1/2}). \tag{S.3.37}$$

Equations (S.3.35)– (S.3.37) give

$$a^{\mathrm{T}} \mathcal{R}_\omega \widehat{z} = o_{L^4}(1), \qquad \widehat{z}^{\mathrm{T}} \mathcal{R}_\omega s_n = o_{L^4}(1),$$

uniformly on Γ. To see the moment orders, put $J = J_n$, $M = M_n$, $e = \epsilon_{m,n}$, and $\mathcal{R} = \mathcal{R}_\omega$. Reverse Woodbury implies

$$M^{\mathrm{T}}\mathcal{R}a = JM^{\mathrm{T}}\check{R}_\omega a, \qquad M^{\mathrm{T}}\mathcal{R}s_n = J\{M^{\mathrm{T}}\check{R}_\omega \check{s}_n - M^{\mathrm{T}}\check{R}_\omega M w\bar{y}\}.$$

The first contraction is $o_{L^4}(1)$, by (S.3.36) and the L^8 bound for J. The first term in braces is $o_{L^5}(1)$, and J has a bounded L^{20} norm, so their product is $o_{L^4}(1)$. The remaining term is $o_{L^4}(1)$, since $\bar{y} = O_{L^8}(n^{-1/2})$ and J, M have every fixed moment bounded. Also, (S.3.33), (S.3.27), and the moment bounds for $M, \bar{y}$ give $s_n = O_{L^8}(1)$: split $\bar{y}$ into its projected Gaussian mean and the independent residual mean. Finally, $e = O_{L^{16}}(n^{-1/2})$, and the resolvents are uniformly bounded; $\widehat{z} = M\mathbf{1}_2 + e$ gives the two displayed L^4 limits. Put

$$S_* = C_n + \widehat{z}\widehat{z}^{\mathrm{T}}, \qquad Q_*(\omega) = \{S_* + (\gamma + \omega)I_p\}^{-1}, \qquad r_* = \widehat{z}\bar{y} + s_n.$$

Sherman–Morrison gives

$$Q_*(\omega) = \mathcal{R}_\omega - \frac{\mathcal{R}_\omega \widehat{z}\widehat{z}^{\mathrm{T}}\mathcal{R}_\omega}{1 + \widehat{z}^{\mathrm{T}}\mathcal{R}_\omega\widehat{z}}.$$

The denominator has modulus at least one. Write $A = a^{\mathrm{T}}\mathcal{R}_\omega\widehat{z}$, $B = \widehat{z}^{\mathrm{T}}\mathcal{R}_\omega s_n$, and $H = \widehat{z}^{\mathrm{T}}\mathcal{R}_\omega\widehat{z}$. The estimates above give $A, B = o_{L^4}(1)$; $\bar{y} = O_{L^8}(n^{-1/2})$ and $H = O_{L^8}(1)$, so $B + \bar{y}H = o_{L^4}(1)$. The exact rank-one corrections are

$$a^{\mathrm{T}}Q_*(\omega)r_* - a^{\mathrm{T}}\mathcal{R}_\omega s_n = \frac{A(\bar{y} - B)}{1 + H},$$
$$r_*^{\mathrm{T}}Q_*(\omega)r_* - s_n^{\mathrm{T}}\mathcal{R}_\omega s_n = 2\bar{y}B + \bar{y}^2 H - \frac{(B + \bar{y}H)^2}{1 + H}.$$

Hölder's inequality therefore gives, uniformly on Γ,

$$a^{\mathrm{T}}Q_*(\omega)r_* - a^{\mathrm{T}}\mathcal{R}_\omega s_n = o_{L^2}(1), \qquad r_*^{\mathrm{T}}Q_*(\omega)r_* - s_n^{\mathrm{T}}\mathcal{R}_\omega s_n = o_{L^2}(1).$$

For the Gram and score remainders, consider i.i.d. endpoints (X_i) and a real $d_1 \times d_2$ matrix-valued kernel $\mathcal{K}$ that is canonical in both endpoints. Introduce an independent copy (X'_j) and, for disjoint index sets I, J, put

$$\mathsf{Y}_j(\mathcal{K}) = \sum_{i \in I} \mathcal{K}(\mathsf{X}_i, \mathsf{X}'_j).$$

Let $\mathbb{E}_j$ denote expectation over X'_j, conditional on $(\mathsf{X}_i)_{i \in I}$. For $q \geq 4$, define

$$\mathfrak{S}_q(\mathcal{K}; I, J) = \left\| \left\{ \sum_{j \in J} \mathbb{E}_j \mathsf{Y}_j(\mathcal{K}) \mathsf{Y}_j(\mathcal{K})^{\mathrm{T}} \right\}^{1/2} \right\|_{L^q, \mathrm{op}} + \left\| \left\{ \sum_{j \in J} \mathbb{E}_j \mathsf{Y}_j(\mathcal{K})^{\mathrm{T}} \mathsf{Y}_j(\mathcal{K}) \right\}^{1/2} \right\|_{L^q, \mathrm{op}},$$

$$\mathfrak{M}_q(\mathcal{K}; I, J) = \left\| \max_{j \in J} \|\mathsf{Y}_j(\mathcal{K})\|_{\mathrm{op}} \right\|_{L^q},$$

and let $\mathfrak{S}_q(\mathcal{K})$ and $\mathfrak{M}_q(\mathcal{K})$ be the suprema over the splits used by the decoupling argument. For an independent endpoint X', define the row and column same-endpoint contractions

$$A^{\mathrm{row}}_{\mathcal{K}}(x) = \mathbb{E}_{\mathsf{X}'}[\mathcal{K}(x, \mathsf{X}')\mathcal{K}(x, \mathsf{X}')^{\mathrm{T}}], \quad L^{\mathrm{row}}_{\mathcal{K}}(x, x') = \mathbb{E}_{\mathsf{X}'}[\mathcal{K}(x, \mathsf{X}')\mathcal{K}(x', \mathsf{X}')^{\mathrm{T}}],$$
$$A^{\mathrm{col}}_{\mathcal{K}}(x) = \mathbb{E}_{\mathsf{X}'}[\mathcal{K}(x, \mathsf{X}')^{\mathrm{T}}\mathcal{K}(x, \mathsf{X}')], \quad L^{\mathrm{col}}_{\mathcal{K}}(x, x') = \mathbb{E}_{\mathsf{X}'}[\mathcal{K}(x, \mathsf{X}')^{\mathrm{T}}\mathcal{K}(x', \mathsf{X}')].$$

Factor reversal changes a row contraction into a column contraction but does not reverse the endpoints. Define the endpoint-reversed contractions separately by

$$A^{\mathrm{row},\leftarrow}_{\mathcal{K}}(x) = \mathbb{E}_{\mathsf{X}'}[\mathcal{K}(\mathsf{X}', x)\mathcal{K}(\mathsf{X}', x)^{\mathrm{T}}], \quad L^{\mathrm{row},\leftarrow}_{\mathcal{K}}(x, x') = \mathbb{E}_{\mathsf{X}'}[\mathcal{K}(\mathsf{X}', x)\mathcal{K}(\mathsf{X}', x')^{\mathrm{T}}],$$
$$A^{\mathrm{col},\leftarrow}_{\mathcal{K}}(x) = \mathbb{E}_{\mathsf{X}'}[\mathcal{K}(\mathsf{X}', x)^{\mathrm{T}}\mathcal{K}(\mathsf{X}', x)], \quad L^{\mathrm{col},\leftarrow}_{\mathcal{K}}(x, x') = \mathbb{E}_{\mathsf{X}'}[\mathcal{K}(\mathsf{X}', x)^{\mathrm{T}}\mathcal{K}(\mathsf{X}', x')].$$

Forward Banach-valued decoupling (Minsker & Wei, 2019, Lemma 5.2) requires no symmetry of the kernel. Conditional on one side of the decoupled split, the resulting matrices are independent

and centered, so the rectangular moment inequality (Dirksen et al., 2022, Theorem 10) applies. With $q = \lceil 4\log(2 + d_1 + d_2)\rceil$, these two results give

$$\left\| \frac{1}{n^2} \sum_{i \neq j} \mathcal{K}(\mathsf{X}_i, \mathsf{X}_j) \right\|_{L^q,\mathrm{op}} \leq \frac{Cq}{n^2} \{\mathfrak{S}_q(\mathcal{K}) + \mathfrak{M}_q(\mathcal{K})\}.$$

Here off sets the diagonal entries of a square matrix to zero. Coordinate multiplication for $\mathcal{K}_\eta = \mathrm{off}(\eta_{ij}\eta_{ij}^{\mathrm{T}})$ gives

$$[A^{\mathrm{row}}_{\mathcal{K}_\eta}(x)]_{k\ell} = \mathbf{1}_{\{k=\ell\}} \tau_1(x_k) \sum_{m \neq k} \tau_1(x_m),$$

$$[L^{\mathrm{row}}_{\mathcal{K}_\eta}(x, x')]_{k\ell} = \mathbf{1}_{\{k=\ell\}} \rho_1(x_k, x'_k) \sum_{m \neq k} \rho_1(x_m, x'_m),$$

where the two endpoint contractions, for $\xi \sim \mathcal{N}(0, 1)$, are

$$\tau_1(x) = \mathbb{E}_\xi h_0(x, \xi)^2, \qquad \rho_1(x, x') = \mathbb{E}_\xi\{h_0(x, \xi)h_0(x', \xi)\},$$

$$\tau_2(y) = \mathbb{E}_\xi h_0(\xi, y)^2, \qquad \rho_2(y, y') = \mathbb{E}_\xi\{h_0(\xi, y)h_0(\xi, y')\}.$$

The displayed formulas describe $A^{\mathrm{row}}_{\mathcal{K}_\eta}, L^{\mathrm{row}}_{\mathcal{K}_\eta}$ and use τ_1, ρ_1. Since $\mathcal{K}_\eta(x, y) = \mathrm{off}(\eta(x, y)\eta(x, y)^{\mathrm{T}})$ is symmetric as a matrix, factor reversal gives the same column contractions and does not change τ_1, ρ_1. The endpoint-reversed contractions $A^{\mathrm{row},\leftarrow}_{\mathcal{K}_\eta}, L^{\mathrm{row},\leftarrow}_{\mathcal{K}_\eta}$, and their column counterparts, are formed from $\mathcal{K}_\eta(\mathsf{X}', x)$; their coordinate formulas use τ_2, ρ_2. Scalar Rosenthal and the maximum over p coordinates, followed by forward canonical Banach-valued decoupling (Minsker & Wei, 2019, Lemma 5.2) and, after conditioning, the independent centered rectangular-matrix inequality (Dirksen et al., 2022, Theorem 10), yield, at $q = \lceil 4\log(2p)\rceil$,

$$\mathfrak{S}_q(\mathcal{K}_\eta) + \mathfrak{M}_q(\mathcal{K}_\eta) \leq Cq^C(n\sqrt{p} + p).$$

After division by n^2,

$$\left\| \frac{1}{n^2} \sum_{i \neq j} \mathcal{K}_\eta(g_i, g_j) \right\|_{L^4,\mathrm{op}} \leq Cn^{-1/2} \log^C n.$$

For the endpoint–canonical cross kernel, define

$$u(x) = f_1(x), \qquad v(y) = f_2(y), \qquad \eta(x, y) = \sigma\{\lambda x + (1 - \lambda)y\} - u(x) - v(y),$$

coordinatewise. For $g' \sim \mathcal{N}(0, I_p)$, set

$$D_1(x) = \mathbb{E}_{g'}[(u(x) + v(g'))\eta(x, g')^{\mathrm{T}}],$$

$$D_2(y) = \mathbb{E}_{g'}[(u(g') + v(y))\eta(g', y)^{\mathrm{T}}],$$

$$\mathcal{K}_\times(x, y) = (u(x) + v(y))\eta(x, y)^{\mathrm{T}} - D_1(x) - D_2(y).$$

Feature-coordinate independence makes D_1, D_2 centered and diagonal. For the decoupled samples $(g_i)_{i\in I}$ and an independent copy $(g'_j)_{j\in J}$, put

$$U^\eta_i = \sum_{j\in J} \eta(g_i, g'_j), \qquad V^\eta_j = \sum_{i\in I} \eta(g_i, g'_j).$$

Then the canonical sum splits exactly into

$$\sum_{i\in I, j\in J} \mathcal{K}_\times(g_i, g'_j) = \left\{ \sum_{i\in I} u(g_i)(U^\eta_i)^{\mathrm{T}} - |I| \sum_{j\in J} D_2(g'_j) \right\} + \left\{ \sum_{j\in J} v(g'_j)(V^\eta_j)^{\mathrm{T}} - |J| \sum_{i\in I} D_1(g_i) \right\}.$$

Conditionally on $(g_i)_{i\in I}$, the summands indexed by j in the second brace are independent and centered. Conditionally on $(g'_j)_{j\in J}$, the summands indexed by i in the first brace are independent and centered. For each brace, both the row and column conditional variance operators are diagonal

under its respective conditioning and have $L^{q/2}$-operator norm $O(q^C n^2 p)$; the maximal summand has L^q-operator norm $O(q^C p\sqrt{n})$. The cited rectangular estimate therefore bounds each brace by $O(q^C n\sqrt{p})$. Forward decoupling and division by n^2 give the same $O(n^{-1/2}\log^C n)$ operator norm rate.

The endpoint contribution to the full Gram matrix satisfies

$$\frac{1}{n^2}\sum_{i,j}(u_i+v_j)(u_i+v_j)^{\mathrm{T}} = C_n - \kappa(\lambda)I_p + (M_n\mathbf{1}_2)(M_n\mathbf{1}_2)^{\mathrm{T}}.$$

Using $z_{ij} = u_i + v_j + \eta_{ij}$, the definitions of $\mathcal{K}_\times, D_1, D_2$, and (S.3.37), we therefore obtain

$$\begin{aligned} S - S_* = {} & \frac{1}{n^2}\sum_{i\neq j}\{\eta_{ij}\eta_{ij}^{\mathrm{T}} - \kappa(\lambda)I_p\} - \frac{\kappa(\lambda)}{n}I_p \\ & + \frac{1}{n^2}\sum_{i\neq j}\{\mathcal{K}_\times(g_i,g_j) + \mathcal{K}_\times(g_i,g_j)^{\mathrm{T}}\} \\ & + \frac{2(n-1)}{n^2}\sum_i\{D_1(g_i)+D_2(g_i)\} \\ & + \frac{1}{n^2}\sum_i\{\eta_{ii}\eta_{ii}^{\mathrm{T}} + (u_i+v_i)\eta_{ii}^{\mathrm{T}} + \eta_{ii}(u_i+v_i)^{\mathrm{T}}\} \\ & - (M_n\mathbf{1}_2)\epsilon_{m,n}^{\mathrm{T}} - \epsilon_{m,n}(M_n\mathbf{1}_2)^{\mathrm{T}} - \epsilon_{m,n}\epsilon_{m,n}^{\mathrm{T}}. \end{aligned} \tag{S.3.38}$$

The factor 2 in the third line includes the transposed cross terms: D_1, D_2 are diagonal. The correction $-\kappa(\lambda)I_p/n$ is due to the $n(n-1)$ off-diagonal sample pairs.

For the first line, the bound for $\mathcal{K}_\eta$ controls the off-diagonal matrix entries. For a diagonal entry, apply the scalar Hoeffding decomposition to $h_0(x,y)^2 - \kappa(\lambda)$. At moment order q, its two first-projection averages have norm $O(q^C n^{-1/2})$. Forward decoupling followed by scalar Rosenthal in each endpoint gives $O(q^C/n)$ for the canonical remainder. All scalar kernel moments grow polynomially in q, since the activation is Lipschitz and the inputs are Gaussian. Taking $q = \lceil 4\log(2p)\rceil$ and bounding the L^4 norm of the coordinate maximum by $p^{1/q}$ times the largest coordinate L^q norm gives

$$\left\|\max_{k\le p}\left|\frac{1}{n^2}\sum_{i\neq j}(\eta_{ij,k}^2 - \kappa(\lambda))\right|\right\|_{L^4} \le Cn^{-1/2}\log^C n.$$

The same scalar moment argument, applied to the independent centered diagonal matrices $D_1(g_i), D_2(g_i)$, yields

$$\left\|\frac{1}{n}\sum_i D_1(g_i)\right\|_{L^4,\mathrm{op}} + \left\|\frac{1}{n}\sum_i D_2(g_i)\right\|_{L^4,\mathrm{op}} \le Cn^{-1/2}\log^C n.$$

For the diagonal sample pairs, each of the two $p\times n$ arrays $[u_1+v_1,\dots,u_n+v_n]$ and $[\eta_{11},\dots,\eta_{nn}]$ has independent centered sub-Gaussian entries. Centering here follows from oddness. For either array, a unit bilinear form has a dimension-independent sub-Gaussian tail. A fixed net of each unit sphere, a union bound, and tail integration give

$$\|[u_1+v_1,\dots,u_n+v_n]\|_{L^8,\mathrm{op}} + \|[\eta_{11},\dots,\eta_{nn}]\|_{L^8,\mathrm{op}} \le C(\sqrt{n}+\sqrt{p}).$$

Bounding each matrix product by the product of these operator norms and applying Hölder's inequality shows that the fourth line of (S.3.38) is $O(n^{-1})$ in L^4-operator norm. Finally, $\|M_n\|_{L^{16}} = O(1)$ and (S.3.37) bound the three empirical mean products by $O(n^{-1/2})$ in L^4-operator norm. Combining these bounds with the canonical cross-kernel estimate proves

$$\|S - S_*\|_{L^4,\mathrm{op}} \le Cn^{-1/2}\log^C n. \tag{S.3.39}$$

For the score remainder, let $G\in\mathbb{R}^{n\times p}$ have rows g_i^{T}, and write

$$y_i^{\|} = g_i^{\mathrm{T}}a, \qquad y_i = y_i^{\|} + \zeta_i, \qquad \bar{y}_{\|} = \frac{1}{n}\sum_i y_i^{\|}, \qquad \bar{\zeta} = \frac{1}{n}\sum_i \zeta_i,$$

and put

$$s_{\|,n} = \frac{1}{n}[F_1 w, \dots, F_n w] H_n G a, \qquad s_{\zeta,n} = \frac{1}{n}[F_1 w, \dots, F_n w] H_n \zeta.$$

Thus

$$r_* = \widehat{z}\,\bar{y} + s_{\|,n} + s_{\zeta,n}.$$

Direct expansion with the ordered-pair normalization n^{-2} gives

$$r_\| - (\widehat{z}\,\bar{y}_\| + s_{\|,n}) = \frac{1}{n^2}\sum_{i\neq j} \eta_{ij}\{\lambda y_i^\| + (1-\lambda) y_j^\| - \bar{y}_\|\} + \frac{1}{n^2}\sum_i \eta_{ii}(y_i^\| - \bar{y}_\|),$$

and

$$r_\zeta - (\widehat{z}\,\bar{\zeta} + s_{\zeta,n}) = \frac{1}{n^2}\sum_{i\neq j} \eta_{ij}\{\lambda \zeta_i + (1-\lambda)\zeta_j - \bar{\zeta}\} + \frac{1}{n^2}\sum_i \eta_{ii}(\zeta_i - \bar{\zeta}).$$

For the projected-label off-diagonal kernel

$$K_{ij} = \eta_{ij}\{\lambda y_i^\| + (1-\lambda) y_j^\|\},$$

its two endpoint projections have coordinates

$$(\pi_1(x))_k = (1-\lambda) a_k \mathbb{E}_\xi\{h_0(x_k,\xi)\xi\}, \qquad (\pi_2(y))_k = \lambda a_k \mathbb{E}_\xi\{h_0(\xi, y_k)\xi\}.$$

Their $L^4(\ell_2)$ norms are bounded because $\|a\| = O(1)$ and the scalar Lipschitz-Gaussian moments are bounded. After subtracting these projections, the remaining kernel π_{12} is canonical in both endpoints. For a decoupled split, its two conditional square functions have L^q norm $O(q^C n\sqrt{p})$, and its maximal term has L^q norm $O(q^C\sqrt{np})$. Indeed, the scalar conditional energy consists of one independent sum of $O(p)$-moment terms and one scalar canonical order-two sum with the same $O(np)$ energy scale. Forward decoupling (Minsker & Wei, 2019, Lemma 5.2), followed after conditioning by (Dirksen et al., 2022, Theorem 10), therefore gives

$$\left\| \frac{1}{n^2}\sum_{i\neq j} \pi_{12}(g_i, g_j) \right\|_{L^4} \le C n^{-1/2} \log^C n.$$

The two first-projection averages are $O_{L^4}(n^{-1/2})$ by the independent centered real-Hilbert inequality (Pinelis, 1994, Theorem 5.3, equation (5.8)). The canonical mean satisfies

$$\left\| \frac{1}{n^2}\sum_{i\neq j} \eta_{ij} \right\|_{L^8} \le C\frac{\sqrt{p}}{n}\log^C n, \qquad \|\bar{y}_\|\|_{L^8} = O(n^{-1/2}),$$

so the empirical mean correction is $O_{L^4}(n^{-1}\log^C n)$. Finally, separate the mean from the centered part and apply Hilbert Rosenthal at moment order four to obtain

$$\left\| \frac{1}{n^2}\sum_i \eta_{ii}(y_i^\| - \bar{y}_\|) \right\|_{L^4} = O(n^{-1}).$$

Consequently,

$$\|r_\| - (\widehat{z}\,\bar{y}_\| + s_{\|,n})\|_{L^4} \le C n^{-1/2}\log^C n.$$

For the residual off-diagonal term without $\bar{\zeta}$, regroup by label:

$$\frac{1}{n^2}\sum_{i\neq j} \eta_{ij}\{\lambda\zeta_i + (1-\lambda)\zeta_j\} = \sum_i \zeta_i b_i,$$

$$b_i = \frac{1}{n^2}\sum_{j\neq i}\{\lambda\eta_{ij} + (1-\lambda)\eta_{ji}\}.$$

Conditionally on g_i, the partner-index summands in each orientation are independent centered real-Hilbert vectors. Applying (Pinelis, 1994, Theorem 5.3, equation (5.8)) on each conditioning fiber gives

$$\|b_i\|_{L^4} \le C\frac{\sqrt{p}}{n^{3/2}}.$$

Conditioning next on the projected Gaussian design, the variables $\zeta_i b_i$ are independent and centered. A second application of the same inequality yields

$$\left\|\sum_i \zeta_i b_i\right\|_{L^4} \le C\frac{\sqrt{p}}{n} = O(n^{-1/2}).$$

The scalar eighth-moment graph expansion for the canonical mean gives

$$\left\|\frac{1}{n^2}\sum_{i\neq j}\eta_{ij}\right\|_{L^8} \le C\frac{\sqrt{p}}{n}, \qquad \|\bar{\zeta}\|_{L^8} = O(n^{-1/2}),$$

so its product with $\bar{\zeta}$ is $O_{L^4}(n^{-1})$. Conditionally on the design, the diagonal residual sum is an independent centered Hilbert sum and satisfies

$$\left\|\frac{1}{n^2}\sum_i \eta_{ii}\zeta_i\right\|_{L^4} = O(n^{-1});$$

the separated empirical mean part has the same order. Hence

$$\|r_\zeta - (\widehat{z}\,\bar{\zeta} + s_{\zeta,n})\|_{L^4} = O(n^{-1/2}).$$

Adding the projected and residual components proves

$$r = r_* + e_{r,n}, \qquad \|e_{r,n}\|_{L^4} \le Cn^{-1/2}\log^C n.$$

Set

$$A_i = F_i w(g_i^{\mathrm{T}} a), \qquad B_i = F_i w\zeta_i.$$

The fixed Gaussian moments, independence of ζ_i, and the uniform residual eighth-moment bound give

$$\mathbb{E}\|A_i\|^8 = O(p^4), \qquad \mathbb{E}\|B_i\|^8 = O(p^4).$$

Split the A_i's into their centered parts and their $O(1)$-norm mean; the B_i's are already centered. Hilbert Rosenthal at moment order eight (Pinelis, 1994, Theorem 5.3, equation (5.8)) bounds their raw averages in L^8. The centering identities are

$$s_{\|,n} = \frac{1}{n}\sum_i A_i - M_n w\bar{y}_\|, \qquad s_{\zeta,n} = \frac{1}{n}\sum_i B_i - M_n w\bar{\zeta}.$$

The first correction is $O(n^{-1/2})$ in L^8 by the design-side sixteenth-moment bounds for M_n and $\bar{y}_\|$. The second has the same bound by independence of M_n and $\bar{\zeta}$, using only the residual eighth moment. Hence

$$\|s_{\|,n}\|_{L^8} + \|s_{\zeta,n}\|_{L^8} = O(1).$$

In the decomposition $\bar{y} = \bar{y}_\| + \bar{\zeta}$, the Gaussian projected-signal mean satisfies $\|\bar{y}_\|\|_{L^{16}} = O(n^{-1/2})$, while the standing residual assumption gives $\|\bar{\zeta}\|_{L^8} = O(n^{-1/2})$. The design-dependent product is bounded by

$$\|\widehat{z}\,\bar{y}_\|\|_{L^8} \le \|\widehat{z}\|_{L^{16}}\|\bar{y}_\|\|_{L^{16}} = O(1).$$

Since $\widehat{z}$ depends only on the projected Gaussian design and is independent of $\bar{\zeta}$,

$$\|\widehat{z}\,\bar{\zeta}\|_{L^8} = \|\widehat{z}\|_{L^8}\|\bar{\zeta}\|_{L^8} = O(1).$$

Together with the preceding L^8 bounds for $s_{\|,n}$ and $s_{\zeta,n}$, this proves

$$\|r_*\|_{L^8} + \sup_{\omega\in\Gamma}\|Q_*(\omega)r_*\|_{L^8} = O(1). \tag{S.3.40}$$

Combining the score remainder with (S.3.39) gives

$$S = S_* + E_{S,n}, \qquad r = r_* + e_{r,n}, \qquad \|E_{S,n}\|_{L^4,\mathrm{op}} + \|e_{r,n}\|_{L^4} \le Cn^{-1/2}\log^C n. \tag{S.3.41}$$

In particular, for

$$Q_I(\omega) = \{S + (\gamma+\omega)I_p\}^{-1},$$

the exact resolvent expansion is

$$Q_I(\omega) - Q_*(\omega) = -Q_I(\omega)E_{S,n}Q_*(\omega). \tag{S.3.42}$$

We now take the clean labels $y_i = x_i^{\mathrm{T}}\beta^*$. Then $r = r_\mu$, $\zeta_i = x_i^{\mathrm{T}}\beta_\perp^*$, and $\rho_n^2 = \|a\|^2 + \|\beta_\perp^*\|^2 = \|\beta^*\|^2$. Expanding the linear and quadratic score forms and applying (S.3.40)– (S.3.42) gives

$$\begin{aligned} \sup_{\omega\in\Gamma} \|a^{\mathrm{T}}Q_I(\omega)r_\mu - \theta_n(\omega)q_n(\omega)\|a\|^2\|_{L^2} &\longrightarrow 0, \\ \sup_{\omega\in\Gamma} \|r_\mu^{\mathrm{T}}Q_I(\omega)r_\mu - \{\theta_n(\omega)^2 q_n(\omega)\|a\|^2 + \|\beta^*\|^2 c_{B,n}(\omega)\}\|_{L^2} &\longrightarrow 0. \end{aligned} \tag{S.3.43}$$

□

S.3.6 Supporting lemmas for the fixed-ridge proofs

S.3.6.1 Fixed-point stability from the reduced gap

Lemma S.3.10 Under the standing conditions and (S.1.7), after increasing n_0 if necessary, for every $n \ge n_0$ the two-scalar fixed-point equations defining $\bar{Q}$ and $\bar{v}$ have a unique real solution in a fixed neighborhood of $(t_{*,n}, 0)$. This solution satisfies

$$t = t_{*,n} + O(n^{-1}), \qquad s = O(n^{-1}), \tag{S.3.44}$$

with deterministic uniform constants. In particular, (t, s) remains in a fixed compact subset of $[0,\infty)^2$, and

$$\inf_{n\ge n_0} s_{\min}\Big[I_4 + t(I_2 \otimes V) + \beta_\sigma^2 s J_4\Big] > 0, \qquad \inf_{n\ge n_0} s_{\min}\Big[K(m(t,s)) + \gamma I_p\Big] > 0. \tag{S.3.45}$$

Moreover,

$$\inf_{n\ge n_0} s_{\min}(L) > 0, \tag{S.3.46}$$

and

$$\inf_{n\ge n_0} \beta_\sigma^2(1 - \mathbf{1}_4^{\mathrm{T}}\bar{v}) > 0. \tag{S.3.47}$$

Proof of Lemma S.3.10. The matrix V is the covariance matrix of $(f_1(G), f_2(G))^{\mathrm{T}}$, hence $V \ge 0$. To verify the identity for $\kappa(\lambda)$, put

$$Y = \sigma(\lambda G_1 + (1-\lambda)G_2), \qquad u_1 = f_1(G_1) - \beta_\sigma, \qquad u_2 = f_2(G_2) - \beta_\sigma.$$

Then $Y - \beta_\sigma = u_1 + u_2 + h_0$. Independence gives $\mathbb{E}[u_1 u_2] = 0$, while

$$\mathbb{E}[h_0 \mid G_1] = 0, \qquad \mathbb{E}[h_0 \mid G_2] = 0.$$

Thus u_1, u_2, h_0 are pairwise orthogonal in L^2, and

$$\nu_\alpha - \beta_\sigma^2 = \mathbb{E}u_1^2 + \mathbb{E}u_2^2 + \mathbb{E}h_0^2 = \operatorname{tr} V + \mathbb{E}h_0^2.$$

Consequently, $\kappa(\lambda) = \mathbb{E}h_0^2 \ge 0$. If $\vartheta_1, \vartheta_2 \ge 0$ are the eigenvalues of V, the right-hand side of (S.1.10), minus c_n, is

$$F_n(\tau) = (\gamma + \kappa(\lambda))\tau + \sum_{r=1}^{2} \frac{\vartheta_r \tau}{1 + \vartheta_r \tau} - c_n.$$

It satisfies

$$F_n'(\tau) = \gamma + \kappa(\lambda) + \sum_{r=1}^{2} \frac{\vartheta_r}{(1+\vartheta_r\tau)^2} = \gamma + \kappa(\lambda) + \operatorname{tr}\{V(I_2 + \tau V)^{-2}\} \ge \gamma_-. \tag{S.3.48}$$

Since $F_n(0) = -c_n < 0$ and $F_n(\tau) \to \infty$, the positive root $t_{*,n}$ exists and is unique. For fixed $0 < c_- < c < c_+ < \infty$, one has, for all large n,

$$\frac{c_-}{\gamma_+ + \kappa(\lambda) + \operatorname{tr} V} \le t_{*,n} \le \frac{c_+}{\gamma_-}. \tag{S.3.49}$$

On the free-coordinate line $\varsigma = 0$, we repeatedly use two identities. Since $v = \operatorname{vec}(V)$, with the coordinate order used in the definition of m,

$$m(\tau, 0) = \tau \operatorname{vec}\{(I_2 + \tau V)^{-1} V\}.$$

Writing $C_\tau = (I_2 + \tau V)^{-1}$, this gives

$$\nu_\alpha - \beta_\sigma^2 - v^{\mathrm{T}} m(\tau, 0) = \kappa(\lambda) + \operatorname{tr}(V C_\tau), \tag{S.3.50}$$

$$1 - \mathbf{1}_4^{\mathrm{T}} m(\tau, 0) = \mathbf{1}_2^{\mathrm{T}} C_\tau \mathbf{1}_2 - 1. \tag{S.3.51}$$

The first follows from

$$v^{\mathrm{T}} m(\tau, 0) = \tau \operatorname{tr}(V^2 C_\tau) = \operatorname{tr} V - \operatorname{tr}(V C_\tau),$$

and the identity $\tau C_\tau V = I_2 - C_\tau$ gives the second:

$$\mathbf{1}_4^{\mathrm{T}} m(\tau, 0) = \tau \mathbf{1}_2^{\mathrm{T}} C_\tau V \mathbf{1}_2 = 2 - \mathbf{1}_2^{\mathrm{T}} C_\tau \mathbf{1}_2.$$

Define, in a neighborhood of $x_{*,n} = (t_{*,n}, 0)$,

$$\xi(\tau, \varsigma) = \gamma + \nu_\alpha - \beta_\sigma^2 - v^{\mathrm{T}} m(\tau, \varsigma), \qquad \delta(\tau, \varsigma) = \beta_\sigma^2 \{1 - \mathbf{1}_4^{\mathrm{T}} m(\tau, \varsigma)\}.$$

At the base point, (S.1.10), (S.3.50), and (S.3.51) give

$$\xi_{*,n} := \xi(t_{*,n}, 0) = \gamma + \kappa(\lambda) + \operatorname{tr}(V C_{*,n}) = \frac{c_n}{t_{*,n}}, \qquad \delta_{*,n} := \delta(t_{*,n}, 0) = \beta_\sigma^2 \Delta_{*,n}. \tag{S.3.52}$$

Consequently, uniformly in n,

$$0 < c_\xi \le \xi_{*,n} \le C_\xi, \qquad \delta_{*,n} \ge \beta_\sigma^2 \Delta_0 > 0. \tag{S.3.53}$$

The first matrix in the definition of $m(t_{*,n}, 0)$ is $I_4 + t_{*,n}(I_2 \otimes V) \succeq I_4$. Hence continuity, the compactness in (S.3.49), and (S.3.53) provide a fixed real neighborhood $\mathcal{N}$ of all $x_{*,n}$ on which m has uniformly bounded first two derivatives and

$$\xi(\tau, \varsigma) \ge \frac{c_\xi}{2}, \qquad \delta(\tau, \varsigma) \ge \frac{1}{2} \beta_\sigma^2 \Delta_0. \tag{S.3.54}$$

Choose the neighborhood small enough that the inverse defining m is also uniformly bounded there.

Let

$$P = \frac{1}{p} \mathbf{1}\mathbf{1}^{\mathrm{T}}, \qquad P_\perp = I_p - P.$$

On $\mathcal{N}$,

$$K(m(\tau, \varsigma)) + \gamma I_p = \xi(\tau, \varsigma) I_p + \delta(\tau, \varsigma) \mathbf{1}\mathbf{1}^{\mathrm{T}},$$

and therefore

$$\bar{Q}(\tau, \varsigma) = \xi(\tau, \varsigma)^{-1} P_\perp + \{\xi(\tau, \varsigma) + p\delta(\tau, \varsigma)\}^{-1} P.$$

Define the two free-coordinate maps by

$$\begin{aligned} \mathcal{T}_n(\tau, \varsigma) &:= \frac{p-1}{n\xi(\tau, \varsigma)} + \frac{1}{n\{\xi(\tau, \varsigma) + p\delta(\tau, \varsigma)\}}, \\ \mathcal{S}_n(\tau, \varsigma) &:= \frac{p}{n\{\xi(\tau, \varsigma) + p\delta(\tau, \varsigma)\}}. \end{aligned} \tag{S.3.55}$$

The actual fixed-point values satisfy

$$t = \mathcal{T}_n(t, s), \qquad s = \mathcal{S}_n(t, s).$$

Put

$$\mathcal{G}_n(\tau, \varsigma) = \begin{pmatrix} \tau - \mathcal{T}_n(\tau, \varsigma) \\ \varsigma - \mathcal{S}_n(\tau, \varsigma) \end{pmatrix}.$$

At $x_{*,n}$, equations (S.1.10) and (S.3.52) give the exact residual

$$\mathcal{G}_n(x_{*,n}) = \begin{pmatrix} \dfrac{1}{n\xi_{*,n}} - \dfrac{1}{n(\xi_{*,n} + p\delta_{*,n})} \\ -\dfrac{p}{n(\xi_{*,n} + p\delta_{*,n})} \end{pmatrix}.$$

Thus

$$\|\mathcal{G}_n(x_{*,n})\| \le \frac{C}{n}. \tag{S.3.56}$$

Let

$$H_{*,n} = \xi_{*,n} + p\delta_{*,n}.$$

Set

$$a_{\tau,*,n} = a_\tau(t_{*,n}, 0), \quad a_{\varsigma,*,n} = a_\varsigma(t_{*,n}, 0),$$
$$b_{\tau,*,n} = b_\tau(t_{*,n}, 0), \quad b_{\varsigma,*,n} = b_\varsigma(t_{*,n}, 0).$$

Direct differentiation of (S.3.55) gives the exact Jacobian

$$\mathcal{J}_n := D_{(\tau,\varsigma)}\mathcal{G}_n(x_{*,n}) = \begin{pmatrix} 1 - \dfrac{p-1}{n}\dfrac{a_{\tau,*,n}}{\xi^2_{*,n}} - \dfrac{1}{n}\dfrac{a_{\tau,*,n} + p\beta_\sigma^2 b_{\tau,*,n}}{H^2_{*,n}} & -\dfrac{p-1}{n}\dfrac{a_{\varsigma,*,n}}{\xi^2_{*,n}} - \dfrac{1}{n}\dfrac{a_{\varsigma,*,n} + p\beta_\sigma^2 b_{\varsigma,*,n}}{H^2_{*,n}} \\ -\dfrac{p}{n}\dfrac{a_{\tau,*,n} + p\beta_\sigma^2 b_{\tau,*,n}}{H^2_{*,n}} & 1 - \dfrac{p}{n}\dfrac{a_{\varsigma,*,n} + p\beta_\sigma^2 b_{\varsigma,*,n}}{H^2_{*,n}} \end{pmatrix}.$$

Since $H_{*,n} \ge c_H p$ for some $c_H > 0$, and m_τ, m_ς are uniformly bounded, this becomes

$$\mathcal{J}_n = \begin{pmatrix} F_n'(t_{*,n})/\xi_{*,n} + O(n^{-1}) & O(1) \\ O(n^{-1}) & 1 + O(n^{-1}) \end{pmatrix}. \tag{S.3.57}$$

The upper-left entry follows from

$$\partial_\tau \xi(\tau, 0) = -v^{\mathrm{T}} m_\tau(\tau, 0) = -a_\tau(\tau, 0), \qquad \frac{c_n}{\xi^2_{*,n}} = \frac{t_{*,n}}{\xi_{*,n}},$$

which give

$$1 - \frac{c_n a_\tau(t_{*,n}, 0)}{\xi^2_{*,n}} = \frac{\xi_{*,n} - t_{*,n} a_\tau(t_{*,n}, 0)}{\xi_{*,n}} = \frac{F_n'(t_{*,n})}{\xi_{*,n}}.$$

The derivatives of the mean-direction denominator contribute $O(n^{-1})$ to the lower row and $O(n^{-1})$ to the upper-left entry of (S.3.57). By (S.3.48) and the uniform bounds on $\xi_{*,n}$, the first diagonal entry in (S.3.57) is bounded away from zero. The entries of $\mathcal{J}_n$ are uniformly bounded and

$$\det \mathcal{J}_n = \frac{F_n'(t_{*,n})}{\xi_{*,n}} + O(n^{-1})$$

is bounded away from zero. Hence

$$\sup_{n \ge n_0} \|\mathcal{J}_n^{-1}\| < \infty. \tag{S.3.58}$$

All second derivatives of $\mathcal{G}_n$ are uniformly bounded on a smaller fixed neighborhood of $x_{*,n}$. To see this for the mean-direction terms, set $H = \xi + p\delta$. On that neighborhood, $|H| \ge c_H p$, while its first and second derivatives are $O(p)$; hence

$$\partial_a H^{-1} = -H_a H^{-2}, \qquad \partial_{ab} H^{-1} = 2H_a H_b H^{-3} - H_{ab} H^{-2}$$

are $O(p^{-1})$. The corresponding derivatives of ξ^{-1} are uniformly bounded by (S.3.54).

For the quantitative implicit-function argument, define

$$\mathcal{N}_n(x) = x - \mathcal{J}_n^{-1}\mathcal{G}_n(x).$$

Let $C_J = \sup_{n \ge n_0} \|\mathcal{J}_n^{-1}\|$, let M bound the second derivatives of $\mathcal{G}_n$, and choose a fixed $r_0 > 0$, inside the preceding neighborhood, such that

$$C_J M r_0 \le \frac{1}{2}.$$

On the fixed ball

$$B_n = \{x : \|x - x_{*,n}\| \le r_0\},$$

Taylor's theorem gives, for all sufficiently large n,

$$\sup_{x \in B_n} \|D\mathcal{N}_n(x)\| \le C_J M r_0 \le \frac{1}{2}.$$

Moreover, (S.3.56) gives

$$\|\mathcal{N}_n(x_{*,n}) - x_{*,n}\| \le \frac{C}{n} \le \frac{r_0}{2}$$

for all large n. For every $x \in B_n$,

$$\begin{aligned}\|\mathcal{N}_n(x) - x_{*,n}\| &\le \|\mathcal{N}_n(x) - \mathcal{N}_n(x_{*,n})\| + \|\mathcal{N}_n(x_{*,n}) - x_{*,n}\| \\ &\le \frac{1}{2} r_0 + \frac{1}{2} r_0 = r_0.\end{aligned}$$

Hence $\mathcal{N}_n$ maps B_n into itself and is a contraction there. Its unique fixed point $x_n = (t, s)$ satisfies $\mathcal{G}_n(x_n) = 0$. Finally,

$$\begin{aligned}\|x_n - x_{*,n}\| &\le \|\mathcal{N}_n(x_n) - \mathcal{N}_n(x_{*,n})\| + \|\mathcal{N}_n(x_{*,n}) - x_{*,n}\| \\ &\le \frac{1}{2}\|x_n - x_{*,n}\| + \frac{C}{n},\end{aligned}$$

so $\|x_n - x_{*,n}\| \le 2C/n$, which proves (S.3.44). Since the right-hand sides of (S.3.55) are positive on this neighborhood, $t, s \ge 0$.

It remains to verify the four stability conclusions. First,

$$I_4 + t(I_2 \otimes V) + \beta_\sigma^2 s J_4 \succeq I_4,$$

which proves the first bound in (S.3.45). Next, (S.3.44) and (S.3.52) imply

$$\xi(t, s) \ge \frac{\gamma_-}{2}, \qquad \delta(t, s) \ge \frac{1}{2}\beta_\sigma^2 \Delta_0$$

for all large n. The two eigenvalues of $K(m(t, s)) + \gamma I_p$ are

$$\xi(t, s) \quad \text{on } \mathbf{1}^\perp, \qquad \xi(t, s) + p\delta(t, s) \quad \text{on } \operatorname{span}\{\mathbf{1}\}.$$

This proves the second bound in (S.3.45), as well as (S.3.47).

To identify the Jacobian, differentiate the free-coordinate matrix

$$\bar{Q}(\tau, \varsigma) = \{K(m(\tau, \varsigma)) + \gamma I_p\}^{-1}$$

and evaluate at $(\tau, \varsigma) = (t, s)$ to obtain

$$\partial_\tau \bar{Q}(\tau, \varsigma)\big|_{(\tau,\varsigma)=(t,s)} = \bar{Q}\{a_\tau I_p + \beta_\sigma^2 b_\tau \mathbf{1}\mathbf{1}^\top\}\bar{Q}, \qquad \partial_\varsigma \bar{Q}(\tau, \varsigma)\big|_{(\tau,\varsigma)=(t,s)} = \bar{Q}\{a_\varsigma I_p + \beta_\sigma^2 b_\varsigma \mathbf{1}\mathbf{1}^\top\}\bar{Q}.$$

Taking the normalized trace and the normalized $\mathbf{1}$-quadratic form gives the identity

$$D_{(\tau,\varsigma)}\mathcal{G}_n(t, s) = L,$$

with L as in (S.1.14). The uniform second-derivative bound and $\|(t, s) - x_{*,n}\| = O(n^{-1})$ give

$$\|L - \mathcal{J}_n\| = O(n^{-1}).$$

Together with (S.3.58), this proves (S.3.46). □

From now on, (t, s), $\bar{v} = m(t, s)$, and $\bar{Q}$ refer to the local solution given by Lemma S.3.10. All asymptotic statements omit the finitely many indices below n_0.

S.3.6.2 Mean-spike control

Lemma S.3.11 Assume $\beta_\sigma \neq 0$. For $|\mathcal{I}| \leq 2$, retain the original n^{-2} normalization and define

$$S^{(\mathcal{I})} = \frac{1}{n^2} \sum_{a,b \notin \mathcal{I}} z_{ab} z_{ab}^\top, \qquad Q_I^{(\mathcal{I})}(\omega) = \{S^{(\mathcal{I})} + (\gamma + \omega) I_p\}^{-1}.$$

There is a fixed closed complex disk $\mathcal{U}_{\mathrm{ms}}$ about zero, independent of n, such that $\sup_{\omega \in \mathcal{U}_{\mathrm{ms}}} |\omega| \leq \gamma_-/4$ and

$$\sup_{|\mathcal{I}| \leq 2} \sup_{\omega \in \mathcal{U}_{\mathrm{ms}}} \mathbb{E}\left[|\mathbf{1}^\top Q_I^{(\mathcal{I})}(\omega)\mathbf{1}|^4 + \|Q_I^{(\mathcal{I})}(\omega)\mathbf{1}\|^8 \right] = O(1). \tag{S.3.59}$$

Proof. Fix $\mathcal{I} \subset [n]$ with $q = |\mathcal{I}| \leq 2$, put $J = [n] \setminus \mathcal{I}$ and $m = |J| = n - q$, and define

$$S^{(\mathcal{I})} = \frac{1}{n^2} \sum_{a,b \in J} z_{ab} z_{ab}^\top, \qquad Q^{(\mathcal{I})} = (S^{(\mathcal{I})} + \gamma I_p)^{-1}, \qquad \widehat{z}^{(\mathcal{I})} = \frac{1}{n^2} \sum_{a,b \in J} z_{ab}.$$

For one feature coordinate, set

$$h(x, y) = \sigma(\lambda x + (1 - \lambda) y), \qquad \psi(x) = f_1(x) + f_2(x) - 2\beta_\sigma,$$

and

$$h_0(x, y) = h(x, y) - f_1(x) - f_2(y) + \beta_\sigma.$$

For independent standard Gaussians G_1, G_2,

$$\mathbb{E}[h_0(G_1, G_2) \mid G_1] = \mathbb{E}[h_0(G_1, G_2) \mid G_2] = 0, \qquad \mathbb{E}\psi(G_1) = 0. \tag{S.3.60}$$

The diagonal value is not canonical:

$$\mathbb{E}h_0(G, G) = \mu_0(\sigma) - \beta_\sigma. \tag{S.3.61}$$

The definition of h_0 gives the exact decomposition

$$h(x, y) = \beta_\sigma + \{f_1(x) - \beta_\sigma\} + \{f_2(y) - \beta_\sigma\} + h_0(x, y). \tag{S.3.62}$$

Summing (S.3.62) over $a, b \in J$ gives, coordinatewise,

$$\widehat{z}_k^{(\mathcal{I})} = \frac{m^2}{n^2}\beta_\sigma + \frac{m}{n^2} \sum_{a \in J} \psi(g_{ak}) + \frac{1}{n^2} \sum_{a,b \in J} h_0(g_{ak}, g_{bk}). \tag{S.3.63}$$

Taking expectations in (S.3.63) and using (S.3.60)–(S.3.61) yields

$$\beta_{n,\mathcal{I}} := \mathbb{E}\widehat{z}_k^{(\mathcal{I})} = \frac{m(m-1)}{n^2}\beta_\sigma + \frac{m}{n^2}\mu_0(\sigma), \tag{S.3.64}$$

because the $m(m-1)$ off-diagonal h_0-terms have zero mean and the m diagonal terms have mean $\mu_0(\sigma) - \beta_\sigma$. Subtracting (S.3.64) from (S.3.63) gives

$$\begin{aligned} \delta_k^{(\mathcal{I})} &:= \widehat{z}_k^{(\mathcal{I})} - \beta_{n,\mathcal{I}} \\ &= \frac{m}{n^2} \sum_{a \in J} \psi(g_{ak}) + \frac{1}{n^2} \sum_{\substack{a,b \in J \\ a \neq b}} h_0(g_{ak}, g_{bk}) \\ &\quad + \frac{1}{n^2} \sum_{a \in J} \{h_0(g_{ak}, g_{ak}) - (\mu_0(\sigma) - \beta_\sigma)\}. \end{aligned}$$

The three terms A_k, B_k, C_k are, respectively, an average of independent centered variables, a canonical ordered U-statistic of order two, and a diagonal correction:

$$\begin{aligned} A_k &= \frac{m}{n^2} \sum_{a \in J} \psi(g_{ak}), \\ B_k &= \frac{1}{n^2} \sum_{\substack{a,b \in J \\ a \neq b}} h_0(g_{ak}, g_{bk}), \\ C_k &= \frac{1}{n^2} \sum_{a \in J} \{h_0(g_{ak}, g_{ak}) - (\mu_0(\sigma) - \beta_\sigma)\}. \end{aligned}$$

Thus $\delta_k^{(I)} = A_k + B_k + C_k$. Put

$$h_{\mathrm{diag}}(x) = h_0(x,x) - \{\mu_0(\sigma) - \beta_\sigma\}$$

and, for $r \in \{2,4,8\}$, define

$$\mu_{\psi,r} = \mathbb{E}|\psi(G)|^r, \qquad \mu_{\mathrm{diag},r} = \mathbb{E}|h_{\mathrm{diag}}(G)|^r, \qquad M_r = \mathbb{E}|h_0(G_1,G_2)|^r.$$

Lipschitz continuity and Gaussian moments imply that these quantities are finite. For iid centered real variables $X_1,\dots,X_m$, with $\mu_r = \mathbb{E}|X_1|^r$, we use

$$\begin{aligned}
\mathbb{E}\left|\sum_{a=1}^m X_a\right|^2 &= m\mu_2,\\
\mathbb{E}\left|\sum_{a=1}^m X_a\right|^4 &= m\mu_4 + 3m(m-1)\mu_2^2,\\
\mathbb{E}\left|\sum_{a=1}^m X_a\right|^8 &\le C\{m\mu_8 + (m\mu_2)^4\}.
\end{aligned} \tag{S.3.65}$$

The last line is the order-eight Rosenthal inequality.

By (S.3.60), applying (S.3.65) to $X_a = \psi(g_{ak})$ gives

$$\mathbb{E}|A_k|^2 = \frac{m^2}{n^4}\, m\mu_{\psi,2} = O\!\left(\frac{m^3}{n^4}\right) = O(n^{-1}), \tag{S.3.66}$$

$$\mathbb{E}|A_k|^4 = \frac{m^4}{n^8}\{m\mu_{\psi,4} + 3m(m-1)\mu_{\psi,2}^2\} = O\!\left(\frac{m^6}{n^8}\right) = O(n^{-2}). \tag{S.3.67}$$

For B_k, let

$$\mathcal{E}_J = \{(a,b) \in J^2 : a \neq b\}, \qquad K_{ab} = h_0(g_{ak}, g_{bk}).$$

Then

$$B_k = \frac{1}{n^2}\sum_{(a,b)\in\mathcal{E}_J} K_{ab}, \qquad \mathbb{E}[K_{ab} \mid g_{ak}] = \mathbb{E}[K_{ab} \mid g_{bk}] = 0.$$

For $(a_\ell, b_\ell) \in \mathcal{E}_J$, $1 \le \ell \le s$, set

$$d_v^{(s)} = \sum_{\ell=1}^s \{\mathbf{1}_{\{a_\ell = v\}} + \mathbf{1}_{\{b_\ell = v\}}\}.$$

If $d_v^{(s)} = 1$, let ℓ_v be the unique factor containing v. With $\mathcal{F}_{-v,k} = \sigma(g_{uk} : u \in J \setminus \{v\})$, conditioning on all Gaussian variables except g_{vk} gives

$$\mathbb{E}\prod_{\ell=1}^s K_{a_\ell b_\ell} = \mathbb{E}\left[\prod_{\ell\neq\ell_v} K_{a_\ell b_\ell}\, \mathbb{E}\Big[K_{a_{\ell_v} b_{\ell_v}} \mid \mathcal{F}_{-v,k}\Big]\right] = 0. \tag{S.3.68}$$

The last equality follows from (S.3.60).

For $s = 2$, (S.3.68) leaves only $(c,d) = (a,b)$ and $(c,d) = (b,a)$. Hence

$$\begin{aligned}
\mathbb{E}|B_k|^2 &= \frac{1}{n^4}\sum_{(a,b),(c,d)\in\mathcal{E}_J} \mathbb{E}[K_{ab}K_{cd}]\\
&= \frac{m(m-1)}{n^4}\left\{M_2 + \mathbb{E}[h_0(G_1,G_2)h_0(G_2,G_1)]\right\}.
\end{aligned} \tag{S.3.69}$$

By Cauchy–Schwarz and exchangeability,

$$|\mathbb{E}[h_0(G_1,G_2)h_0(G_2,G_1)]| \le M_2.$$

For $s = 4$, the exact expansion is

$$\mathbb{E}|B_k|^4 = \frac{1}{n^8}\sum_{\substack{(a_\ell,b_\ell)\in\mathcal{E}_J\\ 1\le\ell\le 4}} \mathbb{E}\prod_{\ell=1}^4 K_{a_\ell b_\ell}. \tag{S.3.70}$$

Every nonzero summand in (S.3.70) satisfies $d_v^{(4)} \geq 2$ for each occurring index v, by (S.3.68). Since $\sum_v d_v^{(4)} = 8$, the number r of distinct indices lies in $\{2, 3, 4\}$. Therefore the number $C_{4,m}$ of possibly nonzero summands satisfies

$$C_{4,m} \leq \sum_{r=2}^{4} \binom{m}{r} [r(r-1)]^4 \leq \sum_{r=2}^{4} \binom{m}{r} r^8 \leq C m^4. \tag{S.3.71}$$

For each such summand, Hölder's inequality gives

$$\left| \mathbb{E} \prod_{\ell=1}^{4} K_{a_\ell b_\ell} \right| \leq \prod_{\ell=1}^{4} \{\mathbb{E}|K_{a_\ell b_\ell}|^4\}^{1/4} = M_4.$$

Combining (S.3.69), (S.3.70), and (S.3.71) yields

$$\mathbb{E}|B_k|^2 \leq C \frac{m^2}{n^4} = O(n^{-2}), \qquad \mathbb{E}|B_k|^4 \leq C \frac{m^4}{n^8} = O(n^{-4}). \tag{S.3.72}$$

Finally, (S.3.61) shows that the variables $(h_{\mathrm{diag}}(g_{ak}))_{a \in J}$ are iid and centered. Applying (S.3.65) once more gives

$$\begin{aligned} \mathbb{E}|C_k|^2 &= \frac{m}{n^4} \mu_{\mathrm{diag},2} = O\left(\frac{m}{n^4}\right) = O(n^{-3}), \\ \mathbb{E}|C_k|^4 &= \frac{m\mu_{\mathrm{diag},4} + 3m(m-1)\mu_{\mathrm{diag},2}^2}{n^8} = O\left(\frac{m^2}{n^8}\right) = O(n^{-6}). \end{aligned} \tag{S.3.73}$$

Using $|x+y+z|^2 \leq 3(|x|^2 + |y|^2 + |z|^2)$ and $|x+y+z|^4 \leq 27(|x|^4 + |y|^4 + |z|^4)$ in (S.3.66)–(S.3.73) therefore proves, uniformly over $|\mathcal{I}| \leq 2$,

$$\mathbb{E}|\delta_k^{(\mathcal{I})}|^2 \leq \frac{C}{n}, \qquad \mathbb{E}|\delta_k^{(\mathcal{I})}|^4 \leq \frac{C}{n^2}. \tag{S.3.74}$$

For $k \leq p$, let $\mathcal{H}_k = \sigma(g_{ak} : a \in J)$. Conditional on the orthonormal feature matrix W,

$$\mathrm{Cov}(g_{ak}, g_{b\ell} \mid W) = \delta_{ab}\delta_{k\ell}, \qquad ((g_{ak})_{a \in J})_{k=1}^{p} \mid W \sim \bigotimes_{k=1}^{p} \mathcal{N}(0, I_m). \tag{S.3.75}$$

The conditional law in (S.3.75) does not depend on W. Since $\delta_k^{(\mathcal{I})}$ is $\mathcal{H}_k$-measurable and centered, the variables $\delta_1^{(\mathcal{I})}, \ldots, \delta_p^{(\mathcal{I})}$ are iid and centered, both conditionally on W and unconditionally.

Let

$$e = \frac{\mathbf{1}}{\sqrt{p}}, \qquad P_\perp = I_p - ee^\top, \qquad \delta^{(\mathcal{I})} = (\delta_1^{(\mathcal{I})}, \ldots, \delta_p^{(\mathcal{I})})^\top.$$

Since $\widehat{z}^{(\mathcal{I})} = \beta_{n,\mathcal{I}} \sqrt{p}\, e + \delta^{(\mathcal{I})}$, its orthogonal decomposition is

$$\begin{aligned} \widehat{z}^{(\mathcal{I})} &= \mathfrak{a}_{\mathcal{I}} e + b_{\mathcal{I}} \\ &= \left\{ \beta_{n,\mathcal{I}} \sqrt{p} + \frac{1}{\sqrt{p}} \sum_{k=1}^{p} \delta_k^{(\mathcal{I})} \right\} e + P_\perp \delta^{(\mathcal{I})}. \end{aligned} \tag{S.3.76}$$

Consequently,

$$b_{\mathcal{I}} = P_\perp \delta^{(\mathcal{I})}, \qquad \mathfrak{a}_{\mathcal{I}} - \beta_{n,\mathcal{I}} \sqrt{p} = \frac{1}{\sqrt{p}} \sum_{k=1}^{p} \delta_k^{(\mathcal{I})}.$$

Orthogonal projection does not increase the Euclidean norm, so independence of the original coordinates and (S.3.74) give

$$\mathbb{E}\|b_{\mathcal{I}}\|^2 \leq \mathbb{E}\|\delta^{(\mathcal{I})}\|^2 = p\, \mathbb{E}|\delta_1^{(\mathcal{I})}|^2 = O(p/n) = O(1), \tag{S.3.77}$$

$$\begin{aligned} \mathbb{E}\|b_{\mathcal{I}}\|^4 &\leq \mathbb{E}\|\delta^{(\mathcal{I})}\|^4 \\ &= p\, \mathbb{E}|\delta_1^{(\mathcal{I})}|^4 + p(p-1)(\mathbb{E}|\delta_1^{(\mathcal{I})}|^2)^2 \\ &= O(p/n^2) + O(p^2/n^2) = O(1). \end{aligned} \tag{S.3.78}$$

For the e-component, the centered-sum identities yield

$$\mathbb{E}|\mathfrak{a}_{\mathcal{I}} - \beta_{n,\mathcal{I}}\sqrt{p}|^2 = \frac{1}{p}\, p\, \mathbb{E}|\delta_1^{(\mathcal{I})}|^2 = O(n^{-1}), \tag{S.3.79}$$

$$\begin{aligned}\mathbb{E}|\mathfrak{a}_{\mathcal{I}} - \beta_{n,\mathcal{I}}\sqrt{p}|^4 &= \frac{1}{p^2}\left\{p\,\mathbb{E}|\delta_1^{(\mathcal{I})}|^4 + 3p(p-1)(\mathbb{E}|\delta_1^{(\mathcal{I})}|^2)^2\right\} \\ &= O((pn^2)^{-1}) + O(n^{-2}) = O(n^{-2}).\end{aligned} \tag{S.3.80}$$

Equations (S.3.77)–(S.3.80) give

$$\|b_{\mathcal{I}}\|_{L^4} = O(1), \qquad \|\mathfrak{a}_{\mathcal{I}} - \beta_{n,\mathcal{I}}\sqrt{p}\|_{L^4} = O(n^{-1/2}).$$

Because $q = |\mathcal{I}| \le 2$ and $m = n - q$,

$$\begin{aligned}\beta_{n,\mathcal{I}} &= \beta_\sigma + \frac{\mu_0(\sigma) - (2q+1)\beta_\sigma}{n} \\ &\quad + \frac{q(q+1)\beta_\sigma - q\mu_0(\sigma)}{n^2} = \beta_\sigma + O(n^{-1})\end{aligned}$$

uniformly over $|\mathcal{I}| \le 2$. Since $\beta_\sigma \neq 0$,

$$\inf_{|\mathcal{I}|\le 2} |\beta_{n,\mathcal{I}}| \ge \frac{3}{4}|\beta_\sigma| \tag{S.3.81}$$

for all sufficiently large n. Define

$$\mathcal{G}_{\mathcal{I}} = \left\{|\mathfrak{a}_{\mathcal{I}} - \beta_{n,\mathcal{I}}\sqrt{p}| \le \frac{1}{4}|\beta_\sigma|\sqrt{p}\right\}. \tag{S.3.82}$$

By (S.3.81), on $\mathcal{G}_{\mathcal{I}}$,

$$|\mathfrak{a}_{\mathcal{I}}| \ge |\beta_{n,\mathcal{I}}|\sqrt{p} - |\mathfrak{a}_{\mathcal{I}} - \beta_{n,\mathcal{I}}\sqrt{p}| \ge \frac{1}{2}|\beta_\sigma|\sqrt{p}.$$

Using (S.3.80), Markov's inequality gives

$$\mathbb{P}(\mathcal{G}_{\mathcal{I}}^c) \le \frac{4^4\,\mathbb{E}|\mathfrak{a}_{\mathcal{I}} - \beta_{n,\mathcal{I}}\sqrt{p}|^4}{|\beta_\sigma|^4 p^2} \le \frac{C}{n^2 p^2}. \tag{S.3.83}$$

Set $c_{\mathcal{I}} = m^2/n^2$. Weighted centering gives

$$S^{(\mathcal{I})} - c_{\mathcal{I}}^{-1}\widehat{z}^{(\mathcal{I})}(\widehat{z}^{(\mathcal{I})})^{\mathsf{T}} = \frac{1}{n^2}\sum_{a,b\in J}\left(z_{ab} - c_{\mathcal{I}}^{-1}\widehat{z}^{(\mathcal{I})}\right)\left(z_{ab} - c_{\mathcal{I}}^{-1}\widehat{z}^{(\mathcal{I})}\right)^{\mathsf{T}} \ge 0. \tag{S.3.84}$$

Since $c_{\mathcal{I}} \le 1$, (S.3.84) and Loewner monotonicity imply

$$\begin{aligned}S^{(\mathcal{I})} &\ge c_{\mathcal{I}}^{-1}\widehat{z}^{(\mathcal{I})}(\widehat{z}^{(\mathcal{I})})^{\mathsf{T}} \ge \widehat{z}^{(\mathcal{I})}(\widehat{z}^{(\mathcal{I})})^{\mathsf{T}}, \\ 0 \le Q^{(\mathcal{I})} &\le \{\gamma I_p + \widehat{z}^{(\mathcal{I})}(\widehat{z}^{(\mathcal{I})})^{\mathsf{T}}\}^{-1}.\end{aligned} \tag{S.3.85}$$

For the rank-one inverse, Sherman–Morrison gives

$$\{\gamma I_p + \widehat{z}^{(\mathcal{I})}(\widehat{z}^{(\mathcal{I})})^{\mathsf{T}}\}^{-1} = \frac{1}{\gamma}I_p - \frac{1}{\gamma}\frac{\widehat{z}^{(\mathcal{I})}(\widehat{z}^{(\mathcal{I})})^{\mathsf{T}}}{\gamma + \|\widehat{z}^{(\mathcal{I})}\|^2}. \tag{S.3.86}$$

Since $e^{\mathsf{T}}\widehat{z}^{(\mathcal{I})} = \mathfrak{a}_{\mathcal{I}}$ and $\|\widehat{z}^{(\mathcal{I})}\|^2 = \mathfrak{a}_{\mathcal{I}}^2 + \|b_{\mathcal{I}}\|^2$, (S.3.85) and (S.3.86) yield

$$\begin{aligned}U_{\mathcal{I}} &:= \mathbf{1}^{\mathsf{T}}Q^{(\mathcal{I})}\mathbf{1} \\ &\le p\, e^{\mathsf{T}}\left\{\widehat{z}^{(\mathcal{I})}(\widehat{z}^{(\mathcal{I})})^{\mathsf{T}} + \gamma I_p\right\}^{-1} e \\ &= \frac{p}{\gamma}\frac{\gamma + \|b_{\mathcal{I}}\|^2}{\gamma + \mathfrak{a}_{\mathcal{I}}^2 + \|b_{\mathcal{I}}\|^2}.\end{aligned} \tag{S.3.87}$$

On $\mathcal{G}_{\mathcal{I}}$, use $|\mathfrak{a}_{\mathcal{I}}|^2 \ge |\beta_\sigma|^2 p/4$ in (S.3.87); on its complement, use $Q^{(\mathcal{I})} \le \gamma^{-1} I_p$. Thus

$$U_{\mathcal{I}}\mathbf{1}_{\mathcal{G}_{\mathcal{I}}} \le \frac{4}{\gamma|\beta_\sigma|^2}(\gamma + \|b_{\mathcal{I}}\|^2), \qquad 0 \le U_{\mathcal{I}} \le \frac{p}{\gamma}.$$

Consequently, for $r \in \{2, 4\}$,

$$\mathbb{E}U_{\mathcal{I}}^r \le C_r \mathbb{E}(\gamma + \|b_{\mathcal{I}}\|^2)^r + \left(\frac{p}{\gamma}\right)^r \mathbb{P}(\mathcal{G}_{\mathcal{I}}^c). \tag{S.3.88}$$

Taking $r = 2$, and using (S.3.77), (S.3.78), and (S.3.83), gives

$$\begin{aligned}\mathbb{E}U_{\mathcal{I}}^2 &\le C\mathbb{E}(\gamma + \|b_{\mathcal{I}}\|^2)^2 + \frac{p^2}{\gamma^2}\mathbb{P}(\mathcal{G}_{\mathcal{I}}^c) \\ &\le C\{\gamma^2 + 2\gamma\mathbb{E}\|b_{\mathcal{I}}\|^2 + \mathbb{E}\|b_{\mathcal{I}}\|^4\} + \frac{C}{n^2} = O(1),\end{aligned} \tag{S.3.89}$$

uniformly for $|\mathcal{I}| \le 2$. Moreover, $0 \le Q^{(\mathcal{I})} \le \gamma^{-1} I_p$, so

$$V_{\mathcal{I}} := \mathbf{1}^{\mathrm{T}}(Q^{(\mathcal{I})})^2\mathbf{1} = \|Q^{(\mathcal{I})}\mathbf{1}\|^2 \le \gamma^{-1} U_{\mathcal{I}}. \tag{S.3.90}$$

Equations (S.3.89) and (S.3.90) give

$$\mathbb{E}V_{\mathcal{I}} \le \gamma^{-1}(\mathbb{E}U_{\mathcal{I}}^2)^{1/2} = O(1), \qquad \mathbb{E}V_{\mathcal{I}}^2 \le \gamma^{-2}\mathbb{E}U_{\mathcal{I}}^2 = O(1).$$

We use these lower-order bounds below.

The order-eight inequality in (S.3.65), applied to the sums defining A_k and C_k, gives

$$\begin{aligned}\mathbb{E}|A_k|^8 &\le C\left(\frac{m}{n^2}\right)^8 \{m\mu_{\psi,8} + (m\mu_{\psi,2})^4\} = O\left(\frac{m^{12}}{n^{16}}\right) = O(n^{-4}), \\ \mathbb{E}|C_k|^8 &\le \frac{C}{n^{16}}\{m\mu_{\mathrm{diag},8} + (m\mu_{\mathrm{diag},2})^4\} = O\left(\frac{m^4}{n^{16}}\right) = O(n^{-12}).\end{aligned}$$

For B_k, the exact eighth-moment expansion is

$$\mathbb{E}|B_k|^8 = \frac{1}{n^{16}} \sum_{\substack{(a_\ell, b_\ell) \in \mathcal{E}_J \\ 1 \le \ell \le 8}} \mathbb{E}\prod_{\ell=1}^{8} K_{a_\ell b_\ell}. \tag{S.3.91}$$

By (S.3.68), a nonzero summand in (S.3.91) satisfies $d_v^{(8)} \ge 2$ for every occurring v. Since $\sum_v d_v^{(8)} = 16$, it contains at most eight distinct indices, and

$$C_{8,m} \le \sum_{r=2}^{8} \binom{m}{r} [r(r-1)]^8 \le Cm^8.$$

For each summand, Hölder's inequality gives

$$\left|\mathbb{E}\prod_{\ell=1}^{8} K_{a_\ell b_\ell}\right| \le \prod_{\ell=1}^{8} \{\mathbb{E}|K_{a_\ell b_\ell}|^8\}^{1/8} = M_8.$$

Hence

$$\begin{aligned}\mathbb{E}|B_k|^8 &\le \frac{M_8 C_{8,m}}{n^{16}} \le C\frac{m^8}{n^{16}} = O(n^{-8}), \\ \mathbb{E}|\delta_k^{(\mathcal{I})}|^8 &\le 3^7\{\mathbb{E}|A_k|^8 + \mathbb{E}|B_k|^8 + \mathbb{E}|C_k|^8\} \le Cn^{-4}.\end{aligned} \tag{S.3.92}$$

The deterministic inequality $(\sum_{k=1}^p x_k^2)^4 \le p^3 \sum_{k=1}^p |x_k|^8$, (S.3.76), and (S.3.92) give

$$\begin{aligned}\mathbb{E}\|b_{\mathcal{I}}\|^8 &\le \mathbb{E}\left(\sum_{k=1}^{p} |\delta_k^{(\mathcal{I})}|^2\right)^4 \\ &\le p^3 \sum_{k=1}^{p} \mathbb{E}|\delta_k^{(\mathcal{I})}|^8 \le C\left(\frac{p}{n}\right)^4 = O(1).\end{aligned} \tag{S.3.93}$$

For the mean component, the order-eight Rosenthal inequality, (S.3.75), (S.3.74), and (S.3.92) give

$$\begin{aligned}
&\mathbb{E}\left|\mathfrak{a}_{\mathcal{I}} - \beta_{n,\mathcal{I}}\sqrt{p}\right|^8 \\
&\quad = \frac{1}{p^4}\mathbb{E}\left|\sum_{k=1}^{p}\delta_k^{(\mathcal{I})}\right|^8 \\
&\quad \le \frac{C}{p^4}\left\{p\,\mathbb{E}|\delta_1^{(\mathcal{I})}|^8 + (p\,\mathbb{E}|\delta_1^{(\mathcal{I})}|^2)^4\right\} = O(n^{-4}).
\end{aligned} \tag{S.3.94}$$

Thus

$$\mathbb{E}\|b_{\mathcal{I}}\|^8 = O(1), \qquad \mathbb{E}\left|\mathfrak{a}_{\mathcal{I}} - \beta_{n,\mathcal{I}}\sqrt{p}\right|^8 = O(n^{-4}).$$

Markov's inequality, (S.3.82), and (S.3.94) yield

$$\mathbb{P}(\mathcal{G}_{\mathcal{I}}^c) \le \frac{4^8\,\mathbb{E}|\mathfrak{a}_{\mathcal{I}} - \beta_{n,\mathcal{I}}\sqrt{p}|^8}{|\beta_\sigma|^8 p^4} \le \frac{C}{n^4 p^4}. \tag{S.3.95}$$

Taking $r = 4$ in (S.3.88), and using (S.3.93) and (S.3.95), gives

$$\mathbb{E}U_{\mathcal{I}}^4 \le C\{\gamma^4 + \mathbb{E}\|b_{\mathcal{I}}\|^8\} + \frac{p^4}{\gamma^4}\mathbb{P}(\mathcal{G}_{\mathcal{I}}^c) = O(1), \qquad \mathbb{E}V_{\mathcal{I}}^4 \le \gamma^{-4}\mathbb{E}U_{\mathcal{I}}^4 = O(1).$$

Thus, uniformly over $|\mathcal{I}| \le 2$,

$$\mathbb{E}U_{\mathcal{I}}^4 + \mathbb{E}V_{\mathcal{I}}^4 = O(1).$$

For $\omega \in \mathcal{U}_{\mathrm{ms}}$, put

$$Q = Q^{(\mathcal{I})}, \qquad Q_\omega = Q_{\mathcal{I}}^{(\mathcal{I})}(\omega).$$

Since $\|\omega Q\| \le 1/4$,

$$Q_\omega = (I_p + \omega Q)^{-1}Q, \qquad \|Q_\omega \mathbf{1}\| \le \frac{4}{3}\|Q\mathbf{1}\|.$$

The resolvent identity also gives

$$\begin{aligned}
|\mathbf{1}^{\mathsf{T}}Q_\omega\mathbf{1}| &\le U_{\mathcal{I}} + |\omega|\,\|Q\mathbf{1}\|\,\|Q_\omega\mathbf{1}\| \\
&\le U_{\mathcal{I}} + CV_{\mathcal{I}} \le CU_{\mathcal{I}},
\end{aligned}$$

where $V_{\mathcal{I}} \le \gamma^{-1}U_{\mathcal{I}}$ by (S.3.90). Hence

$$\mathbb{E}|\mathbf{1}^{\mathsf{T}}Q_\omega\mathbf{1}|^4 + \mathbb{E}\|Q_\omega\mathbf{1}\|^8 \le C\{\mathbb{E}U_{\mathcal{I}}^4 + \mathbb{E}V_{\mathcal{I}}^4\} = O(1)$$

uniformly over $|\mathcal{I}| \le 2$ and $\omega \in \mathcal{U}_{\mathrm{ms}}$. This proves (S.3.59). □

S.3.6.3 Uniform first-order deterministic equivalent

Lemma S.3.12 Assume the standing conditions for a nonzero feature mean, including (S.1.7); in particular, $\beta_\sigma \neq 0$. Define

$$Q_I(\omega) = (S + (\gamma + \omega)I_p)^{-1}.$$

For $i \neq j$, use $S^{(ij)}$ from Lemma S.3.22 and write

$$t_{I,n}(\omega) := \frac{1}{n}\operatorname{tr} Q_I(\omega), \qquad s_{I,n}(\omega) := \frac{1}{n}\mathbf{1}^{\mathsf{T}}Q_I(\omega)\mathbf{1},$$

$$Q_I^{(ij)}(\omega) := \{S^{(ij)} + (\gamma + \omega)I_p\}^{-1}, \qquad t_{I,n}^{(ij)}(\omega) := \frac{1}{n}\operatorname{tr} Q_I^{(ij)}(\omega), \qquad s_{I,n}^{(ij)}(\omega) := \frac{1}{n}\mathbf{1}^{\mathsf{T}}Q_I^{(ij)}(\omega)\mathbf{1}.$$

Here the roman subscript I records the scalar perturbation direction $D = I_p$. There are two concentric closed complex disks $\mathcal{U} \Subset \mathcal{U}_+$ about zero, independent of n, on which $\mathrm{Re}(\gamma + \omega) \ge \gamma_-/2$. On $\mathcal{U}_+$ let

$$\bar{Q}_I(\omega) = \left[K(\bar{v}_I(\omega)) + (\gamma + \omega)I_p\right]^{-1}, \qquad \bar{v}_I(\omega) = m(t_I(\omega), s_I(\omega)),$$

where

$$t_I(\omega) = \frac{1}{n}\operatorname{tr}\bar{Q}_I(\omega), \qquad s_I(\omega) = \frac{1}{n}\mathbf{1}^{\mathsf{T}}\bar{Q}_I(\omega)\mathbf{1}.$$

Take the analytic fixed-point branch extending the solution at $\omega = 0$. Then

$$\sup_{\omega \in \mathcal{U}} \left\| \mathbb{E}Q_I(\omega) - \bar{Q}_I(\omega) \right\| \longrightarrow 0. \tag{S.3.96}$$

Proof. Condition on W. The vectors $g_i = W^\top x_i$ are independent $\mathcal{N}(0, I_p)$. Choose the disks so that $\mathcal{U}_+ \subset \mathcal{U}_{\mathrm{ms}}$. Fix their radii after imposing all the restrictions below. Whenever we shrink $\mathcal{U}_+$, we preserve $\mathcal{U} \Subset \mathcal{U}_+$.

Step 1: resolvent bounds and analytic continuation.

(a) Random resolvent bound. Choose the outer disk small enough that $|\omega| \le \gamma_-/4$ on $\mathcal{U}_+$. Then

$$\sup_{\omega \in \mathcal{U}_+} \max \left\{ \|Q_I(\omega)\|, \max_{i \neq j} \|Q_I^{(ij)}(\omega)\| \right\} \le \frac{1}{\gamma_- - |\omega|} \le \frac{4}{3\gamma_-}. \tag{S.3.97}$$

(b) Deterministic analytic branch. For all sufficiently large n, $p \ge 2$. Define

$$\xi(\tau, \varsigma, \omega) := \gamma + \omega + \nu_\alpha - \beta_\sigma^2 - v^\top m(\tau, \varsigma), \qquad \delta(\tau, \varsigma) := \beta_\sigma^2 \{1 - \mathbf{1}_4^\top m(\tau, \varsigma)\}.$$

Then

$$K(m(\tau, \varsigma)) + (\gamma + \omega) I_p = \xi(\tau, \varsigma, \omega) I_p + \delta(\tau, \varsigma) \mathbf{1}\mathbf{1}^\top. \tag{S.3.98}$$

Recall that $P = p^{-1}\mathbf{1}\mathbf{1}^\top$ and $P_\perp = I_p - P$. Whenever the matrix in (S.3.98) is invertible, its inverse has the exact decomposition

$$\left[K(m(\tau, \varsigma)) + (\gamma + \omega) I_p \right]^{-1} = \xi(\tau, \varsigma, \omega)^{-1} P_\perp + \{\xi(\tau, \varsigma, \omega) + p\delta(\tau, \varsigma)\}^{-1} P. \tag{S.3.99}$$

Consequently, the two components of the deterministic fixed-point map are

$$\begin{aligned} \mathcal{T}_n(\tau, \varsigma, \omega) &= \frac{p-1}{n\xi(\tau, \varsigma, \omega)} + \frac{1}{n\{\xi(\tau, \varsigma, \omega) + p\delta(\tau, \varsigma)\}}, \\ \mathcal{S}_n(\tau, \varsigma, \omega) &= \frac{p}{n\{\xi(\tau, \varsigma, \omega) + p\delta(\tau, \varsigma)\}}. \end{aligned} \tag{S.3.100}$$

Write the corresponding residual as

$$\mathcal{G}_n(\tau, \varsigma, \omega) := \begin{pmatrix} \tau - \mathcal{T}_n(\tau, \varsigma, \omega) \\ \varsigma - \mathcal{S}_n(\tau, \varsigma, \omega) \end{pmatrix}.$$

Thus, with $x_n^0 = (t_I(0), s_I(0))$, one has $\mathcal{G}_n(x_n^0, 0) = 0$. By the choice of the branch, $x_n^0 = (t, s)$.

Since $m(t_I(0), s_I(0)) = \bar{v}$ at $\omega = 0$, (S.3.47) gives

$$c_\delta := \inf_{n \ge n_0} |\delta(t_I(0), s_I(0))| > 0.$$

Lemma S.3.10 and the compactness of x_n^0 yield constants $r_0, r_\omega, c_\xi, C_\xi > 0$, independent of n, such that m is analytic with uniformly bounded first two derivatives and

$$\begin{aligned} &|\tau - t_I(0)| + |\varsigma - s_I(0)| \le r_0, \quad |\omega| \le r_\omega, \\ &c_\xi \le |\xi(\tau, \varsigma, \omega)| \le C_\xi, \qquad |\delta(\tau, \varsigma)| \ge \frac{c_\delta}{2}, \end{aligned} \tag{S.3.101}$$

with deterministic constants independent of n.

The lower bound on δ implies, for all sufficiently large n,

$$|\xi(\tau, \varsigma, \omega) + p\delta(\tau, \varsigma)| \ge p|\delta(\tau, \varsigma)| - |\xi(\tau, \varsigma, \omega)| \ge \frac{c_\delta}{4} p.$$

Hence, for every multi-index α in $(\tau, \varsigma, \omega)$ with $|\alpha| \le 2$,

$$\left| \partial^\alpha \xi^{-1} \right| \le C, \qquad \left| \partial^\alpha (\xi + p\delta)^{-1} \right| \le \frac{C}{p}. \tag{S.3.102}$$

Indeed, if $h = \xi + p\delta$ and $\zeta_1, \zeta_2 \in \{\tau, \varsigma, \omega\}$, then

$$\partial_{\zeta_1} h^{-1} = -h_{\zeta_1} h^{-2}, \qquad \partial_{\zeta_1\zeta_2} h^{-1} = 2h_{\zeta_1} h_{\zeta_2} h^{-3} - h_{\zeta_1\zeta_2} h^{-2}.$$

Here $|h| \geq cp$, while $|h_{\zeta_1}| + |h_{\zeta_1\zeta_2}| \leq Cp$, so all three orders $0, 1, 2$ are $O(p^{-1})$. The same formulas with $h = \xi$ give the first bound in (S.3.102). Because $p/n = O(1)$, the explicit formulas (S.3.100) and (S.3.102) imply

$$\sup_{n\geq n_0} \sup_{\substack{|\tau - t_I(0)|+|\varsigma - s_I(0)|\leq r_0 \\ |\omega|\leq r_\omega}} \Big(\|D\mathcal{G}_n(\tau, \varsigma, \omega)\| + \|D^2\mathcal{G}_n(\tau, \varsigma, \omega)\| \Big) < \infty. \tag{S.3.103}$$

In particular, integration along the scalar ω-segment gives the uniform residual perturbation estimate

$$\sup_{n\geq n_0} \sup_{\substack{|\tau - t_I(0)|+|\varsigma - s_I(0)|\leq r_0 \\ |\omega|\leq r_\omega}} \|\mathcal{G}_n(\tau, \varsigma, \omega) - \mathcal{G}_n(\tau, \varsigma, 0)\| \leq C|\omega|. \tag{S.3.104}$$

At the base point,

$$\partial_\tau \{K(m(\tau, \varsigma))\}|_{(\tau,\varsigma)=(t,s)} = -a_\tau I_p - \beta_\sigma^2 b_\tau \mathbf{1}\mathbf{1}^\top, \qquad \partial_\varsigma \{K(m(\tau, \varsigma))\}\Big|_{(\tau,\varsigma)=(t,s)} = -a_\varsigma I_p - \beta_\sigma^2 b_\varsigma \mathbf{1}\mathbf{1}^\top,$$

so direct differentiation of the two fixed-point equations gives

$$D_{(\tau,\varsigma)}\mathcal{G}_n(x_n^0, 0) = L.$$

The stability gap (S.3.46) therefore gives $C_L := \sup_{n\geq n_0} \|L^{-1}\| < \infty$. Let M bound the second derivatives in (S.3.103). Choose $0 < \rho < r_0$ and $0 < r_* < r_\omega$, independently of n, with $\sqrt{2}\,\rho < r_0$, so that

$$C_L M(\rho + r_*) \leq \frac{1}{2}, \qquad C_L C r_* \leq \frac{\rho}{2}, \tag{S.3.105}$$

where C is the constant in (S.3.104). For $|\omega| \leq r_*$ define, on the closed complex ball $B_\rho(x_n^0) = \{x : \|x - x_n^0\|_2 \leq \rho\}$,

$$\Psi_{n,\omega}(x) := x - L^{-1}\mathcal{G}_n(x, \omega).$$

Since

$$D_x\Psi_{n,\omega}(x) = L^{-1}\{L - D_x\mathcal{G}_n(x, \omega)\},$$

the mean-value estimate and the first inequality in (S.3.105) give

$$\sup_{x\in B_\rho(x_n^0)} \|D_x\Psi_{n,\omega}(x)\| \leq C_L M(\rho + r_*) \leq \frac{1}{2}.$$

Furthermore, (S.3.104) and $\mathcal{G}_n(x_n^0, 0) = 0$ yield

$$\|\Psi_{n,\omega}(x_n^0) - x_n^0\| \leq C_L \|\mathcal{G}_n(x_n^0, \omega)\| \leq \frac{\rho}{2}.$$

Consequently $\Psi_{n,\omega}$ maps $B_\rho(x_n^0)$ into itself and is a contraction there. Banach's fixed-point theorem gives a unique solution

$$\omega \longmapsto (t_I(\omega), s_I(\omega)), \qquad |\omega| \leq r_*,$$

inside this ball. Applying the analytic implicit function theorem at these solutions gives one analytic branch: uniqueness ensures that the local branches agree on overlapping balls. Moreover,

$$\left\| L^{-1}\{D_x\mathcal{G}_n(t_I(\omega), s_I(\omega), \omega) - L\} \right\| \leq \frac{1}{2},$$

so the Neumann series gives

$$\sup_{n\geq n_0} \sup_{|\omega|\leq r_*} \left\| \{D_x\mathcal{G}_n(t_I(\omega), s_I(\omega), \omega)\}^{-1} \right\| \leq 2C_L.$$

Differentiating $\mathcal{G}_n(t_I(\omega), s_I(\omega), \omega) = 0$ gives

$$\frac{d}{d\omega}\begin{pmatrix} t_I(\omega) \\ s_I(\omega) \end{pmatrix} = -\left[D_{(\tau,\varsigma)}\mathcal{G}_n(t_I(\omega), s_I(\omega), \omega)\right]^{-1} \partial_\omega \mathcal{G}_n(t_I(\omega), s_I(\omega), \omega),$$

and hence

$$|t_I(\omega) - t_I(0)| + |s_I(\omega) - s_I(0)| \le C|\omega|. \tag{S.3.106}$$

Since m is uniformly Lipschitz on the neighborhood in (S.3.101),

$$\|\bar{v}_I(\omega) - \bar{v}_I(0)\| \le C|\omega|, \qquad |\delta(t_I(\omega), s_I(\omega)) - \delta(t_I(0), s_I(0))| \le C|\omega|.$$

Shrinking $\mathcal{U}_+ \Subset \{|\omega| < r_*\}$ finally gives

$$\inf_{n\ge n_0} \inf_{\omega\in\mathcal{U}_+} \left|\beta_\sigma^2\{1 - \mathbf{1}_4^{\mathrm{T}}\bar{v}_I(\omega)\}\right| \ge \frac{c_\delta}{2}. \tag{S.3.107}$$

Thus the branch is analytic on $\mathcal{U}_+$, its Jacobian is uniformly invertible, and (S.3.107) holds there.

Equations (S.3.99), (S.3.101), and (S.3.107) also give

$$\sup_{n\ge n_0} \sup_{\omega\in\mathcal{U}_+} \left\{\|\bar{Q}_I(\omega)\| + |t_I(\omega)| + n|s_I(\omega)|\right\} \le C.$$

Equation (S.3.97) and

$$\partial_\omega^k Q_I(\omega) = (-1)^k k! Q_I(\omega)^{k+1},$$

provide an integrable majorant, so expectation, complex differentiation, and Cauchy integration may be interchanged.

Step 2: concentration of $n^{-1}\operatorname{tr} Q_I$. Put

$$t_n^\circ(\omega) = \mathbb{E}t_{I,n}(\omega), \qquad s_n^\circ(\omega) = \mathbb{E}s_{I,n}(\omega).$$

Write $g_i = W^{\mathrm{T}}x_i$ and, for $a, b \in [n]$, put

$$\mathcal{R}_k = \{(a,b) \in [n]^2 : a = k \text{ or } b = k\}.$$

Thus $|\mathcal{R}_k| = 2n - 1$. If e_ℓ denotes the ℓth Euclidean basis vector of $\mathbb{R}^p$, then, at every point at which σ is differentiable,

$$\begin{aligned} \partial_{g_{k\ell}} z_{ab} &= d_{ab;k\ell}\,\mathrm{e}_\ell, \\ d_{ab;k\ell} &= (\lambda\mathbf{1}_{\{a=k\}} + (1-\lambda)\mathbf{1}_{\{b=k\}})\sigma'(\lambda g_{a\ell} + (1-\lambda)g_{b\ell}), \qquad |d_{ab;k\ell}| \le L_\sigma. \end{aligned} \tag{S.3.108}$$

For $a = b = k$, the coefficient in parentheses equals one. The derivative vanishes outside $\mathcal{R}_k$. Since

$$S = \frac{1}{n^2}\sum_{a,b=1}^n z_{ab}z_{ab}^{\mathrm{T}},$$

equation (S.3.108) gives

$$\partial_{g_{k\ell}} S = \frac{1}{n^2}\sum_{(a,b)\in\mathcal{R}_k} d_{ab;k\ell}(\mathrm{e}_\ell z_{ab}^{\mathrm{T}} + z_{ab}\mathrm{e}_\ell^{\mathrm{T}}).$$

From $\partial Q_I = -Q_I(\partial S)Q_I$, cyclicity of the trace, and the complex symmetry $Q_I(\omega)^{\mathrm{T}} = Q_I(\omega)$, we obtain the exact derivative

$$\begin{aligned} \partial_{g_{k\ell}} t_{I,n}(\omega) &= -\frac{1}{n}\operatorname{tr}\Big(Q_I(\omega)(\partial_{g_{k\ell}} S)Q_I(\omega)\Big) \\ &= -\frac{2}{n^3}\sum_{(a,b)\in\mathcal{R}_k} d_{ab;k\ell}\,\mathrm{e}_\ell^{\mathrm{T}}Q_I(\omega)^2 z_{ab}. \end{aligned}$$

Consequently, Cauchy–Schwarz over the $2n-1$ affected augmented rows yields

$$\left|\partial_{g_{k\ell}} t_{I,n}(\omega)\right|^2 \le \frac{C|\mathcal{R}_k|}{n^6}\sum_{(a,b)\in\mathcal{R}_k}\left|\mathrm{e}_\ell^{\mathrm{T}}Q_I(\omega)^2 z_{ab}\right|^2 \le \frac{C}{n^5}\sum_{(a,b)\in\mathcal{R}_k}\left|\mathrm{e}_\ell^{\mathrm{T}}Q_I(\omega)^2 z_{ab}\right|^2. \tag{S.3.109}$$

When (S.3.109) is summed over k, each ordered off-diagonal pair (a,b) occurs only for $k = a$ and $k = b$, while a diagonal pair occurs only once. Therefore, after also summing over ℓ,

$$\begin{aligned} \sum_{k=1}^n\sum_{\ell=1}^p \mathbb{E}\left|\partial_{g_{k\ell}} t_{I,n}(\omega)\right|^2 &\le \frac{C}{n^5}\sum_{a,b=1}^n \mathbb{E}\sum_{\ell=1}^p\left|\mathrm{e}_\ell^{\mathrm{T}}Q_I(\omega)^2 z_{ab}\right|^2 \\ &= \frac{C}{n^5}\sum_{a,b=1}^n \mathbb{E}\|Q_I(\omega)^2 z_{ab}\|^2 \\ &\le \frac{C}{n^5}\sum_{a,b=1}^n \mathbb{E}\|z_{ab}\|^2 \le \frac{Cn^2p}{n^5} = O\Big(\frac{p}{n^3}\Big) = O(n^{-2}). \end{aligned} \tag{S.3.110}$$

The penultimate inequality in (S.3.110) uses the uniform bound (S.3.97); the last inequality follows from $\sup_{a,b} \mathbb{E}\|z_{ab}\|^2 \le Cp$ and $p/n \to c \in (0,\infty)$.

Conditional on W, the variables $(g_{k\ell})_{k,\ell}$ are independent standard Gaussians, and their conditional law does not depend on W. The conditional law of the entire array entering $t_{I,n}(\omega)$ is therefore independent of W, and hence

$$\mathbb{E}[t_{I,n}(\omega) \mid W] = \mathbb{E}t_{I,n}(\omega) = t_n^\circ(\omega).$$

Thus the total variance equals the expected conditional variance. The Gaussian Sobolev chain rule applies because σ is Lipschitz (equivalently, one may first smooth σ and then pass to the limit). Applying the conditional Gaussian Poincare inequality separately to the real and imaginary parts of $t_{I,n}(\omega)$, and then averaging over W, gives

$$\begin{aligned}\mathbb{E}|t_{I,n}(\omega) - t_n^\circ(\omega)|^2 &= \mathrm{Var}(\mathrm{Re}\, t_{I,n}(\omega)) + \mathrm{Var}(\mathrm{Im}\, t_{I,n}(\omega)) \\ &\le \sum_{k=1}^{n}\sum_{\ell=1}^{p} \mathbb{E}\left(|\partial_{g_{k\ell}} \mathrm{Re}\, t_{I,n}(\omega)|^2 + |\partial_{g_{k\ell}} \mathrm{Im}\, t_{I,n}(\omega)|^2\right) \\ &= \sum_{k=1}^{n}\sum_{\ell=1}^{p} \mathbb{E}|\partial_{g_{k\ell}} t_{I,n}(\omega)|^2. \end{aligned} \tag{S.3.111}$$

Combining (S.3.111) with (S.3.110) proves

$$\sup_{\omega\in\mathcal{U}_+} \mathbb{E}\left|t_{I,n}(\omega) - t_n^\circ(\omega)\right|^2 = O(n^{-2}). \tag{S.3.112}$$

Thus the centered normalized trace is $O_{L^2}(n^{-1})$, uniformly in $\omega \in \mathcal{U}_+$.

The resolvent identity gives

$$Q_I(\omega) - Q_I(\omega') = -(\omega - \omega')Q_I(\omega)Q_I(\omega'),$$

and hence, uniformly in n,

$$|t_{I,n}(\omega) - t_{I,n}(\omega')| \le \frac{p}{n}\|Q_I(\omega)\|\|Q_I(\omega')\||\omega - \omega'| \le C|\omega - \omega'|. \tag{S.3.113}$$

The bound (S.3.113) also holds for t_n°. A two-dimensional net $\mathcal{N}_n$ of $\mathcal{U}$, with mesh $n^{-1/4}$, has $O(n^{1/2})$ points. For every fixed $\varepsilon > 0$, Chebyshev's inequality and a union bound give

$$\mathbb{P}\left(\max_{\omega\in\mathcal{N}_n} |t_{I,n}(\omega) - t_n^\circ(\omega)| > \varepsilon\right) \le \frac{C|\mathcal{N}_n|}{\varepsilon^2 n^2} = O(n^{-3/2}),$$

while Lipschitz interpolation contributes $O(n^{-1/4})$. Hence

$$\sup_{\omega\in\mathcal{U}} |t_{I,n}(\omega) - t_n^\circ(\omega)| \xrightarrow{\mathbb{P}} 0.$$

For $i \ne j$, introduce the intermediate resolvent

$$\widetilde{Q}_{ij}(\omega) := \left\{S^{(ij)} + U_{ij}U_{ij}^{\mathsf{T}} + (\gamma + \omega)I_p\right\}^{-1}.$$

It contains the rank-four contribution of the rows having endpoint i or j, but not the Gram remainder R_{ij}. Woodbury's identity gives the exact formula

$$\widetilde{Q}_{ij} = Q_I^{(ij)} - Q_I^{(ij)}U_{ij}\left(I_4 + U_{ij}^{\mathsf{T}}Q_I^{(ij)}U_{ij}\right)^{-1}U_{ij}^{\mathsf{T}}Q_I^{(ij)}.$$

The 4×4 inverse exists because both matrices defining $Q_I^{(ij)}$ and $\widetilde{Q}_{ij}$ are invertible; equivalently, this follows from the matrix determinant lemma:

$$\det\left((Q_I^{(ij)})^{-1} + U_{ij}U_{ij}^{\mathsf{T}}\right) = \det\left((Q_I^{(ij)})^{-1}\right)\det\left(I_4 + U_{ij}^{\mathsf{T}}Q_I^{(ij)}U_{ij}\right) \ne 0.$$

In particular,

$$\mathrm{rank}\left(\widetilde{Q}_{ij} - Q_I^{(ij)}\right) \le 4.$$

All three matrices Q_I, $Q_I^{(ij)}$, and $\widetilde{Q}_{ij}$ are resolvents of a positive-semidefinite matrix at the scalar shift $\gamma + \omega$. Hence (S.3.97) gives, uniformly in i, j and $\omega \in \mathcal{U}$,

$$\|Q_I\| + \|Q_I^{(ij)}\| + \|\widetilde{Q}_{ij}\| \le C.$$

Using the standard trace–rank bound $|\operatorname{tr}(\cdot)| \le \|\cdot\|_* \le \operatorname{rank}(\cdot)\|\cdot\|$, we therefore obtain

$$\frac{1}{n}\left|\operatorname{tr}\{Q_I^{(ij)} - \widetilde{Q}_{ij}\}\right| \le \frac{4}{n}\|Q_I^{(ij)} - \widetilde{Q}_{ij}\| \le \frac{C}{n}. \tag{S.3.114}$$

The exact Gram decomposition (S.3.264) also gives

$$Q_I(\omega)^{-1} = \widetilde{Q}_{ij}(\omega)^{-1} + R_{ij}.$$

Thus the resolvent identity is

$$Q_I - \widetilde{Q}_{ij} = -Q_I R_{ij} \widetilde{Q}_{ij}.$$

Consequently,

$$\frac{1}{n}\left|\operatorname{tr}\{Q_I - \widetilde{Q}_{ij}\}\right| \le \frac{p}{n}\|Q_I\|\,\|R_{ij}\|\,\|\widetilde{Q}_{ij}\| \le C\frac{p}{n}\|R_{ij}\|. \tag{S.3.115}$$

Equations (S.3.114) and (S.3.115), together with the triangle inequality, give

$$\max_{i\neq j}\sup_{\omega\in\mathcal{U}}\left|t_{I,n}^{(ij)}(\omega) - t_{I,n}(\omega)\right| \le \frac{C}{n} + C\frac{p}{n}\max_{i\neq j}\|R_{ij}\|. \tag{S.3.116}$$

The constants are independent of (i, j, ω), and R_{ij} does not depend on ω. The $O(n^{-1})$ term comes from the rank-four update. Since $p/n = O(1)$ and $\max_{i\neq j}\|R_{ij}\| = o_{\mathbb{P}}(1)$ by (S.3.280), this proves

$$\max_{i\neq j}\sup_{\omega\in\mathcal{U}}\left|t_{I,n}^{(ij)}(\omega) - t_{I,n}(\omega)\right| = o_{\mathbb{P}}(1).$$

Squaring (S.3.116) and using the moment remainder estimate (S.3.263) also gives

$$\mathbb{E}\left[\max_{i\neq j}\sup_{\omega\in\mathcal{U}}\left|t_{I,n}^{(ij)}(\omega) - t_{I,n}(\omega)\right|^2\right] \le \frac{C}{n^2} + C\left(\frac{p}{n}\right)^2\mathbb{E}\max_{i\neq j}\|R_{ij}\|^2 = o(1). \tag{S.3.117}$$

Since the maximum in (S.3.117) dominates every fixed pair and every ω, equations (S.3.112) and (S.3.117) imply

$$\begin{aligned}&\sup_{i\neq j}\sup_{\omega\in\mathcal{U}}\mathbb{E}|t_{I,n}^{(ij)}(\omega) - t_n^{\circ}(\omega)|^2\\ &\quad\le 2\mathbb{E}\left[\max_{a\neq b}\sup_{\omega'\in\mathcal{U}}\left|t_{I,n}^{(ab)}(\omega') - t_{I,n}(\omega')\right|^2\right] + 2\sup_{\omega\in\mathcal{U}}\mathbb{E}\left|t_{I,n}(\omega) - t_n^{\circ}(\omega)\right|^2 = o(1).\end{aligned}$$

Equation (S.3.59) gives

$$\sup_{i\neq j}\sup_{\omega\in\mathcal{U}}\mathbb{E}|\mathbf{1}^{\mathrm{T}}Q_I^{(ij)}(\omega)\mathbf{1}|^2 \le C, \qquad \sup_{\omega\in\mathcal{U}}\mathbb{E}|\mathbf{1}^{\mathrm{T}}Q_I(\omega)\mathbf{1}|^2 \le C.$$

Since $s_{I,n}^{(ij)} = n^{-1}\mathbf{1}^{\mathrm{T}}Q_I^{(ij)}\mathbf{1}$, (S.3.59) gives

$$\begin{aligned}&\sup_{i\neq j}\sup_{\omega\in\mathcal{U}}\mathbb{E}|s_{I,n}^{(ij)}(\omega)|^2\\ &\quad= \frac{1}{n^2}\sup_{i\neq j}\sup_{\omega\in\mathcal{U}}\mathbb{E}|\mathbf{1}^{\mathrm{T}}Q_I^{(ij)}(\omega)\mathbf{1}|^2 \le \frac{C}{n^2}.\end{aligned}$$

Hence $s_{I,n}^{(ij)} = O_{L^2}(n^{-1})$ uniformly. For the deterministic scalar, linearity of expectation and Jensen's inequality give

$$\begin{aligned}|s_n^{\circ}(\omega)|^2 &= \frac{1}{n^2}|\mathbb{E}[\mathbf{1}^{\mathrm{T}}Q_I(\omega)\mathbf{1}]|^2\\ &\le \frac{1}{n^2}\mathbb{E}|\mathbf{1}^{\mathrm{T}}Q_I(\omega)\mathbf{1}|^2 \le \frac{C}{n^2}.\end{aligned}$$

Thus

$$\sup_{i\neq j}\sup_{\omega\in\mathcal{U}}\mathbb{E}|s_{I,n}^{(ij)}(\omega)|^2 + \sup_{\omega\in\mathcal{U}}|s_n^{\circ}(\omega)|^2 = O(n^{-2}). \tag{S.3.118}$$

For every fixed (i, j, ω),

$$\mathbb{E}|s_{I,n}^{(ij)}(\omega) - s_n^\circ(\omega)|^2 \le 2\mathbb{E}|s_{I,n}^{(ij)}(\omega)|^2 + 2|s_n^\circ(\omega)|^2.$$

Taking the two suprema and using (S.3.118) yields

$$\sup_{i\neq j}\sup_{\omega\in\mathcal{U}} \mathbb{E}|s_{I,n}^{(ij)}(\omega) - s_n^\circ(\omega)|^2 = O(n^{-2}) = o(1).$$

Because J_4 and $\mathbf{1}_4$ have fixed norm,

$$\sup_{i\neq j}\sup_{\omega\in\mathcal{U}} \mathbb{E}\left\|\beta_\sigma^2\{s_{I,n}^{(ij)}(\omega) - s_n^\circ(\omega)\}J_4\right\|^2 + \sup_{i\neq j}\sup_{\omega\in\mathcal{U}} \mathbb{E}\left\|\beta_\sigma^2\{s_{I,n}^{(ij)}(\omega) - s_n^\circ(\omega)\}\mathbf{1}_4\right\|^2 = O(n^{-2}),$$

Consequently,

$$\sup_{i\neq j}\sup_{\omega\in\mathcal{U}} \mathbb{E}\left\{|t_{I,n}^{(ij)}(\omega) - t_n^\circ(\omega)|^2 + |s_{I,n}^{(ij)}(\omega) - s_n^\circ(\omega)|^2\right\} = o(1),$$

Step 3: approximation of $\widehat{v}_{ij}$. For $i \neq j$ abbreviate

$$B_{ij} = Q_I^{(ij)}(\omega), \qquad M_{ij} = U_{ij}^{\mathrm{T}} B_{ij} U_{ij},$$

$$h_{ij} = \frac{1}{\sqrt{n}} U_{ij}^{\mathrm{T}} B_{ij} z_{ij}, \qquad \widehat{v}_{ij} = (I_4 + M_{ij})^{-1} h_{ij},$$

and set

$$M_n^\circ = t_n^\circ(I_2 \otimes V) + \beta_\sigma^2 s_n^\circ J_4, \quad h_n^\circ = t_n^\circ v + \beta_\sigma^2 s_n^\circ \mathbf{1}_4.$$

Define

$$\Delta_{ij}^M := M_{ij} - t_{I,n}^{(ij)}(\omega)(I_2 \otimes V) - \beta_\sigma^2 s_{I,n}^{(ij)}(\omega) J_4,$$
$$\Delta_{ij}^h := h_{ij} - t_{I,n}^{(ij)}(\omega) v - \beta_\sigma^2 s_{I,n}^{(ij)}(\omega)\mathbf{1}_4.$$

Condition on

$$\mathcal{F}_{ij} := \sigma(g_k : k \notin \{i, j\}), \qquad \mathbb{E}_{ij}[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{ij}], \qquad \mathbb{P}_{ij}(\cdot) := \mathbb{P}(\cdot \mid \mathcal{F}_{ij}).$$

Then B_{ij} is $\mathcal{F}_{ij}$-measurable and (g_i, g_j) is independent of $\mathcal{F}_{ij}$. Index the four columns of U_{ij} by

$$\mathcal{A}_{ij} = \{(i, 1), (i, 2), (j, 1), (j, 2)\},$$

and recall the endpoint feature vectors $m_{ur} = f_r(g_u)$. For $\rho = (u, r), \rho' = (u', r') \in \mathcal{A}_{ij}$, the normalizations in the definitions of U_{ij}, M_{ij}, and h_{ij} give exactly

$$[M_{ij}]_{\rho,\rho'} = \frac{1}{n} m_{ur}^{\mathrm{T}} B_{ij} m_{u'r'}, \qquad [h_{ij}]_\rho = \frac{1}{n} m_{ur}^{\mathrm{T}} B_{ij} z_{ij}.$$

Independence across feature coordinates gives

$$\mathbb{E}_{ij}[m_{ur,k} m_{u'r',k}] = \beta_\sigma^2 + \mathbf{1}_{\{u=u'\}} V_{rr'}, \qquad \mathrm{Cov}_{ij}(m_{ur,k}, z_{ij,k}) = v_\rho.$$

Since B_{ij} is $\mathcal{F}_{ij}$-measurable, summing over its entries gives

$$\mathbb{E}_{ij}[M_{ij}]_{\rho,\rho'} = t_{I,n}^{(ij)}(\omega)\mathbf{1}_{\{u=u'\}} V_{rr'} + \beta_\sigma^2 s_{I,n}^{(ij)}(\omega),$$
$$\mathbb{E}_{ij}[h_{ij}]_\rho = t_{I,n}^{(ij)}(\omega) v_\rho + \beta_\sigma^2 s_{I,n}^{(ij)}(\omega),$$

where v_ρ denotes the corresponding entry of $v = (v_{11}, v_{12}, v_{12}, v_{22})^{\mathrm{T}}$. Hence

$$[\Delta_{ij}^M]_{\rho,\rho'} = [M_{ij}]_{\rho,\rho'} - \mathbb{E}_{ij}[M_{ij}]_{\rho,\rho'}, \qquad [\Delta_{ij}^h]_\rho = [h_{ij}]_\rho - \mathbb{E}_{ij}[h_{ij}]_\rho.$$

Conditionally on $\mathcal{F}_{ij}$, the feature-coordinate blocks are independent and uniformly sub-Gaussian. Moreover, uniformly in i, j and $\omega \in \mathcal{U}$,

$$\|B_{ij}\| \le C, \qquad \|B_{ij}\|_{\mathrm{F}} \le C\sqrt{p}, \qquad \|B_{ij}\mathbf{1}\| \le C\sqrt{p}. \tag{S.3.119}$$

For an entry of Δ_{ij}^{M}, set $\bar{m}_{ur} = m_{ur} - \beta_\sigma \mathbf{1}$. Then

$$\begin{aligned} &m_{ur}^{\mathrm{T}} B_{ij} m_{u'r'} - \mathbb{E}_{ij}[m_{ur}^{\mathrm{T}} B_{ij} m_{u'r'}] \\ &\quad = \bar{m}_{ur}^{\mathrm{T}} B_{ij} \bar{m}_{u'r'} - \mathbb{E}_{ij}[\bar{m}_{ur}^{\mathrm{T}} B_{ij} \bar{m}_{u'r'}] + \beta_\sigma \mathbf{1}^{\mathrm{T}} B_{ij} \bar{m}_{u'r'} + \beta_\sigma \bar{m}_{ur}^{\mathrm{T}} B_{ij} \mathbf{1}. \end{aligned}$$

The two centered endpoint vectors need not be independent when they come from the same endpoint, but the symmetry of B_{ij} gives

$$\bar{m}_{ur}^{\mathrm{T}} B_{ij} \bar{m}_{u'r'} = \frac{1}{4}\left\{(\bar{m}_{ur} + \bar{m}_{u'r'})^{\mathrm{T}} B_{ij} (\bar{m}_{ur} + \bar{m}_{u'r'}) - (\bar{m}_{ur} - \bar{m}_{u'r'})^{\mathrm{T}} B_{ij} (\bar{m}_{ur} - \bar{m}_{u'r'})\right\}.$$

For complex ω, write

$$B_{ij} = B_{ij}^{\mathrm{R}} + i B_{ij}^{\mathrm{I}}.$$

Since B_{ij} is a resolvent of a real symmetric matrix, B_{ij}^{R} and B_{ij}^{I} are real symmetric and satisfy

$$\|B_{ij}^{\mathrm{R}}\| + \|B_{ij}^{\mathrm{I}}\| \le C, \qquad \|B_{ij}^{\mathrm{R}}\|_{\mathrm{F}} + \|B_{ij}^{\mathrm{I}}\|_{\mathrm{F}} \le C\sqrt{p}, \qquad \|B_{ij}^{\mathrm{R}}\mathbf{1}\| + \|B_{ij}^{\mathrm{I}}\mathbf{1}\| \le C\sqrt{p}.$$

For complex ω, we apply Hanson–Wright and the sub-Gaussian linear-form inequality separately to B_{ij}^{R} and B_{ij}^{I} and combine the bounds by a union bound. The vectors in the quadratic forms have independent, centered, uniformly sub-Gaussian feature coordinates. Hence, for constants $C, c_0 > 0$ and every $r_{\mathrm{tail}} > 0$,

$$\begin{aligned} &\mathbb{P}_{ij}\left(|\bar{m}_{ur}^{\mathrm{T}} B_{ij} \bar{m}_{u'r'} - \mathbb{E}_{ij}[\bar{m}_{ur}^{\mathrm{T}} B_{ij} \bar{m}_{u'r'}]| > n r_{\mathrm{tail}}\right) \\ &\qquad \le C \exp\left[-c_0 \min\left\{\frac{n^2 r_{\mathrm{tail}}^2}{\|B_{ij}\|_{\mathrm{F}}^2}, \frac{n r_{\mathrm{tail}}}{\|B_{ij}\|}\right\}\right], \\ &\mathbb{P}_{ij}\left(|\mathbf{1}^{\mathrm{T}} B_{ij} \bar{m}_{u'r'}| > n r_{\mathrm{tail}}\right) + \mathbb{P}_{ij}\left(|\bar{m}_{ur}^{\mathrm{T}} B_{ij} \mathbf{1}| > n r_{\mathrm{tail}}\right) \\ &\qquad \le C \exp\left[-c_0 \frac{n^2 r_{\mathrm{tail}}^2}{\|B_{ij}\mathbf{1}\|^2}\right]. \end{aligned}$$

Since $[\Delta_{ij}^{M}]_{\rho,\rho'}$ contains an additional factor n^{-1}, (S.3.119) and $p/n = O(1)$ give

$$\mathbb{P}_{ij}\left(|[\Delta_{ij}^{M}]_{\rho,\rho'}| > r_{\mathrm{tail}}\right) \le C \exp\{-c_0 n \min(r_{\mathrm{tail}}^2, r_{\mathrm{tail}})\}. \tag{S.3.120}$$

For an entry of Δ_{ij}^{h}, both $\bar{m}_{ur}$ and $z_{ij} - \beta_\sigma \mathbf{1}$ have conditional mean zero. Their two coordinates within one feature block may be dependent, but the pairs $([\bar{m}_{ur}]_k, [z_{ij} - \beta_\sigma \mathbf{1}]_k)$ are independent across k. The complete centered decomposition is

$$\begin{aligned} &m_{ur}^{\mathrm{T}} B_{ij} z_{ij} - \mathbb{E}_{ij}[m_{ur}^{\mathrm{T}} B_{ij} z_{ij}] \\ &\quad = \bar{m}_{ur}^{\mathrm{T}} B_{ij} (z_{ij} - \beta_\sigma \mathbf{1}) - \mathbb{E}_{ij}\left[\bar{m}_{ur}^{\mathrm{T}} B_{ij} (z_{ij} - \beta_\sigma \mathbf{1})\right] \\ &\qquad + \beta_\sigma \mathbf{1}^{\mathrm{T}} B_{ij} (z_{ij} - \beta_\sigma \mathbf{1}) + \beta_\sigma \bar{m}_{ur}^{\mathrm{T}} B_{ij} \mathbf{1}. \end{aligned} \tag{S.3.121}$$

The bilinear term is polarized exactly as

$$\begin{aligned} &\bar{m}_{ur}^{\mathrm{T}} B_{ij} (z_{ij} - \beta_\sigma \mathbf{1}) \\ &\quad = \frac{1}{4}\Big\{[\bar{m}_{ur} + (z_{ij} - \beta_\sigma \mathbf{1})]^{\mathrm{T}} B_{ij} [\bar{m}_{ur} + (z_{ij} - \beta_\sigma \mathbf{1})] \\ &\qquad - [\bar{m}_{ur} - (z_{ij} - \beta_\sigma \mathbf{1})]^{\mathrm{T}} B_{ij} [\bar{m}_{ur} - (z_{ij} - \beta_\sigma \mathbf{1})]\Big\}. \end{aligned} \tag{S.3.122}$$

Since $[\Delta_{ij}^{h}]_\rho$ is n^{-1} times the first centered quantity, Hanson–Wright for the two quadratic terms in (S.3.122) and the sub-Gaussian linear-form inequality for the last two terms in (S.3.121) give

$$\mathbb{P}_{ij}\left(|[\Delta_{ij}^{h}]_\rho| > r_{\mathrm{tail}}\right) \le C \exp\{-c_0 n \min(r_{\mathrm{tail}}^2, r_{\mathrm{tail}})\}. \tag{S.3.123}$$

Equations (S.3.120) and (S.3.123) apply to the sixteen entries of Δ_{ij}^{M} and the four entries of Δ_{ij}^{h}. For complex ω, combine the real and imaginary estimates using

$$|Z| > r_{\mathrm{tail}} \quad \Longrightarrow \quad \max\{|\operatorname{Re} Z|, |\operatorname{Im} Z|\} > r_{\mathrm{tail}}/\sqrt{2}.$$

Consequently, for every such scalar entry Δ,

$$\mathbb{P}_{ij}(|\Delta| > r_{\text{tail}}) \le C \exp\{-c_0 n \min(r_{\text{tail}}^2, r_{\text{tail}})\},$$

with deterministic constants independent of (i, j, ω) and of the conditioning variables.

The tail-to-moment identity now gives

$$\begin{aligned}\mathbb{E}_{ij}|\Delta|^2 &= 2\int_0^\infty r_{\text{tail}}\,\mathbb{P}_{ij}(|\Delta| > r_{\text{tail}})\,dr_{\text{tail}} \\ &\le 2C\int_0^1 r_{\text{tail}} e^{-c_0 n r_{\text{tail}}^2}\,dr_{\text{tail}} + 2C\int_1^\infty r_{\text{tail}} e^{-c_0 n r_{\text{tail}}}\,dr_{\text{tail}} \\ &= \frac{C}{c_0 n}(1 - e^{-c_0 n}) + 2Ce^{-c_0 n}\left(\frac{1}{c_0 n} + \frac{1}{c_0^2 n^2}\right) \le \frac{C}{n}.\end{aligned} \tag{S.3.124}$$

Finally, $\|\Delta_{ij}^M\|^2 \le \|\Delta_{ij}^M\|_{\text{F}}^2$, and the squared Euclidean norm of a four-vector is the sum of its four coordinate squares. Summing (S.3.124) over the sixteen matrix entries and four vector entries, and then using the tower property, yields

$$\begin{aligned}&\sup_{i\neq j}\sup_{\omega\in\mathcal{U}} \mathbb{E}\left(\|\Delta_{ij}^M\|^2 + \|\Delta_{ij}^h\|^2\right) \\ &\le \sup_{i\neq j}\sup_{\omega\in\mathcal{U}} \mathbb{E}\left[\sum_{\rho,\rho'\in\mathcal{A}_{ij}} |[\Delta_{ij}^M]_{\rho,\rho'}|^2 + \sum_{\rho\in\mathcal{A}_{ij}} |[\Delta_{ij}^h]_\rho|^2\right] \le \frac{C}{n}.\end{aligned}$$

By their definitions, the following decompositions are exact:

$$\begin{aligned}M_{ij} - M_n^\circ &= \Delta_{ij}^M + \{t_{I,n}^{(ij)}(\omega) - t_n^\circ(\omega)\}(I_2 \otimes V) + \beta_\sigma^2\{s_{I,n}^{(ij)}(\omega) - s_n^\circ(\omega)\}J_4, \\ h_{ij} - h_n^\circ &= \Delta_{ij}^h + \{t_{I,n}^{(ij)}(\omega) - t_n^\circ(\omega)\}v + \beta_\sigma^2\{s_{I,n}^{(ij)}(\omega) - s_n^\circ(\omega)\}\mathbf{1}_4.\end{aligned}$$

The vectors and matrices $I_2 \otimes V, J_4, v, \mathbf{1}_4$ have fixed dimension and bounded norm. Hence (S.3.112), (S.3.117), and (S.3.118), together with $\|a + b + c\|^2 \le 3(\|a\|^2 + \|b\|^2 + \|c\|^2)$, give

$$\sup_{i\neq j}\sup_{\omega\in\mathcal{U}} \mathbb{E}(\|M_{ij} - M_n^\circ\|^2 + \|h_{ij} - h_n^\circ\|^2) = o(1). \tag{S.3.125}$$

To interpolate (S.3.125) to L^4, it suffices to establish uniform L^8 bounds for M_{ij} and h_{ij}. The following estimates reduce these bounds to the sixteenth moments of U_{ij} and z_{ij}:

$$\|M_{ij}\| = \|U_{ij}^{\text{T}} B_{ij} U_{ij}\| \le C\|U_{ij}\|^2, \qquad \|h_{ij}\| \le \frac{C}{\sqrt{n}}\|U_{ij}\|\,\|z_{ij}\|. \tag{S.3.126}$$

The four columns of $\sqrt{n}\,U_{ij}$ are $m_{ur} = f_r(g_u)$. Lipschitz continuity and Gaussian input imply, uniformly in every scalar coordinate,

$$\mathbb{E}|[m_{ur}]_k|^{16} \le C, \qquad \mathbb{E}|[z_{ij}]_k|^{16} \le C.$$

For example, this follows from $|\sigma(x)| \le |\sigma(0)| + L_\sigma|x|$; the same property holds for f_1, f_2 by conditional Jensen's inequality or directly from their Lipschitz constants. The elementary inequality

$$\left(\sum_{k=1}^p a_k\right)^8 \le p^7 \sum_{k=1}^p a_k^8, \qquad a_k \ge 0,$$

therefore gives

$$\mathbb{E}\|m_{ur}\|^{16} = O(p^8), \qquad \mathbb{E}\|z_{ij}\|^{16} = O(p^8).$$

Moreover,

$$\|U_{ij}\|^2 \le \|U_{ij}\|_{\text{F}}^2 = \frac{1}{n}\sum_{(u,r)\in\mathcal{A}_{ij}} \|m_{ur}\|^2.$$

Since $\mathcal{A}_{ij}$ contains only four elements,

$$\begin{aligned}\mathbb{E}\|U_{ij}\|^{16} &\le \frac{C}{n^8}\sum_{(u,r)\in\mathcal{A}_{ij}} \mathbb{E}\|m_{ur}\|^{16} \\ &\le C\left(\frac{p}{n}\right)^8 = O(1).\end{aligned} \tag{S.3.127}$$

Raising (S.3.126) to the eighth power and applying Cauchy–Schwarz to their product yields

$$\begin{aligned}\mathbb{E}\|M_{ij}\|^8 &\le C\mathbb{E}\|U_{ij}\|^{16} = O(1),\\ \mathbb{E}\|h_{ij}\|^8 &\le \frac{C}{n^4}\mathbb{E}(\|U_{ij}\|^8\|z_{ij}\|^8)\\ &\le \frac{C}{n^4}(\mathbb{E}\|U_{ij}\|^{16})^{1/2}(\mathbb{E}\|z_{ij}\|^{16})^{1/2} = O\Big((p/n)^4\Big) = O(1).\end{aligned} \tag{S.3.128}$$

On the larger disk $\mathcal{U}_+$, the definition of t_n°, the trace inequality, and (S.3.97) give

$$\begin{aligned}\sup_{\omega\in\mathcal{U}_+}|t_n^\circ(\omega)| &= \sup_{\omega\in\mathcal{U}_+}\frac{1}{n}\,|\mathrm{tr}\,\mathbb{E}Q_I(\omega)|\\ &\le \frac{1}{n}\sup_{\omega\in\mathcal{U}_+}\mathbb{E}|\,\mathrm{tr}\,Q_I(\omega)|\\ &\le \frac{p}{n}\sup_{\omega\in\mathcal{U}_+}\mathbb{E}\|Q_I(\omega)\| \le C,\end{aligned}$$

where $p/n = O(1)$. Equation (S.3.97) also gives, on the larger disk,

$$\begin{aligned}\sup_{\omega\in\mathcal{U}_+}|s_n^\circ(\omega)| &= \sup_{\omega\in\mathcal{U}_+}\frac{1}{n}\left|\mathbf{1}^{\mathrm{T}}\mathbb{E}Q_I(\omega)\mathbf{1}\right|\\ &\le \frac{\|\mathbf{1}\|^2}{n}\sup_{\omega\in\mathcal{U}_+}\mathbb{E}\|Q_I(\omega)\|\\ &= \frac{p}{n}\sup_{\omega\in\mathcal{U}_+}\mathbb{E}\|Q_I(\omega)\| \le C.\end{aligned}$$

On the smaller disk $\mathcal{U}$, equation (S.3.118) gives

$$\sup_{\omega\in\mathcal{U}}|s_n^\circ(\omega)| \le \frac{C}{n}.$$

Since V, v, β_σ depend only on σ and λ, their norms are independent of n. Using

$$M_n^\circ = t_n^\circ(I_2\otimes V) + \beta_\sigma^2 s_n^\circ J_4, \qquad h_n^\circ = t_n^\circ v + \beta_\sigma^2 s_n^\circ \mathbf{1}_4,$$

gives

$$\begin{aligned}\sup_{\omega\in\mathcal{U}_+}\|M_n^\circ(\omega)\| &\le \sup_{\omega\in\mathcal{U}_+}|t_n^\circ(\omega)|\,\|I_2\otimes V\| + \beta_\sigma^2\sup_{\omega\in\mathcal{U}_+}|s_n^\circ(\omega)|\,\|J_4\| \le C,\\ \sup_{\omega\in\mathcal{U}_+}\|h_n^\circ(\omega)\| &\le \sup_{\omega\in\mathcal{U}_+}|t_n^\circ(\omega)|\,\|v\| + \beta_\sigma^2\sup_{\omega\in\mathcal{U}_+}|s_n^\circ(\omega)|\,\|\mathbf{1}_4\| \le C.\end{aligned} \tag{S.3.129}$$

These estimates uniformly bound $t_n^\circ, s_n^\circ, M_n^\circ, h_n^\circ$. Combining them with the eighth-power triangle inequality and (S.3.128)–(S.3.129) gives

$$\sup_{n\ge n_0}\sup_{i\ne j,\,\omega\in\mathcal{U}}\mathbb{E}\left(\|M_{ij}-M_n^\circ\|^8 + \|h_{ij}-h_n^\circ\|^8\right) \le C. \tag{S.3.130}$$

For either $X = \|M_{ij}-M_n^\circ\|$ or $X = \|h_{ij}-h_n^\circ\|$, interpolation between (S.3.125) and (S.3.130) gives

$$\|X\|_{L^4} \le \|X\|_{L^2}^{1/3}\|X\|_{L^8}^{2/3}, \qquad \mathbb{E}X^4 \le (\mathbb{E}X^2)^{2/3}(\mathbb{E}X^8)^{1/3}. \tag{S.3.131}$$

In (S.3.131), the L^2 factor tends to zero uniformly, while the L^8 factor remains uniformly bounded. Therefore

$$\sup_{i\ne j,\,\omega\in\mathcal{U}}\mathbb{E}\left(\|M_{ij}-M_n^\circ\|^4 + \|h_{ij}-h_n^\circ\|^4\right) = o(1). \tag{S.3.132}$$

At $\omega = 0$, $t_n^\circ(0), s_n^\circ(0) \ge 0$, while $V \ge 0$ because it is a covariance matrix. Hence

$$M_n^\circ(0) \ge 0, \qquad I_4 + M_n^\circ(0) \ge I_4.$$

The scalar resolvent identity gives the exact relations

$$\begin{aligned}t_n^\circ(\omega) - t_n^\circ(0) &= -\frac{\omega}{n}\,\mathrm{tr}\,\mathbb{E}[Q_I(\omega)Q_I(0)],\\ s_n^\circ(\omega) - s_n^\circ(0) &= -\frac{\omega}{n}\mathbf{1}^{\mathrm{T}}\mathbb{E}[Q_I(\omega)Q_I(0)]\mathbf{1}.\end{aligned}$$

Consequently, (S.3.97) and $p/n = O(1)$ give

$$|t_n^\circ(\omega) - t_n^\circ(0)| \le |\omega| \frac{p}{n} \mathbb{E}[\|Q_I(\omega)\| \, \|Q_I(0)\|] \le C|\omega|,$$
$$|s_n^\circ(\omega) - s_n^\circ(0)| \le |\omega| \frac{\|\mathbf{1}\|^2}{n} \mathbb{E}[\|Q_I(\omega)\| \, \|Q_I(0)\|] \le C|\omega|.$$

It follows that

$$\|M_n^\circ(\omega) - M_n^\circ(0)\| \le C|\omega|.$$

Since $s_{\min}(I_4 + M_n^\circ(0)) \ge 1$, shrinking $\mathcal{U}_+$ while preserving $\mathcal{U} \Subset \mathcal{U}_+$ and applying singular-value perturbation gives

$$\inf_{\omega \in \mathcal{U}_+} s_{\min}(I_4 + M_n^\circ(\omega)) \ge \frac{1}{2}. \tag{S.3.133}$$

Define

$$v_n^\circ := (I_4 + M_n^\circ)^{-1} h_n^\circ = m(t_n^\circ, s_n^\circ).$$

Equation (S.3.133) gives

$$\sup_{\omega \in \mathcal{U}_+} \|(I_4 + M_n^\circ(\omega))^{-1}\| = \sup_{\omega \in \mathcal{U}_+} \frac{1}{s_{\min}(I_4 + M_n^\circ(\omega))} \le 2. \tag{S.3.134}$$

Combining (S.3.134) with the bound for h_n° in (S.3.129) yields

$$\sup_{\omega \in \mathcal{U}_+} \|v_n^\circ(\omega)\| \le \sup_{\omega \in \mathcal{U}_+} \|(I_4 + M_n^\circ(\omega))^{-1}\| \sup_{\omega \in \mathcal{U}_+} \|h_n^\circ(\omega)\| \le C.$$

Woodbury's identity gives

$$(I_4 + M_{ij})^{-1} = I_4 - U_{ij}^{\mathrm{T}} \widetilde{Q}_{ij} U_{ij}.$$

The intermediate resolvent obeys the same bound as (S.3.97). In particular,

$$\|(I_4 + M_{ij})^{-1}\| \le 1 + C\|U_{ij}\|^2.$$

The sixteenth-moment estimate (S.3.127) therefore implies

$$\sup_{n \ge n_0} \sup_{i \ne j, \, \omega \in \mathcal{U}} \mathbb{E}\|(I_4 + M_{ij})^{-1}\|^4 \le C. \tag{S.3.135}$$

The exact difference identity is

$$\widehat{v}_{ij} - v_n^\circ = (I_4 + M_{ij})^{-1} \left\{ (h_{ij} - h_n^\circ) - (M_{ij} - M_n^\circ) v_n^\circ \right\}.$$

The uniform bound on v_n°, together with (S.3.132), shows that $\sup_{i \ne j} \sup_{\omega \in \mathcal{U}} \mathbb{E}\|(h_{ij} - h_n^\circ) - (M_{ij} - M_n^\circ) v_n^\circ\|^4 = o(1)$. Applying Hölder's inequality and the uniform fourth-moment bound for $(I_4 + M_{ij})^{-1}$ from (S.3.135) yields

$$\mathbb{E}\|\widehat{v}_{ij} - v_n^\circ\|^2 \le \left(\mathbb{E}\|(I_4 + M_{ij})^{-1}\|^4\right)^{1/2} \left(\mathbb{E}\|(h_{ij} - h_n^\circ) - (M_{ij} - M_n^\circ) v_n^\circ\|^4\right)^{1/2} = o(1)$$

uniformly, and therefore

$$\sup_{i \ne j} \sup_{\omega \in \mathcal{U}} \mathbb{E}\|\widehat{v}_{ij} - v_n^\circ\|^2 = o(1).$$

Step 4: scalar equation for t_n°. For every realization,

$$Q_I\{S + (\gamma + \omega) I_p\} = I_p.$$

Taking normalized traces and expectations gives

$$\frac{p}{n} = (\gamma + \omega) t_n^\circ + \frac{1}{n^3} \sum_{i,j=1}^{n} \mathbb{E}[z_{ij}^{\mathrm{T}} Q_I(\omega) z_{ij}]. \tag{S.3.136}$$

The n diagonal pairs contribute $O(n^{-1})$ uniformly on $\mathcal{U}$, because (S.3.97) gives

$$\frac{1}{n^3} \sum_{i=1}^{n} \mathbb{E}|z_{ii}^{\mathrm{T}} Q_I(\omega) z_{ii}| \le \frac{C}{n^3} \sum_{i=1}^{n} \mathbb{E}\|z_{ii}\|^2 \le \frac{Cnp}{n^3} = O(n^{-1}).$$

For $i \neq j$, use the intermediate resolvent

$$\widetilde{Q}_{ij}(\omega) = \{S^{(ij)} + U_{ij}U_{ij}^{\mathrm{T}} + (\gamma + \omega)I_p\}^{-1}.$$

Since $S^{(ij)} + U_{ij}U_{ij}^{\mathrm{T}}$ is real symmetric and positive semidefinite, for every $\omega \in \mathcal{U}$,

$$\|\widetilde{Q}_{ij}(\omega)\| = \max_{\lambda \in \mathrm{spec}(S^{(ij)} + U_{ij}U_{ij}^{\mathrm{T}})} \frac{1}{|\lambda + \gamma + \omega|} \leq \frac{1}{\gamma_- - |\omega|} \leq \frac{4}{3\gamma_-}.$$

The decomposition $S = S^{(ij)} + U_{ij}U_{ij}^{\mathrm{T}} + R_{ij}$, the resolvent identity, and the Woodbury identity give exactly

$$Q_I - \widetilde{Q}_{ij} = -Q_I R_{ij} \widetilde{Q}_{ij}, \qquad \widetilde{Q}_{ij} z_{ij} = B_{ij}(z_{ij} - \sqrt{n}\, U_{ij}\widehat{v}_{ij}).$$

The moment estimate (S.3.263) and $\mathbb{E}\|z_{ij}\|^4 = O(p^2)$ imply

$$\begin{aligned} &\frac{1}{n^3} \sum_{i \neq j} \mathbb{E}\left| z_{ij}^{\mathrm{T}}(Q_I - \widetilde{Q}_{ij}) z_{ij} \right| \\ &\qquad \leq C \sup_{i \neq j} (\mathbb{E}\|R_{ij}\|^2)^{1/2} \frac{n(n-1)p}{n^3} = o(1). \end{aligned} \tag{S.3.137}$$

The matrix B_{ij} is complex symmetric, and $(\sqrt{n}\, U_{ij})^{\mathrm{T}} B_{ij} z_{ij} = n h_{ij}$. Consequently,

$$z_{ij}^{\mathrm{T}} B_{ij}(\sqrt{n}\, U_{ij})(\widehat{v}_{ij} - v_n^\circ) = n h_{ij}^{\mathrm{T}}(\widehat{v}_{ij} - v_n^\circ),$$

so

$$\begin{aligned} &\frac{1}{n^3} \sum_{i \neq j} \mathbb{E}\left| z_{ij}^{\mathrm{T}} B_{ij}(\sqrt{n}\, U_{ij})(\widehat{v}_{ij} - v_n^\circ) \right| \\ &\qquad \leq \frac{1}{n^3} \sum_{i \neq j} n (\mathbb{E}\|h_{ij}\|^2)^{1/2} (\mathbb{E}\|\widehat{v}_{ij} - v_n^\circ\|^2)^{1/2} = o(1). \end{aligned} \tag{S.3.138}$$

The estimates (S.3.137) and (S.3.138) are uniform on $\mathcal{U}$. Together with the exact Woodbury formula they yield

$$\frac{1}{n^3} \sum_{i \neq j} \mathbb{E}[z_{ij}^{\mathrm{T}} Q_I z_{ij}] = \frac{1}{n^3} \sum_{i \neq j} \mathbb{E}\Big[z_{ij}^{\mathrm{T}} B_{ij}(z_{ij} - \sqrt{n}\, U_{ij} v_n^\circ) \Big] + o(1),$$

uniformly for $\omega \in \mathcal{U}$.

Conditionally on $\mathcal{F}_{ij}$, the matrix B_{ij} is fixed and (g_i, g_j) is independent of it. Under the standing conditioning on W, $v_n^\circ(\omega)$ is deterministic because t_n° and s_n° are expectations over the full sample. For every deterministic $u \in \mathbb{C}^4$, coordinate independence gives

$$\mathbb{E}_{ij}[(z_{ij} - \sqrt{n}\, U_{ij} u) z_{ij}^{\mathrm{T}}] = K(u). \tag{S.3.139}$$

For each feature coordinate k, the definitions of β_σ, ν_α, and v give

$$\mathbb{E}_{ij} z_{ij,k} = \beta_\sigma, \qquad \mathbb{E}_{ij}(\sqrt{n}\, U_{ij})_{k,\cdot}^{\mathrm{T}} = \beta_\sigma \mathbf{1}_4, \qquad \mathbb{E}_{ij} z_{ij,k}^2 = \nu_\alpha,$$

and

$$\mathbb{E}_{ij}[(\sqrt{n}\, U_{ij})_{k,\cdot}^{\mathrm{T}} z_{ij,k}] = v + \beta_\sigma^2 \mathbf{1}_4. \tag{S.3.140}$$

The four entries of the vector in (S.3.140) are

$$\begin{aligned} &\mathbb{E}[f_1(G_1)\sigma(\lambda G_1 + (1-\lambda)G_2)] = v_{11} + \beta_\sigma^2, \quad \mathbb{E}[f_2(G_1)\sigma(\lambda G_1 + (1-\lambda)G_2)] = v_{12} + \beta_\sigma^2, \\ &\mathbb{E}[f_1(G_2)\sigma(\lambda G_1 + (1-\lambda)G_2)] = v_{12} + \beta_\sigma^2, \quad \mathbb{E}[f_2(G_2)\sigma(\lambda G_1 + (1-\lambda)G_2)] = v_{22} + \beta_\sigma^2. \end{aligned}$$

For example, $\mathbb{E}[f_1(G_1)\sigma(\lambda G_1 + (1-\lambda)G_2)] = \mathbb{E} f_1(G_1)^2$ by conditioning on G_1; the other three identities follow in the same way by conditioning on the appropriate endpoint. This verifies (S.3.140) in the coordinate order defining v. Write the two coefficients of $K(u)$ as

$$\alpha_K(u) = \nu_\alpha - \beta_\sigma^2 - v^{\mathrm{T}} u, \qquad \delta(u) = \beta_\sigma^2(1 - \mathbf{1}_4^{\mathrm{T}} u).$$

For two distinct feature coordinates $k \neq \ell$, the Gaussian coordinate blocks (g_{ik}, g_{jk}) and $(g_{i\ell}, g_{j\ell})$ are independent. Hence

$$\begin{aligned}\mathbb{E}_{ij}[(z_{ij,k} - (\sqrt{n}\,U_{ij})_{k,\cdot}u)z_{ij,\ell}] &= \mathbb{E}_{ij}(z_{ij,k} - (\sqrt{n}\,U_{ij})_{k,\cdot}u)\,\mathbb{E}_{ij}z_{ij,\ell} \\ &= \delta(u).\end{aligned}$$

For $k = \ell$, equation (S.3.140) yields

$$\begin{aligned}\mathbb{E}_{ij}[(z_{ij,k} - (\sqrt{n}\,U_{ij})_{k,\cdot}u)z_{ij,k}] &= \nu_\alpha - (v + \beta_\sigma^2\mathbf{1}_4)^{\mathrm{T}}u \\ &= \delta(u) + \nu_\alpha - \beta_\sigma^2 - v^{\mathrm{T}}u.\end{aligned}$$

Consequently the common off-diagonal entry is $\delta(u)$, while the diagonal entry minus that value is $\alpha_K(u)$. Therefore

$$K(u) = \alpha_K(u)I_p + \delta(u)\mathbf{1}\mathbf{1}^{\mathrm{T}}. \tag{S.3.141}$$

Since

$$z_{ij}^{\mathrm{T}}B_{ij}(z_{ij} - \sqrt{n}\,U_{ij}u) = \mathrm{tr}\{B_{ij}(z_{ij} - \sqrt{n}\,U_{ij}u)z_{ij}^{\mathrm{T}}\},$$

conditioning on $\mathcal{F}_{ij}$ and using (S.3.139) gives

$$\mathbb{E}_{ij}[z_{ij}^{\mathrm{T}}B_{ij}(z_{ij} - \sqrt{n}\,U_{ij}u)] = \mathrm{tr}\{B_{ij}K(u)\}.$$

All ordered off-diagonal pairs have the same distribution. Taking $u = v_n^\circ$ and then taking expectation over $\mathcal{F}_{ij}$ gives

$$\frac{1}{n^3}\sum_{i\neq j}\mathbb{E}[z_{ij}^{\mathrm{T}}B_{ij}(z_{ij} - \sqrt{n}\,U_{ij}v_n^\circ)] = \frac{n(n-1)}{n^3}\,\mathrm{tr}\{\mathbb{E}B_{12}K(v_n^\circ)\} = \frac{1 - 1/n}{n}\,\mathrm{tr}\{\mathbb{E}B_{12}K(v_n^\circ)\}.$$

Since $B_{12} = Q_I^{(12)}$, equation (S.3.141) gives the exact identity

$$\frac{1}{n}\,\mathrm{tr}\{\mathbb{E}B_{12}K(v_n^\circ)\} = \alpha_K(v_n^\circ)\mathbb{E}t_{I,n}^{(12)} + \delta(v_n^\circ)\mathbb{E}s_{I,n}^{(12)}.$$

From (S.3.117) and Cauchy–Schwarz,

$$\begin{aligned}|\mathbb{E}t_{I,n}^{(12)}(\omega) - t_n^\circ(\omega)| &\le \mathbb{E}\left|\frac{1}{n}\,\mathrm{tr}(B_{12}(\omega) - Q_I(\omega))\right| \\ &\le \left\{\mathbb{E}\left|\frac{1}{n}\,\mathrm{tr}(B_{12}(\omega) - Q_I(\omega))\right|^2\right\}^{1/2} = o(1)\end{aligned} \tag{S.3.142}$$

uniformly on $\mathcal{U}$. For the mean direction, Lemma S.3.11, applied to the empty deletion set and to $\{1, 2\}$ and followed by Cauchy–Schwarz, gives

$$\begin{aligned}|\mathbb{E}s_{I,n}^{(12)}(\omega) - s_n^\circ(\omega)| &\le \frac{1}{n}\mathbb{E}|\mathbf{1}^{\mathrm{T}}B_{12}(\omega)\mathbf{1}| + \frac{1}{n}\mathbb{E}|\mathbf{1}^{\mathrm{T}}Q_I(\omega)\mathbf{1}| \\ &\le \frac{C}{n}.\end{aligned} \tag{S.3.143}$$

The functions α_K and δ are affine, and the uniform bound on v_n° therefore gives $|\alpha_K(v_n^\circ)| + |\delta(v_n^\circ)| \le C$. Together with the uniform bounds on t_n° and s_n°, (S.3.142) and (S.3.143) imply

$$\begin{aligned}&\frac{1 - 1/n}{n}\,\mathrm{tr}\{\mathbb{E}B_{12}K(v_n^\circ)\} \\ &\quad= (1 - 1/n)\{\alpha_K(v_n^\circ)\mathbb{E}t_{I,n}^{(12)} + \delta(v_n^\circ)\mathbb{E}s_{I,n}^{(12)}\} \\ &\quad= (1 - 1/n)\{\alpha_K(v_n^\circ)t_n^\circ + \delta(v_n^\circ)s_n^\circ + o(1)\} \\ &\quad= \alpha_K(v_n^\circ)t_n^\circ + \delta(v_n^\circ)s_n^\circ + o(1),\end{aligned} \tag{S.3.144}$$

uniformly on $\mathcal{U}$. In the last line of (S.3.144), removing the factor $1 - 1/n$ introduces an $O(n^{-1})$ error because the expression in braces is uniformly bounded.

Combining (S.3.144) with (S.3.136) proves

$$\frac{p}{n} = \{\gamma + \omega + \alpha_K(v_n^\circ)\}t_n^\circ + \delta(v_n^\circ)s_n^\circ + o(1) \tag{S.3.145}$$

uniformly on $\mathcal{U}$.

Step 5: identification of t_n°. Define

$$\Phi_n(\tau;\omega) := \tau\left\{\gamma+\omega+\nu_\alpha-\beta_\sigma^2-v^{\mathrm{T}}m(\tau,0)\right\}-\frac{p}{n}. \tag{S.3.146}$$

By (S.3.118), (S.3.133), and the resolvent identity for the 4×4 inverse,

$$m(t_n^\circ, s_n^\circ) = m(t_n^\circ, 0) + O(n^{-1})$$

uniformly on $\mathcal{U}$. The matrix perturbation is $\beta_\sigma^2 s_n^\circ J_4 = O(n^{-1})$, so (S.3.133) also bounds the inverse at $\varsigma = 0$ for all large n; the numerator perturbation $\beta_\sigma^2 s_n^\circ \mathbf{1}_4$ is likewise $O(n^{-1})$. Since t_n°, v_n°, and $\delta(v_n^\circ)$ are uniformly bounded, the scalar equation (S.3.145) gives

$$\sup_{\omega\in\mathcal{U}} |\Phi_n(t_n^\circ(\omega);\omega)| \longrightarrow 0. \tag{S.3.147}$$

Let $\vartheta_1, \vartheta_2 \geq 0$ be the eigenvalues of V. For

$$h_0(x,y) = \sigma(\lambda x + (1-\lambda)y) - f_1(x) - f_2(y) + \beta_\sigma,$$

one has the orthogonal decomposition

$$\sigma(\lambda G_1 + (1-\lambda)G_2) - \beta_\sigma = \{f_1(G_1)-\beta_\sigma\} + \{f_2(G_2)-\beta_\sigma\} + h_0(G_1,G_2).$$

The first two terms are independent, while the conditional expectation of $h_0(G_1,G_2)$ given either G_1 or G_2 is zero. Hence the three terms are pairwise orthogonal in L^2, and

$$\kappa(\lambda) = \nu_\alpha - \beta_\sigma^2 - \operatorname{tr} V = \mathbb{E}h_0(G_1,G_2)^2 \geq 0. \tag{S.3.148}$$

Moreover, $v = \operatorname{vec}(V)$ in the coordinate order used here and $I_2 \otimes V$ is block diagonal. Therefore, whenever $1+\vartheta_r\tau \neq 0$ for $r = 1, 2$,

$$\begin{aligned} v^{\mathrm{T}}m(\tau,0) &= \tau\, v^{\mathrm{T}}[I_4 + \tau(I_2\otimes V)]^{-1}v \\ &= \tau \operatorname{tr}\{V^2(I_2+\tau V)^{-1}\} = \sum_{r=1}^{2}\frac{\tau\vartheta_r^2}{1+\tau\vartheta_r}. \end{aligned} \tag{S.3.149}$$

Substituting (S.3.148) and (S.3.149) into (S.3.146) gives the explicit residual

$$\Phi_n(\tau;\omega) = (\gamma+\omega+\kappa(\lambda))\tau + \sum_{r=1}^{2}\frac{\vartheta_r\tau}{1+\vartheta_r\tau} - \frac{p}{n}. \tag{S.3.150}$$

At $\omega = 0$, equation (S.3.150) has exactly one zero in the closed right half-plane. To see this, let $\tau = x + iy$ with $x \geq 0$. If $\Phi_n(\tau;0) = 0$, then

$$0 = \operatorname{Im}\Phi_n(\tau;0) = y\left\{\gamma+\kappa(\lambda)+\sum_{r=1}^{2}\frac{\vartheta_r}{|1+\vartheta_r\tau|^2}\right\}.$$

The expression in braces is at least γ_-, so $y = 0$. On the nonnegative real axis,

$$\frac{\partial}{\partial\tau}\Phi_n(\tau;0) = \gamma+\kappa(\lambda)+\sum_{r=1}^{2}\frac{\vartheta_r}{(1+\vartheta_r\tau)^2} \geq \gamma_- > 0,$$

while $\Phi_n(0;0) = -p/n < 0$ and $\Phi_n(\tau;0) \to \infty$. Thus there is a unique simple positive zero, denoted $t_{*,n}(0)$. Increasing n_0 if necessary, choose fixed constants $c_-, c_+ > 0$ so that $c_- < p/n < c_+$ for every $n \geq n_0$. The zero satisfies the uniform bounds

$$\frac{c_-}{\gamma_+ + \kappa(\lambda) + \operatorname{tr} V} \leq t_{*,n}(0) \leq \frac{c_+}{\gamma_-}. \tag{S.3.151}$$

By uniqueness and (S.3.148), $t_{*,n}(0) = t_{*,n}$, where the quantity on the right is defined in (S.1.10).

If $\operatorname{Re}\tau \geq 0$, then

$$\left|\frac{\vartheta_r\tau}{1+\vartheta_r\tau}\right| \leq 1,$$

because

$$|1 + \vartheta_r\tau|^2 - \vartheta_r^2|\tau|^2 = 1 + 2\vartheta_r \operatorname{Re}\tau \geq 1.$$

In particular, on the closed right half plane,

$$|1 + \vartheta_r\tau| \geq 1. \tag{S.3.152}$$

Thus the rational function $\Phi_n(\,\cdot\,;\omega)$ has no pole there. If $\operatorname{Re}\tau \geq 0$, $|\omega| \leq \gamma_-/2$, and $\Phi_n(\tau;\omega) = 0$, then

$$|\gamma + \omega + \kappa(\lambda)|\,|\tau| \leq \frac{p}{n} + \sum_{r=1}^{2}\left|\frac{\vartheta_r\tau}{1+\vartheta_r\tau}\right| \leq c_+ + 2.$$

Moreover,

$$|\gamma + \omega + \kappa(\lambda)| \geq \operatorname{Re}(\gamma+\omega) + \kappa(\lambda) \geq \frac{\gamma_-}{2},$$

and therefore every such zero satisfies

$$|\tau| \leq R_0 := \frac{2(c_+ + 2)}{\gamma_-}.$$

Choose $R > R_0$ with a fixed positive margin and also large enough that

$$\sup_{n\geq n_0}\sup_{\omega\in\mathcal{U}_+}\max\{|t_n^\circ(\omega)|, |t_I(\omega)|\} < R.$$

Let

$$\mathcal{H}_R = \{\tau \in \mathbb{C} : \operatorname{Re}\tau > 0,\ |\tau| < R\}.$$

Equation (S.3.152) shows that $\Phi_n(\,\cdot\,;\omega)$ is holomorphic on a neighborhood of $\overline{\mathcal{H}_R}$. At $\omega = 0$, there is no zero on $\partial\mathcal{H}_R$ and exactly one zero inside. Compactness of the parameter set $[c_-, c_+]\times[\gamma_-, \gamma_+]$, together with the uniqueness just proved and continuity, gives

$$\eta := \inf_{n\geq n_0}\inf_{\tau\in\partial\mathcal{H}_R}|\Phi_n(\tau;0)| > 0.$$

If this infimum were zero, a convergent subsequence in $(p/n, \gamma, \tau)$ would produce a boundary zero for a limiting parameter pair, contradicting uniqueness in the closed right half-plane and the strict margin used to choose R. Shrink $\mathcal{U}_+$, while preserving $\mathcal{U} \Subset \mathcal{U}_+$, so that $\sup_{\omega\in\mathcal{U}_+}|\omega|R \leq \eta/2$. Since

$$|\Phi_n(\tau;\omega) - \Phi_n(\tau;0)| = |\omega\tau| \leq \eta/2 < \eta \qquad (\tau \in \partial\mathcal{H}_R),$$

Rouché's theorem shows that (S.3.150) has exactly one zero $t_{*,n}(\omega)$ in $\mathcal{H}_R$, counted with multiplicity, for every $\omega \in \mathcal{U}_+$. The zero is therefore simple. By the analytic implicit function theorem and uniqueness, it is analytic in ω. The zero is also uniformly isolated: for every fixed $\varepsilon > 0$, there is a constant $c_\varepsilon > 0$ such that

$$\inf_{n\geq n_0}\inf_{\omega\in\mathcal{U}}\inf_{\substack{\tau\in\overline{\mathcal{H}_R}\\ |\tau - t_{*,n}(\omega)|\geq\varepsilon}}|\Phi_n(\tau;\omega)| \geq c_\varepsilon. \tag{S.3.153}$$

If this bound failed, there would be sequences n_k, $\omega_k \in \mathcal{U}$, and $\tau_k \in \overline{\mathcal{H}_R}$ such that

$$|\tau_k - t_{*,n_k}(\omega_k)| \geq \varepsilon, \qquad |\Phi_{n_k}(\tau_k;\omega_k)| \longrightarrow 0.$$

After taking a subsequence, the parameter triples $(p/n_k, \gamma, \omega_k)$, the points τ_k, and the roots $t_{*,n_k}(\omega_k)$ all converge. The boundary gap from the preceding Rouché estimate places both point limits in $\mathcal{H}_R$. By continuity, both limits are zeros of the same limiting residual. But Rouché's theorem gives that residual only one zero in $\mathcal{H}_R$. This contradicts their separation by ε.

The empirical trace belongs to $\mathcal{H}_R$: if $\lambda_1(S), \ldots, \lambda_p(S) \geq 0$ are the eigenvalues of S, then

$$\operatorname{Re} t_n^\circ(\omega) = \frac{1}{n}\sum_{k=1}^{p}\mathbb{E}\frac{\lambda_k(S) + \gamma + \operatorname{Re}\omega}{|\lambda_k(S) + \gamma + \omega|^2} > 0.$$

Hence (S.3.147) and (S.3.153) imply

$$\sup_{\omega\in\mathcal{U}}|t_n^\circ(\omega) - t_{*,n}(\omega)| \longrightarrow 0. \tag{S.3.154}$$

For the deterministic branch set

$$\xi_I = \gamma + \omega + \alpha_K(\bar{v}_I), \qquad \delta_I = \delta(\bar{v}_I).$$

Sherman–Morrison gives

$$\bar{Q}_I = \xi_I^{-1} P_\perp + (\xi_I + p\delta_I)^{-1} P. \tag{S.3.155}$$

Equations (S.3.107) and (S.3.155) give

$$s_I = \frac{p}{n(\xi_I + p\delta_I)} = O(n^{-1}). \tag{S.3.156}$$

Thus $\bar{v}_I = m(t_I, 0) + O(n^{-1})$, and taking the trace in (S.3.155) gives the more explicit identity

$$t_I \xi_I - \frac{p}{n} = -\frac{1}{n} + \frac{\xi_I}{n(\xi_I + p\delta_I)} = O(n^{-1}). \tag{S.3.157}$$

Since $\xi_I = \gamma + \omega + \nu_\alpha - \beta_\sigma^2 - v^\top m(t_I, 0) + O(n^{-1})$, (S.3.156) and (S.3.157) imply

$$\sup_{\omega \in \mathcal{U}} |\Phi_n(t_I(\omega); \omega)| = O(n^{-1}). \tag{S.3.158}$$

Since $t_I(0) > 0$, it belongs to $\mathcal{H}_R$. Equations (S.3.153) and (S.3.158) show that $t_I(0) - t_{*,n}(0) \to 0$. The lower bound (S.3.151) therefore places $t_I(0)$ uniformly away from the imaginary axis. By the uniform implicit-function estimate (S.3.106), $\mathcal{U}_+$ may be reduced, independently of n and while preserving $\mathcal{U} \Subset \mathcal{U}_+$, so that $t_I(\omega) \in \mathcal{H}_R$ throughout the disk. Applying (S.3.153) and (S.3.158) once more gives

$$\sup_{\omega \in \mathcal{U}} |t_I(\omega) - t_{*,n}(\omega)| \longrightarrow 0. \tag{S.3.159}$$

Combining (S.3.154) and (S.3.159) yields

$$\sup_{\omega \in \mathcal{U}} |t_n^\circ(\omega) - t_I(\omega)| \longrightarrow 0. \tag{S.3.160}$$

Step 6: operator norm estimate. Let Π be an arbitrary $p \times p$ coordinate-permutation matrix. Under the standing conditioning on W, the vectors $g_i = W^\top x_i$ are independent $\mathcal{N}(0, I_p)$ vectors. Since the same activation is applied to every feature coordinate, jointly over all ordered pairs,

$$(z_{ij})_{i,j} \overset{d}{=} (\Pi z_{ij})_{i,j}.$$

For all sufficiently large n, $p \geq 2$. The distributional invariance implies

$$S \overset{d}{=} \Pi S \Pi^\top, \qquad Q_I(\omega) \overset{d}{=} \Pi Q_I(\omega) \Pi^\top,$$

and hence

$$\Pi\, \mathbb{E} Q_I(\omega)\, \Pi^\top = \mathbb{E} Q_I(\omega) \qquad \text{for every coordinate permutation } \Pi. \tag{S.3.161}$$

Commutation with every coordinate transposition in (S.3.161) forces all diagonal entries of $\mathbb{E}Q_I$ to be equal and all off-diagonal entries to be equal. Hence $\mathbb{E}Q_I$ belongs to $\mathrm{span}\{P_\perp, P\}$. Its two coefficients are already determined by t_n° and s_n°: since $\mathrm{tr}\, P_\perp = p - 1$ and $\mathbf{1}^\top P \mathbf{1} = p$, one has the exact representation

$$\mathbb{E} Q_I(\omega) = \frac{n t_n^\circ(\omega) - \frac{n}{p} s_n^\circ(\omega)}{p - 1} P_\perp + \frac{n}{p} s_n^\circ(\omega) P.$$

The coefficient of P is the eigenvalue in the direction $\mathbf{1}/\sqrt{p}$. Lemma S.3.11 gives, uniformly on $\mathcal{U}$,

$$\begin{aligned} \left| \frac{n}{p} s_n^\circ(\omega) \right| &= \frac{1}{p} \left| \mathbb{E} \mathbf{1}^\top Q_I(\omega) \mathbf{1} \right| \\ &\leq \frac{1}{p} \left\{ \mathbb{E} |\mathbf{1}^\top Q_I(\omega) \mathbf{1}|^2 \right\}^{1/2} \leq \frac{C}{p}. \end{aligned}$$

For $\bar{Q}_I$, (S.3.155) and the definition of s_I give

$$\frac{n}{p} s_I = (\xi_I + p\delta_I)^{-1}.$$

Moreover, (S.3.107) gives, after renaming its lower-bound constant, constants $c_\delta, C_\xi > 0$ independent of n and $\omega \in \mathcal{U}$ such that

$$|\delta_I(\omega)| \geq c_\delta, \qquad |\xi_I(\omega)| \leq C_\xi.$$

Therefore, for all sufficiently large n,

$$|\xi_I + p\delta_I| \geq p|\delta_I| - |\xi_I| \geq pc_\delta - C_\xi \geq \frac{1}{2}pc_\delta,$$

and hence

$$\sup_{\omega\in\mathcal{U}} \left|\frac{n}{p} s_I(\omega)\right| \leq \frac{C}{p}.$$

Therefore

$$\sup_{\omega\in\mathcal{U}} \frac{n}{p}|s_n^\circ(\omega) - s_I(\omega)| = O(p^{-1}) = o(1). \tag{S.3.162}$$

The definitions of $t_I = n^{-1}\operatorname{tr}\bar{Q}_I$ and $s_I = n^{-1}\mathbf{1}^\top\bar{Q}_I\mathbf{1}$ give the deterministic matrix

$$\bar{Q}_I(\omega) = \frac{nt_I(\omega) - \frac{n}{p}s_I(\omega)}{p-1} P_\perp + \frac{n}{p} s_I(\omega) P.$$

Hence the difference between the two coefficients of $P_\perp$ is

$$\frac{nt_n^\circ - \frac{n}{p}s_n^\circ}{p-1} - \frac{nt_I - \frac{n}{p}s_I}{p-1} = \frac{n}{p-1}(t_n^\circ - t_I) - \frac{n}{p(p-1)}(s_n^\circ - s_I). \tag{S.3.163}$$

Since $p/n \to c \in (0,\infty)$, (S.3.160), (S.3.162), and (S.3.163) imply

$$\sup_{\omega\in\mathcal{U}} \left|\frac{nt_n^\circ(\omega) - \frac{n}{p}s_n^\circ(\omega)}{p-1} - \frac{nt_I(\omega) - \frac{n}{p}s_I(\omega)}{p-1}\right| \longrightarrow 0. \tag{S.3.164}$$

Because P and $P_\perp$ are orthogonal projections,

$$\|\mathbb{E}Q_I - \bar{Q}_I\| = \max\left\{\left|\frac{nt_n^\circ - \frac{n}{p}s_n^\circ}{p-1} - \frac{nt_I - \frac{n}{p}s_I}{p-1}\right|, \frac{n}{p}|s_n^\circ - s_I|\right\}.$$

Equations (S.3.162) and (S.3.164) prove (S.3.96).

□

S.3.6.4 Cross-term contraction

Throughout Sections S.3.6.4 and S.3.6.5, we use the local abbreviations

$$\rho_\perp^2 := \rho_{\perp,n}^2, \qquad \rho^2 := \rho_n^2.$$

Here $\rho_\perp^2$ and ρ^2 denote average training second moments. Under the bounded-profile condition (S.3.10), the conditional covariance of the residual vector is the diagonal matrix $D_{\zeta,n}$ defined below.

Lemma S.3.13 Assume the standing conditions for a nonzero feature mean, including (S.1.7); in particular, $\beta_\sigma \neq 0$. Put

$$\ell_{ij}^{\mathrm{proj}} := \{\lambda g_i + (1-\lambda)g_j\}^\top a.$$

Then

$$\frac{1}{n^2}\sum_{i,j=1}^n \ell_{ij}^{\mathrm{proj}}\, a^\top Q z_{ij} - d_a(\alpha^2 - c_B)\, a^\top\bar{Q}a \xrightarrow{\mathbb{P}} 0. \tag{S.3.165}$$

Proof. Condition on W throughout and take $u = a$, which is then deterministic with $\|u\| = O(1)$. Put

$$g_i = W^\top x_i, \qquad y_i^{\|} = g_i^\top a, \qquad \bar{y}_{\|} = \frac{1}{n}\sum_{i=1}^n y_i^{\|},$$

and

$$r_{\|} = \frac{1}{n^2}\sum_{i,j=1}^{n} z_{ij}\{\lambda y_i^{\|} + (1-\lambda)y_j^{\|}\}.$$

By the definition of t_{ij}^{proj}, the left-hand side of (S.3.165), before subtraction of its deterministic term, is exactly $u^{\mathrm{T}}Qr_{\|}$. It therefore suffices to prove

$$\left\|u^{\mathrm{T}}Qr_{\|} - \theta_0 q_0 u^{\mathrm{T}}P_{\perp}a\right\|_{L^2} \longrightarrow 0, \tag{S.3.166}$$

where

$$e = \frac{\mathbf{1}}{\sqrt{p}}, \qquad P_{\perp} = I_p - ee^{\mathrm{T}}, \qquad q_0 = \frac{t_{*,n}}{c_n},$$
$$C_{*,n} = (I_2 + t_{*,n}V)^{-1}, \qquad \theta_0 = d_\alpha w^{\mathrm{T}}C_{*,n}w.$$

To reduce the full-mixup Gram matrix to its two endpoint projections, write

$$b_i = f_1(g_i) - \beta_\sigma\mathbf{1}, \qquad c_i = f_2(g_i) - \beta_\sigma\mathbf{1}, \qquad F_i^{\circ} = [b_i, c_i],$$

let $\mathsf{U} = [b_1, \dots, b_n]$, $\mathsf{V} = [c_1, \dots, c_n]$, and set

$$H_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}},$$

$$C_n = \kappa(\lambda)I_p + \frac{1}{n}\mathsf{U}H_n\mathsf{U}^{\mathrm{T}} + \frac{1}{n}\mathsf{V}H_n\mathsf{V}^{\mathrm{T}}, \qquad s_{\|,n} = \frac{1}{n}\{\lambda\mathsf{U} + (1-\lambda)\mathsf{V}\}H_n y^{\|}.$$

Also put

$$\widehat{z} = \frac{1}{n^2}\sum_{i,j} z_{ij}, \qquad S_* = C_n + \widehat{z}\widehat{z}^{\mathrm{T}}, \qquad Q_* = (S_* + \gamma I_p)^{-1},$$

$$r_{*,\|} = \widehat{z}\,\bar{y}_{\|} + s_{\|,n}.$$

For $i \neq j$, the coordinatewise Hoeffding decomposition is

$$z_{ij} = \beta_\sigma\mathbf{1} + b_i + c_j + \eta_{ij}, \qquad \mathbb{E}[\eta_{ij} \mid g_i] = \mathbb{E}[\eta_{ij} \mid g_j] = 0,$$

where different feature coordinates are independent. We use the following decoupled matrix inequality only for the kernels below. Let $\mathcal{K}(\mathsf{X}_i, \mathsf{X}_j)$ be a centered $d_1 \times d_2$ kernel of iid inputs that is canonical in both endpoints. Introduce an independent copy (X'_j), and, after a split of the two index sets, put

$$\mathsf{Y}_j(\mathcal{K}) = \sum_{i\in I}\mathcal{K}(\mathsf{X}_i, \mathsf{X}'_j), \qquad j \in J,$$

where $|I|, |J| \le n$. Define

$$\mathfrak{S}_q(\mathcal{K}; I, J) = \left\|\left\{\sum_{j\in J}\mathbb{E}_j\mathsf{Y}_j(\mathcal{K})\mathsf{Y}_j(\mathcal{K})^{\mathrm{T}}\right\}^{1/2}\right\|_{L^q,\mathrm{op}}$$
$$+ \left\|\left\{\sum_{j\in J}\mathbb{E}_j\mathsf{Y}_j(\mathcal{K})^{\mathrm{T}}\mathsf{Y}_j(\mathcal{K})\right\}^{1/2}\right\|_{L^q,\mathrm{op}},$$

$$\mathfrak{M}_q(\mathcal{K}; I, J) = \left\|\max_{j\in J}\|\mathsf{Y}_j(\mathcal{K})\|\right\|_{L^q}.$$

Here $\mathbb{E}_j$ integrates only X'_j, conditionally on $(\mathsf{X}_i)_{i\in I}$. Set

$$\mathfrak{S}_q(\mathcal{K}) = \sup_{I,J}\mathfrak{S}_q(\mathcal{K}; I, J), \qquad \mathfrak{M}_q(\mathcal{K}) = \sup_{I,J}\mathfrak{M}_q(\mathcal{K}; I, J),$$

where the suprema range over the splits used in the decoupling argument. If X' is an independent input, define

$$A_{\mathcal{K}}(x) = \mathbb{E}_{\mathsf{X}'}[\mathcal{K}(x, \mathsf{X}')\mathcal{K}(x, \mathsf{X}')^{\mathrm{T}}],$$
$$L_{\mathcal{K}}(x, x') = \mathbb{E}_{\mathsf{X}'}[\mathcal{K}(x, \mathsf{X}')\mathcal{K}(x', \mathsf{X}')^{\mathrm{T}}],$$

and define $\widetilde{A}_{\mathcal{K}}, \widetilde{L}_{\mathcal{K}}$ by reversing the two factors. Expanding $\mathsf{Y}_j\mathsf{Y}_j^{\mathsf{T}}$ and $\mathsf{Y}_j^{\mathsf{T}}\mathsf{Y}_j$ gives

$$\sum_{j\in J}\mathbb{E}_j\mathsf{Y}_j\mathsf{Y}_j^{\mathsf{T}} = |J|\left\{\sum_{i\in I}A_{\mathcal{K}}(\mathsf{X}_i) + \sum_{\substack{i,k\in I\\ i\neq k}}L_{\mathcal{K}}(\mathsf{X}_i,\mathsf{X}_k)\right\},$$
$$\sum_{j\in J}\mathbb{E}_j\mathsf{Y}_j^{\mathsf{T}}\mathsf{Y}_j = |J|\left\{\sum_{i\in I}\widetilde{A}_{\mathcal{K}}(\mathsf{X}_i) + \sum_{\substack{i,k\in I\\ i\neq k}}\widetilde{L}_{\mathcal{K}}(\mathsf{X}_i,\mathsf{X}_k)\right\}. \tag{S.3.167}$$

Canonicality also gives, for an independent X_k,

$$\mathbb{E}_{\mathsf{X}_k}[L_{\mathcal{K}}(\mathsf{X}_i,\mathsf{X}_k)\mid \mathsf{X}_i] = \mathbb{E}_{\mathsf{X}_k}[\widetilde{L}_{\mathcal{K}}(\mathsf{X}_i,\mathsf{X}_k)\mid \mathsf{X}_i] = 0.$$

For background on decoupling and noncommutative Rosenthal inequalities, see (de la Peña & Giné, 1999; Junge & Xu, 2003). Further background and alternative comparisons appear in (de la Peña & Montgomery-Smith, 1995; Junge & Xu, 2008). Here we use the forward canonical Banach-valued decoupling in (Minsker & Wei, 2019, Lemma 5.2), which requires no symmetry of the kernel. Conditional on one side of the decoupled bipartition, the resulting matrices are independent and centered, so the rectangular L^q moment estimate (Dirksen et al., 2022, Theorem 10) applies. Taking

$$q \geq \max\{4, \log(2 + d_1 + d_2)\}$$

makes its dimensional factor $O(\sqrt{q})$, and the Schatten-to-operator comparison is

$$\|A\|_{\mathrm{op}} \leq \|A\|_{S_q} \leq (d_1 + d_2)^{1/q}\|A\|_{\mathrm{op}}.$$

The random-bipartition identity contributes only a universal factor and gives

$$\left\|\frac{1}{n^2}\sum_{i\neq j}\mathcal{K}(\mathsf{X}_i,\mathsf{X}_j)\right\|_{L^q,\mathrm{op}} \leq \frac{Cq}{n^2}\{\mathfrak{S}_q(\mathcal{K}) + \mathfrak{M}_q(\mathcal{K})\}. \tag{S.3.168}$$

To obtain this bound, decouple and split the indices, condition on $(\mathsf{X}_i)_{i\in I}$, and apply Rosenthal to the independent centered matrices $\mathsf{Y}_j(\mathcal{K})$. Averaging over the split gives (S.3.168). We take

$$q = q_{\log} := \lceil 4\log(2p)\rceil,$$

so the dimension factor is absorbed and the $L^{q_{\log}}$ estimate controls the required L^4 norm.

Let

$$\mathrm{off}(A) = A - \mathrm{diag}(A), \qquad \mathcal{K}_\eta(g_i, g_j) = \mathrm{off}(\eta_{ij}\eta_{ij}^{\mathsf{T}}).$$

The kernel $\mathcal{K}_\eta$ is canonical in both endpoints. Coordinate independence and the fixed Gaussian moments imply, at $q = q_{\log}$,

$$\left\|\mathbb{E}\mathcal{K}_\eta\mathcal{K}_\eta^{\mathsf{T}}\right\| + \left\|\mathbb{E}\mathcal{K}_\eta^{\mathsf{T}}\mathcal{K}_\eta\right\| \leq Cp,$$
$$\|\mathcal{K}_\eta(g_1,g_2)\|_{L^{q_{\log}},\mathrm{op}} \leq Cq_{\log}^C p.$$

For this kernel, write

$$\eta_{ij,r} = h_0(g_{ir}, g_{jr}), \qquad \tau(x) = \mathbb{E}_G h_0(x,G)^2, \qquad \rho(x,x') = \mathbb{E}_G[h_0(x,G)h_0(x',G)].$$

Since $\mathbb{E}_G h_0(x,G) = 0$, direct multiplication of $\mathrm{off}(\eta\eta^{\mathsf{T}})$ gives, for $k,\ell \leq p$,

$$[A_{\mathcal{K}_\eta}(x)]_{k\ell} = \mathbf{1}_{\{k=\ell\}}\,\tau(x_k)\sum_{m\neq k}\tau(x_m),$$
$$[L_{\mathcal{K}_\eta}(x,x')]_{k\ell} = \mathbf{1}_{\{k=\ell\}}\,\rho(x_k,x_k')\sum_{m\neq k}\rho(x_m,x_m').$$

Indeed, an off-diagonal entry is a sum of expectations $\eta_k\eta_\ell\eta_m^2$. If $m \notin \{k,\ell\}$, the expectation is zero by centering and independence of the two coordinates; if $m = k$ or $m = \ell$, one of the removed diagonal entries is present. The same coordinate argument proves the second formula. The matrices are symmetric, so the two reversed products obey identical formulas.

Gaussian moment bounds give $\|\tau(G)\|_{L^s} + \|\rho(G, G')\|_{L^s} \le C s^C$. Because $q_{\log} \ge 4 \log p$, the maximum over p coordinates costs a bounded factor in $L^{q_{\log}/2}$. Minkowski and scalar Rosenthal therefore give

$$\left\| \max_{k \le p} \sum_{i \in I} \tau(g_{ik}) \sum_{m \ne k} \tau(g_{im}) \right\|_{L^{q_{\log}/2}} \le C q_{\log}^C np.$$

Moreover, $\mathbb{E}_{G'} \rho(x, G') = 0$. For fixed r,

$$\ell_r(x, x') = \rho(x_r, x'_r) \sum_{m \ne r} \rho(x_m, x'_m)$$

is canonical in both endpoints and has L^s-norm $O(s^C \sqrt{p})$. In a decoupled split, scalar Rosenthal gives the conditional square-function bound $O(s^C n \sqrt{p})$ and the maximal summand bound $O(s^C \sqrt{np})$. Applying (S.3.168) in dimension one and then taking the maximum over r yields

$$\left\| \max_{r \le p} \left| \sum_{\substack{i,j \in I \\ i \ne j}} \rho(g_{ir}, g_{jr}) \sum_{m \ne r} \rho(g_{im}, g_{jm}) \right| \right\|_{L^{q_{\log}/2}} \le C q_{\log}^C n \sqrt{p}.$$

Combining the two diagonal formulas proves, uniformly in I,

$$\left\| \sum_{i \in I} A_{\mathcal{K}_\eta}(g_i) + \sum_{\substack{i,k \in I \\ i \ne k}} L_{\mathcal{K}_\eta}(g_i, g_k) \right\|_{L^{q_{\log}/2}, \mathrm{op}} + \left\| \sum_{i \in I} \widetilde{A}_{\mathcal{K}_\eta}(g_i) + \sum_{\substack{i,k \in I \\ i \ne k}} \widetilde{L}_{\mathcal{K}_\eta}(g_i, g_k) \right\|_{L^{q_{\log}/2}, \mathrm{op}} \le C q_{\log}^C np. \tag{S.3.169}$$

Conditioning instead on g'_j, matrix Rosenthal for the independent centered matrices $\{\mathcal{K}_\eta(g_i, g'_j) : i \in I\}$, followed by the maximum over $j \in J$, gives

$$\mathfrak{M}_{q_{\log}}(\mathcal{K}_\eta) \le C q_{\log}^C \{ \sqrt{np} + p \}. \tag{S.3.170}$$

Combining (S.3.167), (S.3.169), and (S.3.170) yields

$$\mathfrak{S}_{q_{\log}}(\mathcal{K}_\eta) + \mathfrak{M}_{q_{\log}}(\mathcal{K}_\eta) \le C q_{\log}^C \{ n \sqrt{p} + p \}.$$

Consequently, (S.3.168), $p/n = O(1)$, and monotonicity of L^q-norms give

$$\left\| \frac{1}{n^2} \sum_{i \ne j} \mathcal{K}_\eta(g_i, g_j) \right\|_{L^4, \mathrm{op}} \le C n^{-1/2} \log^C n. \tag{S.3.171}$$

For the diagonal coordinates, the scalar Hoeffding decomposition of $\eta_{ij,k}^2 - \kappa(\lambda)$, scalar Rosenthal, and a union bound over $k \le p$ give

$$\left\| \max_{k \le p} \left| \frac{1}{n^2} \sum_{i \ne j} (\eta_{ij,k}^2 - \kappa(\lambda)) \right| \right\|_{L^4} \le C n^{-1/2} \log^C n.$$

For $(b_i + c_j)\eta_{ij}^{\mathrm{T}}$, define

$$D_1(x) = \mathbb{E}_{G'}[(b(x) + c(G'))\eta(x, G')^{\mathrm{T}}], \qquad D_2(y) = \mathbb{E}_G[(b(G) + c(y))\eta(G, y)^{\mathrm{T}}],$$

and put

$$\mathcal{K}_\times(x, y) = (b(x) + c(y))\eta(x, y)^{\mathrm{T}} - D_1(x) - D_2(y).$$

Conditional centering of η removes the $b(x)$-part of D_1 and the $c(y)$-part of D_2. Feature-coordinate independence shows that every off-diagonal entry of D_1, D_2 vanishes, while integrating the remaining endpoint shows that each diagonal projection is centered. Thus $\mathcal{K}_\times$ is canonical in both endpoints. We estimate it using a two-sample representation, without an entrywise Schur bound. Introduce independent endpoint samples $(G_i)_{i \in I}$ and $(G'_j)_{j \in J}$, all distributed as $\mathcal{N}(0, I_p)$, and write

$$b(x) = f_1(x) - \beta_\sigma \mathbf{1}, \qquad c(y) = f_2(y) - \beta_\sigma \mathbf{1}, \qquad \eta(x, y)_r = h_0(x_r, y_r).$$

Let

$$D_1(x) = \mathbb{E}_{G'}[(b(x) + c(G'))\eta(x, G')^{\mathrm{T}}],$$
$$D_2(y) = \mathbb{E}_{G}[(b(G) + c(y))\eta(G, y)^{\mathrm{T}}],$$

Both D_1, D_2 are diagonal and centered. If

$$U_i^\eta = \sum_{j\in J} \eta(G_i, G'_j), \qquad V_j^\eta = \sum_{i\in I} \eta(G_i, G'_j),$$

then the decoupled canonical sum has the exact two-channel decomposition

$$\sum_{\substack{i\in I\\ j\in J}} \mathcal{K}_\times(G_i, G'_j) = \left\{ \sum_{i\in I} b(G_i)(U_i^\eta)^{\mathrm{T}} - |I| \sum_{j\in J} D_2(G'_j) \right\} + \left\{ \sum_{j\in J} c(G'_j)(V_j^\eta)^{\mathrm{T}} - |J| \sum_{i\in I} D_1(G_i) \right\}. \tag{S.3.172}$$

Conditionally on $(G_i)_{i\in I}$, the matrices

$$A_j = c(G'_j)(V_j^\eta)^{\mathrm{T}} - \mathbb{E}_{G'_j}[c(G'_j)(V_j^\eta)^{\mathrm{T}} \mid (G_i)_{i\in I}]$$

are independent and centered; their sum is the second brace in (S.3.172). Centering decreases both conditional second moments in the positive-semidefinite order. The two uncentered variance matrices are

$$\mathbb{E}_j[\|V_j^\eta\|^2 c(G'_j)c(G'_j)^{\mathrm{T}}], \qquad \mathbb{E}_j[\|c(G'_j)\|^2 V_j^\eta (V_j^\eta)^{\mathrm{T}}].$$

Both are diagonal. In an off-diagonal entry of the first matrix, expand $\|V_j^\eta\|^2 = \sum_r (V_{j,r}^\eta)^2$. If $r \notin \{k, \ell\}$, independence leaves the factors $\mathbb{E}c_k = \mathbb{E}c_\ell = 0$; if $r = k$ or $r = \ell$, the other coordinate remains centered. The second matrix is diagonal by the same argument, using the conditional centering of each $V_{j,r}^\eta$.

For every coordinate,

$$V_{j,r}^\eta = \sum_{i\in I} h_0(G_{i,r}, G'_{j,r}).$$

Scalar Rosenthal and the canonical contraction give

$$\|\mathbb{E}_j(V_{j,r}^\eta)^2\|_{L^{q_{\log}/2}} \le C q_{\log}^C n, \qquad \|V_{j,r}^\eta\|_{L^{2q_{\log}}} \le C q_{\log}^C \sqrt{n}.$$

Summing the p matched coordinates in each diagonal entry, applying Hölder's inequality, and taking the maximum over the diagonal entries gives

$$\left\| \sum_{j\in J} \mathbb{E}_j A_j A_j^{\mathrm{T}} \right\|_{L^{q_{\log}/2},\mathrm{op}} + \left\| \sum_{j\in J} \mathbb{E}_j A_j^{\mathrm{T}} A_j \right\|_{L^{q_{\log}/2},\mathrm{op}} \le C q_{\log}^C n^2 p, \tag{S.3.173}$$

$$\left\| \max_{j\in J} \|A_j\| \right\|_{L^{q_{\log}}} \le C q_{\log}^C p \sqrt{n}. \tag{S.3.174}$$

Summing over $|J| \le n$ proves (S.3.173). Moreover,

$$\|c(G'_j)(V_j^\eta)^{\mathrm{T}}\| = \|c(G'_j)\| \, \|V_j^\eta\|, \qquad \||\|c(G'_j)\| \, \|V_j^\eta\|\|_{L^{q_{\log}}} \le C q_{\log}^C p \sqrt{n},$$

and $n^{1/q_{\log}} = O(1)$ permits the maximum over j. This proves (S.3.174). Conditional matrix Rosenthal now yields

$$\left\| \sum_{j\in J} A_j \right\|_{L^{q_{\log}},\mathrm{op}} \le C q_{\log}^C n \sqrt{p},$$

where p/n is bounded above and below. Conditioning on the second endpoint sample and interchanging b, U^η, I with c, V^η, J gives the same estimate for the first brace in (S.3.172). Decoupling and averaging over the random split therefore give

$$\left\| \frac{1}{n^2} \sum_{i\ne j} \mathcal{K}_\times(g_i, g_j) \right\|_{L^4,\mathrm{op}} \le C n^{-1/2} \log^C n. \tag{S.3.175}$$

The transpose has the same operator norm. Since D_1, D_2 are diagonal and centered, scalar Rosenthal and a maximum over their coordinates also give

$$\left\|\frac{1}{n}\sum_i D_1(g_i)\right\|_{L^4,\mathrm{op}} + \left\|\frac{1}{n}\sum_i D_2(g_i)\right\|_{L^4,\mathrm{op}} \le Cn^{-1/2}\log^C n.$$

Absorb these terms into $E_{S,n}$. Expanding the remaining non-canonical part and collecting the empirical feature means then gives

$$\kappa(\lambda)I_p + \frac{1}{n}\mathsf{U}H_n\mathsf{U}^\top + \frac{1}{n}\mathsf{V}H_n\mathsf{V}^\top + \widehat{zz}^\top.$$

The factor $1-n^{-1}$ multiplying the off-diagonal residual covariance differs from one by $O(n^{-1})$, and this deterministic discrepancy is absorbed into $E_{S,n}$. The n diagonal pairs have weight n^{-1}. Their centered part is $O_{L^4,\mathrm{op}}(n^{-1/2}\log^C n)$. Their deterministic covariance part has operator norm $O(n^{-1})$ and is also included in $E_{S,n}$; the only order-one operator norm contribution is

$$d_n\mathbf{1}\mathbf{1}^\top, \qquad d_n = \frac{n-1}{n^2}\{\mu_0(\sigma)-\beta_\sigma\}^2.$$

Indeed, with $\beta_n = (1-n^{-1})\beta_\sigma + n^{-1}\mu_0(\sigma)$,

$$\left(1-\frac{1}{n}\right)\beta_\sigma^2 + \frac{1}{n}\mu_0(\sigma)^2 - \beta_n^2 = \frac{n-1}{n^2}\{\mu_0(\sigma)-\beta_\sigma\}^2.$$

For the score, put

$$K_{ij}^{\|} = \eta_{ij}\{\lambda y_i^{\|} + (1-\lambda)y_j^{\|}\}, \qquad q_{ii} = \beta_\sigma\mathbf{1} + b_i + c_i.$$

Direct expansion of $r_{\|}$ and $r_{*,\|}$ gives the exact identity

$$\begin{aligned} r_{\|} - r_{*,\|} = {} & \frac{1}{n^2}\sum_{i\neq j}\eta_{ij}\{\lambda y_i^{\|} + (1-\lambda)y_j^{\|} - \bar{y}_{\|}\} \\ & + \frac{1}{n^2}\sum_i (z_{ii}-q_{ii})(y_i^{\|}-\bar{y}_{\|}). \end{aligned} \tag{S.3.176}$$

For the vector kernel $(g_i, g_j) \mapsto \eta_{ij}$, (S.3.167), $\mathbb{E}\eta_{12}\eta_{12}^\top = \kappa(\lambda)I_p$, and scalar Rosenthal for the remaining canonical contractions give

$$\mathfrak{S}_{q_{\log}}(\eta) + \mathfrak{M}_{q_{\log}}(\eta) \le Cq_{\log}^C n\sqrt{p}.$$

Hence the vector canonical mean satisfies

$$\left\|\frac{1}{n^2}\sum_{i\neq j}\eta_{ij}\right\|_{L^8} \le Cn^{-1/2}\log^C n,$$

by the Hilbert space version of (S.3.168). Since $\|\bar{y}_{\|}\|_{L^8} = O(n^{-1/2})$, the part $-\bar{y}_{\|}n^{-2}\sum_{i\neq j}\eta_{ij}$ in the first line of (S.3.176) is $O_{L^4}(n^{-1}\log^C n)$.

Define the two first projections and the canonical remainder by

$$\pi_{1,\|}(g_i) = \mathbb{E}[K_{ij}^{\|} \mid g_i], \qquad \pi_{2,\|}(g_j) = \mathbb{E}[K_{ij}^{\|} \mid g_j],$$

$$\pi_{12,\|}(g_i, g_j) = K_{ij}^{\|} - \pi_{1,\|}(g_i) - \pi_{2,\|}(g_j).$$

Here $\mathbb{E}K_{ij}^{\|} = 0$, and $\pi_{12,\|}$ is canonical in both endpoints. Conditional canonicality of η_{ij}, coordinate independence, and $\|a\| = O(1)$ give

$$\|\pi_{1,\|}\|_{L^4(\ell_2)} + \|\pi_{2,\|}\|_{L^4(\ell_2)} \le C, \qquad \|\pi_{12,\|}\|_{L^4(\ell_2)} \le C\sqrt{p}.$$

Thus the two first projections contribute $O_{L^4}(n^{-1/2})$. For the canonical remainder,

$$\|\pi_{12,\|}\|_{L^{q_{\log}}(\ell_2)} \le Cq_{\log}^C\sqrt{p}.$$

For a decoupled split, set

$$\mathsf{Y}_j^\pi = \sum_{i\in I} \pi_{12,\|}(G_i, G'_j), \qquad a_\pi(x) = \mathbb{E}_G\|\pi_{12,\|}(x, G)\|^2,$$

$$\ell_\pi(x, x') = \mathbb{E}_G\langle \pi_{12,\|}(x, G), \pi_{12,\|}(x', G)\rangle.$$

Then

$$\sum_{j\in J} \mathbb{E}_j\|\mathsf{Y}_j^\pi\|^2 = |J|\left\{\sum_{i\in I} a_\pi(G_i) + \sum_{\substack{i,k\in I\\ i\neq k}} \ell_\pi(G_i, G_k)\right\}. \tag{S.3.177}$$

The kernel ℓ_π is canonical. Expanding $y^{\|} = \sum_r a_r g_r$ and using $\sum_r a_r^2 = O(1)$ give

$$\|a_\pi(G_1)\|_{L^{q_{\log}/2}} + \|\ell_\pi(G_1, G_2)\|_{L^{q_{\log}/2}} \le C q_{\log}^C p.$$

Apply scalar Rosenthal to the first sum, and scalar order-two decoupling followed by Rosenthal to the second, to obtain

$$\left\|\sum_{i\in I} a_\pi(G_i) + \sum_{\substack{i,k\in I\\ i\neq k}} \ell_\pi(G_i, G_k)\right\|_{L^{q_{\log}/2}} \le C q_{\log}^C np. \tag{S.3.178}$$

Since

$$\left\|\sum_{j\in J} \mathbb{E}_j \mathsf{Y}_j^\pi (\mathsf{Y}_j^\pi)^{\mathrm{T}}\right\| \le \sum_{j\in J} \mathbb{E}_j\|\mathsf{Y}_j^\pi\|^2,$$

(S.3.177)–(S.3.178) control both conditional square functions. Conditioning on G'_j and using Hilbert space Rosenthal for the independent centered vectors $\{\pi_{12,\|}(G_i, G'_j) : i \in I\}$ also gives

$$\mathfrak{M}_{q_{\log}}(\pi_{12,\|}) \le C q_{\log}^C \sqrt{np}.$$

Therefore

$$\mathfrak{S}_{q_{\log}}(\pi_{12,\|}) + \mathfrak{M}_{q_{\log}}(\pi_{12,\|}) \le C q_{\log}^C n\sqrt{p}.$$

Apply the Hilbert space version of (S.3.168) to the canonical remainder and scalar Rosenthal to the two first projections. Together with the preceding empirical mean correction and the bound

$$\left\|\frac{1}{n^2}\sum_i (z_{ii} - q_{ii})(y_i^{\|} - \bar{y}_{\|})\right\|_{L^4} \le \frac{C}{n}$$

for the second line of (S.3.176), these estimates give

$$\begin{aligned} S &= S_* + d_n \mathbf{1}\mathbf{1}^{\mathrm{T}} + E_{S,n}, & d_n &= \frac{n-1}{n^2}\{\mu_0(\sigma) - \beta_\sigma\}^2 = O(n^{-1}),\\ r_{\|} &= r_{*,\|} + e_{\|,n}, & \|E_{S,n}\|_{L^4,\mathrm{op}} + \|e_{\|,n}\|_{L^4} &\le C n^{-1/2}\log^C n. \end{aligned} \tag{S.3.179}$$

The resolvent identity and (S.3.179) yield

$$\begin{aligned} u^{\mathrm{T}} Q r_{\|} - u^{\mathrm{T}} Q_* r_{*,\|} = {} & u^{\mathrm{T}} Q e_{\|,n} - u^{\mathrm{T}} Q E_{S,n} Q_* r_{*,\|}\\ & - d_n (u^{\mathrm{T}} Q \mathbf{1})(\mathbf{1}^{\mathrm{T}} Q_* r_{*,\|}). \end{aligned}$$

The fixed Gaussian moments give

$$\|\widehat{z}\|_{L^8} = O(\sqrt{p}), \qquad \|\bar{y}_{\|}\|_{L^8} = O(n^{-1/2}), \qquad \|s_{\|,n}\|_{L^4} = O(1),$$

and hence $\|r_{*,\|}\|_{L^4} = O(1)$. The ridge bound then gives

$$\|u^{\mathrm{T}} Q e_{\|,n}\|_{L^2} + \|u^{\mathrm{T}} Q E_{S,n} Q_* r_{*,\|}\|_{L^2} \le C n^{-1/2}\log^C n.$$

Moreover, Lemma S.3.11 and the ridge bound imply

$$\|u^{\mathrm{T}} Q \mathbf{1}\|_{L^4} \le C, \qquad \|\mathbf{1}^{\mathrm{T}} Q_* r_{*,\|}\|_{L^4} \le \sqrt{p}\,\gamma_-^{-1}\|r_{*,\|}\|_{L^4} \le C\sqrt{n}.$$

Since $d_n = O(n^{-1})$,

$$\left\|u^{\mathrm{T}} Q r_{\|} - u^{\mathrm{T}} Q_* r_{*,\|}\right\|_{L^2} \le C n^{-1/2} \log^C n. \tag{S.3.180}$$

For the centered covariance channel, define

$$\check{C}_n = \kappa(\lambda) I_p + \frac{1}{n}\sum_{i=1}^{n} F_i^{\circ}(F_i^{\circ})^{\mathrm{T}}, \qquad \check{s}_{\|,n} = \frac{1}{n}\sum_{i=1}^{n} F_i^{\circ} w\, y_i^{\|},$$

$$\check{R} = (\check{C}_n + \gamma I_p)^{-1}.$$

After deleting sample i, put

$$\check{R}_i = \left\{\kappa(\lambda) I_p + \frac{1}{n}\sum_{k\neq i} F_k^{\circ}(F_k^{\circ})^{\mathrm{T}} + \gamma I_p\right\}^{-1},$$

$$T_i = \frac{1}{n}(F_i^{\circ})^{\mathrm{T}} \check{R}_i F_i^{\circ}, \qquad C_i = (I_2 + T_i)^{-1}.$$

The pair $(F_i^{\circ}, y_i^{\|})$ is independent of $\check{R}_i$. Woodbury's identity gives

$$u^{\mathrm{T}} \check{R} \check{s}_{\|,n} = \frac{1}{n}\sum_{i=1}^{n} y_i^{\|} w^{\mathrm{T}} C_i (F_i^{\circ})^{\mathrm{T}} \check{R}_i u. \tag{S.3.181}$$

Conditional quadratic-form concentration, a rank-two trace replacement, and the reduced fixed-point equation yield

$$\left\{\frac{1}{n}\sum_i \|T_i - t_{*,n} V\|_{L^8}^8\right\}^{1/8} + \left\{\frac{1}{n}\sum_i \|C_i - C_{*,n}\|_{L^8}^8\right\}^{1/8} \le C n^{-1/2} \log^C n. \tag{S.3.182}$$

To verify this estimate, condition on $\check{R}_i$. Then

$$\left\|T_i - \frac{\operatorname{tr}\check{R}_i}{n} V\right\|_{L^8} \le \frac{C}{n}\{\operatorname{tr}(\check{R}_i \check{R}_i^*)\}^{1/2} \le C n^{-1/2}.$$

Set

$$\check{t}_n = \mathbb{E}\frac{1}{n}\operatorname{tr}\check{R}.$$

The rank-two resolvent identity gives

$$\left|\mathbb{E}\frac{1}{n}\operatorname{tr}\check{R}_i - \check{t}_n\right| \le \frac{C}{n}.$$

An explicit gradient calculation gives the trace-fluctuation estimate

$$\left\|\frac{1}{n}\operatorname{tr}\check{R}_i - \mathbb{E}\frac{1}{n}\operatorname{tr}\check{R}_i\right\|_{L^8} \le C n^{-1},$$

To verify it, let $\dot{F}_{k\ell}^{\circ} = \partial_{g_{k\ell}} F_k^{\circ}$. Then $\dot{F}_{k\ell}^{\circ}$ has only its ℓth row nonzero and has uniformly bounded norm. For $k \neq i$,

$$\partial_{g_{k\ell}}\left(\frac{1}{n}\operatorname{tr}\check{R}_i\right) = -\frac{2}{n^2}\operatorname{tr}\left((F_k^{\circ})^{\mathrm{T}} \check{R}_i^2 \dot{F}_{k\ell}^{\circ}\right),$$

whereas the derivative is zero for $k = i$. Hence

$$\begin{aligned}\sum_{k,\ell}\left|\partial_{g_{k\ell}}\left(\frac{1}{n}\operatorname{tr}\check{R}_i\right)\right|^2 &\le \frac{C}{n^4}\sum_{k\neq i}\operatorname{tr}((F_k^{\circ})^{\mathrm{T}}\check{R}_i^4 F_k^{\circ}) \\ &= \frac{C}{n^3}\operatorname{tr}\left[\check{R}_i^4\left\{\frac{1}{n}\sum_{k\neq i} F_k^{\circ}(F_k^{\circ})^{\mathrm{T}}\right\}\right] \le \frac{C}{n^2}.\end{aligned}$$

The last inequality uses $p/n = O(1)$, the ridge bound, and $\sup_{x\ge 0} x(x+\gamma_-)^{-4} < \infty$. The Gaussian L^8-Poincaré inequality proves the displayed trace-fluctuation estimate. Therefore

$$\|T_i - \check{t}_n V\|_{L^8} \le C n^{-1/2}.$$

Taking the normalized trace in $\check{R}(\check{C}_n + \gamma I_p) = I_p$, using

$$\check{R}F_i^\circ = \check{R}_i F_i^\circ C_i, \qquad (F_i^\circ)^{\mathrm{T}} \check{R}_i F_i^\circ = nT_i,$$

first gives the exact random identity

$$c_n = (\gamma + \kappa(\lambda))\frac{1}{n}\operatorname{tr}\check{R} + \frac{1}{n}\sum_{i=1}^{n}\operatorname{tr}(T_i C_i). \tag{S.3.183}$$

Taking expectation in (S.3.183) yields

$$c_n = (\gamma + \kappa(\lambda))\check{t}_n + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\operatorname{tr}(T_i C_i).$$

The Lipschitz continuity of $T \mapsto T(I_2 + T)^{-1}$ and the preceding L^8 bounds then give

$$c_n = (\gamma + \kappa(\lambda))\check{t}_n + \operatorname{tr}\left[\check{t}_n V(I_2 + \check{t}_n V)^{-1}\right] + O(n^{-1/2}).$$

The derivative bound (S.3.48) now yields

$$|\check{t}_n - t_{*,n}| \le Cn^{-1/2}.$$

Thus the first term in (S.3.182) has the stated rate. Finally,

$$C_i - C_{*,n} = -C_i(T_i - t_{*,n}V)C_{*,n}.$$

This proves (S.3.182).

The conditional Gaussian moment bounds give, uniformly for deterministic $\|b\| \le C$,

$$\sup_i \|y_i^{\|}(F_i^\circ)^{\mathrm{T}}\check{R}_i b\|_{L^4} \le C.$$

Consequently, Hölder's inequality and (S.3.182) show that replacing C_i by $C_{*,n}$ in (S.3.181) changes the average by $o_{L^2}(1)$.

Gaussian integration by parts gives the exact covariance

$$\mathbb{E}[y_i^{\|}F_i^\circ] = d_\alpha a w^{\mathrm{T}}.$$

Thus, with

$$X_i = y_i^{\|}w^{\mathrm{T}}C_{*,n}(F_i^\circ)^{\mathrm{T}}\check{R}_i u - \theta_0 a^{\mathrm{T}}\check{R}_i u,$$

one has

$$\mathbb{E}[X_i \mid (g_k)_{k\neq i}, W] = 0, \qquad \sup_i \|X_i\|_{L^4} \le C.$$

For $i \neq j$, let

$$\mathcal{G}_{ij} = \sigma\{W, g_k : k \notin \{i, j\}\},$$

let $\check{R}_{ij}$ be the resolvent with both samples deleted, and let $X_i^{(ij)}$ be obtained by replacing $\check{R}_i$ with $\check{R}_{ij}$. Conditional on $\mathcal{G}_{ij}$, $X_i^{(ij)}$ and $X_j^{(ij)}$ are independent and centered. Moreover,

$$\check{R}_i - \check{R}_{ij} = -\frac{1}{n}\check{R}_i F_j^\circ (F_j^\circ)^{\mathrm{T}}\check{R}_{ij}.$$

Consequently,

$$X_i - X_i^{(ij)} = -\frac{1}{n}\left\{y_i^{\|}w^{\mathrm{T}}C_{*,n}(F_i^\circ)^{\mathrm{T}}\check{R}_i F_j^\circ - \theta_0 a^{\mathrm{T}}\check{R}_i F_j^\circ\right\}(F_j^\circ)^{\mathrm{T}}\check{R}_{ij}u.$$

The conditional fourth-moment bounds

$$\|\check{R}_{ij}F_i^\circ\|_{L^4} = O(\sqrt{n}), \qquad \|(F_j^\circ)^{\mathrm{T}}\check{R}_{ij}F_i^\circ\|_{L^4} = O(\sqrt{n}),$$

together with

$$\|(F_j^\circ)^{\mathrm{T}}\check{R}_{ij}u\|_{L^8} + \|b^{\mathrm{T}}\check{R}_i F_j^\circ\|_{L^8} \le C, \qquad \|y_i^{\|}(F_i^\circ)^{\mathrm{T}}\check{R}_i F_j^\circ\|_{L^4} \le C\sqrt{n},$$

for deterministic bounded b, give

$$\max_{i\neq j}\left\{\|X_i - X_i^{(ij)}\|_{L^2} + \|X_j - X_j^{(ij)}\|_{L^2}\right\} \le Cn^{-1/2}.$$

Therefore

$$|\mathbb{E}[X_i X_j]| \le Cn^{-1/2} \qquad (i \ne j),$$

and

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right|^2 \le \frac{C}{n} + Cn^{-1/2} = o(1). \tag{S.3.184}$$

The same rank-two identity gives

$$\left\|\frac{1}{n}\sum_i a^{\mathrm{T}}\check{R}_i u - a^{\mathrm{T}}\check{R}u\right\|_{L^2} = O(n^{-1}).$$

It follows that

$$u^{\mathrm{T}}\check{R}\check{s}_{\|,n} - \theta_0 u^{\mathrm{T}}\check{R}a \xrightarrow{L^2} 0. \tag{S.3.185}$$

We also need the deterministic bilinear form of $\check{R}$. For deterministic bounded b, d, set

$$Y_i = b^{\mathrm{T}}\check{R}_i F_i^{\circ} C_{*,n}(F_i^{\circ})^{\mathrm{T}} d - \mathrm{tr}(VC_{*,n})\, b^{\mathrm{T}}\check{R}_i d.$$

Then $\mathbb{E}[Y_i \mid (g_k)_{k\ne i}, W] = 0$. For the common leave-two calculation, let $Y_i^{(ij)}$ be defined by replacing $\check{R}_i$ with $\check{R}_{ij}$. Then

$$\mathbb{E}[Y_i^{(ij)} Y_j^{(ij)} \mid \mathcal{G}_{ij}] = 0$$

and

$$\begin{aligned} Y_i - Y_i^{(ij)} = &-\frac{1}{n}(b^{\mathrm{T}}\check{R}_i F_j^{\circ})((F_j^{\circ})^{\mathrm{T}}\check{R}_{ij}F_i^{\circ})C_{*,n}((F_i^{\circ})^{\mathrm{T}}d) \\ &+ \frac{\mathrm{tr}(VC_{*,n})}{n}(b^{\mathrm{T}}\check{R}_i F_j^{\circ})((F_j^{\circ})^{\mathrm{T}}\check{R}_{ij}d). \end{aligned}$$

Thus

$$\sup_i \|Y_i\|_{L^8} \le C, \qquad \max_{i\ne j}\|Y_i - Y_i^{(ij)}\|_{L^2} \le Cn^{-1/2}.$$

Expanding the off-diagonal covariance as in (S.3.184) therefore gives

$$\mathbb{E}\left|\frac{1}{n}\sum_i Y_i\right|^2 \le \frac{C}{n} + Cn^{-1/2} = o(1).$$

Taking the (b, d) bilinear form in $\check{R}(\check{C}_n + \gamma I_p) = I_p$, using

$$\check{R}F_i^{\circ} = \check{R}_i F_i^{\circ} C_i,$$

gives the exact identity

$$b^{\mathrm{T}}d = (\gamma + \kappa(\lambda))b^{\mathrm{T}}\check{R}d + \frac{1}{n}\sum_i b^{\mathrm{T}}\check{R}_i F_i^{\circ} C_i (F_i^{\circ})^{\mathrm{T}} d.$$

By (S.3.182) and the conditional Gaussian moment bounds,

$$\begin{aligned} &\frac{1}{n}\sum_i b^{\mathrm{T}}\check{R}_i F_i^{\circ} C_i (F_i^{\circ})^{\mathrm{T}} d \\ &\quad = \frac{1}{n}\sum_i Y_i + \mathrm{tr}(VC_{*,n})\frac{1}{n}\sum_i b^{\mathrm{T}}\check{R}_i d + o_{L^2}(1). \end{aligned}$$

The rank-two identity also gives, uniformly in i,

$$\|b^{\mathrm{T}}(\check{R}_i - \check{R})d\|_{L^2} = \left\|\frac{1}{n} b^{\mathrm{T}}\check{R}_i F_i^{\circ} C_i (F_i^{\circ})^{\mathrm{T}}\check{R}_i d\right\|_{L^2} \le \frac{C}{n}.$$

Together with the preceding bound on $n^{-1}\sum_i Y_i$, this proves

$$b^{\mathrm{T}}d = \{\gamma + \kappa(\lambda) + \mathrm{tr}(VC_{*,n})\}b^{\mathrm{T}}\check{R}d + o_{L^2}(1).$$

The reduced fixed-point equation gives

$$\gamma + \kappa(\lambda) + \operatorname{tr}(VC_{*,n}) = \frac{c_n}{t_{*,n}} = \frac{1}{q_0}.$$

Hence

$$b^{\mathsf{T}}\check{R}d - q_0 b^{\mathsf{T}}d \xrightarrow{L^2} 0 \tag{S.3.186}$$

for every deterministic bounded pair of sequences $b = b_n, d = d_n$. Combining (S.3.185) and (S.3.186) gives

$$u^{\mathsf{T}}\check{R}\check{s}_{\|,n} - \theta_0 q_0 u^{\mathsf{T}}a \xrightarrow{L^2} 0.$$

To restore empirical centering, write

$$\bar{b} = \frac{1}{n}\sum_i b_i, \qquad \bar{c} = \frac{1}{n}\sum_i c_i, \qquad M_n = [\bar{b}, \bar{c}].$$

Then

$$C_n = \check{C}_n - M_n M_n^{\mathsf{T}}, \qquad s_{\|,n} = \check{s}_{\|,n} - M_n w\, \bar{y}_{\|}.$$

Let

$$\mathcal{R} = (C_n + \gamma I_p)^{-1}, \qquad K_n = M_n^{\mathsf{T}}\check{R}M_n, \qquad J_n = (I_2 - K_n)^{-1}.$$

The reverse Woodbury identity gives

$$J_n = I_2 + M_n^{\mathsf{T}}\mathcal{R}M_n,$$

and the fixed Gaussian moments give $\|M_n\|_{L^{16}} \le C$. Hence $\|J_n\|_{L^8} \le C$. For a deterministic bounded vector b,

$$g_n(b) := M_n^{\mathsf{T}}\check{R}b = \frac{1}{n}\sum_i C_i (F_i^{\circ})^{\mathsf{T}}\check{R}_i b.$$

Put

$$Z_i(b) = C_{*,n}(F_i^{\circ})^{\mathsf{T}}\check{R}_i b, \qquad Z_i^{(ij)}(b) = C_{*,n}(F_i^{\circ})^{\mathsf{T}}\check{R}_{ij} b.$$

Equation (S.3.182) and Hölder's inequality give

$$\left\| g_n(b) - \frac{1}{n}\sum_i Z_i(b) \right\|_{L^2} \le Cn^{-1/2}\log^C n.$$

Conditionally on $\mathcal{G}_{ij}$, $Z_i^{(ij)}(b)$ and $Z_j^{(ij)}(b)$ are independent and centered. Moreover,

$$Z_i(b) - Z_i^{(ij)}(b) = -\frac{1}{n} C_{*,n}(F_i^{\circ})^{\mathsf{T}}\check{R}_i F_j^{\circ}((F_j^{\circ})^{\mathsf{T}}\check{R}_{ij} b),$$

so

$$\max_{i\neq j} \|Z_i(b) - Z_i^{(ij)}(b)\|_{L^2} \le Cn^{-1/2}.$$

Expanding the pair covariance gives

$$\mathbb{E}\left\| \frac{1}{n}\sum_i Z_i(b) \right\|^2 \le \frac{C}{n} + Cn^{-1/2}.$$

Together with the quantitative replacement above, this yields

$$\mathbb{E}\|g_n(b)\|^2 \le Cn^{-1/2} = o(1) \tag{S.3.187}$$

The fixed Gaussian moments also give $\sup_n \|g_n(b)\|_{L^{16}} \le C$. Interpolation therefore shows that

$$\|g_n(b)\|_{L^q} = o(1), \qquad 2 \le q < 16,$$

for every deterministic bounded sequence $b = b_n$; the constants depend only on an upper bound for $\|b\|$. Put $h_{\|,n} = M_n^{\mathsf{T}}\check{R}\check{s}_{\|,n}$. If $\mathcal{X}_i = F_i^{\circ} w\, y_i^{\|}$, then the $\mathcal{X}_i$ are iid,

$$\|\mathbb{E}\mathcal{X}_i\| \le C, \qquad \mathbb{E}\|\mathcal{X}_i\|^2 \le Cp, \qquad \|\mathcal{X}_i\|_{L^{16}} \le C\sqrt{p}.$$

The Hilbert space Rosenthal inequality and $p/n = O(1)$ therefore give

$$\|\check{s}_{\|,n}\|_{L^{16}} = \left\| \frac{1}{n} \sum_i \mathcal{X}_i \right\|_{L^{16}} \le C.$$

Together with $\|M_n\|_{L^{16}} \le C$ and the ridge bound, this yields

$$\|h_{\|,n}\|_{L^8} \le \gamma_-^{-1} \|M_n\|_{L^{16}} \|\check{s}_{\|,n}\|_{L^{16}} \le C.$$

Woodbury's identity now yields the exact equality

$$\begin{aligned} u^{\mathrm{T}} \mathcal{R} s_{\|,n} - u^{\mathrm{T}} \check{R} \check{s}_{\|,n} \\ = g_n(u)^{\mathrm{T}} J_n h_{\|,n} - \bar{y}_{\|} g_n(u)^{\mathrm{T}} J_n w. \end{aligned}$$

Since $\|\bar{y}_{\|}\|_{L^8} = O(n^{-1/2})$, interpolation and (S.3.187) imply

$$\|g_n(u)^{\mathrm{T}} J_n h_{\|,n}\|_{L^2} \le \|g_n(u)\|_{L^6} \|J_n\|_{L^6} \|h_{\|,n}\|_{L^6} = o(1),$$

and we bound the term containing $\bar{y}_{\|}$ with the same Hölder inequality. Therefore

$$u^{\mathrm{T}} \mathcal{R} s_{\|,n} - \theta_0 q_0 u^{\mathrm{T}} a \xrightarrow{L^2} 0. \tag{S.3.188}$$

Similarly, for deterministic bounded b, d,

$$b^{\mathrm{T}} (\mathcal{R} - \check{R}) d = g_n(b)^{\mathrm{T}} J_n g_n(d) = o_{L^2}(1),$$

and hence

$$b^{\mathrm{T}} \mathcal{R} d - q_0 b^{\mathrm{T}} d \xrightarrow{L^2} 0. \tag{S.3.189}$$

To restore the rank-one empirical feature mean, put

$$A_z = \widehat{z}^{\mathrm{T}} \mathcal{R} \widehat{z}, \qquad b_z = \widehat{z}^{\mathrm{T}} \mathcal{R} s_{\|,n}.$$

With β_n as above, the endpoint Hoeffding decomposition gives

$$\widehat{z} = \beta_n \sqrt{p}\, e + \delta_n, \qquad \|\delta_n\|_{L^4} = O(1). \tag{S.3.190}$$

Equations (S.3.188), (S.3.189), and (S.3.190) imply

$$u^{\mathrm{T}} \mathcal{R} s_{\|,n} - \theta_0 q_0 u^{\mathrm{T}} a \xrightarrow{L^2} 0, \tag{S.3.191}$$

$$p^{-1/2} u^{\mathrm{T}} \mathcal{R} \widehat{z} - \beta_\sigma q_0 u^{\mathrm{T}} e \xrightarrow{L^2} 0, \tag{S.3.192}$$

$$p^{-1}(1 + A_z) - \beta_\sigma^2 q_0 \xrightarrow{L^2} 0, \tag{S.3.193}$$

$$p^{-1/2} b_z - \beta_\sigma \theta_0 q_0 e^{\mathrm{T}} a \xrightarrow{L^2} 0. \tag{S.3.194}$$

For instance,

$$p^{-1/2} \delta_n^{\mathrm{T}} \mathcal{R} s_{\|,n} = o_{L^2}(1)$$

by the ridge bound, $\|\delta_n\|_{L^4} = O(1)$, and $\|s_{\|,n}\|_{L^4} = O(1)$; all other terms containing $\delta_n/\sqrt{p}$ satisfy the same type of bound.

Sherman–Morrison gives the exact scalar identity

$$u^{\mathrm{T}} Q_* r_{*,\|} = u^{\mathrm{T}} \mathcal{R} s_{\|,n} + \frac{u^{\mathrm{T}} \mathcal{R} \widehat{z}}{1 + A_z} \{\bar{y}_{\|} - b_z\}. \tag{S.3.195}$$

On the real resolvent axis,

$$\frac{|u^{\mathrm{T}} \mathcal{R} \widehat{z}|}{1 + A_z} \le \{u^{\mathrm{T}} \mathcal{R} u\}^{1/2} \frac{A_z^{1/2}}{1 + A_z} \le C.$$

Thus the term containing $\bar{y}_{\|}$ is $O_{L^2}(n^{-1/2})$. Further,

$$\left| \frac{(u^{\mathrm{T}} \mathcal{R} \widehat{z}) b_z}{1 + A_z} \right| \le \{u^{\mathrm{T}} \mathcal{R} u\}^{1/2} \{s_{\|,n}^{\mathrm{T}} \mathcal{R} s_{\|,n}\}^{1/2},$$

so these ratios are uniformly integrable in L^2. Since $\beta_\sigma \neq 0$ and q_0 is bounded away from zero, (S.3.192)–(S.3.194) give

$$\frac{(u^{\mathrm{T}}\mathcal{R}\widehat{z})b_z}{1+A_z} = \frac{(p^{-1/2}u^{\mathrm{T}}\mathcal{R}\widehat{z})(p^{-1/2}b_z)}{p^{-1}(1+A_z)},$$

and hence

$$\frac{(u^{\mathrm{T}}\mathcal{R}\widehat{z})b_z}{1+A_z} - \theta_0 q_0 (u^{\mathrm{T}}e)(e^{\mathrm{T}}a) \xrightarrow{L^2} 0. \tag{S.3.196}$$

Combining (S.3.195), (S.3.191), and (S.3.196) proves

$$u^{\mathrm{T}}Q_* r_{*,\|} - \theta_0 q_0 u^{\mathrm{T}}P_\perp a \xrightarrow{L^2} 0.$$

Together with (S.3.180), this proves (S.3.166).

To express the result in the notation of the statement, use the fixed-point stability estimate

$$t = t_{*,n} + O(n^{-1}), \qquad s = O(n^{-1}).$$

Since $\bar{v} = m(t,s)$,

$$\bar{v} = t_{*,n}\,\mathrm{vec}(C_{*,n}V) + O(n^{-1}), \qquad c_B = t_{*,n}w^{\mathrm{T}}C_{*,n}Vw + O(n^{-1}) = w^{\mathrm{T}}(I_2 - C_{*,n})w + O(n^{-1}).$$

As $w^{\mathrm{T}}w = \alpha^2$,

$$d_\alpha(\alpha^2 - c_B) = \theta_0 + O(n^{-1}). \tag{S.3.197}$$

Moreover, the projector representation (S.3.155), the deterministic trace identity (S.3.157), and $t = t_{*,n} + O(n^{-1})$ imply

$$\|\bar{Q} - q_0 P_\perp\| = O(n^{-1}). \tag{S.3.198}$$

Indeed, with

$$\xi = \gamma + \nu_\alpha - \beta_\sigma^2 - \nu^{\mathrm{T}}\bar{v}, \qquad \delta = \beta_\sigma^2(1 - \mathbf{1}_4^{\mathrm{T}}\bar{v}),$$

the identity $\bar{Q}_I(0) = \bar{Q}$ and (S.3.47) give

$$\bar{Q} = \xi^{-1}P_\perp + (\xi + p\delta)^{-1}ee^{\mathrm{T}}, \qquad |(\xi + p\delta)^{-1}| \le Cp^{-1}.$$

Furthermore,

$$t\xi - c_n = O(n^{-1})$$

by (S.3.157); hence

$$\xi^{-1} = \frac{t}{c_n} + O(n^{-1}) = q_0 + O(n^{-1}).$$

Therefore

$$d_\alpha(\alpha^2 - c_B)u^{\mathrm{T}}\bar{Q}a - \theta_0 q_0 u^{\mathrm{T}}P_\perp a = O(n^{-1}).$$

Combining this identity with (S.3.166) proves (S.3.165). □

S.3.6.5 Quadratic-term reductions

Lemma S.3.14 Define

$$\phi_i = \frac{1}{n}\sum_{j=1}^n \{\lambda z_{ij} + (1-\lambda)z_{ji}\}, \qquad \widehat{z} = \frac{1}{n^2}\sum_{i,j=1}^n z_{ij}, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i.$$

Then

$$\widehat{z} = \frac{1}{n}\sum_{i=1}^n \phi_i, \qquad r = \frac{1}{n}\sum_{i=1}^n \phi_i y_i, \tag{S.3.199}$$

If $\Phi \in \mathbb{R}^{n\times p}$ has rows ϕ_i^{T}, then

$$\frac{1}{n}\Phi^{\mathrm{T}}\Phi \preceq S.$$

Proof. The coefficient of y_i in

$$\frac{1}{n^2}\sum_{a,b=1}^{n} z_{ab}\{\lambda y_a + (1-\lambda)y_b\}$$

is $n^{-1}\phi_i$, which proves the second identity in (S.3.199). Averaging ϕ_i over i and reindexing the second double sum gives the first identity.

For the matrix inequality, let $Z \in \mathbb{R}^{n^2\times p}$ have rows z_{ij}^{T}. The full-pair label-mixing matrix M from (S.5.1) satisfies

$$\Phi = \frac{1}{n}\mathsf{M}^{\mathrm{T}}Z.$$

The bound $\mathsf{M}\mathsf{M}^{\mathrm{T}} \preceq nI_{n^2}$ in that identity gives

$$\frac{1}{n}\Phi^{\mathrm{T}}\Phi = \frac{1}{n^3}Z^{\mathrm{T}}\mathsf{M}\mathsf{M}^{\mathrm{T}}Z \preceq \frac{1}{n^2}Z^{\mathrm{T}}Z = S.$$

□

Lemma S.3.15 Assume $\beta_\sigma \neq 0$ and (S.1.7). Let

$$H_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}},$$

and define the centered endpoint projections

$$u_i = f_1(g_i) - \beta_\sigma\mathbf{1}, \qquad v_i = f_2(g_i) - \beta_\sigma\mathbf{1}.$$

Let $U = [u_1,\dots,u_n]$, $V_n = [v_1,\dots,v_n]$, and put

$$\kappa(\lambda) = \nu_\alpha - \beta_\sigma^2 - \operatorname{tr} V,$$

$$C_n = \kappa(\lambda)I_p + \frac{1}{n}UH_nU^{\mathrm{T}} + \frac{1}{n}V_nH_nV_n^{\mathrm{T}}, \qquad s_n = \frac{1}{n}\{\lambda U + (1-\lambda)V_n\}H_ny.$$

For ω on a fixed sufficiently small circle Γ, write

$$Q_I(\omega) = (S + (\gamma+\omega)I_p)^{-1},$$

$$S_* = C_n + \widehat{zz}^{\mathrm{T}}, \qquad Q_*(\omega) = \{S_* + (\gamma+\omega)I_p\}^{-1}, \qquad r_* = \widehat{z}\bar{y} + s_n.$$

Then

$$\sup_{\omega\in\Gamma}\left\|r^{\mathrm{T}}Q_I(\omega)r - r_*^{\mathrm{T}}Q_*(\omega)r_*\right\|_{L^2} = o(1). \tag{S.3.200}$$

Moreover,

$$\left\|\mathbf{1}^{\mathrm{T}}Qr - \mathbf{1}^{\mathrm{T}}Q_*(0)r_*\right\|_{L^2} = O(n^{-1/2}\log^C n). \tag{S.3.201}$$

Proof. For $i \neq j$, write

$$z_{ij} = \beta_\sigma\mathbf{1} + u_i + v_j + \eta_{ij}, \qquad \eta_{ij} = z_{ij} - f_1(g_i) - f_2(g_j) + \beta_\sigma\mathbf{1}.$$

Then

$$\mathbb{E}[\eta_{ij} \mid g_i] = \mathbb{E}[\eta_{ij} \mid g_j] = 0, \qquad \mathbb{E}\eta_{ij}\eta_{ij}^{\mathrm{T}} = \kappa(\lambda)I_p.$$

The Gram-matrix estimates from the proof of Lemma S.3.13 remain valid because they do not involve the labels. Specifically, (S.3.171) controls the off-diagonal part of $\eta_{ij}\eta_{ij}^{\mathrm{T}}$, and (S.3.175) controls the endpoint–canonical cross terms. Together with the scalar diagonal estimate used in the derivation of (S.3.171), they give

$$S = C_n + \widehat{zz}^{\mathrm{T}} + d_n\mathbf{1}\mathbf{1}^{\mathrm{T}} + E_{S,n}, \qquad \|E_{S,n}\|_{L^4,\mathrm{op}} \le Cn^{-1/2}\log^C n, \tag{S.3.202}$$

where

$$d_n = \frac{n-1}{n^2}\{\mu_0(\sigma) - \beta_\sigma\}^2.$$

Indeed, if

$$\beta_n = \left(1 - \frac{1}{n}\right)\beta_\sigma + \frac{1}{n}\mu_0(\sigma),$$

then

$$\left(1-\frac{1}{n}\right)\beta_\sigma^2+\frac{1}{n}\mu_0(\sigma)^2-\beta_n^2=d_n.$$

The residual-covariance discrepancy and the remaining deterministic part of the diagonal rows have operator norm $O(n^{-1})$ and are included in $E_{S,n}$. Here

$$d_n=O(n^{-1}),\qquad \|d_n\mathbf{1}\mathbf{1}^{\mathrm{T}}\|=d_np=O(1),$$

so this rank-one term is not absorbed into $E_{S,n}$.

For the score, put

$$q_{ii}=\beta_\sigma\mathbf{1}+u_i+v_i.$$

The endpoint identity in Lemma S.3.14 gives the exact decomposition

$$\begin{aligned}r-r_*&=\frac{1}{n^2}\sum_{i\neq j}\eta_{ij}\{\widetilde{y}_{ij}-\bar{y}\}\\&\quad+\frac{1}{n^2}\sum_i(z_{ii}-q_{ii})(y_i-\bar{y}).\end{aligned}\tag{S.3.203}$$

For $y_i^{\|}=g_i^{\mathrm{T}}a$, the contribution of $y^{\|}$ is precisely the score error bounded in (S.3.179). For the residual labels, put

$$T_{\zeta,n}=\frac{1}{n^2}\sum_{i\neq j}\eta_{ij}\{\lambda\zeta_i+(1-\lambda)\zeta_j\},\qquad M_{\eta,n}=\frac{1}{n^2}\sum_{i\neq j}\eta_{ij}.$$

The kernel defining $T_{\zeta,n}$ is canonical in the independent endpoints (g_i,ζ_i). Indeed, if

$$H_{ij}:=\eta_{ij}\{\lambda\zeta_i+(1-\lambda)\zeta_j\},$$

then independence of ζ_k from the Gaussian array and $\mathbb{E}[\eta_{ij}\mid g_i]=\mathbb{E}[\eta_{ij}\mid g_j]=0$ give

$$\begin{aligned}&\mathbb{E}[H_{ij}\mid g_i,\zeta_i]=\lambda\zeta_i\mathbb{E}[\eta_{ij}\mid g_i]+(1-\lambda)\mathbb{E}[\eta_{ij}\zeta_j\mid g_i]=0,\\&\mathbb{E}[H_{ij}\mid g_j,\zeta_j]=0.\end{aligned}$$

Because the residuals need not be identically distributed, regroup the residual part as

$$T_{\zeta,n}=\sum_{i=1}^n\zeta_ib_i(G),\qquad b_i(G)=\frac{1}{n^2}\sum_{j\neq i}\{\lambda\eta_{ij}+(1-\lambda)\eta_{ji}\}.$$

Conditional on g_i, the summands defining b_i are independent and centered as j varies. Hilbert space Rosenthal, together with the fixed Gaussian moment bounds $\|\eta_{ij}\|_{L^2}=O(\sqrt{p})$ and $\|\eta_{ij}\|_{L^4}=O(\sqrt{p})$, therefore gives, uniformly in i,

$$\|b_i(G)\|_{L^4}\le\frac{C}{n^2}\{\sqrt{np}+n^{1/4}\sqrt{p}\}\le C\frac{\sqrt{p}}{n^{3/2}}.\tag{S.3.204}$$

Now condition on the entire Gaussian array. Since the ζ_i are independent, centered, and have uniformly bounded fourth moments, a second application of Hilbert space Rosenthal yields

$$\|T_{\zeta,n}\|_{L^4}\le C\left\{\left(\sum_i\|b_i\|_{L^4}^2\right)^{1/2}+\left(\sum_i\|b_i\|_{L^4}^4\right)^{1/4}\right\}\le C\frac{\sqrt{p}}{n}.$$

The canonical Gaussian endpoint estimate above, which does not involve the labels, also gives

$$\|M_{\eta,n}\|_{L^8}\le C\frac{\sqrt{p}}{n}.$$

Moreover, scalar Rosenthal for independent, not necessarily identically distributed variables and the uniform eighth-moment bound give

$$\|\bar{\zeta}\|_{L^8}\le C\left\{\left(\frac{1}{n^2}\sum_i\mathbb{E}\zeta_i^2\right)^{1/2}+\left(\frac{1}{n^8}\sum_i\mathbb{E}|\zeta_i|^8\right)^{1/8}\right\}\le Cn^{-1/2},\qquad\left\|\frac{1}{n^2}\sum_i(z_{ii}-q_{ii})(\zeta_i-\bar{\zeta})\right\|_{L^4}\le Cn^{-1}.$$

For the last bound, split off the term containing $\bar{\zeta}$, condition on the Gaussian array in the remaining sum, and apply the same Rosenthal inequality with coefficients $n^{-2}(z_{ii} - q_{ii})$; their fixed fourth moments are $O(p^2)$. The separated term is $O_{L^4}(\sqrt{p}\, n^{-3/2}) = O(n^{-1})$ because $p \asymp n$. Substitution in (S.3.203), together with (S.3.179), proves

$$r = r_* + e_{r,n}, \qquad \|e_{r,n}\|_{L^4} \le C n^{-1/2} \log^C n. \tag{S.3.205}$$

For the moment bounds used below, decompose

$$r_* = \widehat{z}\,\bar{y}_{\|} + \widehat{z}\,\bar{\zeta} + s_{n,\|} + s_{n,\zeta},$$

and put

$$D_{\mathrm{ep},n} = \lambda U + (1-\lambda) V_n, \qquad s_{n,\|} = \frac{1}{n} D_{\mathrm{ep},n} H_n y^{\|}, \qquad s_{n,\zeta} = \frac{1}{n} D_{\mathrm{ep},n} H_n \zeta.$$

With $\bar{u} = n^{-1}\sum_i u_i$ and $\bar{v} = n^{-1}\sum_i v_i$, the endpoint Hoeffding decomposition is

$$\widehat{z} = \beta_n \mathbf{1} + \bar{u} + \bar{v} + \varepsilon_{m,n}, \qquad \|\bar{u}\|_{L^{16}} + \|\bar{v}\|_{L^{16}} \le C,$$

$$\|\varepsilon_{m,n}\|_{L^{16}} \le C n^{-1/2} \log^C n.$$

The fixed-order vector Hoeffding bound for the canonical endpoint remainder gives the last estimate. The fixed-order Gaussian matrix and scalar moment bounds give

$$\|\widehat{z}\|_{L^{16}} \le C\sqrt{p}, \qquad \|\bar{y}_{\|}\|_{L^{16}} \le C n^{-1/2},$$

$$\big\| \|D_{\mathrm{ep},n}\|_{\mathrm{op}} \big\|_{L^{16}} \le C(\sqrt{n} + \sqrt{p}), \qquad \big\| \|y^{\|}\| \big\|_{L^{16}} \le C\sqrt{n}.$$

Indeed, the first bound follows by applying the endpoint Hoeffding decomposition at the fixed exponent sixteen, while the third is the standard operator norm estimate for a matrix with independent centered Lipschitz-Gaussian columns. Hence, since $p \asymp n$,

$$\|\widehat{z}\,\bar{y}_{\|}\|_{L^8} \le C, \qquad \|s_{n,\|}\|_{L^8} \le \frac{1}{n} \|\|D_{\mathrm{ep},n}\|_{\mathrm{op}}\|_{L^{16}} \|\|y^{\|}\|\|_{L^{16}} \le C.$$

Moreover, conditionally on W, $\widehat{z}$ and $\bar{\zeta}$ are independent, so

$$\|\widehat{z}\,\bar{\zeta}\|_{L^8} \le \|\widehat{z}\|_{L^8} \|\bar{\zeta}\|_{L^8} \le C.$$

Conditionally on the Gaussian sample, $s_{n,\zeta}$ is a sum of independent centered vectors. Set

$$c_i = \frac{1}{n} [D_{\mathrm{ep},n} H_n]_{\cdot i}.$$

Conditional Hilbert space Rosenthal gives

$$\{\mathbb{E}_\zeta \|s_{n,\zeta}\|^8\}^{1/8} \le C \left\{ \left(\sum_i \|c_i\|^2 \right)^{1/2} + \left(\sum_i \|c_i\|^8 \right)^{1/8} \right\} \le \frac{C}{n} \|D_{\mathrm{ep},n}\|_F.$$

The last quantity is bounded in L^8 by the fixed Gaussian moments, so $\|s_{n,\zeta}\|_{L^8} \le C$. This calculation uses only $\sup_i \mathbb{E}|\zeta_i|^8 < \infty$, and the vector Rosenthal inequality used here allows independent summands with different distributions. The ridge bound now proves

$$\|r_*\|_{L^8} + \sup_{\omega \in \Gamma} \|Q_*(\omega) r_*\|_{L^8} \le C. \tag{S.3.206}$$

Choose Γ inside the identity-perturbation disk of Lemma S.3.11, with $\mathrm{Re}(\gamma + \omega) \ge \gamma_-/2$. The resolvent identity gives the exact equality

$$\begin{aligned} r^{\mathrm{T}} Q_I(\omega) r - r_*^{\mathrm{T}} Q_*(\omega) r_* = {} & 2 e_{r,n}^{\mathrm{T}} Q_I(\omega) r_* + e_{r,n}^{\mathrm{T}} Q_I(\omega) e_{r,n} \\ & - r_*^{\mathrm{T}} Q_I(\omega) E_{S,n} Q_*(\omega) r_* \\ & - d_n \{ r_*^{\mathrm{T}} Q_I(\omega) \mathbf{1} \} \{ \mathbf{1}^{\mathrm{T}} Q_*(\omega) r_* \}. \end{aligned}$$

The first three terms are $O_{L^2}(n^{-1/2} \log^C n)$ by (S.3.202)–(S.3.206) and the ridge bound. For the last term, (S.3.59) and the ridge bound give

$$\|r_*^{\mathrm{T}} Q_I(\omega) \mathbf{1}\|_{L^4} \le C, \qquad \|\mathbf{1}^{\mathrm{T}} Q_*(\omega) r_*\|_{L^4} \le C\sqrt{p}.$$

Since $d_n = O(n^{-1})$ and $p \asymp n$, the last term is $O_{L^2}(n^{-1/2})$. This proves (S.3.200).

At $\omega = 0$, the corresponding linear identity is

$$\mathbf{1}^{\mathrm{T}} Q r - \mathbf{1}^{\mathrm{T}} Q_*(0) r_* = \mathbf{1}^{\mathrm{T}} Q e_{r,n} - \mathbf{1}^{\mathrm{T}} Q E_{S,n} Q_*(0) r_* - d_n (\mathbf{1}^{\mathrm{T}} Q \mathbf{1})(\mathbf{1}^{\mathrm{T}} Q_*(0) r_*).$$

Here (S.3.59) implies

$$\|Q\mathbf{1}\|_{L^8} + \|\mathbf{1}^{\mathrm{T}} Q \mathbf{1}\|_{L^4} \le C.$$

The first two terms are $O_{L^2}(n^{-1/2} \log^C n)$; the last one is $O_{L^2}(n^{-1/2})$ by the preceding ridge estimate. This proves (S.3.201). □

Lemma S.3.16 Assume $\beta_\sigma \neq 0$ and (S.1.7). Let

$$c_n = \frac{p}{n}, \qquad e = \frac{\mathbf{1}}{\sqrt{p}}, \qquad P_\perp = I_p - ee^{\mathrm{T}}, \qquad \overline{\sigma}^2_{\varepsilon,n} = \frac{1}{n}\sum_i \sigma^2_{n,i}, \qquad \rho^2 = \rho_n^2 = \|\beta^*\|^2 + \overline{\sigma}^2_{\varepsilon,n}.$$

For real ω near zero, let $t_\omega := t_{*,n}(\omega) > 0$ denote the unique solution of

$$c_n = t_\omega \left[\gamma + \omega + \kappa(\lambda) + \mathrm{tr}\{V(I_2 + t_\omega V)^{-1}\}\right]. \tag{S.3.207}$$

Let the same notation denote its analytic continuation to a fixed Cauchy circle Γ in an n-independent neighborhood of zero. Write $t_0 := t_\omega|_{\omega=0} = t_{*,n}$, and set

$$C_{2,\omega} = (I_2 + t_\omega V)^{-1}, \qquad q_\omega = \frac{t_\omega}{c_n},$$

$$c_{B,\omega} = w^{\mathrm{T}}(I_2 - C_{2,\omega})w, \qquad \theta_\omega = d_\alpha w^{\mathrm{T}} C_{2,\omega} w.$$

Here the subscript 0 in t_0 means $\omega = 0$ at fixed positive ridge; it does not denote the zero-ridge root in (S.1.8). Then

$$\sup_{\omega\in\Gamma} \left\| r^{\mathrm{T}} Q_I(\omega) r - \mathcal{E}_\omega \right\|_{L^2} = o(1), \tag{S.3.208}$$

where

$$\mathcal{E}_\omega = \theta_\omega^2 q_\omega \|P_\perp a\|^2 + \rho^2 c_{B,\omega}.$$

Moreover,

$$\mathbb{E}|\mathbf{1}^{\mathrm{T}} Q r|^2 = O(n^{-1}\log^C n). \tag{S.3.209}$$

Proof. For real $\gamma + \omega > 0$, the solution is unique because the derivative of the right-hand side of (S.3.207) with respect to t_ω is

$$\gamma + \omega + \kappa(\lambda) + \mathrm{tr}\{V(I_2 + t_\omega V)^{-2}\} > 0.$$

At $\omega = 0$ this derivative is at least γ_-. Since c_n remains in a compact subset of $(0, \infty)$, so does t_0. The analytic implicit function theorem gives an $\varepsilon > 0$, independent of n, such that

$$\inf_{\substack{|\omega|\le\varepsilon \\ n\ge n_0}} \left|\gamma + \omega + \kappa(\lambda) + \mathrm{tr}\{V(I_2 + t_\omega V)^{-2}\}\right| \ge \frac{\gamma_-}{2}.$$

Choose Γ inside this disk.

Put

$$F_i^\circ = [u_i, v_i], \qquad \check{C}_n = \kappa(\lambda) I_p + \frac{1}{n}\sum_i F_i^\circ (F_i^\circ)^{\mathrm{T}},$$

$$\check{s}_n = \frac{1}{n}\sum_i F_i^\circ w y_i, \qquad \check{R}_\omega = \{\check{C}_n + (\gamma + \omega) I_p\}^{-1}.$$

After deleting sample i, define

$$\check{R}_i = \left\{\kappa(\lambda) I_p + \frac{1}{n}\sum_{k\neq i} F_k^\circ (F_k^\circ)^{\mathrm{T}} + (\gamma + \omega) I_p\right\}^{-1},$$

$$T_i = \frac{1}{n}(F_i^\circ)^{\mathrm{T}}\check{R}_i F_i^\circ, \qquad C_i = (I_2 + T_i)^{-1}, \qquad \check{s}^{(i)} = \frac{1}{n}\sum_{k\neq i} F_k^\circ w y_k.$$

The pair $(\check{R}_i, \check{s}^{(i)})$ is independent of (F_i°, y_i). Woodbury gives

$$(F_i^\circ)^{\mathrm{T}}\check{R}_\omega = C_i(F_i^\circ)^{\mathrm{T}}\check{R}_i,$$

$$(F_i^\circ)^{\mathrm{T}}\check{R}_\omega \check{s}_n = C_i(F_i^\circ)^{\mathrm{T}}\check{R}_i\check{s}^{(i)} + (I_2 - C_i)wy_i.$$

Since $\mathrm{Re}(\gamma + \omega) > 0$,

$$\mathrm{Re}\,\check{R}_i \geq 0, \qquad \mathrm{Re}\,T_i = \frac{1}{n}(F_i^\circ)^{\mathrm{T}}\,\mathrm{Re}(\check{R}_i)F_i^\circ \geq 0.$$

Therefore

$$s_{\min}(I_2 + T_i) \geq 1, \qquad \|C_i\| \leq 1.$$

The analytic branch $C_{2,\omega}$ is also uniformly bounded on Γ. In particular,

$$\sup_i \sup_{\omega\in\Gamma} \|C_i\|_{L^q} + \sup_{\omega\in\Gamma} \|C_{2,\omega}\| \leq C_q \qquad (q < \infty). \tag{S.3.210}$$

Consequently,

$$\begin{aligned}\check{s}_n^{\mathrm{T}}\check{R}_\omega\check{s}_n = \frac{1}{n}\sum_i y_i w^{\mathrm{T}} C_i(F_i^\circ)^{\mathrm{T}}\check{R}_i\check{s}^{(i)} \\ + \frac{1}{n}\sum_i y_i^2 w^{\mathrm{T}}(I_2 - C_i)w.\end{aligned} \tag{S.3.211}$$

Put

$$\check{t}_\omega = \mathbb{E}\frac{1}{n}\,\mathrm{tr}\,\check{R}_\omega, \qquad \check{t}_{i,\omega} = \frac{1}{n}\,\mathrm{tr}\,\check{R}_i.$$

Conditionally on $\check{R}_i$, the rows of F_i° are independent centered two-dimensional vectors with covariance V. Let $\mathbb{E}_i^g$ integrate only g_i, conditionally on the remaining variables. Then

$$\left\{\mathbb{E}_i^g \|T_i - \check{t}_{i,\omega} V\|^8\right\}^{1/8} \leq \frac{C}{n}\{\mathrm{tr}(\check{R}_i\check{R}_i^*)\}^{1/2} \leq Cn^{-1/2}. \tag{S.3.212}$$

The rank-two resolvent identity gives

$$|\mathbb{E}\check{t}_{i,\omega} - \check{t}_\omega| \leq Cn^{-1}.$$

An explicit gradient calculation gives the sharper n^{-1} fluctuation rate for the normalized trace. If $\dot{F}_{k\ell}^\circ = \partial_{g_{k\ell}} F_k^\circ$, then, for $k \neq i$,

$$\partial_{g_{k\ell}}\check{t}_{i,\omega} = -\frac{2}{n^2}\,\mathrm{tr}\{(F_k^\circ)^{\mathrm{T}}\check{R}_i^2\dot{F}_{k\ell}^\circ\},$$

whereas the derivative is zero for $k = i$. Since $\dot{F}_{k\ell}^\circ$ has only its ℓ-th row nonzero,

$$\sum_{k,\ell} |\partial_{g_{k\ell}}\check{t}_{i,\omega}|^2 \leq \frac{C}{n^4}\sum_{k\neq i}\mathrm{tr}\{(F_k^\circ)^{\mathrm{T}}\check{R}_i^2\check{R}_i^{*2}F_k^\circ\}.$$

To bound the resulting sum, put

$$B_i = \frac{1}{n}\sum_{k\neq i} F_k^\circ(F_k^\circ)^{\mathrm{T}}.$$

Since $\check{R}_i = \{B_i + (\kappa(\lambda) + \gamma + \omega)I_p\}^{-1}$,

$$\begin{aligned}&\sum_{k\neq i}\mathrm{tr}\{(F_k^\circ)^{\mathrm{T}}\check{R}_i^2\check{R}_i^{*2}F_k^\circ\} \\ &\quad = n\,\mathrm{tr}\{B_i\check{R}_i^2\check{R}_i^{*2}\} = n\sum_{\ell=1}^p \frac{\lambda_\ell(B_i)}{|\lambda_\ell(B_i) + \kappa(\lambda) + \gamma + \omega|^4} \leq Cnp = O(n^2).\end{aligned}$$

Thus the gradient square is at most Cn^{-2}. Gaussian L^q-Poincaré therefore yields, for every fixed $q < \infty$,

$$\|\check{t}_{i,\omega} - \mathbb{E}\check{t}_{i,\omega}\|_{L^q} \leq C_q n^{-1}. \tag{S.3.213}$$

Taking the normalized trace in $\check{R}_\omega\{\check{C}_n + (\gamma+\omega)I_p\} = I_p$ gives the exact identity

$$c_n = (\gamma+\omega+\kappa(\lambda))\frac{1}{n}\operatorname{tr}\check{R}_\omega + \frac{1}{n}\sum_i \operatorname{tr}(T_i C_i).$$

After taking expectation, (S.3.212)–(S.3.213), (S.3.210), and the resolvent identity

$$T(I_2+T)^{-1} - T'(I_2+T')^{-1} = (I_2+T)^{-1}(T-T')(I_2+T')^{-1}$$

give

$$c_n = (\gamma+\omega+\kappa(\lambda))\check{t}_\omega + \operatorname{tr}\Big[\check{t}_\omega V(I_2+\check{t}_\omega V)^{-1}\Big] + O(n^{-1/2}).$$

For comparison with the complex analytic root, define

$$\Phi_{n,\omega}(z) = z\Big[\gamma+\omega+\kappa(\lambda)+\operatorname{tr}\{V(I_2+zV)^{-1}\}\Big] - c_n.$$

At $\omega = 0$, both $\check{t}_0$ and t_0 are real and the monotonicity bound $\partial_z\Phi_{n,0}(z) \geq \gamma_-$ on the positive real axis gives $|\check{t}_0 - t_0| \leq Cn^{-1/2}$. The resolvent identity and the analytic implicit function theorem give, uniformly in n,

$$|\check{t}_\omega - \check{t}_0| \leq C|\omega|, \qquad |t_\omega - t_0| \leq C|\omega|.$$

After shrinking the disk, both quantities lie in the same convex implicit-function neighborhood, on which

$$\operatorname{Re}\partial_z\Phi_{n,\omega}(z) \geq \frac{\gamma_-}{4}.$$

Since $\Phi_{n,\omega}(t_\omega) = 0$, the line-segment identity

$$\Phi_{n,\omega}(\check{t}_\omega) = (\check{t}_\omega - t_\omega)\int_0^1 \partial_z\Phi_{n,\omega}\{t_\omega + s(\check{t}_\omega - t_\omega)\}\,ds$$

and the preceding approximate fixed-point equation give

$$|\check{t}_\omega - t_\omega| \leq Cn^{-1/2}.$$

Since

$$C_i - C_{2,\omega} = -C_i(T_i - t_\omega V)C_{2,\omega},$$

we obtain, uniformly on Γ,

$$\left\{\frac{1}{n}\sum_i \|T_i - t_\omega V\|_{L^8}^8\right\}^{1/8} + \left\{\frac{1}{n}\sum_i \|C_i - C_{2,\omega}\|_{L^8}^8\right\}^{1/8} \leq Cn^{-1/2}. \tag{S.3.214}$$

For the anisotropic resolvent estimate, let b, d be deterministic vectors of bounded norm and put $\vartheta_\omega = \operatorname{tr}(VC_{2,\omega})$. Taking the (b,d)-bilinear form in the resolvent equation gives

$$b^{\mathrm{T}}d = (\gamma+\omega+\kappa(\lambda))b^{\mathrm{T}}\check{R}_\omega d + \frac{1}{n}\sum_i b^{\mathrm{T}}\check{R}_i F_i^\circ C_i (F_i^\circ)^{\mathrm{T}} d. \tag{S.3.215}$$

Define

$$Y_i(b,d) = b^{\mathrm{T}}\check{R}_i F_i^\circ C_{2,\omega}(F_i^\circ)^{\mathrm{T}}d - \vartheta_\omega b^{\mathrm{T}}\check{R}_i d.$$

Then $\mathbb{E}[Y_i(b,d) \mid (g_k)_{k\neq i}, W] = 0$. For $i \neq j$, let $\check{R}_{ij}$ delete both endpoints, put

$$\mathcal{G}_{ij} = \sigma\{W, g_k : k \notin \{i,j\}\},$$

and let $Y_i^{(ij)}$ be obtained by replacing $\check{R}_i$ with $\check{R}_{ij}$. Conditionally on $\mathcal{G}_{ij}$, $Y_i^{(ij)}$ and $Y_j^{(ij)}$ are independent and centered. The exact return is

$$\begin{aligned} Y_i - Y_i^{(ij)} = &-\frac{1}{n}(b^{\mathrm{T}}\check{R}_i F_j^\circ)((F_j^\circ)^{\mathrm{T}}\check{R}_{ij}F_i^\circ)C_{2,\omega}((F_i^\circ)^{\mathrm{T}}d) \\ &+ \frac{\vartheta_\omega}{n}(b^{\mathrm{T}}\check{R}_i F_j^\circ)((F_j^\circ)^{\mathrm{T}}\check{R}_{ij}d). \end{aligned} \tag{S.3.216}$$

To bound the factor containing $\check{R}_i F_j^\circ$, apply Woodbury's identity once more:

$$\check{R}_i F_j^\circ = \check{R}_{ij}F_j^\circ C_{j|i}, \qquad C_{j|i} = \left(I_2 + \frac{1}{n}(F_j^\circ)^{\mathrm{T}}\check{R}_{ij}F_j^\circ\right)^{-1}.$$

The positive Hermitian part of $\check{R}_{ij}$ gives $\|C_{j|i}\| \le 1$. Conditionally on $\mathcal{G}_{ij}$, $\check{R}_{ij}$ is fixed and the two feature blocks are independent. The Gaussian moment bounds therefore give, for every fixed $q \ge 2$,

$$\|b^{\mathrm{T}}\check{R}_{ij}F_j^{\circ}\|_{L^q} + \|(F_j^{\circ})^{\mathrm{T}}\check{R}_{ij}d\|_{L^q} + \|(F_i^{\circ})^{\mathrm{T}}d\|_{L^q} \le C_q,$$
$$\|(F_j^{\circ})^{\mathrm{T}}\check{R}_{ij}F_i^{\circ}\|_{L^q} \le C_q\sqrt{p}.$$

We use the Euclidean norm for vectors and the operator norm for matrices. Conditioning on the variables remaining after deleting i also gives a uniform L^8 bound for Y_i. Apply Hölder's inequality to (S.3.216), with three L^6 factors in the first term and two L^4 factors in the second. Uniformly for $\omega \in \Gamma$, this gives

$$\sup_i \|Y_i\|_{L^8} \le C, \qquad \max_{i\ne j} \|Y_i - Y_i^{(ij)}\|_{L^2} \le \frac{C(\sqrt{p}+1)}{n} \le Cn^{-1/2}.$$

Write $\Delta Y_i = Y_i - Y_i^{(ij)}$. Conditional independence and centering give

$$\mathbb{E}\left[Y_i^{(ij)}\overline{Y_j^{(ij)}} \mid \mathcal{G}_{ij}\right] = 0.$$

Consequently,

$$|\mathbb{E}[Y_i\overline{Y_j}]| \le \|\Delta Y_i\|_{L^2}\|Y_j\|_{L^2} + \|Y_i^{(ij)}\|_{L^2}\|\Delta Y_j\|_{L^2}$$
$$\le Cn^{-1/2}.$$

The diagonal terms are bounded by C, and Minkowski's inequality gives $\|n^{-1}\sum_i Y_i(b,d)\|_{L^8} \le C$. Thus the covariance expansion and interpolation between L^2 and L^8 yield

$$\mathbb{E}\left|\frac{1}{n}\sum_i Y_i(b,d)\right|^2 \le \frac{C}{n} + Cn^{-1/2} = o(1), \qquad \sup_{\omega\in\Gamma}\left\|\frac{1}{n}\sum_i Y_i(b,d)\right\|_{L^4} = o(1). \tag{S.3.217}$$

Replacing C_i by $C_{2,\omega}$ in (S.3.215) gives the explicit error bound

$$\left\|\frac{1}{n}\sum_i b^{\mathrm{T}}\check{R}_iF_i^{\circ}(C_i - C_{2,\omega})(F_i^{\circ})^{\mathrm{T}}d\right\|_{L^4}$$
$$\le \frac{1}{n}\sum_i \|C_i - C_{2,\omega}\|_{L^8(\mathrm{op})}\|b^{\mathrm{T}}\check{R}_iF_i^{\circ}\|_{L^{16}(\ell_2)}\|(F_i^{\circ})^{\mathrm{T}}d\|_{L^{16}(\ell_2)}$$
$$\le C\left\{\frac{1}{n}\sum_i \|C_i - C_{2,\omega}\|_{L^8}^8\right\}^{1/8} \le Cn^{-1/2}.$$

The first inequality follows from the operator norm bound and Hölder's inequality with $1/4 = 1/8 + 1/16 + 1/16$. Since $\check{R}_i$ is independent of F_i° and has uniformly bounded operator norm, conditioning on the leave-one variables gives a constant bound for both projected-feature L^{16} norms. These moments involve only the Gaussian design. The last line follows from the power-mean inequality and (S.3.214). All constants are uniform for $\omega \in \Gamma$. The rank-two identity also gives

$$\sup_i \|b^{\mathrm{T}}(\check{R}_i - \check{R}_\omega)d\|_{L^4} \le Cn^{-1}.$$

Substitution of these estimates and (S.3.214) into (S.3.215) gives

$$b^{\mathrm{T}}d = \{\gamma + \omega + \kappa(\lambda) + \vartheta_\omega\}b^{\mathrm{T}}\check{R}_\omega d + o_{L^4}(1).$$

The reduced fixed-point equation identifies the coefficient with $1/q_\omega$. Hence, uniformly on Γ,

$$b^{\mathrm{T}}\check{R}_\omega d - q_\omega b^{\mathrm{T}}d \xrightarrow{L^4} 0. \tag{S.3.218}$$

The score contraction holds for every deterministic bounded vector u. Woodbury gives

$$u^{\mathrm{T}}\check{R}_\omega \check{s}_n = \frac{1}{n}\sum_i y_i w^{\mathrm{T}}C_i(F_i^{\circ})^{\mathrm{T}}\check{R}_i u.$$

Gaussian integration by parts and the independence of ζ_i give

$$\mathbb{E}[y_i F_i^\circ] = d_\alpha a w^\top.$$

This covariance is the same for every i, so we center each leave-one variable at this common value. Moreover, Hölder's inequality and (S.3.214) give

$$\left\| \frac{1}{n} \sum_i y_i w^\top (C_i - C_{2,\omega})(F_i^\circ)^\top \check{R}_i u \right\|_{L^2} \le \frac{1}{n} \sum_i \|y_i\|_{L^8} \|C_i - C_{2,\omega}\|_{L^8} \|(F_i^\circ)^\top \check{R}_i u\|_{L^4} \le C n^{-1/2}.$$

After replacing C_i by $C_{2,\omega}$, define

$$X_i(u) = y_i w^\top C_{2,\omega} (F_i^\circ)^\top \check{R}_i u - \theta_\omega a^\top \check{R}_i u.$$

The variable $X_i(u)$ is conditionally centered after deleting i. With the same common leave-two resolvent,

$$X_i(u) - X_i^{(ij)}(u) = -\frac{1}{n} \left\{ y_i w^\top C_{2,\omega} (F_i^\circ)^\top \check{R}_i F_j^\circ - \theta_\omega a^\top \check{R}_i F_j^\circ \right\} \times (F_j^\circ)^\top \check{R}_{ij} u.$$

Conditional fourth-moment bounds give

$$\sup_i \|X_i(u)\|_{L^4} \le C, \qquad \max_{i \ne j} \|X_i(u) - X_i^{(ij)}(u)\|_{L^2} \le C n^{-1/2}.$$

The covariance expansion from (S.3.217) follows from the common leave-two argument. The rank-two replacement

$$\left\| \frac{1}{n} \sum_i a^\top \check{R}_i u - a^\top \check{R}_\omega u \right\|_{L^2} \le C n^{-1},$$

follows directly from the rank-two resolvent identity. For the covariance expansion, set $\Delta X_i = X_i - X_i^{(ij)}$. Conditional centering of the common leave-two variables gives

$$|\mathbb{E}[X_i \overline{X_j}]| \le \|\Delta X_i\|_{L^2} \|X_j\|_{L^2} + \|X_i^{(ij)}\|_{L^2} \|\Delta X_j\|_{L^2} \le C n^{-1/2}.$$

Thus

$$\mathbb{E} \left| \frac{1}{n} \sum_i X_i(u) \right|^2 \le \frac{C}{n} + C n^{-1/2} = o(1),$$

and consequently

$$u^\top \check{R}_\omega \check{s}_n - \theta_\omega q_\omega u^\top a \xrightarrow{L^2} 0. \tag{S.3.219}$$

To evaluate the first average in (S.3.211), define

$$Z_i = \{ y_i w^\top C_{2,\omega} (F_i^\circ)^\top - \theta_\omega a^\top \} \check{R}_i \check{s}^{(i)}.$$

The corresponding replacement error satisfies

$$\left\| \frac{1}{n} \sum_i y_i w^\top (C_i - C_{2,\omega})(F_i^\circ)^\top \check{R}_i \check{s}^{(i)} \right\|_{L^2} \le \frac{1}{n} \sum_i \|y_i\|_{L^8} \|C_i - C_{2,\omega}\|_{L^8} \|(F_i^\circ)^\top \check{R}_i \check{s}^{(i)}\|_{L^4} \le C n^{-1/2}. \tag{S.3.220}$$

Let $\check{s}^{(ij)}$ delete i, j, and define $Z_i^{(ij)}$ with $\check{R}_{ij}, \check{s}^{(ij)}$. Put

$$\mathcal{H}_{ij} = \sigma\{W, (g_k, \zeta_k) : k \notin \{i, j\}\}.$$

Conditionally on $\mathcal{H}_{ij}$, $Z_i^{(ij)}$ and $Z_j^{(ij)}$ are independent and separately centered. The identities

$$\check{s}^{(i)} = \check{s}^{(ij)} + \frac{1}{n} F_j^\circ w y_j, \qquad \check{R}_i - \check{R}_{ij} = -\frac{1}{n} \check{R}_i F_j^\circ (F_j^\circ)^\top \check{R}_{ij}$$

give

$$Z_i - Z_i^{(ij)} = \frac{1}{n}\{y_i w^{\mathsf{T}} C_{2,\omega}(F_i^\circ)^{\mathsf{T}} - \theta_\omega a^{\mathsf{T}}\}\check{R}_i F_j^\circ\{w y_j - (F_j^\circ)^{\mathsf{T}}\check{R}_{ij}\check{s}^{(ij)}\}.$$

Thus

$$\sup_i \|Z_i\|_{L^4} \le C, \qquad \max_{i\neq j}\|Z_i - Z_i^{(ij)}\|_{L^2} \le Cn^{-1/2},$$

Writing $\Delta Z_i = Z_i - Z_i^{(ij)}$, conditional centering with respect to $\mathcal{H}_{ij}$ gives

$$|\mathbb{E}[Z_i\overline{Z_j}]| \le \|\Delta Z_i\|_{L^2}\|Z_j\|_{L^2} + \|Z_i^{(ij)}\|_{L^2}\|\Delta Z_j\|_{L^2} \le Cn^{-1/2}.$$

Hence

$$\mathbb{E}\left|\frac{1}{n}\sum_i Z_i\right|^2 \le \frac{C}{n} + Cn^{-1/2} = o(1).$$

Moreover,

$$\check{R}_\omega\check{s}_n - \check{R}_i\check{s}^{(i)} = \frac{1}{n}\check{R}_i F_i^\circ C_i\{w y_i - (F_i^\circ)^{\mathsf{T}}\check{R}_i\check{s}^{(i)}\}. \tag{S.3.221}$$

Equations (S.3.219) and (S.3.221) identify the coefficient explicitly. Indeed, (S.3.220) and the definition of Z_i give

$$\begin{aligned}&\frac{1}{n}\sum_i y_i w^{\mathsf{T}} C_i(F_i^\circ)^{\mathsf{T}}\check{R}_i\check{s}^{(i)}\\ &\quad= \frac{1}{n}\sum_i Z_i + \theta_\omega\frac{1}{n}\sum_i a^{\mathsf{T}}\check{R}_i\check{s}^{(i)} + o_{L^2}(1).\end{aligned} \tag{S.3.222}$$

Taking the a-bilinear form in (S.3.221), averaging over i, and applying the fixed-order cavity moment bounds gives

$$\begin{aligned}&\left\|\frac{1}{n}\sum_i a^{\mathsf{T}}(\check{R}_\omega\check{s}_n - \check{R}_i\check{s}^{(i)})\right\|_{L^2}\\ &\quad\le \frac{1}{n^2}\sum_i \|a^{\mathsf{T}}\check{R}_i F_i^\circ\|_{L^8}\|C_i\|_{L^8}\|w y_i - (F_i^\circ)^{\mathsf{T}}\check{R}_i\check{s}^{(i)}\|_{L^4} \le \frac{C}{n}.\end{aligned}$$

Thus

$$\left\|\frac{1}{n}\sum_i a^{\mathsf{T}}\check{R}_i\check{s}^{(i)} - a^{\mathsf{T}}\check{R}_\omega\check{s}_n\right\|_{L^2} \le Cn^{-1}. \tag{S.3.223}$$

The first term on the right-hand side of (S.3.222) is $o_{L^2}(1)$ by the covariance estimate above, while (S.3.219) with $u = a$ and (S.3.223) show that the second one converges to $\theta_\omega^2 q_\omega\|a\|^2$. Therefore

$$\frac{1}{n}\sum_i y_i w^{\mathsf{T}} C_i(F_i^\circ)^{\mathsf{T}}\check{R}_i\check{s}^{(i)} - \theta_\omega^2 q_\omega\|a\|^2 \xrightarrow{L^2} 0.$$

For the second average, (S.3.214) and Hölder's inequality give

$$\left\|\frac{1}{n}\sum_i y_i^2 w^{\mathsf{T}}(C_i - C_{2,\omega})w\right\|_{L^2} \le \frac{1}{n}\sum_i \|y_i\|_{L^8}^2\|C_i - C_{2,\omega}\|_{L^4} \le Cn^{-1/2}.$$

The following is the only part of the label-energy contraction that can depend on a deterministic variance profile. Lemma S.3.23, with

$$C_{i,\omega} = C_i, \qquad C_\omega = C_{2,\omega}, \qquad \mathcal{F}_n = \sigma(W, G), \qquad \upsilon_{i,n} = \varepsilon_i, \qquad d_{i,n} = \sigma_{n,i}^2,$$

applies by (S.3.210) and (S.3.214), with $r_n = Cn^{-1/2}$. It gives

$$\sup_{\omega\in\Gamma}\left\|\frac{1}{n}\sum_i \varepsilon_i^2 w^{\mathsf{T}}(I_2 - C_i)w - \overline{\sigma}_{\varepsilon,n}^2 c_{B,\omega}\right\|_{L^2} \le Cn^{-1/2}.$$

Thus the bounded variance profile contributes only through its empirical mean.

The squares of the full labels also satisfy the corresponding triangular-array law of large numbers. The y_i are independent and, by the uniform eighth-moment assumption, have uniformly bounded fourth moments. Moreover,

$$\mathbb{E}y_i^2 = \|\beta^*\|^2 + \sigma_{n,i}^2.$$

Consequently,

$$\left\| \frac{1}{n}\sum_i y_i^2 - \rho^2 \right\|_{L^2}^2 = \frac{1}{n^2}\sum_i \mathrm{Var}(y_i^2) \le \frac{C}{n}, \qquad \rho^2 = \|\beta^*\|^2 + \overline{\sigma}_{\varepsilon,n}^2.$$

Combining this estimate with the preceding C_i-replacement yields

$$\frac{1}{n}\sum_i y_i^2 w^{\mathrm{T}}(I_2 - C_i)w - \rho^2 c_{B,\omega} \xrightarrow{L^2} 0.$$

Substitution in (S.3.211) proves

$$\check{s}_n^{\mathrm{T}}\check{R}_\omega\check{s}_n - \{\theta_\omega^2 q_\omega\|a\|^2 + \rho^2 c_{B,\omega}\} \xrightarrow{L^2} 0. \tag{S.3.224}$$

For the centered quantities, put

$$\bar{u} = \frac{1}{n}\sum_i u_i, \qquad \bar{v} = \frac{1}{n}\sum_i v_i, \qquad M_n = [\bar{u}, \bar{v}].$$

Then

$$C_n = \check{C}_n - M_n M_n^{\mathrm{T}}, \qquad s_n = \check{s}_n - M_n w\bar{y}.$$

Define

$$\mathcal{R}_\omega = \{C_n + (\gamma+\omega)I_p\}^{-1}, \qquad \mathcal{R}_0 = \mathcal{R}_\omega\big|_{\omega=0},$$
$$K_n = M_n^{\mathrm{T}}\check{R}_\omega M_n, \qquad J_n = (I_2 - K_n)^{-1}.$$

The reverse Woodbury identities are

$$\mathcal{R}_\omega = \check{R}_\omega + \check{R}_\omega M_n J_n M_n^{\mathrm{T}}\check{R}_\omega, \qquad J_n = I_2 + M_n^{\mathrm{T}}\mathcal{R}_\omega M_n. \tag{S.3.225}$$

Since $\|M_n\|$ has bounded moments of every fixed order, the ridge bound implies the same for J_n. Set

$$g_n = M_n^{\mathrm{T}}\check{R}_\omega e, \qquad h_n = M_n^{\mathrm{T}}\check{R}_\omega\check{s}_n, \qquad \Delta_i = C_i - C_{2,\omega}.$$

The leave-one identities give the exact decompositions

$$g_n = C_{2,\omega}G_e + \frac{1}{n}\sum_i \Delta_i (F_i^\circ)^{\mathrm{T}}\check{R}_i e, \tag{S.3.226}$$

$$h_n = C_{2,\omega}G_s + (I_2 - C_{2,\omega})w\bar{y} + \frac{1}{n}\sum_i \Delta_i (F_i^\circ)^{\mathrm{T}}\check{R}_i\check{s}^{(i)} - \frac{1}{n}\sum_i \Delta_i w y_i, \tag{S.3.227}$$

where

$$G_e = \frac{1}{n}\sum_i (F_i^\circ)^{\mathrm{T}}\check{R}_i e, \qquad G_s = \frac{1}{n}\sum_i (F_i^\circ)^{\mathrm{T}}\check{R}_i\check{s}^{(i)}.$$

With $\mathbb{E}_i^g$ as above, if b_i is measurable after endpoint i has been deleted, then

$$\left\{\mathbb{E}_i^g \|\Delta_i (F_i^\circ)^{\mathrm{T}} b_i\|^8\right\}^{1/8} \le C\{n^{-1/2} + |\check{t}_{i,\omega} - t_\omega|\}\|b_i\|. \tag{S.3.228}$$

Indeed, use

$$T_i - t_\omega V = \{T_i - \check{t}_{i,\omega}V\} + (\check{t}_{i,\omega} - t_\omega)V$$

and $\Delta_i = -C_i(T_i - t_\omega V)C_{2,\omega}$; the first term is a centered quadratic form multiplied by a linear form in F_i°, and its conditional eighth moment is at most $Cn^{-4}\|b_i\|^8$. For $b_i = \check{R}_i e$, integration of (S.3.228) and (S.3.213) gives an $O(n^{-1/2})$ bound. For $b_i = \check{R}_i\check{s}^{(i)}$, split $\check{s}^{(i)}$ into its Gaussian signal part and its part linear in $(\zeta_k)_{k\neq i}$. Conditionally on the Gaussian sample, Rosenthal's inequality for independent

triangular arrays controls the latter part using only $\sup_i \mathbb{E}|\zeta_i|^8 < \infty$, and then integration of the Gaussian coefficients gives the same bound. This Hölder step uses the fixed-order L^{16} version of (S.3.213). The additional moment is needed only for the Gaussian trace. Finally, conditionally on all variables except (g_i, ζ_i),

$$\mathbb{E}_{g_i,\zeta_i}\{\|\Delta_i\|^8|y_i|^8\} \le C\mathbb{E}_{g_i}\{\|\Delta_i\|^8|g_i^{\mathrm{T}}a|^8\} + C\mathbb{E}_{g_i}\|\Delta_i\|^8\,\mathbb{E}|\zeta_i|^8,$$

so

$$\sup_i \|\Delta_i w y_i\|_{L^8} \le Cn^{-1/2}. \tag{S.3.229}$$

Write

$$\check{s}_n = \check{s}_{g,n} + \check{s}_{\zeta,n}, \qquad h_n = h_{g,n} + h_{\zeta,n},$$

according to $y_i = g_i^{\mathrm{T}}a + \zeta_i$. The quantities g_n and $h_{g,n}$ are functions only of the Gaussian array. Put

$$B_n = \frac{1}{n}\sum_i F_i^\circ (F_i^\circ)^{\mathrm{T}}, \qquad y_{k,g} = g_k^{\mathrm{T}}a, \qquad \dot{F}^\circ_{k\ell} = \partial_{g_{k\ell}} F_k^\circ.$$

The matrix $\dot{F}^\circ_{k\ell}$ has only its ℓ-th row nonzero, and that row is uniformly bounded. Since

$$\partial\check{R}_\omega = -\check{R}_\omega(\partial\check{C}_n)\check{R}_\omega, \qquad \partial\check{C}_n = \frac{1}{n}\{(\partial F_k^\circ)(F_k^\circ)^{\mathrm{T}} + F_k^\circ(\partial F_k^\circ)^{\mathrm{T}}\},$$

and

$$\partial_{g_{k\ell}} M_n = \frac{1}{n}\dot{F}^\circ_{k\ell}, \qquad \partial_{g_{k\ell}}\check{s}_{g,n} = \frac{1}{n}\{\dot{F}^\circ_{k\ell} w\, y_{k,g} + F_k^\circ w\, a_\ell\},$$

direct differentiation gives

$$\partial_{g_{k\ell}} g_n = \frac{1}{n}(\dot{F}^\circ_{k\ell})^{\mathrm{T}}\check{R}_\omega e - \frac{1}{n}M_n^{\mathrm{T}}\check{R}_\omega\{\dot{F}^\circ_{k\ell}(F_k^\circ)^{\mathrm{T}} + F_k^\circ(\dot{F}^\circ_{k\ell})^{\mathrm{T}}\}\check{R}_\omega e, \tag{S.3.230}$$

$$\begin{aligned}\partial_{g_{k\ell}} h_{g,n} &= \frac{1}{n}(\dot{F}^\circ_{k\ell})^{\mathrm{T}}\check{R}_\omega\check{s}_{g,n} - \frac{1}{n}M_n^{\mathrm{T}}\check{R}_\omega\{\dot{F}^\circ_{k\ell}(F_k^\circ)^{\mathrm{T}} + F_k^\circ(\dot{F}^\circ_{k\ell})^{\mathrm{T}}\}\check{R}_\omega\check{s}_{g,n} \\ &\quad + \frac{1}{n}M_n^{\mathrm{T}}\check{R}_\omega\{\dot{F}^\circ_{k\ell} w\, y_{k,g} + F_k^\circ w\, a_\ell\}.\end{aligned} \tag{S.3.231}$$

For every vector b, the spectral calculus for $\check{R}_\omega = \{B_n + (\kappa(\lambda) + \gamma + \omega)I_p\}^{-1}$ gives

$$\sum_k \|(F_k^\circ)^{\mathrm{T}}\check{R}_\omega b\|^2 = nb^*\check{R}_\omega^* B_n\check{R}_\omega b \le Cn\|b\|^2,$$

$$\sum_k \|M_n^{\mathrm{T}}\check{R}_\omega F_k^\circ\|_F^2 = n\,\mathrm{tr}\{M_n^{\mathrm{T}}\check{R}_\omega B_n\check{R}_\omega^* M_n\} \le Cn\|M_n\|_F^2.$$

Using these two contractions term by term in (S.3.230)–(S.3.231) yields the pathwise bound

$$\begin{aligned}&\sum_{k,\ell}\left(\|\partial_{g_{k\ell}} g_n\|^2 + \|\partial_{g_{k\ell}} h_{g,n}\|^2\right) \\ &\qquad \le \frac{C}{n}(1 + \|M_n\|_F^2)\left\{1 + \|\check{s}_{g,n}\|^2 + \frac{1}{n}\sum_k y_{k,g}^2\right\}.\end{aligned}$$

The last term in (S.3.231) uses $\sum_\ell a_\ell^2 = \|a\|^2 = O(1)$. The quantities in the last display have bounded moments of every fixed order. Consequently,

$$\left\|\sum_{k,\ell}\left(\|\partial_{g_{k\ell}} g_n\|^2 + \|\partial_{g_{k\ell}} h_{g,n}\|^2\right)\right\|_{L^{q/2}} \le \frac{C_q}{n}, \qquad 2 \le q \le 16.$$

For complex ω, apply the same estimate to the real and imaginary parts.

Gaussian L^q-Poincaré, for $2 \le q \le 16$, now gives

$$\|g_n - \mathbb{E}g_n\|_{L^q} + \|h_{g,n} - \mathbb{E}h_{g,n}\|_{L^q} \le C_q n^{-1/2}.$$

In the signal versions of (S.3.226)–(S.3.227), conditional centering gives $\mathbb{E}G_e = \mathbb{E}G_{s,g} = 0$ and $\mathbb{E}\bar{y}_g = 0$. Equations (S.3.228) and (S.3.229) therefore imply

$$\|\mathbb{E}g_n\| + \|\mathbb{E}h_{g,n}\| \le Cn^{-1/2}.$$

Conditionally on the Gaussian sample,

$$h_{\zeta,n} = \frac{1}{n}\sum_i M_n^{\mathrm{T}}\check{R}_\omega F_i^\circ w\,\zeta_i$$

is a sum of independent centered two-dimensional vectors, and

$$\begin{aligned}\sum_i \left\|\frac{1}{n}M_n^{\mathrm{T}}\check{R}_\omega F_i^\circ w\right\|^2 &\le \frac{C}{n}\operatorname{tr}\left[M_n^{\mathrm{T}}\check{R}_\omega B_n \check{R}_\omega^* M_n\right]\\ &\le \frac{C}{n}\|M_n\|_F^2.\end{aligned}$$

The last inequality follows from $\sup_{x\ge 0} x/|x+\kappa(\lambda)+\gamma+\omega|^2 < C$. Conditional vector Rosenthal gives, with $c_i = n^{-1}M_n^{\mathrm{T}}\check{R}_\omega F_i^\circ w$,

$$\{\mathbb{E}_\zeta\|h_{\zeta,n}\|^8\}^{1/8} \le C\left\{\left(\sum_i\|c_i\|^2\right)^{1/2} + \left(\sum_i\|c_i\|^8\right)^{1/8}\right\} \le Cn^{-1/2}\|M_n\|_F.$$

This argument gives

$$\|g_n\|_{L^q} \le C_q n^{-1/2} \quad (2\le q\le 16), \qquad \|h_n\|_{L^8} \le Cn^{-1/2}. \tag{S.3.232}$$

Direct substitution in (S.3.225) gives

$$\begin{aligned}s_n^{\mathrm{T}}\mathcal{R}_\omega s_n - \check{s}_n^{\mathrm{T}}\check{R}_\omega\check{s}_n &= h_n^{\mathrm{T}}J_n h_n - 2\bar{y}\,w^{\mathrm{T}}J_n h_n + \bar{y}^2 w^{\mathrm{T}}K_n J_n w,\\ e^{\mathrm{T}}\mathcal{R}_\omega e - e^{\mathrm{T}}\check{R}_\omega e &= g_n^{\mathrm{T}}J_n g_n,\\ e^{\mathrm{T}}\mathcal{R}_\omega s_n - e^{\mathrm{T}}\check{R}_\omega\check{s}_n &= g_n^{\mathrm{T}}J_n h_n - \bar{y}\,g_n^{\mathrm{T}}J_n w.\end{aligned}$$

All three right-hand sides are $O_{L^2}(n^{-1})$. Combining them with (S.3.218), (S.3.219), and (S.3.224) gives, uniformly on Γ,

$$\begin{aligned}e^{\mathrm{T}}\mathcal{R}_\omega e - q_\omega &\xrightarrow{L^2} 0,\\ e^{\mathrm{T}}\mathcal{R}_\omega s_n - \theta_\omega q_\omega e^{\mathrm{T}}a &\xrightarrow{L^2} 0,\\ s_n^{\mathrm{T}}\mathcal{R}_\omega s_n - \{\theta_\omega^2 q_\omega\|a\|^2 + \rho^2 c_{B,\omega}\} &\xrightarrow{L^2} 0.\end{aligned} \tag{S.3.233}$$

Write $\eta_n = e^{\mathrm{T}}a$ and

$$\beta_n = \left(1-\frac{1}{n}\right)\beta_\sigma + \frac{1}{n}\mu_0(\sigma), \qquad \widehat{z} = \beta_n\sqrt{p}\,e + \delta_n.$$

The endpoint Hoeffding decomposition gives, for every fixed $q \le 16$,

$$\delta_n = M_n\begin{pmatrix}1\\1\end{pmatrix} + \varepsilon_{m,n}, \qquad \|\delta_n\|_{L^q} \le C_q, \qquad \|\varepsilon_{m,n}\|_{L^q} \le C_q n^{-1/2}\log^C n. \tag{S.3.234}$$

Set

$$A_z = \widehat{z}^{\mathrm{T}}\mathcal{R}_\omega\widehat{z}, \qquad b_z = \widehat{z}^{\mathrm{T}}\mathcal{R}_\omega s_n.$$

Equation (S.3.233) and (S.3.234) imply

$$\begin{pmatrix}p^{-1}(1+A_z) & p^{-1/2}b_z\\ p^{-1/2}b_z & s_n^{\mathrm{T}}\mathcal{R}_\omega s_n\end{pmatrix} - \begin{pmatrix}\beta_\sigma^2 q_\omega & \beta_\sigma\theta_\omega q_\omega\eta_n\\ \beta_\sigma\theta_\omega q_\omega\eta_n & \theta_\omega^2 q_\omega\|a\|^2 + \rho^2 c_{B,\omega}\end{pmatrix} \xrightarrow{L^2} 0. \tag{S.3.235}$$

To justify division by $1+A_z$, first note that if $z_\omega = \gamma+\omega$, the spectral decomposition of C_n gives

$$\operatorname{Re}A_z \ge \frac{\gamma_-/2}{(\|C_n\| + \gamma_+ + \sup_\Gamma|\omega|)^2}|\widehat{z}|^2.$$

On $\mathcal{H}_n = \{\|\delta_n\| \le |\beta_n|\sqrt{p}/2\}$, $\|\widehat{z}\|^2 \ge \beta_n^2 p/4$. By (S.3.234), $\mathbb{P}(\mathcal{H}_n^c) \le Cp^{-8}$, while the fixed-ratio sample-covariance estimate gives $\|\|C_n\|\|_{L^q} \le C_q$. Since $|1 + A_z| \ge 1 + \mathrm{Re}\, A_z$, splitting over $\mathcal{H}_n$ and its complement yields

$$\sup_{\omega\in\Gamma} \left\| \frac{p}{1 + A_z} \right\|_{L^8} \le C. \tag{S.3.236}$$

To see the corresponding bound for b_z, let

$$A_w = [F_1^\circ w, \dots, F_n^\circ w].$$

The residual-label part of s_n is $s_{n,\zeta} = n^{-1} A_w H_n \zeta$. Conditionally on the Gaussian sample,

$$\widehat{z}^{\mathrm{T}} \mathcal{R}_\omega s_{n,\zeta} = c^{\mathrm{T}} \zeta, \qquad c = \frac{1}{n} H_n A_w^{\mathrm{T}} \mathcal{R}_\omega^{\mathrm{T}} \widehat{z},$$

and

$$\begin{aligned} \|c\|^2 &\le \frac{1}{n^2} \widehat{z}^{\mathrm{T}} \mathcal{R}_\omega^* A_w H_n A_w^{\mathrm{T}} \mathcal{R}_\omega \widehat{z} \\ &\le \frac{C}{n} \widehat{z}^{\mathrm{T}} \mathcal{R}_\omega^* B_n \mathcal{R}_\omega \widehat{z} \le \frac{C}{n} (1 + \|M_n\|^2) |\widehat{z}\|^2. \end{aligned}$$

Here $B_n = C_n - \kappa(\lambda) I_p + M_n M_n^{\mathrm{T}}$, and the last inequality uses the ridge bound and $\sup_{x\ge 0} x/|x+\gamma+\omega|^2 < C$ on the centered covariance part. Conditional scalar Rosenthal and $\sup_i \mathbb{E}|\zeta_i|^8 < \infty$ therefore give a uniform L^8 bound for the residual-label part. The signal part is $O_{L^8}(\sqrt{p})$: the fixed Gaussian moment bounds and the ridge bound give

$$\|\widehat{z}^{\mathrm{T}} \mathcal{R}_\omega s_{n,\|}\|_{L^8} \le C \|\widehat{z}\|_{L^{16}} \|s_{n,\|}\|_{L^{16}} \le C\sqrt{p}.$$

The L^{16} bound for $s_{n,\|}$ follows from the Hilbert space Rosenthal inequality and the Gaussian moments of the sample-centering term. Consequently,

$$\sup_{\omega\in\Gamma} \|p^{-1/2} b_z\|_{L^8} \le C. \tag{S.3.237}$$

Sherman–Morrison gives

$$s_n^{\mathrm{T}} Q_*(\omega) s_n = s_n^{\mathrm{T}} \mathcal{R}_\omega s_n - \frac{b_z^2}{1 + A_z}. \tag{S.3.238}$$

Equations (S.3.235)–(S.3.237) show that the right-hand side converges in probability to

$$\theta_\omega^2 q_\omega (\|a\|^2 - \eta_n^2) + \rho^2 c_{B,\omega}.$$

The convergence is in L^2: the left-hand side of (S.3.238) is uniformly bounded in L^4, since $\|Q_*(\omega)\| \le 2\gamma_-^{-1}$ and $\|s_n\|_{L^8} \le C$. Finally,

$$r_*^{\mathrm{T}} Q_*(\omega) r_* = s_n^{\mathrm{T}} Q_*(\omega) s_n + 2\bar{y} \frac{b_z}{1 + A_z} + \bar{y}^2 \frac{A_z}{1 + A_z}.$$

The inverse-moment bounds imply

$$\left\| \frac{b_z}{1 + A_z} \right\|_{L^4} \le Cp^{-1/2}, \qquad \left| \frac{A_z}{1 + A_z} \right| \le 2.$$

Since $\|\bar{y}\|_{L^4} = O(n^{-1/2})$, both terms containing $\bar{y}$ are $o_{L^2}(1)$, uniformly on Γ. Combining the last display with (S.3.200) proves (S.3.208).

For the estimate in the $\mathbf{1}$-direction, evaluate all quantities below at $\omega = 0$ and put $d_2 = (1, 1)^{\mathrm{T}}$. From (S.3.225),

$$M_n^{\mathrm{T}} \mathcal{R}_0 = J_n M_n^{\mathrm{T}} \check{R}_0.$$

Since $s_n = \check{s}_n - M_n w \bar{y}$, this gives the exact identity

$$M_n^{\mathrm{T}} \mathcal{R}_0 s_n = J_n \{h_n - K_n w \bar{y}\}.$$

The bounds above imply

$$\|K_n\|_{L^{16}} + \|J_n\|_{L^{16}} \le C, \qquad \|h_n\|_{L^8} + \|\bar{y}\|_{L^8} \le Cn^{-1/2},$$

and the ridge estimate gives $\|\mathcal{R}_0 s_n\|_{L^8} \le C$. Together with (S.3.234), Hölder's inequality therefore yields

$$\|\delta_n^{\mathrm{T}} \mathcal{R}_0 s_n\|_{L^4} \le C n^{-1/2} \log^C n.$$

In particular,

$$\mathbb{E}|\delta_n^{\mathrm{T}} \mathcal{R}_0 s_n|^2 = O(n^{-1} \log^C n). \tag{S.3.239}$$

Set

$$D_n = 1 + \widehat{z}^{\mathrm{T}} \mathcal{R}_0 \widehat{z}, \qquad \mathrm{r}_n = \frac{\mathbf{1}^{\mathrm{T}} \mathcal{R}_0 \widehat{z}}{D_n},$$
$$u_n = \mathbf{1}^{\mathrm{T}} \mathcal{R}_0 \delta_n, \qquad v_n = \delta_n^{\mathrm{T}} \mathcal{R}_0 \delta_n.$$

Since $\widehat{z} = \beta_n \mathbf{1} + \delta_n$, direct algebra gives

$$\mathrm{r}_n - \frac{1}{\beta_n} = -\frac{1 + \beta_n u_n + v_n}{\beta_n D_n}, \qquad 1 - \beta_n \mathrm{r}_n = \frac{1 + \beta_n u_n + v_n}{D_n}. \tag{S.3.240}$$

On $\mathcal{H}_n = \{\|\delta_n\| \le |\beta_n| \sqrt{p}/2\}$, positivity of $\mathcal{R}_0$ gives

$$D_n \ge \frac{|\widehat{z}\|^2}{\|C_n\| + \gamma} \ge \frac{\beta_n^2 p}{4(\|C_n\| + \gamma)}.$$

On $\mathcal{H}_n^c$ use $D_n \ge 1$. Equation (S.3.234) gives $\mathbb{P}(\mathcal{H}_n^c) \le C p^{-8}$; hence

$$\left\| \frac{p}{D_n} \right\|_{L^8} \le C. \tag{S.3.241}$$

Furthermore,

$$M_n^{\mathrm{T}} \mathcal{R}_0 e = J_n g_n.$$

Since $\delta_n = M_n d_2 + \varepsilon_{m,n}$,

$$u_n = \sqrt{p}\, d_2^{\mathrm{T}} J_n g_n + \sqrt{p}\, e^{\mathrm{T}} \mathcal{R}_0 \varepsilon_{m,n}.$$

The fixed-exponent version of the Gaussian cavity estimate (S.3.232) gives $\|g_n\|_{L^{16}} \le C n^{-1/2}$. Therefore

$$\|u_n\|_{L^8} \le C \log^C n, \qquad \|v_n\|_{L^8} \le C.$$

Since $\beta_n - \beta_\sigma = O(n^{-1})$ and $\beta_\sigma \ne 0$, (S.3.240) and (S.3.241) give

$$\left\| \mathrm{r}_n - \frac{1}{\beta_\sigma} \right\|_{L^4} + \|1 - \beta_n \mathrm{r}_n\|_{L^4} \le C p^{-1} \log^C n.$$

Using $\widehat{z} = \beta_n \mathbf{1} + \delta_n$ once more,

$$\mathbf{1}^{\mathrm{T}} \mathcal{R}_0 s_n - \mathrm{r}_n \widehat{z}^{\mathrm{T}} \mathcal{R}_0 s_n + \frac{1}{\beta_\sigma} \delta_n^{\mathrm{T}} \mathcal{R}_0 s_n$$
$$= (1 - \beta_n \mathrm{r}_n) \mathbf{1}^{\mathrm{T}} \mathcal{R}_0 s_n + \left(\frac{1}{\beta_\sigma} - \mathrm{r}_n \right) \delta_n^{\mathrm{T}} \mathcal{R}_0 s_n.$$

The ridge bound gives $\|\mathbf{1}^{\mathrm{T}} \mathcal{R}_0 s_n\|_{L^4} \le C \sqrt{p}$. Consequently,

$$\mathbf{1}^{\mathrm{T}} \mathcal{R}_0 s_n - \mathrm{r}_n \widehat{z}^{\mathrm{T}} \mathcal{R}_0 s_n = -\frac{1}{\beta_\sigma} \delta_n^{\mathrm{T}} \mathcal{R}_0 s_n + O_{L^2}(p^{-1/2} \log^C n). \tag{S.3.242}$$

Sherman–Morrison gives

$$\mathbf{1}^{\mathrm{T}} Q_*(0) r_* = \mathrm{r}_n \bar{y} + \mathbf{1}^{\mathrm{T}} \mathcal{R}_0 s_n - \mathrm{r}_n \widehat{z}^{\mathrm{T}} \mathcal{R}_0 s_n.$$

$$\|\mathrm{r}_n \bar{y}\|_{L^2} \le \|\mathrm{r}_n\|_{L^4} \|\bar{y}\|_{L^4} \le C n^{-1/2}.$$

Combining the last identity with (S.3.201), (S.3.239), and (S.3.242), and using $p \asymp n$, proves (S.3.209). □

Lemma S.3.17 Assume $\beta_\sigma \neq 0$ and (S.1.7). At $\omega = 0$, abbreviate

$$C_2 = (I_2 + t_0 V)^{-1},$$
$$\theta = d_\alpha w^{\mathrm{T}} C_2 w.$$

Define

$$\tau_0 = \frac{t_0}{\gamma + \kappa(\lambda) + \mathrm{tr}(VC_2^2)}, \qquad \ell = \tau_0\, w^{\mathrm{T}} V C_2^2 w.$$

Here $t_0, C_2, \theta, \tau_0, \ell$ are finite-dimensional auxiliary quantities, and their dependence on n is suppressed. In the proof of Theorem S.3.5, they are identified respectively with $t_{*,n}, C_{*,n}, \theta_n, \tau_n, \ell_n$. Then

$$r^{\mathrm{T}} Q^2 r - \mathcal{A}_{\mathrm{bulk}} \xrightarrow{\mathbb{P}} 0, \tag{S.3.243}$$

where

$$\mathcal{A}_{\mathrm{bulk}} = \theta^2 \frac{\tau_0}{c_n} \|P_\perp a\|^2 + \left\{ \rho^2 - 2 d_\alpha \theta \frac{t_0}{c_n} \|P_\perp a\|^2 \right\} \ell. \tag{S.3.244}$$

Equivalently, if

$$\eta_n = e^{\mathrm{T}} a, \qquad r_{\perp,n}^2 = \|P_\perp a\|^2, \qquad \rho_\perp^2 = \rho^2 - \|a\|^2,$$

then

$$\mathcal{A}_{\mathrm{bulk}} = K_{\perp,n} r_{\perp,n}^2 + \ell(\eta_n^2 + \rho_\perp^2), \qquad K_{\perp,n} = \theta^2 \frac{\tau_0}{c_n} + \left(1 - 2 d_\alpha \theta \frac{t_0}{c_n}\right) \ell. \tag{S.3.245}$$

Proof. Let a dot denote differentiation with respect to ω at zero. Differentiating (S.3.207) gives

$$0 = \dot{t}_0 \left[\gamma + \kappa(\lambda) + \mathrm{tr}(VC_2) - t_0\, \mathrm{tr}(V^2 C_2^2) \right] + t_0.$$

Since V commutes with C_2 and $C_2 - t_0 V C_2^2 = C_2^2$, the expression in brackets equals

$$\gamma + \kappa(\lambda) + \mathrm{tr}(VC_2^2).$$

Hence

$$\dot{t}_0 = -\tau_0, \qquad \dot{C}_{2,0} = -C_2 V C_2 \dot{t}_0 = \tau_0 V C_2^2.$$

It follows that

$$\dot{q}_0 = -\frac{\tau_0}{c_n}, \qquad \dot{c}_{B,0} = -\ell, \qquad \dot{\theta}_0 = d_\alpha \ell,$$

and consequently

$$-\dot{\mathcal{E}}_0 = \theta^2 \frac{\tau_0}{c_n} \|P_\perp a\|^2 + \left\{ \rho^2 - 2 d_\alpha \theta \frac{t_0}{c_n} \|P_\perp a\|^2 \right\} \ell = \mathcal{A}_{\mathrm{bulk}}.$$

Put

$$D_n(\omega) = r^{\mathrm{T}} Q_I(\omega) r - \mathcal{E}_\omega$$

and let r_Γ be the radius of Γ. Both terms are analytic inside Γ, and Cauchy's formula together with (S.3.208) gives

$$\|D_n'(0)\|_{L^2} \leq r_\Gamma^{-1} \sup_{\omega \in \Gamma} \|D_n(\omega)\|_{L^2} = o(1).$$

Since

$$-\partial_\omega \{ r^{\mathrm{T}} Q_I(\omega) r \}\big|_{\omega=0} = r^{\mathrm{T}} Q^2 r,$$

we have

$$-D_n'(0) = r^{\mathrm{T}} Q^2 r - \mathcal{A}_{\mathrm{bulk}},$$

which proves (S.3.243). Since $\rho^2 = r_{\perp,n}^2 + \eta_n^2 + \rho_\perp^2$, collecting the coefficients of these three deterministic energies in (S.3.244) gives (S.3.245). □

Lemma S.3.18 Let

$$\kappa_\sigma = \nu - \mu_0(\sigma)^2, \qquad q_1 = Q\mathbf{1}, \qquad b = \Phi q_1,$$

and define

$$T_{a,n} = \frac{1}{n} b^{\mathsf{T}} G a, \qquad T_{\perp,n} = \frac{1}{n} b^{\mathsf{T}} \zeta, \qquad \mathcal{V}_{\phi,n} = \frac{1}{n}\|b\|^2 = q_1^{\mathsf{T}} \left(\frac{1}{n}\Phi^{\mathsf{T}}\Phi\right) q_1.$$

Then

$$A = \kappa_\sigma r^{\mathsf{T}} Q^2 r + A_{11,n}, \qquad A_{11,n} = \mu_0(\sigma)^2 (T_{a,n} + T_{\perp,n})^2. \tag{S.3.246}$$

If $\beta_\sigma \neq 0$ and (S.1.7) holds, then

$$A_{11,n} = O_{\mathbb{P}}(n^{-1} \log^C n) = o_{\mathbb{P}}(1). \tag{S.3.247}$$

Let

$$\mathcal{G}_W = \sigma(W, g_1, \ldots, g_n).$$

Conditionally on $\mathcal{G}_W$,

$$\mathbb{E}_\zeta[A_{11,n} \mid \mathcal{G}_W] = \mu_0(\sigma)^2 \left\{ T_{a,n}^2 + \frac{1}{n^2} b^{\mathsf{T}} D_{\zeta,n} b \right\}. \tag{S.3.248}$$

Furthermore,

$$0 \le \mathcal{V}_{\phi,n} \le q_1^{\mathsf{T}} S q_1 = \mathbf{1}^{\mathsf{T}} Q \mathbf{1} - \gamma \mathbf{1}^{\mathsf{T}} Q^2 \mathbf{1} \le \mathbf{1}^{\mathsf{T}} Q \mathbf{1}. \tag{S.3.249}$$

If the residual variables are Gaussian, then, conditionally on $\mathcal{G}_W$,

$$A_{11,n} \overset{d}{=} \mu_0(\sigma)^2 \left(T_{a,n} + \frac{\sqrt{b^{\mathsf{T}} D_{\zeta,n} b}}{n} Z \right)^2, \qquad Z \sim \mathcal{N}(0,1), \tag{S.3.250}$$

where Z is independent of $\mathcal{G}_W$.

Proof. Since $W^{\mathsf{T}} W = I_p$, the coordinates of $W^{\mathsf{T}} x_{\text{new}}$ are iid standard Gaussian and

$$\mathbb{E}\left[\sigma(W^{\mathsf{T}} x_{\text{new}}) \sigma(W^{\mathsf{T}} x_{\text{new}})^{\mathsf{T}}\right] = \kappa_\sigma I_p + \mu_0(\sigma)^2 \mathbf{1}\mathbf{1}^{\mathsf{T}}.$$

This proves the first equality in (S.3.246). By Lemma S.3.14,

$$\mathbf{1}^{\mathsf{T}} Q r = q_1^{\mathsf{T}} r = \frac{1}{n} \sum_i (q_1^{\mathsf{T}} \phi_i) y_i = T_{a,n} + T_{\perp,n},$$

which proves the second equality.

Conditionally on $\mathcal{G}_W$, the vector b is fixed, while the entries of ζ are independent and centered. Therefore

$$\mathbb{E}_\zeta T_{\perp,n} = 0, \qquad \mathbb{E}_\zeta T_{\perp,n}^2 = \frac{1}{n^2} b^{\mathsf{T}} D_{\zeta,n} b \le \frac{\|D_{\zeta,n}\|_{\text{op}}}{n} \mathcal{V}_{\phi,n}, \tag{S.3.251}$$

which proves (S.3.248). The matrix order in Lemma S.3.14 gives $\mathcal{V}_{\phi,n} \le q_1^{\mathsf{T}} S q_1$. The bound

$$QSQ = Q - \gamma Q^2$$

gives (S.3.249). If ζ is Gaussian, then $T_{\perp,n}$, conditionally on $\mathcal{G}_W$, is exactly Gaussian with the variance in (S.3.251), which proves (S.3.250). Finally, when $\beta_\sigma \neq 0$ and (S.1.7) holds, (S.3.209) and (S.3.246) imply

$$\mathbb{E} A_{11,n} = O(n^{-1} \log^C n).$$

Markov's inequality then proves (S.3.247) on this branch. □

S.3.7 Supporting lemmas for zero ridge and Haar averaging

S.3.7.1 Noncentered ridge-to-ridgeless continuity

Lemma S.3.19 Under the assumptions of Theorem S.3.8, define

$$\widetilde{S}_n = S - \widetilde{z}\widetilde{z}^{\mathsf{T}}.$$

There are events $\mathcal{G}_n$ with $\mathbb{P}(\mathcal{G}_n) \to 1$ on which

$$\lambda_{\min}(\widetilde{S}_n) \geq \kappa(\lambda)/2, \qquad \lambda_{\min}(S) \geq \kappa(\lambda)/2. \tag{S.3.252}$$

For $\gamma \geq 0$, write $Q_\gamma = (S + \gamma I_p)^{-1}$ on $\mathcal{G}_n$. For every fixed $\gamma_1 < \infty$,

$$\sup_{0\leq\gamma\leq\gamma_1} \left\{ \|Q_\gamma\|_{\mathrm{op}} + \|Q_\gamma r\| + \|Q_\gamma \mathbf{1}\| + |\mathbf{1}^{\mathrm{T}} Q_\gamma r| \right\} = O_{\mathbb{P}}(1) \tag{S.3.253}$$

on these events. Consequently, for every $\varepsilon > 0$,

$$\lim_{\gamma\downarrow 0} \limsup_{n\to\infty} \mathbb{P}\{|\,\mathrm{Risk}_{\gamma,n} - \mathrm{Risk}_{0,n}\,| > \varepsilon\} = 0. \tag{S.3.254}$$

Proof. The ridge-free expansion (S.3.202) gives

$$S = C_n + \widehat{z}\widehat{z}^{\mathrm{T}} + d_n \mathbf{1}\mathbf{1}^{\mathrm{T}} + E_{S,n}, \qquad \|E_{S,n}\|_{\mathrm{op}} = o_{\mathbb{P}}(1),$$

where

$$C_n = \kappa(\lambda) I_p + \frac{1}{n} U H_n U^{\mathrm{T}} + \frac{1}{n} V_n H_n V_n^{\mathrm{T}} \geq \kappa(\lambda) I_p, \qquad d_n \geq 0.$$

Thus, on

$$\mathcal{G}_n = \{\|E_{S,n}\|_{\mathrm{op}} \leq \kappa(\lambda)/2\},$$

both inequalities in (S.3.252) hold, and $\mathbb{P}(\mathcal{G}_n) \to 1$.

On $\mathcal{G}_n$, put

$$\widetilde{Q}_\gamma = (\widetilde{S}_n + \gamma I_p)^{-1}.$$

Then

$$\sup_{0\leq\gamma\leq\gamma_1} \{\|\widetilde{Q}_\gamma\|_{\mathrm{op}} + \|Q_\gamma\|_{\mathrm{op}}\} \leq \frac{2}{\kappa(\lambda)}.$$

Sherman–Morrison gives

$$Q_\gamma = \widetilde{Q}_\gamma - \frac{\widetilde{Q}_\gamma \widehat{z}\widehat{z}^{\mathrm{T}} \widetilde{Q}_\gamma}{1 + \widehat{z}^{\mathrm{T}} \widetilde{Q}_\gamma \widehat{z}}. \tag{S.3.255}$$

By (S.3.234),

$$\widehat{z} = \beta_n \mathbf{1} + \delta_n, \qquad \beta_n \to \beta_\sigma \neq 0, \qquad \|\delta_n\| = O_{\mathbb{P}}(1).$$

Let $A_\gamma = \widehat{z}^{\mathrm{T}} \widetilde{Q}_\gamma \widehat{z} \geq 0$. Equation (S.3.255) yields

$$Q_\gamma \widehat{z} = \frac{\widetilde{Q}_\gamma \widehat{z}}{1 + A_\gamma}.$$

Moreover,

$$\|\widetilde{Q}_\gamma \widehat{z}\|^2 \leq \|\widetilde{Q}_\gamma\|_{\mathrm{op}} A_\gamma,$$

so

$$\sup_{0\leq\gamma\leq\gamma_1} \|Q_\gamma \widehat{z}\| \leq \sqrt{\frac{2}{\kappa(\lambda)}} \sup_{x\geq 0} \frac{\sqrt{x}}{1+x} \leq C.$$

Since

$$Q_\gamma \mathbf{1} = \beta_n^{-1} \{Q_\gamma \widehat{z} - Q_\gamma \delta_n\},$$

we obtain the uniform bound for $Q_\gamma \mathbf{1}$. The ridge-free parts of (S.3.206) and (S.3.205) give

$$\|r\|_{L^4} \leq \|r_*\|_{L^4} + \|r - r_*\|_{L^4} \leq C.$$

The remaining bounds in (S.3.253) follow.

On $\mathcal{G}_n$, the resolvent identity gives

$$Q_\gamma - Q_0 = -\gamma Q_\gamma Q_0, \qquad \|Q_\gamma - Q_0\|_{\mathrm{op}} \leq \frac{4\gamma}{\kappa(\lambda)^2}.$$

It follows that

$$\|Q_\gamma r - Q_0 r\| = O_{\mathbb{P}}(\gamma),$$

and, since Q_γ and Q_0 commute,

$$|\mathbf{1}^\top(Q_\gamma - Q_0)r| = \gamma|(Q_\gamma \mathbf{1})^\top Q_0 r| = O_{\mathbb{P}}(\gamma).$$

Moreover,

$$|r^\top Q_\gamma^2 r - r^\top Q_0^2 r| \le \{\|Q_\gamma r\| + \|Q_0 r\|\}\|(Q_\gamma - Q_0)r\| = O_{\mathbb{P}}(\gamma),$$
$$|(\mathbf{1}^\top Q_\gamma r)^2 - (\mathbf{1}^\top Q_0 r)^2| \le \{|\mathbf{1}^\top Q_\gamma r| + |\mathbf{1}^\top Q_0 r|\}|\mathbf{1}^\top(Q_\gamma - Q_0)r| = O_{\mathbb{P}}(\gamma).$$

To average the risk over the training noise, we use the moment bounds underlying these estimates. On the design event $\mathcal{G}_n$,

$$\sup_{0\le\gamma\le\gamma_1} \|Q_\gamma \mathbf{1}\| \le C(1 + \|\delta_n\|), \qquad \|\delta_n\|_{L^4} \le C.$$

The second bound follows from (S.3.234). For $\mathcal{M}_n$ in (S.3.5), conditional Jensen's inequality and the displayed raw-score bound give $\mathbb{E}\mathcal{M}_n^2 \le \mathbb{E}\|r\|^4 \le C$. Apply (S.1.4) to the realized-risk difference and then condition on X, W. The resolvent bounds above give

$$\mathbf{1}_{\mathcal{G}_n}|\,\mathrm{Risk}_{\gamma,n} - \mathrm{Risk}_{0,n}\,| \le C\gamma\left[\{1 + (1 + \|\delta_n\|)^2\}\mathcal{M}_n + \|a\|\sqrt{\mathcal{M}_n}\right].$$

The expression in brackets has bounded expectation by Cauchy–Schwarz. Markov's inequality therefore bounds the probability of a risk difference exceeding ε by $\mathbb{P}(\mathcal{G}_n^c) + C\gamma/\varepsilon$. First letting $n \to \infty$ and then $\gamma \downarrow 0$ proves (S.3.254). □

S.3.7.2 Finite-dimensional Haar centering and deconditioning

Recall that $\kappa_\sigma = \nu$ in the odd branch. Define

$$(k_n, T_n) = \begin{cases} (p-1,\ \|P_\perp W^\top \beta^*\|^2), & \beta_\sigma \ne 0, \\ (p,\ \|W^\top \beta^*\|^2), & \sigma \text{ odd}. \end{cases}$$

The average training variance $\overline{\sigma}_{\varepsilon,n}^2$ and the test variance σ_{test}^2 are deterministic, so neither depends on the Haar frame. All moment bounds are uniform in the deterministic frame U. We can therefore remove the conditioning on W without introducing another random coefficient. In either branch, the conditional deterministic equivalent has the common form

$$\overline{\mathrm{Risk}}_n(W) = \mathfrak{L}\{\|\beta^*\|^2 + \overline{\sigma}_{\varepsilon,n}^2\} + (\mathfrak{K} - \mathfrak{L})T_n + \sigma_{\mathrm{test}}^2 - \overline{\sigma}_{\varepsilon,n}^2. \tag{S.3.256}$$

Indeed, in the noncentered branch

$$(e^\top a)^2 + \rho_{\perp,n}^2 = \|\beta^*\|^2 - T_n + \overline{\sigma}_{\varepsilon,n}^2,$$

while in the odd branch $\rho_{\perp,n}^2 = \|\beta^*\|^2 - T_n + \overline{\sigma}_{\varepsilon,n}^2$.

Lemma S.3.20 Assume the design conditions and Assumption 1, fix $\lambda \in (0,1]$ and $\gamma > 0$, and let (β_n^*) be the deterministic teacher sequence satisfying the standing norm bound. In the noncentered branch, also assume (S.3.10) and (S.1.7). Then, for every $\delta > 0$,

$$\sup_{U\in\mathrm{St}(d_n,p_n)} \mathbb{P}_X\left(\left|\mathrm{Risk}_n(U) - \overline{\mathrm{Risk}}_n^{\mathrm{nz}}(U)\right| > \delta\right) \longrightarrow 0, \tag{S.3.257}$$

where $\overline{\mathrm{Risk}}_n^{\mathrm{nz}}(U)$ is the limit-coefficient conditional centering in (S.3.22), evaluated at

$$a = U^T\beta^*, \qquad r_{\perp,n}^2 = \|P_\perp a\|^2, \qquad \rho_{\perp,n}^2 = \|\beta^*\|^2 - \|a\|^2 + \overline{\sigma}_{\varepsilon,n}^2.$$

If σ is odd, then, for every $\delta > 0$,

$$\sup_{U\in\mathrm{St}(d_n,p_n)} \mathbb{P}_X\left(\left|\mathrm{Risk}_n(U) - \overline{\mathrm{Risk}}_n^{\mathrm{odd}}(U)\right| > \delta\right) \longrightarrow 0, \tag{S.3.258}$$

where $\overline{\mathrm{Risk}}_n^{\mathrm{odd}}(U)$ is defined in (S.2.2).

Proof. We begin with the sequential form. Write the dimensions and teacher as d_n, p_n, β_n^*, and suppress these subscripts elsewhere when the meaning is clear. Consider the noncentered branch and let

$$U_n \in \mathrm{St}(d_n, p_n)$$

be an arbitrary deterministic triangular array. Write

$$a_n = U_n^{\mathsf{T}}\beta_n^*, \qquad e_{p_n} = \frac{\mathbf{1}_{p_n}}{\sqrt{p_n}}, \qquad \rho_{\perp,n}^2 = \|\beta_n^*\|^2 - \|a_n\|^2 + \overline{\sigma}_{\varepsilon,n}^2.$$

The teacher norm bound and the bounded variance profile imply that

$$\left(\|a_n\|^2, e_{p_n}^{\mathsf{T}} a_n, \rho_{\perp,n}^2\right)$$

is a bounded sequence of triples.

Suppose the sequential assertion fails. Then there are $\delta, \epsilon > 0$ and a subsequence (n_j) such that

$$\mathbb{P}_X\left(\left|\mathrm{Risk}_{n_j}(U_{n_j}) - \overline{\mathrm{Risk}}_{n_j}^{\mathrm{nz}}(U_{n_j})\right| > \delta\right) \geq \epsilon.$$

Boundedness allows us to pass to a further subsequence on which

$$\|a_{n_j}\|^2 \longrightarrow r_a^2, \qquad e_{p_{n_j}}^{\mathsf{T}} a_{n_j} \longrightarrow \eta, \qquad \rho_{\perp,n_j}^2 \longrightarrow \rho_{\perp,\infty}^2.$$

This subsequence still satisfies the standing assumptions and the mean-gap condition. Applying Theorem S.3.6 to the deterministic frame sequence (U_{n_j}) then gives

$$\mathrm{Risk}_{n_j}(U_{n_j}) \xrightarrow{\mathbb{P}} \mathfrak{R}(r_a^2 - \eta^2) + \mathfrak{L}(\eta^2 + \rho_{\perp,\infty}^2) + \sigma_{\mathrm{test}}^2 - \overline{\sigma}_{\varepsilon}^2.$$

The exact identity

$$\overline{\mathrm{Risk}}_n^{\mathrm{nz}}(U) = \mathfrak{R}\left\{\|U^{\mathsf{T}}\beta^*\|^2 - (e^{\mathsf{T}}U^{\mathsf{T}}\beta^*)^2\right\} + \mathfrak{L}\left\{(e^{\mathsf{T}}U^{\mathsf{T}}\beta^*)^2 + \rho_{\perp,n}^2\right\} + \sigma_{\mathrm{test}}^2 - \overline{\sigma}_{\varepsilon,n}^2$$

shows that $\overline{\mathrm{Risk}}_{n_j}^{\mathrm{nz}}(U_{n_j})$ converges to the same deterministic limit. Their difference therefore converges to zero in probability, contradicting the preceding lower bound. Thus, for every deterministic triangular array (U_n),

$$\mathbb{P}_X\left(\left|\mathrm{Risk}_n(U_n) - \overline{\mathrm{Risk}}_n^{\mathrm{nz}}(U_n)\right| > \delta\right) \longrightarrow 0.$$

The sequential conclusions imply the two uniform statements. For either branch, write

$$R_n(U) = \mathbb{P}_X\left(\left|\mathrm{Risk}_n(U) - \overline{\mathrm{Risk}}_n(U)\right| > \delta\right),$$

where $\overline{\mathrm{Risk}}_n$ denotes the corresponding branch-specific equivalent. If $\sup_U R_n(U) \not\to 0$, there would be $\epsilon > 0$ and a subsequence (n_j) on which $\sup_U R_{n_j}(U) \geq \epsilon$. Choose deterministic frames $U_{n_j} \in \mathrm{St}(d_{n_j}, p_{n_j})$ within $1/j$ of the corresponding suprema. Then, for all large j,

$$R_{n_j}(U_{n_j}) \geq \sup_{U \in \mathrm{St}(d_{n_j}, p_{n_j})} R_{n_j}(U) - \frac{1}{j} \geq \frac{\epsilon}{2}.$$

Choose arbitrary deterministic orthonormal frames at indices outside this subsequence. The resulting full triangular array contradicts the corresponding sequential conclusion. This proves (S.3.257) and (S.3.258). □

Proposition S.3.21 Assume the branch-specific conditions of Lemma S.3.20, let the teacher sequence be deterministic, let Haar W be independent of the sample, and impose either (S.1.7) in the noncentered branch or oddness of σ in the centered branch. No deterministic limits for $\|a\|^2$, $e^{\mathsf{T}}a$, or $\rho_{\perp,n}^2$ are assumed here. Define

$$\overline{\mathrm{Risk}}_n^{\mathrm{Haar}} := \mathfrak{L}\{\|\beta^*\|^2 + \overline{\sigma}_{\varepsilon,n}^2\} + (\mathfrak{R} - \mathfrak{L})\frac{k_n}{d}\|\beta^*\|^2 + \sigma_{\mathrm{test}}^2 - \overline{\sigma}_{\varepsilon,n}^2. \tag{S.3.259}$$

This is the exact Haar expectation of the conditional equivalent (S.3.256), and

$$\mathrm{Var}_W\{\overline{\mathrm{Risk}}_n(W)\} = (\mathfrak{K} - \mathfrak{Q})^2\|\beta^*\|^4 \frac{2k_n(d-k_n)}{d^2(d+2)}. \tag{S.3.260}$$

Moreover, with probability taken jointly over W, X,

$$\mathrm{Risk}_n - \overline{\mathrm{Risk}}_n^{\mathrm{Haar}} \xrightarrow{\mathbb{P}} 0. \tag{S.3.261}$$

Proof. Equations (S.3.259) and (S.3.260) follow immediately from Lemma S.5.1. To pass from a fixed frame to the joint probability space, let $\mu^{\mathrm{Haar}}_{d_n,p_n}$ be Haar probability measure on $\mathrm{St}(d_n, p_n)$. In the noncentered branch, (S.3.257) gives, for every $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}_{W,X}\left(|\,\mathrm{Risk}_n(W) - \overline{\mathrm{Risk}}_n(W)| > \delta\right) \\
&\quad = \int \mathbb{P}_X\left(|\,\mathrm{Risk}_n(U) - \overline{\mathrm{Risk}}_n^{\mathrm{nz}}(U)| > \delta\right) d\mu^{\mathrm{Haar}}_{d_n,p_n}(U) \\
&\quad \le \sup_{U\in\mathrm{St}(d_n,p_n)} \mathbb{P}_X\left(|\,\mathrm{Risk}_n(U) - \overline{\mathrm{Risk}}_n^{\mathrm{nz}}(U)| > \delta\right) \longrightarrow 0.
\end{aligned}$$

The same argument applies to the odd branch, using (S.3.258). Since the limiting risk weights are fixed and $\sup_n \|\beta^*\| < \infty$, (S.3.260) is $O(d^{-1})$. Moreover, $p/n \to c > 0$ and $p \le d$ imply $d \to \infty$. Chebyshev's inequality therefore gives

$$\overline{\mathrm{Risk}}_n(W) - \overline{\mathrm{Risk}}_n^{\mathrm{Haar}} \xrightarrow{\mathbb{P}} 0.$$

Combining the last two displays proves (S.3.261). □

S.3.8 Technical lemmas used in the supporting arguments

S.3.8.1 Dependent full-pair Gram and leave-two-out approximation

Lemma S.3.22 For $i \neq j$, retain the original n^{-2} normalization and define

$$S^{(ij)} = \frac{1}{n^2} \sum_{a,b\notin\{i,j\}} z_{ab} z_{ab}^{\mathrm{T}},$$

$$U_{ij} = \frac{1}{\sqrt{n}}[f_1(g_i), f_2(g_i), f_1(g_j), f_2(g_j)],$$

Then

$$S = S^{(ij)} + U_{ij}U_{ij}^{\mathrm{T}} + R_{ij}, \qquad i \neq j, \tag{S.3.262}$$

and

$$\mathbb{E}\max_{i\neq j}\|R_{ij}\|^2 = o(1). \tag{S.3.263}$$

Proof. Write $g_i = W^{\mathrm{T}}x_i$. Then the vectors g_i are independent $\mathcal{N}(0, I_p)$.

For the exact leave-two-out decomposition of the Gram matrix, fix $E = \{i, j\}$ with $i \neq j$, put $E^c = [n] \setminus E$, and abbreviate

$$m_{u1} = f_1(g_u), \qquad m_{u2} = f_2(g_u), \qquad u \in [n].$$

For $u \in [n]$ and every $k \neq u$, define the conditionally centered vectors

$$\eta^{(1)}_{uk} = z_{uk} - m_{u1}, \qquad \eta^{(2)}_{uk} = z_{ku} - m_{u2}.$$

For every $k \neq u$,

$$\mathbb{E}[\eta^{(1)}_{uk} \mid g_u] = 0, \qquad \mathbb{E}[\eta^{(2)}_{uk} \mid g_u] = 0.$$

The rows omitted from $S^{(ij)}$ are precisely those with at least one endpoint in E, so

$$S - S^{(ij)} = \frac{1}{n^2}\sum_{u\in E}\sum_{k\in E^c}\{z_{uk}z_{uk}^{\mathrm{T}} + z_{ku}z_{ku}^{\mathrm{T}}\} + \frac{1}{n^2}\sum_{u,v\in E} z_{uv}z_{uv}^{\mathrm{T}}.$$

Substitute $z_{uk} = m_{u1} + \eta_{uk}^{(1)}$ and $z_{ku} = m_{u2} + \eta_{uk}^{(2)}$. Because $|E^c| = n - 2$ and

$$U_{ij}U_{ij}^{\mathrm{T}} = \frac{1}{n}\sum_{u\in E}\sum_{r=1}^{2} m_{ur}m_{ur}^{\mathrm{T}},$$

the coefficient of each $m_{ur}m_{ur}^{\mathrm{T}}$ in the remainder is

$$\frac{|E^c|}{n^2} - \frac{1}{n} = \frac{n-2}{n^2} - \frac{n}{n^2} = -\frac{2}{n^2}.$$

Hence

$$S = S^{(ij)} + U_{ij}U_{ij}^{\mathrm{T}} + R_{ij}, \tag{S.3.264}$$

where

$$\begin{aligned} R_{ij} = {} & \frac{1}{n^2}\sum_{u\in E}\sum_{k\in E^c}\Big[m_{u1}(\eta_{uk}^{(1)})^{\mathrm{T}} + \eta_{uk}^{(1)}m_{u1}^{\mathrm{T}} + \eta_{uk}^{(1)}(\eta_{uk}^{(1)})^{\mathrm{T}}\Big] \\ & + \frac{1}{n^2}\sum_{u\in E}\sum_{k\in E^c}\Big[m_{u2}(\eta_{uk}^{(2)})^{\mathrm{T}} + \eta_{uk}^{(2)}m_{u2}^{\mathrm{T}} + \eta_{uk}^{(2)}(\eta_{uk}^{(2)})^{\mathrm{T}}\Big] \\ & - \frac{2}{n^2}\sum_{u\in E}\sum_{r=1}^{2} m_{ur}m_{ur}^{\mathrm{T}} + \frac{1}{n^2}\sum_{u,v\in E} z_{uv}z_{uv}^{\mathrm{T}}. \end{aligned} \tag{S.3.265}$$

Fix (u, r) and condition on $\{g_u : u \in E\}$. Each $\eta_{uk}^{(r)}$ is then a measurable function of (g_u, g_k), so the independence of $(g_k)_{k\in E^c}$ implies that the family $\{\eta_{uk}^{(r)} : k \in E^c\}$ is conditionally independent. Variables sharing the same k but having different (u, r) may still be dependent. We therefore apply each Bernstein inequality separately for fixed (u, r). Put

$$\mathcal{G}_u = \sigma(g_u), \qquad \mathbb{E}_u[\,\cdot\,] = \mathbb{E}[\,\cdot \mid \mathcal{G}_u], \qquad \mathbb{P}_u(\,\cdot\,) = \mathbb{P}(\,\cdot \mid \mathcal{G}_u).$$

Let L_σ be a Lipschitz constant of σ. Jensen's inequality gives

$$|f_1(x) - f_1(x')| \le \lambda L_\sigma |x - x'|, \qquad |f_2(x) - f_2(x')| \le (1-\lambda)L_\sigma |x - x'|.$$

Put $L_1 = \lambda L_\sigma$ and $L_2 = (1-\lambda)L_\sigma$. For independent standard Gaussians G_1, G_2, the tower property gives

$$\mathbb{E}f_1(G_1) = \mathbb{E}\sigma(\lambda G_1 + (1-\lambda)G_2) = \beta_\sigma, \qquad \mathbb{E}f_2(G_2) = \beta_\sigma.$$

Thus, for each fixed (u, r), the coordinates of $m_{ur} - \beta_\sigma \mathbf{1}$ are independent and centered. They have uniformly bounded sub-Gaussian norms because f_r is L_r-Lipschitz on a standard Gaussian input.

To control $\|m_{ur}\|$, define

$$\Psi_r(g) = \|f_r(g)\|, \qquad g \in \mathbb{R}^p,$$

where f_r acts coordinatewise. The reverse triangle inequality and the coordinatewise Lipschitz bound imply

$$|\Psi_r(g) - \Psi_r(g')| \le \|f_r(g) - f_r(g')\| \le L_r\|g - g'\|.$$

Moreover, conditional Jensen's inequality gives

$$\mathbb{E}f_r(G)^2 \le \mathbb{E}\sigma(\lambda G_1 + (1-\lambda)G_2)^2 = \nu_\alpha,$$

with the roles of G_1, G_2 interchanged when $r = 2$. Hence

$$\mathbb{E}\Psi_r(g_u) \le \{\mathbb{E}\Psi_r(g_u)^2\}^{1/2} = \{p\,\mathbb{E}f_r(G)^2\}^{1/2} \le C\sqrt{p}.$$

The Gaussian concentration inequality for the Lipschitz map Ψ_r therefore gives, for every $t \ge 0$,

$$\mathbb{P}(\|m_{ur}\| > C\sqrt{p} + t) \le 2\exp(-ct^2).$$

Taking $t = C\sqrt{r_{\mathrm{tail}}}$ yields

$$\mathbb{P}\Big(\|m_{ur}\| > C\{\sqrt{p} + \sqrt{r_{\mathrm{tail}}}\}\Big) \le 2e^{-cr_{\mathrm{tail}}}.$$

Taking $r_{\mathrm{tail}} = C_A \log n$, with $cC_A \ge A + 4$, and applying the union bound to the $2n$ pairs (u, r), we obtain, after enlarging C_A,

$$\mathbb{P}\left(\max_{u\le n,\, r\le 2}\|m_{ur}\| > C_A\sqrt{p + \log n}\right) \le 4n^{1-cC_A} \le n^{-A}.$$

This is the first estimate in (S.3.272) below.

For fixed (u, r), condition on g_u and truncate

$$\widetilde{\eta}_{uk}^{(r)} = \eta_{uk}^{(r)} \mathbf{1}_{\{\|\eta_{uk}^{(r)}\| \le B_n\}} - \mathbb{E}_u\Big[\eta_{uk}^{(r)} \mathbf{1}_{\{\|\eta_{uk}^{(r)}\| \le B_n\}}\Big], \qquad B_n = K_A \sqrt{p \log n}.$$

For $u \neq k$ and $r \le 2$, set

$$T_{uk}^{(r)} = \{\|\eta_{uk}^{(r)}\| > B_n\}, \qquad \pi_n^{\mathrm{tail}} = \max_{\substack{u \neq k \\ r \le 2}} \operatorname*{ess\,sup}_{g_u} \mathbb{P}\Big(T_{uk}^{(r)} \mid g_u\Big).$$

Conditionally on g_u, every coordinate of $\eta_{uk}^{(r)}$ is a centered, uniformly Lipschitz function of a standard Gaussian. Gaussian Poincaré and Gaussian moment bounds therefore give, uniformly in the conditioning value,

$$\mathbb{E}_u\|\eta_{uk}^{(r)}\|^2 \le Cp, \qquad \mathbb{E}_u\|\eta_{uk}^{(r)}\| \le C\sqrt{p}, \qquad \max_{\ell \le p} \mathbb{E}_u|\eta_{uk,\ell}^{(r)}|^q \le C_q \tag{S.3.266}$$

for every fixed q. In particular,

$$\mathbb{E}_u\|\eta_{uk}^{(r)}\|^4 \le p \sum_{\ell=1}^{p} \mathbb{E}_u|\eta_{uk,\ell}^{(r)}|^4 \le Cp^2.$$

Since the map from g_k to $\eta_{uk}^{(r)}$ has a uniformly bounded Euclidean Lipschitz constant, Gaussian concentration gives

$$\pi_n^{\mathrm{tail}} \le 2\exp\{-c(B_n - C\sqrt{p})^2\} \le n^{-2A-12}, \tag{S.3.267}$$

after choosing K_A sufficiently large. Consequently the probability that any of the $O(n^2)$ vectors is truncated is at most $2n^2\pi_n^{\mathrm{tail}} \le n^{-2A-9}$.

Fix (u, r). For almost every value of g_u, the variables $\{\widetilde{\eta}_{uk}^{(r)} : k \neq u\}$ are independent and centered under the conditional law $\mathbb{P}_u$. If

$$Y_k = \eta_{uk}^{(r)} \mathbf{1}_{\{\|\eta_{uk}^{(r)}\| \le B_n\}},$$

then

$$\begin{aligned}\mathbb{E}_u\|\widetilde{\eta}_{uk}^{(r)}\|^2 &= \mathbb{E}_u\|Y_k - \mathbb{E}_u Y_k\|^2 \\ &= \mathbb{E}_u\|Y_k\|^2 - \|\mathbb{E}_u Y_k\|^2 \\ &\le \mathbb{E}_u\|\eta_{uk}^{(r)}\|^2 \le Cp\end{aligned}$$

almost surely, uniformly in the conditioning value. Consequently the random conditional variance parameter

$$v_{\mathrm{vec}}(g_u) := \sum_{k \neq u} \mathbb{E}_u\|\widetilde{\eta}_{uk}^{(r)}\|^2$$

satisfies

$$v_{\mathrm{vec}}(g_u) \le Cnp \qquad \text{for almost every } g_u.$$

Moreover,

$$\|\widetilde{\eta}_{uk}^{(r)}\| \le \|Y_k\| + \|\mathbb{E}_u Y_k\| \le B_n + \mathbb{E}_u\|Y_k\| \le 2B_n$$

under the same conditional law.

Apply the independent centered real-Hilbert moment inequality (Pinelis, 1994, Theorem 5.3, equation (5.8)) for almost every g_u on the conditional probability space $(\mathbb{P}_u, \mathbb{E}_u)$, where independence and centering hold on each fiber. For $q \ge 2$,

$$\Big\|\sum_{k \neq u} \widetilde{\eta}_{uk}^{(r)}\Big\|_{L^q(\mathbb{P}_u)} \le C\left\{q\left(\sum_{k \neq u} \mathbb{E}_u\|\widetilde{\eta}_{uk}^{(r)}\|^q\right)^{1/q} + \sqrt{q}\, v_{\mathrm{vec}}(g_u)^{1/2}\right\}.$$

Since $\|\widetilde{\eta}_{uk}^{(r)}\| \le 2B_n$ and $v_{\mathrm{vec}}(g_u) \le Cnp$, this gives

$$\Big\|\sum_{k \neq u} \widetilde{\eta}_{uk}^{(r)}\Big\|_{L^q(\mathbb{P}_u)} \le C\{qn^{1/q}B_n + \sqrt{qnp}\}.$$

Take $q = \lceil (A+3)\log n \rceil$. Because $B_n = K_A\sqrt{p\log n}$ and $p \asymp n$, the right-hand side is at most $C_A\sqrt{np\log n}$. Markov's inequality at exponent q, after enlarging C_A, gives uniformly in the conditioning value

$$\mathbb{P}_u\left(\left\|\sum_{k\neq u}\widetilde{\eta}_{uk}^{(r)}\right\| > C_A\sqrt{np\log n}\right) \le n^{-A-2}.$$

The tower property and the union bound over the $2n$ pairs (u,r) therefore yield

$$\mathbb{P}\left(\max_{u\le n,\, r\le 2}\left\|\sum_{k\neq u}\widetilde{\eta}_{uk}^{(r)}\right\| > C_A\sqrt{np\log n}\right) \le 2n^{-A-1} \le n^{-A}. \tag{S.3.268}$$

Equation (S.3.268) is the truncated analogue of (S.3.273). Below we transfer it to the original $\eta_{uk}^{(r)}$.

For the matrix estimate, put

$$\widehat{X}_k = \eta_{uk}^{(r)}(\eta_{uk}^{(r)})^{\mathsf{T}}\mathbf{1}_{\{\|\eta_{uk}^{(r)}\|\le B_n\}} - \mathbb{E}_u\left[\eta_{uk}^{(r)}(\eta_{uk}^{(r)})^{\mathsf{T}}\mathbf{1}_{\{\|\eta_{uk}^{(r)}\|\le B_n\}}\right].$$

These matrices are conditionally independent, centered, and self-adjoint. To compute the matrix-variance parameter, abbreviate $\eta = \eta_{uk}^{(r)}$ and

$$Y = \eta\eta^{\mathsf{T}}\mathbf{1}_{\{\|\eta\|\le B_n\}}.$$

Since $\widehat{X}_k = Y - \mathbb{E}_u Y$,

$$0 \le \mathbb{E}_u\widehat{X}_k^2 = \mathbb{E}_u Y^2 - (\mathbb{E}_u Y)^2 \preceq \mathbb{E}_u Y^2.$$

For every deterministic unit vector h, the conditional coordinate independence and centering imply

$$\begin{aligned} h^{\mathsf{T}}\mathbb{E}_u Y^2 h &\le \mathbb{E}_u\left[\|\eta\|^2(h^{\mathsf{T}}\eta)^2\right] \\ &= \sum_{\ell=1}^{p}\sum_{q=1}^{p} h_q^2\,\mathbb{E}_u[\eta_\ell^2\eta_q^2] \le Cp. \end{aligned}$$

Indeed, all terms containing $h_q h_{q'}$ with $q \neq q'$ vanish, and the remaining expectations are bounded by the conditional moments in (S.3.266). Taking the supremum over h gives

$$\|\mathbb{E}_u\widehat{X}_k^2\| \le Cp.$$

Moreover,

$$\|\widehat{X}_k\| \le B_n^2 + \mathbb{E}_u\|\eta\|^2 \le CK_A^2 p\log n =: R_{\mathrm{mat}}.$$

Consequently,

$$v_{\mathrm{mat}} := \left\|\sum_{k\neq u}\mathbb{E}_u\widehat{X}_k^2\right\| \le Cnp.$$

The self-adjoint matrix Bernstein inequality (Tropp, 2012) now gives

$$\mathbb{P}_u\left(\left\|\sum_{k\neq u}\widehat{X}_k\right\| > r_{\mathrm{tail}}\right) \le 2p\exp\left[-c\min\left\{\frac{r_{\mathrm{tail}}^2}{v_{\mathrm{mat}}}, \frac{r_{\mathrm{tail}}}{R_{\mathrm{mat}}}\right\}\right]. \tag{S.3.269}$$

After K_A has been fixed in the truncation step, take $r_{\mathrm{tail}} = L_A n(\log n)^2$, with L_A sufficiently large. Since $p/n \to c$, the right-hand side is then at most n^{-A-2}.

The following identities compare the truncated and untruncated sums. Since $\mathbb{E}_u\eta_{uk}^{(r)} = 0$, their vector difference is exactly

$$\sum_{k\neq u}\eta_{uk}^{(r)} - \sum_{k\neq u}\widetilde{\eta}_{uk}^{(r)} = \sum_{k\neq u}\left\{\eta_{uk}^{(r)}\mathbf{1}_{T_{uk}^{(r)}} - \mathbb{E}_u[\eta_{uk}^{(r)}\mathbf{1}_{T_{uk}^{(r)}}]\right\}. \tag{S.3.270}$$

Similarly, writing $\Sigma_{uk}^{(r)} = \mathbb{E}_u[\eta_{uk}^{(r)}(\eta_{uk}^{(r)})^{\mathsf{T}}]$, we have

$$\begin{aligned} &\sum_{k\neq u}\{\eta_{uk}^{(r)}(\eta_{uk}^{(r)})^{\mathsf{T}} - \Sigma_{uk}^{(r)}\} - \sum_{k\neq u}\widehat{X}_k \\ &\qquad = \sum_{k\neq u}\left\{\eta_{uk}^{(r)}(\eta_{uk}^{(r)})^{\mathsf{T}}\mathbf{1}_{T_{uk}^{(r)}} - \mathbb{E}_u[\eta_{uk}^{(r)}(\eta_{uk}^{(r)})^{\mathsf{T}}\mathbf{1}_{T_{uk}^{(r)}}]\right\}. \end{aligned} \tag{S.3.271}$$

Let $\mathcal{T}_n = \bigcup_{u\neq k,\, r\le 2} T_{uk}^{(r)}$. By (S.3.267),

$$\mathbb{P}(\mathcal{T}_n) \le 2n^2\pi_n^{\mathrm{tail}} \le n^{-2A-9}.$$

On $\mathcal{T}_n^c$ the random tail terms on the right-hand sides of (S.3.270) and (S.3.271) vanish. By Cauchy–Schwarz and the uniform second and fourth moments,

$$\begin{aligned}\left\|\mathbb{E}_u[\eta_{uk}^{(r)}\mathbf{1}_{T_{uk}^{(r)}}]\right\| &\le (\mathbb{E}_u\|\eta_{uk}^{(r)}\|^2)^{1/2}(\pi_n^{\mathrm{tail}})^{1/2} \le C\sqrt{p}\,(\pi_n^{\mathrm{tail}})^{1/2},\\ \left\|\mathbb{E}_u[\eta_{uk}^{(r)}(\eta_{uk}^{(r)})^{\mathrm{T}}\mathbf{1}_{T_{uk}^{(r)}}]\right\| &\le \mathbb{E}_u[\|\eta_{uk}^{(r)}\|^2\mathbf{1}_{T_{uk}^{(r)}}]\\ &\le (\mathbb{E}_u\|\eta_{uk}^{(r)}\|^4)^{1/2}(\pi_n^{\mathrm{tail}})^{1/2} \le Cp\,(\pi_n^{\mathrm{tail}})^{1/2}.\end{aligned}$$

Consequently, uniformly in (u,r) and on $\mathcal{T}_n^c$,

$$\begin{aligned}\left\|\sum_{k\neq u}\eta_{uk}^{(r)} - \sum_{k\neq u}\widetilde{\eta}_{uk}^{(r)}\right\| &\le Cn\sqrt{p}\,(\pi_n^{\mathrm{tail}})^{1/2} = o(n^{-A}),\\ \left\|\sum_{k\neq u}\{\eta_{uk}^{(r)}(\eta_{uk}^{(r)})^{\mathrm{T}} - \Sigma_{uk}^{(r)}\} - \sum_{k\neq u}\widehat{X}_k\right\| &\le Cnp\,(\pi_n^{\mathrm{tail}})^{1/2} = o(n^{-A}).\end{aligned}$$

The preceding high-probability bound for the truncated vectors, (S.3.268), and the conditional matrix Bernstein bound (S.3.269) transfer to the corresponding untruncated variables. Therefore, for every fixed $A > 0$, the following bounds hold with probability at least $1 - n^{-A}$:

$$\max_{u\le n,\, r\le 2}\|m_{ur}\| \le C_A\sqrt{p+\log n}, \tag{S.3.272}$$

$$\max_{u\le n,\, r\le 2}\left\|\sum_{k\neq u}\eta_{uk}^{(r)}\right\| \le C_A\sqrt{np}\,(\log n)^{1/2}, \tag{S.3.273}$$

$$\max_{u\le n,\, r\le 2}\left\|\sum_{k\neq u}\left\{\eta_{uk}^{(r)}(\eta_{uk}^{(r)})^{\mathrm{T}} - \mathbb{E}_u[\eta_{uk}^{(r)}(\eta_{uk}^{(r)})^{\mathrm{T}}]\right\}\right\| \le C_A n(\log n)^2. \tag{S.3.274}$$

Here the notation covers both orientations of an endpoint. Conditional on g_u, distinct feature coordinates of $\eta_{uk}^{(r)}$ are independent, so its covariance matrix is diagonal. For each diagonal entry, the scalar Gaussian Poincare inequality and Lipschitz continuity give

$$\mathrm{Var}_u(\eta_{uk,\ell}^{(r)}) \le L_\sigma^2\max\{\lambda^2,(1-\lambda)^2\}.$$

Consequently

$$\max_{\substack{u\neq k\\ r\le 2}}\left\|\mathbb{E}_u[\eta_{uk}^{(r)}(\eta_{uk}^{(r)})^{\mathrm{T}}]\right\| \le C. \tag{S.3.275}$$

For (S.3.265), set

$$\begin{aligned}M_n &= \max_{u,r}\|m_{ur}\|,\\ V_n &= \max_{u,r}\left\|\sum_{k\neq u}\eta_{uk}^{(r)}\right\|,\\ K_n &= \max_{u,r}\left\|\sum_{k\neq u}\{\eta_{uk}^{(r)}(\eta_{uk}^{(r)})^{\mathrm{T}} - \Sigma_{uk}^{(r)}\}\right\|,\\ Z_n &= \max_{u,v}\|z_{uv}\|, \qquad H_n^{\max} = \max_{u\neq k,r}\|\eta_{uk}^{(r)}\|,\end{aligned}$$

where $\Sigma_{uk}^{(r)} = \mathbb{E}_u[\eta_{uk}^{(r)}(\eta_{uk}^{(r)})^{\mathrm{T}}]$. Gaussian concentration and a union bound over the n^2 ordered pairs show that, for every fixed $A > 0$, with probability at least $1 - n^{-A}$,

$$Z_n^2 \le C_A(p+\log n)\log n.$$

Since $\eta_{uk}^{(1)} = z_{uk} - m_{u1}$ and $\eta_{uk}^{(2)} = z_{ku} - m_{u2}$,

$$H_n^{\max} \le Z_n + M_n \le C_A\sqrt{p\log n}.$$

On the common high-probability event given by (S.3.272)–(S.3.275), we therefore have

$$\begin{aligned} M_n &\le C_A\sqrt{p}, & V_n &\le C_A\sqrt{np\log n}, \quad K_n \le C_A n(\log n)^2, \\ Z_n^2 &\le C_A(p+\log n)\log n, & H_n^{\max} &\le C_A\sqrt{p\log n}. \end{aligned} \tag{S.3.276}$$

Fix $E=\{i,j\}$ and $u\in E$, and let v be the other element of E. Because $E^c=[n]\setminus\{u,v\}$,

$$\sum_{k\in E^c}\eta_{uk}^{(r)} = \sum_{k\neq u}\eta_{uk}^{(r)} - \eta_{uv}^{(r)}.$$

Consequently,

$$\max_{E,u,r}\left\|\sum_{k\in E^c}\eta_{uk}^{(r)}\right\| \le V_n + H_n^{\max}. \tag{S.3.277}$$

For the quadratic term, first center it:

$$\sum_{k\in E^c}\eta_{uk}^{(r)}(\eta_{uk}^{(r)})^{\mathrm{T}} = \sum_{k\in E^c}\{\eta_{uk}^{(r)}(\eta_{uk}^{(r)})^{\mathrm{T}} - \Sigma_{uk}^{(r)}\} + \sum_{k\in E^c}\Sigma_{uk}^{(r)}. \tag{S.3.278}$$

Removing the single index v from the first sum in (S.3.278) and using (S.3.275) gives

$$\begin{aligned} \left\|\sum_{k\in E^c}\{\eta_{uk}^{(r)}(\eta_{uk}^{(r)})^{\mathrm{T}} - \Sigma_{uk}^{(r)}\}\right\| &\le K_n + (H_n^{\max})^2 + C, \\ \left\|\sum_{k\in E^c}\Sigma_{uk}^{(r)}\right\| &\le Cn. \end{aligned} \tag{S.3.279}$$

Apply $\|ab^{\mathrm{T}}\| = \|a\|\,\|b\|$ to the two cross terms in (S.3.265). There are two endpoints and two orientations, so (S.3.276), (S.3.277), and (S.3.279) give, uniformly in $i\neq j$,

$$\begin{aligned} \|R_{ij}\| &\le \frac{C}{n^2}\Big[M_n(V_n+H_n^{\max}) + K_n + (H_n^{\max})^2 + n + M_n^2 + Z_n^2\Big] \\ &\le C_A\left\{\sqrt{\frac{\log n}{n}} + \frac{(\log n)^2}{n} + \frac{\log n}{n}\right\}. \end{aligned}$$

Since $p/n\to c$,

$$\max_{i\neq j}\|R_{ij}\| \le C_A\left\{\sqrt{\frac{\log n}{n}} + \frac{(\log n)^2}{n}\right\} = o(1) \tag{S.3.280}$$

on an event whose probability tends to one, proving (S.3.262).

To prove the L^2 bound, use the exact identity (S.3.264). Since at most $4n$ ordered rows are removed, it gives the deterministic estimate

$$\max_{i\neq j}\|R_{ij}\| \le \max_{i\neq j}\|S - S^{(ij)}\| + \max_{i\neq j}\|U_{ij}U_{ij}^{\mathrm{T}}\| \le \frac{4}{n}\left(Z_n^2 + M_n^2\right).$$

For each fixed ordered pair, the map from its one or two Gaussian endpoint vectors to $\|z_{uv}\|$ is uniformly Lipschitz and has mean at most $C\sqrt{p}$; the same holds for $\|m_{ur}\|$. Gaussian concentration and a union bound therefore give, for $x\ge 0$,

$$\begin{aligned} \mathbb{P}\Big[Z_n > C\{\sqrt{p} + \sqrt{\log n} + \sqrt{x}\}\Big] &\le Ce^{-cx}, \\ \mathbb{P}\Big[M_n > C\{\sqrt{p} + \sqrt{\log n} + \sqrt{x}\}\Big] &\le Ce^{-cx}. \end{aligned}$$

Using

$$\mathbb{E}X^{2q} = 2q\int_0^\infty t^{2q-1}\mathbb{P}(X>t)\,dt$$

and splitting the integral at $C(\sqrt{p}+\sqrt{\log n})$ yields, for every fixed integer q,

$$\mathbb{E}Z_n^{2q} + \mathbb{E}M_n^{2q} \le C_q n^q(\log n)^{C_q}.$$

In particular,

$$\mathbb{E}\left(\max_{i\neq j}\|R_{ij}\|\right)^4 \le C(\log n)^C.$$

Let $\mathcal{E}_{n,A}$ be the event in (S.3.280), chosen so that $\mathbb{P}(\mathcal{E}_{n,A}^c) \le n^{-A}$, and put

$$r_n = C_A\left\{\sqrt{\frac{\log n}{n}} + \frac{(\log n)^2}{n}\right\} = o(1).$$

Then Cauchy–Schwarz gives

$$\begin{aligned}\mathbb{E}\max_{i\neq j}\|R_{ij}\|^2 &\le r_n^2 + \left\{\mathbb{E}(\max_{i\neq j}\|R_{ij}\|)^4\right\}^{1/2}\mathbb{P}(\mathcal{E}_{n,A}^c)^{1/2}\\ &\le r_n^2 + C(\log n)^C n^{-A/2} = o(1),\end{aligned}$$

after choosing A sufficiently large. This proves (S.3.263).

□

S.3.8.2 Weighted diagonal self-averaging

Lemma S.3.23 Let Γ be fixed, let $C_\omega \in \mathbb{C}^{2\times 2}$ be deterministic, and let $C_{i,\omega} \in \mathbb{C}^{2\times 2}$ be measurable with respect to a common sigma-field $\mathcal{F}_n$. Suppose that

$$\sup_{\omega\in\Gamma}\left\{\|C_\omega\| + \max_i \|C_{i,\omega}\|_{L^4}\right\} \le C, \qquad \sup_{\omega\in\Gamma}\frac{1}{n}\sum_i \|C_{i,\omega} - C_\omega\|_{L^4} \le r_n \longrightarrow 0. \tag{S.3.281}$$

Let $\upsilon_{1,n},\dots,\upsilon_{n,n}$ be independent of $\mathcal{F}_n$ and of one another, centered, and satisfy

$$d_{i,n} := \mathbb{E}\upsilon_{i,n}^2, \qquad \max_i \mathbb{E}|\upsilon_{i,n}|^4 \le C.$$

Writing $\bar d_n = n^{-1}\sum_i d_{i,n}$, for every deterministic bounded $w \in \mathbb{R}^2$,

$$\sup_{\omega\in\Gamma}\left\|\frac{1}{n}\sum_i d_{i,n}w^{\mathrm{T}}(I_2 - C_{i,\omega})w - \bar d_n w^{\mathrm{T}}(I_2 - C_\omega)w\right\|_{L^2} \le Cr_n, \tag{S.3.282}$$

$$\sup_{\omega\in\Gamma}\left\|\frac{1}{n}\sum_i \upsilon_{i,n}^2 w^{\mathrm{T}}(I_2 - C_{i,\omega})w - \bar d_n w^{\mathrm{T}}(I_2 - C_\omega)w\right\|_{L^2} \le C(r_n + n^{-1/2}). \tag{S.3.283}$$

Proof. For the deterministic assertion, subtract the stated centering and use

$$\frac{1}{n}\sum_i d_{i,n} = \bar d_n.$$

Minkowski's inequality and $d_{i,n} \le (\mathbb{E}|\upsilon_{i,n}|^4)^{1/2} \le C$ give

$$\left\|\frac{1}{n}\sum_i d_{i,n}w^{\mathrm{T}}(C_\omega - C_{i,\omega})w\right\|_{L^2} \le \frac{C}{n}\sum_i \|C_{i,\omega} - C_\omega\|_{L^2} \le Cr_n.$$

For the random assertion, put $a_{i,\omega} = w^{\mathrm{T}}(I_2 - C_{i,\omega})w$. Conditional on $\mathcal{F}_n$, independence and centering give

$$\mathbb{E}_\upsilon\left|\frac{1}{n}\sum_i(\upsilon_{i,n}^2 - d_{i,n})a_{i,\omega}\right|^2 = \frac{1}{n^2}\sum_i \mathrm{Var}(\upsilon_{i,n}^2)|a_{i,\omega}|^2.$$

After taking expectation, the fourth-moment hypothesis and (S.3.281) bound the right-hand side by C/n. Combining this with (S.3.282) proves (S.3.283). □

S.4 Proofs of the Mechanism Results

We first prove Proposition 1.1 from Section 1, then the critical-coordinate identity, oracle comparisons, and finite-sample transfer from Section 4. The arguments cover calibration, population influence, strict oracle separation, and theorem-matched diagnostics.

S.4.1 Linear-model baseline

Proof of Proposition 1.1. Full row rank makes $Xb = y$ feasible for every y. Any such b also fits every mixed pair, since

$$(\lambda x_i + (1-\lambda)x_j)^{\mathrm{T}} b = \lambda y_i + (1-\lambda) y_j.$$

Conversely, the full-pair data include the diagonal pairs $i = j$, whose residuals are $x_i^{\mathrm{T}} b - y_i$. Thus, any predictor with zero mixup loss also satisfies $Xb = y$. Both least-squares objectives attain zero loss and have the same minimizer set $\{b \in \mathbb{R}^d : Xb = y\}$. Its unique element of minimum Euclidean norm is

$$\widehat{\beta}_{\mathrm{mix}}(\lambda) = \widehat{\beta}_{\mathrm{ridgeless}} = X^{\dagger} y = X^{\mathrm{T}}(XX^{\mathrm{T}})^{-1} y.$$

The claimed equality of predictions and risks follows. □

S.4.2 Finite-sample transfer

We first fix the critical-coordinate model and the notation used in the oracle comparisons below. Write $i_n := i_*$ for the deterministic distinguished index in Example 2. We consider the slightly more general variance profile

$$\sigma_{n,i}^2 = \sigma_\varepsilon^2 + n u_n \mathbf{1}_{\{i=i_n\}}, \qquad u_n \geq 0, \quad u_n \longrightarrow u \geq 0, \qquad \sigma_{\mathrm{test}}^2 = \sigma_\varepsilon^2. \tag{S.4.1}$$

The example in the main text corresponds to $u_n = \tau^2$. The noise scales are deterministic, as in Definition 2.1. Set $\rho_{\perp,n}^{2,(0)} = \|\beta_\perp^*\|^2 + \sigma_\varepsilon^2$.

The exact finite-sample identities below do not require the background energies to converge. For fixed limiting risk curves and their oracles, we also assume

$$\|a\|^2 \longrightarrow r_a^2, \qquad \rho_{\perp,n}^{2,(0)} \longrightarrow \rho_{\perp,\infty}^{2,(0)}. \tag{S.4.2}$$

Then $N_n := \|\beta^*\|^2 + \sigma_\varepsilon^2$ converges to $N = r_a^2 + \rho_{\perp,\infty}^{2,(0)}$. For positive ridge, define

$$\begin{gathered} B_\lambda(\gamma) := \mathcal{E}(\lambda,\gamma) r_a^2 + \{\chi(\lambda,\gamma)+1\}\rho_{\perp,\infty}^{2,(0)}, \qquad J_\lambda(\gamma) := \chi(\lambda,\gamma) = \nu\ell, \\ R_\lambda(\gamma;u) := B_\lambda(\gamma) + u J_\lambda(\gamma). \end{gathered} \tag{S.4.3}$$

In the oracle comparisons, r_{eff}^2 is the background effective residual energy: $r_{\mathrm{eff}}^2 = \rho_{\perp,\infty}^{2,(0)} + (1-\mu_1^2/\nu) r_a^2$ when $\nu > 0$. Define

$$D_\lambda = \begin{cases} [0,\infty), & \lambda \in (0,1),\ \sigma \text{ nonaffine}, \\ [0,\infty), & \lambda = 1,\ c \neq 1,\ \nu > 0, \\ (0,\infty), & \text{otherwise}, \end{cases} \tag{S.4.4}$$

$$F_\lambda(u) := \inf_{\gamma \in D_\lambda} R_\lambda(\gamma;u), \qquad F_{\mathrm{ridge}}(u) := F_1(u).$$

Under (S.4.2), the moving equivalents in the main text converge to these fixed curves.

For $s \in \{\mathrm{mix}, \mathrm{end}\}$, let $\lambda_{\mathrm{mix}} = 1/2$ and $\lambda_{\mathrm{end}} = 1$. We use

$$D_s := D_{\lambda_s}, \qquad R_s(\gamma;u) := R_{\lambda_s}(\gamma;u), \qquad F_s(u) := F_{\lambda_s}(u),$$

and abbreviate $R_{\mathrm{mix}} = R_{1/2}$, $R_{\mathrm{end}} = R_1$, $F_{\mathrm{mix}} = F_{1/2}$, $F_{\mathrm{end}} = F_1$, $D_{\mathrm{mix}} = D_{1/2}$, and $D_{\mathrm{end}} = D_1$. The symbol $\ell_{s,\gamma}$ denotes the coefficient ℓ from Section S.1.3 evaluated at (λ_s, γ); when $0 \in D_s$, $\ell_{s,0}$ denotes its finite right limit.

For any $\lambda \in [1/2, 1]$, let $\overline{\mathrm{Risk}}_{\lambda,n}(\gamma; u_n)$ denote the conditional prediction risk $\mathrm{Risk}(\lambda,\gamma)$ under (S.4.1), averaged over the training noises as in Definition 2.4. At $\gamma = 0$, use the minimum-norm Moore–Penrose estimator. For a fixed method s, write $\overline{\mathrm{Risk}}_{s,n} := \overline{\mathrm{Risk}}_{\lambda_s,n}$.

Pointwise deterministic equivalents are not enough to justify tuning at finite n. Appendix S.4.3.1 proves compact containment and uniform convergence, including the possibility that the limiting ridge minimizer is not unique.

Corollary S.4.1 (Finite-sample oracle separation) Fix a finite $u \geq 0$ such that

$$\Delta(u) = F_1(u) - F_{1/2}(u) > 0,$$

let $u_n \to u$, and assume that Theorem S.4.16 applies at u to both the endpoint and midpoint methods. Define the finite-sample oracle values averaged over training noise

$$O_{\mathrm{mid},n} = \inf_{\gamma \geq 0} \overline{\mathrm{Risk}}_{1/2,n}(\gamma; u_n), \qquad O_{\mathrm{end},n} = \inf_{\gamma \geq 0} \overline{\mathrm{Risk}}_{1,n}(\gamma; u_n),$$

$$O_{\mathrm{joint},n} = \inf_{\lambda \in [1/2,1],\, \gamma \geq 0} \overline{\mathrm{Risk}}_{\lambda,n}(\gamma; u_n).$$

Then

$$O_{\mathrm{end},n} - O_{\mathrm{mid},n} \xrightarrow{\mathbb{P}} \Delta(u),$$

and

$$\mathbb{P}\left[O_{\mathrm{joint},n} < O_{\mathrm{end},n} - \frac{\Delta(u)}{2} \right] \longrightarrow 1. \tag{S.4.5}$$

S.4.3 Critical-coordinate identity and finite-sample oracle transfer

Take the baseline labels $y_i^{(0)} = x_i^{\mathrm{T}}\beta^* + \sigma_\varepsilon \xi_i$. Write $y_{i,n}^{\mathrm{sp}} = x_i^{\mathrm{T}}\beta^* + \sigma_{n,i}\xi_i$ and $y_{i,n}^{(0)} = y_i^{(0)}$. For fixed λ, γ, denote the test risks conditional on the realized spiked and baseline labels, evaluated with the same test-noise variance, by $\mathrm{Risk}_n^{\mathrm{real,sp}}$ and $\mathrm{Risk}_n^{\mathrm{real},(0)}$, respectively, and their training noise averages by $\mathrm{Risk}_n^{\mathrm{sp}}$ and $\mathrm{Risk}_n^{(0)}$. Thus $\mathrm{Risk}_n^{\mathrm{sp}} = \overline{\mathrm{Risk}}_{\lambda,n}(\gamma; u_n)$. For the spiked sample, we use the assumptions of Corollary 4.2 without requiring (S.3.10). Put

$$r^{(0)} := \frac{1}{n}\sum_{i=1}^{n} \phi_i y_i^{(0)}, \qquad \widehat{\beta}^{(0)} := Q r^{(0)}.$$

Let $\Phi \in \mathbb{R}^{n \times p}$ have row i equal to ϕ_i^{T}. We use H_σ and the realized-risk notation from (S.1.4).

Proposition S.4.2 (Exact one-coordinate influence identity) Let $b_n = \sigma_{n,i_n} - \sigma_\varepsilon$, let r^{sp} be the score formed from (S.4.1), and put $\widehat{\beta}^{\mathrm{sp}} = Q r^{\mathrm{sp}}$. Then, exactly at every n,

$$r^{\mathrm{sp}} - r^{(0)} = \frac{b_n \xi_{i_n}}{n}\phi_{i_n}, \qquad \widehat{\beta}^{\mathrm{sp}} - \widehat{\beta}^{(0)} = \frac{b_n \xi_{i_n}}{n} Q\phi_{i_n}. \tag{S.4.6}$$

For

$$D_{i,n} := \frac{1}{\sqrt{n}}\phi_i^{\mathrm{T}} Q\{H_\sigma \widehat{\beta}^{(0)} - \mu_1 a\},$$

$$L_{i,n} := \frac{1}{n}\phi_i^{\mathrm{T}} Q H_\sigma Q \phi_i,$$

the realized test risks of the two fitted coefficients, evaluated under any common finite test-noise variance, satisfy

$$\mathrm{Risk}_n^{\mathrm{real,sp}} - \mathrm{Risk}_n^{\mathrm{real},(0)} = 2\frac{b_n}{\sqrt{n}}\xi_{i_n} D_{i_n,n} + \frac{b_n^2}{n}\xi_{i_n}^2 L_{i_n,n}. \tag{S.4.7}$$

Proof. Only the coefficient of y_{i_n} changes, so (S.3.199) gives the score identity. For any coefficient b, Gaussian integration by parts gives the exact test-risk functional

$$\mathcal{E}_n(b; \sigma_{\mathrm{test}}^2) = b^{\mathrm{T}} H_\sigma b - 2\mu_1 a^{\mathrm{T}} b + \|\beta^*\|^2 + \sigma_{\mathrm{test}}^2.$$

Substituting the coefficient increment from (S.4.6) and expanding this quadratic proves (S.4.7). □

The fixed-index argument uses the finite-dimensional coefficients $t_{*,n}, \tau_n, \ell_n$ from (S.1.11).

Lemma S.4.3 (Fixed-index local susceptibility) Retain the design assumptions and Assumption 1. Let σ be odd and Lipschitz, fix $\lambda \in (0, 1]$ and $\gamma > 0$, and let i_n be a deterministic index sequence. Then

$$\frac{1}{n}\phi_{i_n}^{\mathrm{T}} Q^2 \phi_{i_n} - \ell_n \xrightarrow{L^2} 0, \qquad \frac{1}{n}\phi_{i_n}^{\mathrm{T}} Q^2 \phi_{i_n} \xrightarrow{L^2} \ell. \tag{S.4.8}$$

Consequently,

$$L_{i_n,n} \xrightarrow{L^2} \nu\ell. \tag{S.4.9}$$

Proof. Oddness and Gaussian symmetry give $\beta_\sigma = \mu_0 = 0$. The endpoint conditional means f_1, f_2 are odd and centered. Put

$$F_i = [f_1(g_i), f_2(g_i)], \qquad \psi_i = F_i w, \qquad M_n = \frac{1}{n}\sum_{i=1}^{n} F_i.$$

For $i \neq j$, define

$$\eta_{ij} = z_{ij} - f_1(g_i) - f_2(g_j).$$

Then

$$\mathbb{E}(\eta_{ij} \mid g_i) = \mathbb{E}(\eta_{ij} \mid g_j) = 0.$$

Thus the three terms in the scalar decomposition

$$\sigma\{\lambda G_1 + (1-\lambda)G_2\} = f_1(G_1) + f_2(G_2) + h_0(G_1, G_2)$$

are pairwise orthogonal, and

$$\kappa(\lambda) = \nu_\alpha - \operatorname{tr} V = \mathbb{E}h_0(G_1, G_2)^2 \geq 0.$$

Write $\phi_i = F_i w + a_{i,n}$. Expanding the definition of ϕ_i, including its diagonal summand, gives

$$\begin{aligned} a_{i,n} = & -\frac{1}{n}F_i w + \frac{\lambda}{n}\sum_{j\neq i} f_2(g_j) + \frac{1-\lambda}{n}\sum_{j\neq i} f_1(g_j) \\ & + \frac{1}{n}\sum_{j\neq i}\{\lambda\eta_{ij} + (1-\lambda)\eta_{ji}\} + \frac{1}{n}z_{ii}. \end{aligned}$$

Consequently,

$$\phi_i = \psi_i + a_{i,n}, \qquad \sup_i \|a_{i,n}\|_{L^{16}} = O(1), \qquad \sup_i \|\psi_i\|_{L^{16}} = O(\sqrt{p}). \tag{S.4.10}$$

For a fixed endpoint i, the canonical part of $a_{i,n}$ is an n^{-1} sum over the partner index. Its sixteenth moment has n^{-16} normalization. Canonicality makes a term vanish whenever a partner index occurs exactly once, so every surviving tuple contains at most eight distinct partner indices. There are therefore $O(n^8)$ surviving tuples. Hölder's inequality and the fixed Gaussian moments give scalar L^{16}-norm $O(n^{-1/2})$. The first-order endpoint averages have the same scalar order. Minkowski's inequality over $p = O(n)$ coordinates gives the asserted vector L^{16}-bound $O(1)$.

For the full empirical mean, (S.3.37) gives

$$\widehat{z} = M_n \mathbf{1}_2 + \epsilon_{m,n}, \qquad \|\epsilon_{m,n}\|_{L^{16}} = O(n^{-1/2}).$$

The centered endpoint averages satisfy $\|M_n\|_{L^{16},\mathrm{op}} = O(1)$ by the vector Rosenthal inequality and $p = O(n)$. Hence

$$\|M_n\|_{L^{16},\mathrm{op}} + \|\widehat{z}\|_{L^{16}} = O(1).$$

Define

$$\check{C}_n = \kappa(\lambda) I_p + \frac{1}{n}\sum_i F_i F_i^{\mathrm{T}}, \qquad C_n^\circ = \check{C}_n - M_n M_n^{\mathrm{T}}.$$

Use the fixed Cauchy circle Γ from the centered cavity argument leading to (S.3.31), on which $\inf_{\omega\in\Gamma} \operatorname{Re}(\gamma + \omega) > 0$, and put

$$\check{R}_n(\omega) = \{\check{C}_n + (\gamma+\omega) I_p\}^{-1}, \qquad R_n(\omega) = \{C_n^\circ + (\gamma+\omega) I_p\}^{-1},$$

$$Q_I(\omega) = \{S + (\gamma + \omega)I_p\}^{-1}.$$

The centered Gram remainder (S.3.41) gives

$$S = C_n^\circ + \widetilde{zz}^{\mathrm{T}} + E_{S,n}, \qquad \|E_{S,n}\|_{L^4,\mathrm{op}} \le Cn^{-1/2}(\log n)^C.$$

For a fixed index i, let

$$\check{R}_i(\omega) = \left\{\kappa(\lambda)I_p + \frac{1}{n}\sum_{k\neq i} F_k F_k^{\mathrm{T}} + (\gamma+\omega)I_p\right\}^{-1},$$

$$T_i(\omega) = \frac{1}{n}F_i^{\mathrm{T}}\check{R}_i(\omega)F_i, \qquad C_i(\omega) = \{I_2 + T_i(\omega)\}^{-1}.$$

Let $t_n(\omega)$ be the analytic continuation of the finite reduced root and set

$$C_{2,n}(\omega) = \{I_2 + t_n(\omega)V\}^{-1}.$$

These are the same deleted-observation matrices as in (S.3.31), with $C_{2,n}(\omega) = C_{*,n}(\omega)$. That estimate gives

$$\sup_{\omega\in\Gamma}\max_i \|C_i(\omega) - C_{2,n}(\omega)\|_{L^8} = O(n^{-1/2}). \tag{S.4.11}$$

The reverse Woodbury identities are

$$R_n = \check{R}_n + \check{R}_n M_n J_n M_n^{\mathrm{T}}\check{R}_n, \qquad J_n = I_2 + M_n^{\mathrm{T}}R_n M_n.$$

The moment bounds above and the resolvent bound give $\sup_{\omega\in\Gamma}\|J_n(\omega)\|_{L^8} = O(1)$. A leave-one expansion of $M_n^{\mathrm{T}}\check{R}_n\psi_i$, followed by the conditional quadratic-form estimate in the proof of (S.3.31), gives

$$\sup_{\omega\in\Gamma}\max_i \|M_n^{\mathrm{T}}\check{R}_n(\omega)\psi_i\|_{L^8} = O(1).$$

It follows from the reverse Woodbury identity that

$$\sup_{\omega\in\Gamma}\max_{1\le i\le n}\left\|\frac{1}{n}\psi_i^{\mathrm{T}}\{R_n(\omega) - \check{R}_n(\omega)\}\psi_i\right\|_{L^2} = O(n^{-1}).$$

Sherman–Morrison applied to $\widetilde{zz}^{\mathrm{T}}$, whose denominator has real part at least one, similarly gives

$$\sup_{\omega\in\Gamma}\max_{1\le i\le n}\left\|\frac{1}{n}\psi_i^{\mathrm{T}}\left[\{C_n^\circ + \widetilde{zz}^{\mathrm{T}} + (\gamma+\omega)I_p\}^{-1} - R_n(\omega)\right]\psi_i\right\|_{L^2} = O(n^{-1}). \tag{S.4.12}$$

The centered Gram estimate (S.3.41), the resolvent identity, and (S.4.10) show that replacing the last resolvent by $Q_I(\omega)$, and ψ_i by ϕ_i, introduces an $o_{L^2}(1)$ error after division by n. With $Q_*(\omega)$ denoting the surrogate resolvent displayed in (S.4.12), the Gram-replacement error satisfies

$$\sup_{\omega\in\Gamma}\max_i\left\|\frac{1}{n}\phi_i^{\mathrm{T}}\{Q_I(\omega) - Q_*(\omega)\}\phi_i\right\|_{L^2} \le C\max_i\left\|\frac{\|\phi_i\|^2}{n}\right\|_{L^4}\|E_{S,n}\|_{L^4,\mathrm{op}} \longrightarrow 0.$$

Finally, Woodbury gives the exact identity

$$\frac{1}{n}F_i^{\mathrm{T}}\check{R}_n(\omega)F_i = I_2 - C_i(\omega).$$

Combining this identity with the preceding comparisons and (S.4.11) yields, uniformly on Γ,

$$\left\|\frac{1}{n}\phi_{i_n}^{\mathrm{T}}Q_I(\omega)\phi_{i_n} - w^{\mathrm{T}}\{I_2 - C_{2,n}(\omega)\}w\right\|_{L^2} \longrightarrow 0. \tag{S.4.13}$$

The two scalar functions in (S.4.13) are analytic inside Γ. Cauchy's formula and $Q_I'(0) = -Q_I(0)^2$ therefore give

$$\frac{1}{n}\phi_{i_n}^{\mathrm{T}}Q_I(0)^2\phi_{i_n} - \ell_n \xrightarrow{L^2} 0.$$

Equation (S.1.13) gives $\ell_n \to \ell$, completing (S.4.8). By oddness, the fresh-test feature covariance is νI_p, so

$$L_{i,n} = \nu\frac{1}{n}\phi_i^{\mathrm{T}}Q^2\phi_i.$$

This proves (S.4.9). The estimates are uniform in deterministic i. □

Lemma S.4.4 (Fixed-index cross contraction) Retain the positive-ridge odd branch and its assumptions. For the baseline labels and every deterministic index sequence i_n,

$$\max_{1\le i\le n}\|D_{i,n}\|_{L^2}=O(n^{-1/2}),\qquad D_{i_n,n}\xrightarrow{L^2}0.$$

Proof. Condition on a deterministic frame sequence. Write

$$x_i=Wg_i+x_{i,\perp},\qquad h_i=x_{i,\perp}^{\mathrm{T}}\beta_\perp^*.$$

Gaussian orthogonal decomposition shows that the h_i's are centered, independent of the projected design, independent across i, and have uniformly bounded moments. Hence

$$y_i^{(0)}=g_i^{\mathrm{T}}a+h_i+\sigma_\varepsilon\xi_i.$$

The endpoint-score identity gives

$$r^{(0)}=\frac{1}{n}\sum_i\phi_i y_i^{(0)}.$$

We derive its fourth-moment bound directly. Define the centered endpoint blocks

$$F_i=[f_1(g_i),f_2(g_i)],\qquad w=(\lambda,1-\lambda)^{\mathrm{T}},\qquad M_n=\frac{1}{n}\sum_i F_i,$$

and put

$$s_{\|,n}=\frac{1}{n}[F_1w,\dots,F_nw]H_nGa,\qquad s_{\zeta,n}=\frac{1}{n}[F_1w,\dots,F_nw]H_n\zeta,$$

where

$$\zeta_i=h_i+\sigma_\varepsilon\xi_i,\qquad \bar y_\|=\frac{1}{n}\sum_i g_i^{\mathrm{T}}a,\qquad \bar\zeta=\frac{1}{n}\sum_i\zeta_i.$$

The branch-neutral full-pair score remainder proved in (S.3.41) gives the explicit decomposition

$$r^{(0)}=\widehat{z}\,\bar y_\|+s_{\|,n}+\widehat{z}\,\bar\zeta+s_{\zeta,n}+e_{r,n},\qquad \|e_{r,n}\|_{L^4}\le Cn^{-1/2}\log^C n.$$

The vectors $A_i=F_iw(g_i^{\mathrm{T}}a)$ are independent and satisfy

$$\mathbb{E}\|A_i\|^2=O(p),\qquad \mathbb{E}\|A_i\|^4=O(p^2),\qquad \|\mathbb{E}A_i\|=O(1).$$

Apply the fixed-exponent Hilbert space Rosenthal inequality to the centered vectors $A_i-\mathbb{E}A_i$:

$$\left\|\frac{1}{n}\sum_i(A_i-\mathbb{E}A_i)\right\|_{L^4}=O(1).$$

The mean of $n^{-1}\sum_i A_i$ is $\mathbb{E}A_1$, whose norm is $O(1)$; hence

$$\left\|\frac{1}{n}\sum_i A_i\right\|_{L^4}=O(1).$$

Moreover,

$$s_{\|,n}=\frac{1}{n}\sum_i A_i-M_nw\bar y_\|.$$

Since $\|M_n\|_{L^8,\mathrm{op}}=O(1)$ and $\|\bar y_\|\|_{L^8}=O(n^{-1/2})$, the second term is bounded in L^4, and $\|s_{\|,n}\|_{L^4}=O(1)$.

Put $B_i=F_iw\zeta_i$. These vectors are conditionally independent and centered, and

$$s_{\zeta,n}=\frac{1}{n}\sum_i B_i-M_nw\bar\zeta.$$

The uniform eighth moments give

$$\sup_i\mathbb{E}\|B_i\|^2=O(p),\qquad \sup_i\mathbb{E}\|B_i\|^4=O(p^2).$$

Conditional Hilbert space Rosenthal, followed by integration over the projected design, yields

$$\left\| \frac{1}{n} \sum_i B_i \right\|_{L^4} = O(1).$$

The matrix M_n is measurable with respect to the projected design and is independent of $\bar{\zeta}$. Therefore

$$\|M_n w \bar{\zeta}\|_{L^8} \le \|w\| \, \|M_n\|_{L^8,\mathrm{op}} \|\bar{\zeta}\|_{L^8} = O(n^{-1/2}),$$

and consequently $\|s_{\zeta,n}\|_{L^4} = O(1)$. Finally,

$$\|\bar{z}\|_{L^8} = O(\sqrt{p}), \qquad \|\bar{y}_\|\|_{L^8} + \|\bar{\zeta}\|_{L^8} = O(n^{-1/2}),$$

so both empirical mean products are bounded in L^4. Consequently,

$$\sup_n \mathbb{E}\|r^{(0)}\|^4 < \infty.$$

Put

$$g_n^{\mathrm{risk}} = 2\{\nu Q r^{(0)} - \mu_1 a\}.$$

Since $\|Q\|_{\mathrm{op}} \le \gamma^{-1}$ and $\|a\| = O(1)$, the preceding score bound implies

$$\sup_n \mathbb{E}\|g_n^{\mathrm{risk}}\|^2 < \infty.$$

Moreover,

$$D_{i,n} = \frac{1}{2\sqrt{n}} \phi_i^{\mathrm{T}} Q g_n^{\mathrm{risk}}.$$

Let $\mathcal{D}_n$ be generated by the frame and $(g_i, h_i)_{i=1}^n$. Conditionally on $\mathcal{D}_n$,

$$r^{(0)} = r_{\mu,n} + \frac{\sigma_\varepsilon}{n} \Phi^{\mathrm{T}} \xi, \qquad r_{\mu,n} = \frac{1}{n} \sum_i \phi_i (g_i^{\mathrm{T}} a + h_i).$$

Thus $D_{i,n}$ is affine in ξ. Its conditional second moment depends on the standardized noises only through $\mathbb{E}_\xi \xi = 0$ and $\mathbb{E}_\xi \xi \xi^{\mathrm{T}} = I_n$. Consequently, permuting the exchangeable design-side variables in $\mathcal{D}_n$ permutes these conditional second moments. Their unconditional expectations are therefore equal for all i.

Pathwise, Lemma S.3.14 gives $n^{-1}\Phi^{\mathrm{T}}\Phi \le S$, and hence

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n D_{i,n}^2 &= \frac{1}{4n^2} (g_n^{\mathrm{risk}})^{\mathrm{T}} Q \Phi^{\mathrm{T}} \Phi Q g_n^{\mathrm{risk}} \\ &\le \frac{1}{4n} (g_n^{\mathrm{risk}})^{\mathrm{T}} Q S \, Q g_n^{\mathrm{risk}} \\ &\le \frac{1}{16\gamma n} \|g_n^{\mathrm{risk}}\|^2, \end{aligned}$$

because $\sup_{x \ge 0} x/(x+\gamma)^2 = 1/(4\gamma)$. Taking expectations and using equality of the coordinate second moments gives

$$\max_i \mathbb{E} D_{i,n}^2 = \mathbb{E} \frac{1}{n} \sum_i D_{i,n}^2 \le \frac{C}{n}.$$

This proves the deterministic-index conclusion. For an independent random index, condition on its value and average these second-moment bounds to obtain the last assertion. □

Proof of Corollary 4.2. The profile in Example 2 has $\overline{\sigma}_{\varepsilon,n}^2 = \sigma_\varepsilon^2 + \tau^2$ and test variance σ_ε^2. Substituting these values into Theorem 3.1 changes only the training noise term in the homoscedastic equivalent, adding $\tau^2 \chi(\lambda, \gamma)$. □

Realized-noise comparison under (S.4.2). We next derive the fixed limits for both the realized and noise-averaged risks. Oddness gives the exact baseline risk identity

$$\mathrm{Risk}_n^{\mathrm{real},(0)} = \nu\|Qr^{(0)}\|^2 - 2\mu_1 a^{\mathsf{T}} Qr^{(0)} + \|a\|^2 + \rho_{\perp,n}^{2,(0)}.$$

The score-moment calculation in the proof of Lemma S.4.4 gives $\sup_n \mathbb{E}\|r^{(0)}\|^4 < \infty$. The ridge bound and the bounded deterministic energies therefore imply

$$\sup_n \mathbb{E}|\,\mathrm{Risk}_n^{\mathrm{real},(0)}\,|^2 < \infty.$$

We compare the realized baseline risk with its training noise average. Conditional on X, W, write $r^{(0)} = r_\mu + \sigma_\varepsilon \Phi^{\mathsf{T}}\xi/n$. Expanding the preceding exact identity and subtracting its conditional mean gives

$$\begin{aligned}\mathrm{Risk}_n^{\mathrm{real},(0)} - \mathrm{Risk}_n^{(0)} = {}& \frac{2\sigma_\varepsilon}{n}\xi^{\mathsf{T}}\Phi\{\nu Q^2 r_\mu - \mu_1 Qa\} \\ & + \frac{\nu\sigma_\varepsilon^2}{n^2}\left\{\xi^{\mathsf{T}}\Phi Q^2\Phi^{\mathsf{T}}\xi - \mathrm{tr}(\Phi Q^2\Phi^{\mathsf{T}})\right\}.\end{aligned}$$

By Lemma S.3.14, $\Phi^{\mathsf{T}}\Phi/n \preceq S$. Since $\gamma > 0$, the conditional variance of the linear term is bounded by $Cn^{-1}(\|r_\mu\|^2 + \|a\|^2)$. Conditional Jensen's inequality and the preceding fourth-moment bound for $r^{(0)}$ show that its unconditional variance is $O(n^{-1})$. For the quadratic term set $A_n = \nu\sigma_\varepsilon^2\Phi Q^2\Phi^{\mathsf{T}}/n^2$. The same matrix inequality and $p/n = O(1)$ give $\|A_n\|_{\mathrm{op}} = O(n^{-1})$ and $\mathrm{tr}\,A_n = O(1)$, pathwise. Independence and uniformly bounded fourth moments of the standardized noises yield

$$\mathbb{E}_\xi\left|\xi^{\mathsf{T}}A_n\xi - \mathrm{tr}\,A_n\right|^2 \le C\,\mathrm{tr}(A_n^2) = O(n^{-1}).$$

Consequently $\mathrm{Risk}_n^{\mathrm{real},(0)} - \mathrm{Risk}_n^{(0)} = O_{L^2}(n^{-1/2})$. Corollary 4.1 and (S.4.2) give $\mathrm{Risk}_n^{(0)} \to B_\lambda(\gamma)$ in probability. The displayed comparison proves the same limit for $\mathrm{Risk}_n^{\mathrm{real},(0)}$. Uniform integrability then gives

$$\mathrm{Risk}_n^{\mathrm{real},(0)} \longrightarrow B_\lambda(\gamma) \quad \text{in } L^1.$$

The exact coefficient increment is

$$\widehat{\beta}^{\mathrm{sp}} - \widehat{\beta}^{(0)} = \frac{b_n\xi_{i_n}}{n}Q\phi_{i_n}.$$

Proposition S.4.2 therefore gives

$$\mathrm{Risk}_n^{\mathrm{real,sp}} - \mathrm{Risk}_n^{\mathrm{real},(0)} = 2\frac{b_n}{\sqrt{n}}\xi_{i_n}D_{i_n,n} + \frac{b_n^2}{n}\xi_{i_n}^2 L_{i_n,n}.$$

Since $u_n \to u$,

$$\frac{b_n}{\sqrt{n}} = O(1), \qquad \frac{b_n^2}{n} \longrightarrow u.$$

Lemma S.4.4 shows that the first term is $o_{\mathbb{P}}(1)$. The local-susceptibility lemma gives $L_{i_n,n} \to \nu\ell$ in L^2. This quantity depends only on the projected design and the deterministic index, so it is independent of ξ_{i_n}. Hence

$$\xi_{i_n}^2\{L_{i_n,n} - \nu\ell\} \longrightarrow 0 \quad \text{in } L^1.$$

Combining these facts yields

$$\mathrm{Risk}_n^{\mathrm{real,sp}} - B_\lambda(\gamma) - u\nu\ell\,\xi_{i_n}^2 \xrightarrow{\mathbb{P}} 0.$$

Here $J_\lambda(\gamma) = \chi(\lambda,\gamma) = \nu\ell$.

To average over the training noise, let

$$\mathcal{F}_n = \sigma(X, W),$$

and let $\mathbb{E}_\xi$ average only over the standardized observation noises. Conditional Jensen's inequality and the baseline L^1 convergence give

$$\mathbb{E}\left|\mathbb{E}_\xi(\mathrm{Risk}_n^{\mathrm{real},(0)} \mid \mathcal{F}_n) - B_\lambda(\gamma)\right| \longrightarrow 0.$$

The dependence of $D_{i_n,n}$ on ξ_{i_n} is affine. Its coefficient is determined by the observation-noise part $\sigma_\varepsilon n^{-1}\Phi^\top\xi$ of the score, while the orthogonal Gaussian teacher residual is fixed by $\mathcal{F}_n$. Therefore

$$\mathbb{E}_\xi(\xi_{i_n} D_{i_n,n} \mid \mathcal{F}_n) = \frac{\sigma_\varepsilon}{\sqrt{n}} L_{i_n,n}.$$

Taking conditional expectation in the exact risk increment gives

$$\mathrm{Risk}_n^{\mathrm{sp}} - \mathrm{Risk}_n^{(0)} = \frac{b_n^2 + 2\sigma_\varepsilon b_n}{n} L_{i_n,n} = u_n L_{i_n,n}. \tag{S.4.14}$$

Together with $L_{i_n,n} \to \nu\ell$, this proves

$$\mathrm{Risk}_n^{\mathrm{sp}} \xrightarrow{\mathbb{P}} B_\lambda(\gamma) + uJ_\lambda(\gamma) = R_\lambda(\gamma; u).$$

□

S.4.3.1 Uniform control needed for ridge oracles

Throughout this subsection we work in the centered odd critical-observation branch with the deterministic index in (S.4.1) and the separate energy limits in (S.4.2). Here N_n is the risk of the zero predictor under the fixed test distribution. For a fixed method s, write $S_{s,n}$, $Q_{s,n}(\gamma)$, and $\phi_{s,i}$ for the corresponding Gram matrix, resolvent, and endpoint score. More explicitly, if λ_s is its mixing weight and $z_{s,ij} = \sigma(\lambda_s g_i + (1-\lambda_s) g_j)$, then

$$S_{s,n} := \frac{1}{n^2}\sum_{i,j=1}^n z_{s,ij} z_{s,ij}^\top, \qquad \phi_{s,i} := \frac{1}{n}\sum_{j=1}^n \{\lambda_s z_{s,ij} + (1-\lambda_s) z_{s,ji}\},$$

$$r_{s,n} := \frac{1}{n}\sum_{j=1}^n \phi_{s,j} y_{j,n}^{\mathrm{sp}}, \qquad r_{s,n}^{(0)} := \frac{1}{n}\sum_{j=1}^n \phi_{s,j} y_{j,n}^{(0)},$$

and use the zero-ridge convention

$$Q_{s,n}(\gamma) := \begin{cases} (S_{s,n} + \gamma I_p)^{-1}, & \gamma > 0, \\ S_{s,n}^\dagger, & \gamma = 0. \end{cases}$$

In this odd branch the finite local susceptibility is

$$L_{s,i,n}(\gamma) := \frac{\nu}{n} \phi_{s,i}^\top Q_{s,n}(\gamma)^2 \phi_{s,i}.$$

Define

$$r_{\mu,s,n} := \mathbb{E}_\xi(r_{s,n} \mid X, W), \qquad C_{\varepsilon,s,n} := \mathrm{Cov}_\xi(r_{s,n} \mid X, W).$$

Then

$$C_{\varepsilon,s,n} = \frac{1}{n^2}\sum_{j=1}^n \sigma_{n,j}^2 \phi_{s,j}\phi_{s,j}^\top, \tag{S.4.15}$$

Equation (S.1.3) gives

$$\begin{aligned} \overline{\mathrm{Risk}}_{s,n}(\gamma; u_n) = N_n &- 2\mu_1 a^\top Q_{s,n}(\gamma) r_{\mu,s,n} \\ &+ \nu r_{\mu,s,n}^\top Q_{s,n}(\gamma)^2 r_{\mu,s,n} + \nu\,\mathrm{tr}\{Q_{s,n}(\gamma)^2 C_{\varepsilon,s,n}\}. \end{aligned} \tag{S.4.16}$$

Lemma S.4.5 (Score moments and compact-uniform positive-ridge equivalent) For either fixed method,

$$\sup_n \mathbb{E}\left[\|r_{\mu,s,n}\|^4 + \{\mathrm{tr}\, C_{\varepsilon,s,n}\}^2 + \left\{\frac{\|\phi_{s,i_n}\|^2}{n}\right\}^2 + \|r_{s,n}^{(0)}\|^4\right] < \infty. \tag{S.4.17}$$

For every $0 < \gamma_- < \gamma_+ < \infty$,

$$\sup_{\gamma\in[\gamma_-,\gamma_+]} \left|\overline{\mathrm{Risk}}_{s,n}(\gamma; u_n) - R_s(\gamma; u)\right| \xrightarrow{\mathbb{P}} 0, \tag{S.4.18}$$

$$\sup_{\gamma\in[\gamma_-,\gamma_+]} |L_{s,i_n,n}(\gamma) - \nu\ell_{s,\gamma}| \xrightarrow{\mathbb{P}} 0. \tag{S.4.19}$$

Proof. The homoscedastic baseline-score estimate gives $\sup_n \mathbb{E}\|r^{(0)}_{s,n}\|^4 < \infty$. Centering of both noise models and conditional Jensen's inequality give the same bound for $r_{\mu,s,n}$. Lipschitz continuity of the activation and $p \asymp n$ give, uniformly in i,

$$\mathbb{E}\|\phi_{s,i}\|^4 \le Cn^2.$$

Since $\sigma^2_{n,i_n} = \sigma^2_\varepsilon + nu_n$, (S.4.15) gives

$$\operatorname{tr} C_{\varepsilon,s,n} = \frac{\sigma^2_\varepsilon}{n}\sum_{j=1}^{n}\frac{\|\phi_{s,j}\|^2}{n} + u_n\frac{\|\phi_{s,i_n}\|^2}{n}.$$

Cauchy–Schwarz proves (S.4.17). The bounded-profile theorem enters only through the homoscedastic baseline-score bound.

For each fixed positive ridge, Corollary 4.1 and (S.4.2) give convergence of the noise-averaged baseline risk. The exact identity (S.4.14) and Lemma S.4.3 then imply pointwise convergence of the spiked annealed risk and the local coefficient.

For stochastic equicontinuity, $Q'_{s,n} = -Q^2_{s,n}$ and (S.4.16) give, uniformly on the displayed compact interval,

$$\begin{aligned}\left|\frac{d}{d\gamma}\overline{\mathrm{Risk}}_{s,n}(\gamma; u_n)\right| &\le \frac{2|\mu_1|}{\gamma_-^2}\|a\|\|r_{\mu,s,n}\| \\ &\quad + \frac{2\nu}{\gamma_-^3}\{\|r_{\mu,s,n}\|^2 + \operatorname{tr} C_{\varepsilon,s,n}\} = O_{\mathbb{P}}(1),\\ \left|\frac{d}{d\gamma}L_{s,i_n,n}(\gamma)\right| &\le \frac{2\nu}{\gamma_-^3}\frac{\|\phi_{s,i_n}\|^2}{n} = O_{\mathbb{P}}(1).\end{aligned}$$

The scalar root and the limiting coefficients are continuously differentiable at positive ridge. Pointwise convergence on a finite deterministic mesh, together with the tight Lipschitz constants, proves (S.4.18)– (S.4.19). □

Lemma S.4.6 (Uniform extension to zero ridge in the stable cases) For every finite $\Gamma > 0$,

$$\sup_{0\le\gamma\le\Gamma} |\overline{\mathrm{Risk}}_{s,n}(\gamma; u_n) - R_s(\gamma; u)| \xrightarrow{\mathbb{P}} 0,$$

and the analogous uniform convergence holds for $L_{s,i_n,n}(\gamma)$ toward $\nu\ell_{s,\gamma}$, in either of the following cases: (i) s = mix, the weight is interior, and the activation is nonaffine; or (ii) s = end, $c \ne 1$, and $\nu > 0$.

Proof. In case (i), $\kappa(\lambda_s) > 0$ and the interior spectral-floor result gives $\lambda_{\min}(S_{s,n}) \ge \kappa(\lambda_s)/2$ with probability tending to one. In case (ii), the endpoint extreme-eigenvalue result gives a deterministic $a_c > 0$ such that $\lambda^+_{\min}(S_{s,n}) \ge a_c$ with probability tending to one. All response scores, endpoint-score vectors, and the range of $C_{\varepsilon,s,n}$ lie in $\operatorname{ran}(S_{s,n})$, so endpoint null-space components never enter (S.4.16). On the relevant range, the resolvent and squared-resolvent differences between two ridges in $[0, \delta]$ are bounded by a constant times their distance. Together with (S.4.17), this gives a $\delta O_{\mathbb{P}}(1)$ modulus for both random quantities.

For an interior weight, the fixed-point derivative is at least $\kappa(\lambda_s)$, so the deterministic root and coefficients are continuous at zero. At the endpoint with $c \ne 1$, the Marchenko–Pastur formulas give

$$\nu\ell_{\mathrm{end},0} = \begin{cases} c/(1-c), & c < 1,\\ 1/(c-1), & c > 1,\end{cases}$$

The endpoint risk is also continuous at zero under the deterministic right-limit convention. Apply the preceding modulus on $[0, \delta]$ and Lemma S.4.5 on $[\delta, \Gamma]$, first letting $n \to \infty$ and then $\delta \downarrow 0$. □

Lemma S.4.7 (Hard-edge and large-ridge containment) At the endpoint with $c = 1$ and $\nu > 0$,

$$\nu \ell_{\mathrm{end},\gamma} \sim \frac{1}{2}\sqrt{\frac{\nu}{\gamma}} \qquad (\gamma \downarrow 0). \tag{S.4.20}$$

For this endpoint, if $u > 0$, every fixed positive candidate ridge γ_c admits deterministic $0 < \gamma_- < \gamma_c$ and $\eta > 0$ such that, with probability tending to one,

$$\inf_{0 \le \gamma \le \gamma_-} \overline{\mathrm{Risk}}_{\mathrm{end},n}(\gamma; u_n) > \overline{\mathrm{Risk}}_{\mathrm{end},n}(\gamma_c; u_n) + 2\eta. \tag{S.4.21}$$

For either fixed method and every $\Gamma > 0$,

$$\begin{aligned} \sup_{\gamma \ge \Gamma} |\overline{\mathrm{Risk}}_{s,n}(\gamma; u_n) - N_n| \le{}& \frac{2|\mu_1|}{\Gamma}\|a\|\|r_{\mu,s,n}\| \\ &+ \frac{\nu}{\Gamma^2}\{\|r_{\mu,s,n}\|^2 + \operatorname{tr} C_{\varepsilon,s,n}\}. \end{aligned} \tag{S.4.22}$$

Consequently the random large-ridge tail converges to zero in the iterated limit $n \to \infty$, then $\Gamma \to \infty$, and $\sup_{\gamma \ge \Gamma} |R_s(\gamma; u) - N| = O(\Gamma^{-1})$.

Proof. At the endpoint,

$$V = \operatorname{diag}(\nu, 0), \qquad \kappa(1) = 0, \qquad w = (1, 0)^{\mathsf{T}}.$$

At $c = 1$, put $x = \nu t_*$ and $\delta = \gamma/\nu$. The endpoint fixed-point equation gives

$$\delta x(1 + x) = 1, \qquad x = \frac{\sqrt{1 + 4/\delta} - 1}{2}.$$

Substitution into the formulas for τ_{fp} and ℓ gives

$$\nu \ell = \frac{x^2}{2x + 1}.$$

Since $x \sim \delta^{-1/2}$ as $\delta \downarrow 0$, this proves (S.4.20).

After averaging over the noise, the risk is a sum of nonnegative squared-bias and covariance terms. In particular,

$$\overline{\mathrm{Risk}}_{\mathrm{end},n}(\gamma; u_n) \ge \frac{\sigma^2_{n,i_n}}{n} L_{\mathrm{end},i_n,n}(\gamma).$$

Spectral decomposition on $\operatorname{ran}(S_{\mathrm{end,n}})$ shows that the local coefficient is nonincreasing in ridge, including its Moore–Penrose value at zero. Choose γ_- so small that $u\nu\ell_{\mathrm{end},\gamma_-}$ exceeds $R_{\mathrm{end}}(\gamma_c; u)$ by a fixed margin. Convergence of the local coefficient and risk at each fixed positive ridge then proves (S.4.21). This argument does not require a uniform hard-edge local law.

For $\gamma \ge \Gamma$, the bound $\|Q_{s,n}(\gamma)\|_{\mathrm{op}} \le \Gamma^{-1}$ in (S.4.16) gives (S.4.22). Its right-hand side is tight by (S.4.17). The fixed-point equation also gives $t_*(\gamma) \le c/\gamma$, and hence $q = O(\gamma^{-1})$ and $\tau_{\mathrm{fp}}, \ell, K = O(\gamma^{-2})$. Substitution into the limiting curve proves the deterministic tail assertion. □

S.4.4 Ridgeless and oracle comparisons

The fixed risk curves and oracle values are defined in (S.4.3)–(S.4.4). For $s \in \{\mathrm{mix}, \mathrm{end}\}$, define

$$R_s^{(0)}(\gamma) := B_{\lambda_s}(\gamma), \qquad H_s(\gamma) := J_{\lambda_s}(\gamma) = \nu \ell_{s,\gamma}.$$

On a closed domain, their values at zero are their finite right limits. Then

$$R_s(\gamma; u) = R_s^{(0)}(\gamma) + uH_s(\gamma), \qquad F_s(u) = \inf_{\gamma \in D_s} R_s(\gamma; u). \tag{S.4.23}$$

When $\nu > 0$, write

$$\mathcal{B} := \frac{\mu_1^2}{\nu} r_a^2, \qquad \mathcal{E}_{\text{eff}} := r_{\text{eff}}^2.$$

Also let H_c be the Marchenko–Pastur law for the $p \times p$ sample covariance with aspect ratio c, and define, for $\delta > 0$,

$$m_c(\delta) = \int_{[0,\infty)} \frac{1}{x+\delta}\, dH_c(x), \qquad J_{2,c}(\delta) = \int_{[0,\infty)} \frac{1}{(x+\delta)^2}\, dH_c(x) = -m_c'(\delta).$$

The derivative identity is used only for positive δ.

Proposition S.4.8 (Endpoint limiting oracle) Under the bounded-profile centered endpoint assumptions and (S.3.1), suppose $\nu > 0$ and $\sigma_{\text{test}}^2 = \overline{\sigma}_\varepsilon^2 = \sigma_\varepsilon^2$, and put $\delta = \gamma/\nu$. For every $\delta > 0$,

$$R_{\text{end}}^{(0)}(\nu\delta) = \mathcal{E}_{\text{eff}} + \mathcal{B}\delta^2 J_{2,c}(\delta) + c\mathcal{E}_{\text{eff}}\{m_c(\delta) - \delta J_{2,c}(\delta)\}. \tag{S.4.24}$$

If $\mathcal{B}, \mathcal{E}_{\text{eff}} > 0$, the unique minimizer and minimum are

$$\delta_* = \frac{c\mathcal{E}_{\text{eff}}}{\mathcal{B}}, \qquad \gamma_{\text{end}}^* = \nu\frac{c\mathcal{E}_{\text{eff}}}{\mathcal{B}}, \qquad R_{\text{end}}^* = \mathcal{E}_{\text{eff}}\left[1 + c\, m_c\left(\frac{c\mathcal{E}_{\text{eff}}}{\mathcal{B}}\right)\right]. \tag{S.4.25}$$

Under the displayed hypothesis $\mathcal{B}, \mathcal{E}_{\text{eff}} > 0$, the curve has a finite continuous extension to zero ridge when $c \neq 1$. When $c = 1$ under the same hypothesis, its limiting-oracle domain is positive ridge, equivalently the extended convention $R_{\text{end}}(0; 0) = +\infty$.

Proposition S.4.9 (Shifted mean-equivalence) For each fixed (λ, γ) covered by the critical-observation result, $R_\lambda(\gamma; u) + u$ is the homoscedastic limiting loss with both training and test noise variances increased from σ_ε^2 to $\sigma_\varepsilon^2 + u$. In the critical model, the test variance remains σ_ε^2. The equality also holds after taking the infimum over ridge.
More generally, consider a deterministic, uniformly bounded background variance profile whose mean tends to σ_{bg}^2. Keep the test variance at σ_{bg}^2, and add nu_n to the variance of one deterministically chosen training observation. Assume also that the visible and orthogonal teacher energies converge. The shifted noise-averaged limiting curve then depends on the background and critical contributions only through $\sigma_{\text{bg}}^2 + u$.

Proposition S.4.10 (Endpoint oracle under one critical observation) Assume $\nu > 0$. For every $\delta > 0$,

$$R_{\text{end}}(\nu\delta; u) = \mathcal{E}_{\text{eff}} + \mathcal{B}\delta^2 J_{2,c}(\delta) + c(\mathcal{E}_{\text{eff}} + u)\{m_c(\delta) - \delta J_{2,c}(\delta)\}. \tag{S.4.26}$$

If $\mathcal{B} > 0$ and $\mathcal{E}_{\text{eff}} + u > 0$, the unique optimizer is

$$\delta_{\text{end}}^*(u) = \frac{c(\mathcal{E}_{\text{eff}} + u)}{\mathcal{B}}, \qquad \gamma_{\text{end}}^*(u) = \nu\frac{c(\mathcal{E}_{\text{eff}} + u)}{\mathcal{B}}, \tag{S.4.27}$$

and its value is

$$F_{\text{ridge}}(u) = \mathcal{E}_{\text{eff}} + c(\mathcal{E}_{\text{eff}} + u) m_c\left(\frac{c(\mathcal{E}_{\text{eff}} + u)}{\mathcal{B}}\right). \tag{S.4.28}$$

For $c \neq 1$, the curve has a finite continuous zero-ridge limit. Under $\mathcal{B} > 0$ and $\mathcal{E}_{\text{eff}} + u > 0$, adjoining that value does not change the displayed optimizer. For $c = 1$ and $\mathcal{E}_{\text{eff}} + u > 0$, the limiting domain is $D_{\text{end}} = (0, \infty)$, equivalently $R_{\text{end}}(0; u) = +\infty$. When $c = 1$ and $\mathcal{E}_{\text{eff}} + u = 0$, the right limit is finite and equals zero. Finite-sample endpoint oracles still include the Moore–Penrose value at zero.

Proposition S.4.11 (Endpoint-oracle witness) Assume $\nu > 0$, $\mathcal{B} > 0$, and $\mathcal{E}_{\text{eff}} + u > 0$, and put $\gamma_e(u) = \nu c(\mathcal{E}_{\text{eff}} + u)/\mathcal{B}$. For midpoint mixup,

$$F_{\text{end}}(u) - F_{\text{mix}}(u) \geq R_{\text{end}}^{(0)}(\gamma_e(u)) - R_{\text{mix}}^{(0)}(\gamma_e(u)) + u\nu\{\ell_{\text{end},\gamma_e(u)} - \ell_{\text{mix},\gamma_e(u)}\}.$$

At midpoint mixup, let

$$f(x) = \mathbb{E}_G \sigma\Big(\frac{x+G}{2}\Big), \qquad m_{\mathrm{bal}} = \mathbb{E} f(G)^2, \qquad a_V = 2m_{\mathrm{bal}}, \qquad V = m_{\mathrm{bal}} \mathbf{1}_2 \mathbf{1}_2^{\mathsf{T}}.$$

For $c, \delta > 0$, let $x_c(\delta) > 0$ solve

$$c = x_c(\delta)\left\{\delta + \frac{1}{1 + x_c(\delta)}\right\}, \tag{S.4.29}$$

and define

$$G_c(\delta) = \frac{x_c(\delta)}{1 + \delta\{1 + x_c(\delta)\}^2}. \tag{S.4.30}$$

Proposition S.4.12 (Midpoint fixed-ridge attenuation) Suppose σ is odd and nonaffine, $m_{\mathrm{bal}} > 0$, and

$$2m_{\mathrm{bal}} \le \nu \le 4m_{\mathrm{bal}}. \tag{S.4.31}$$

Then $\ell_{\mathrm{mix},\gamma} < \ell_{\mathrm{end},\gamma}$ for every $c, \gamma > 0$, with

$$\ell_{\mathrm{mix},\gamma} = \frac{1}{4m_{\mathrm{bal}}} G_c\Big(\frac{\gamma + \kappa(1/2)}{2m_{\mathrm{bal}}}\Big), \qquad \ell_{\mathrm{end},\gamma} = \frac{1}{\nu} G_c\Big(\frac{\gamma}{\nu}\Big). \tag{S.4.32}$$

Define the dimensionless midpoint moment ratios

$$z = \frac{\nu d_\alpha}{\mu_1 a_V}, \qquad \alpha_0 = \frac{\nu}{2a_V}, \qquad h = z - 1 - \alpha_0, \qquad r_z = \frac{z^2}{4} - \alpha_0.$$

Then put

$$q_z = \frac{(z-2)^2}{4}, \qquad P_c(x) = ch + (2ch - r_z)x + (ch + q_z)x^2.$$

Let

$$t_f = \frac{\mathcal{B}\kappa(1/2)}{a_V c}, \qquad x_e(t) = x_c\Big(\frac{ct}{\mathcal{B}}\Big), \qquad x_{\max} = x_e(\max(\mathcal{E}_{\mathrm{eff}}, t_f)),$$

and

$$x_* = \min\left\{x_{\max}, \left[\frac{r_z - 2ch}{2(ch + q_z)}\right]_+\right\}.$$

Theorem S.4.13 (General midpoint witness certificate) Assume the critical-observation model, an odd nonaffine activation, midpoint mixup, and $r_a^2 > 0$. Suppose further that $m_{\mathrm{bal}} > 0$, $\mu_1 > 0$, $d_\alpha > 0$, and $h > 0$, and that

$$P_c(x_*) > 0. \tag{S.4.33}$$

When $\mathcal{E}_{\mathrm{eff}} < t_f$, suppose additionally

$$F_{\mathrm{end}}(0) > R_{\mathrm{mix}}(0; 0). \tag{S.4.34}$$

Then (4.3) holds for every finite $u \ge 0$.

For the boundary certificate, put

$$x_h = x_c\Big(\frac{\kappa(1/2)}{a_V}\Big), \qquad G_h = G_c\Big(\frac{\kappa(1/2)}{a_V}\Big),$$

and define

$$\begin{aligned} D_{\mathrm{bd}}(t) = {} & t\{1 + x_e(t)\} - (\mathcal{B} + t)(1 + \alpha_0 G_h) \\ & - \mathcal{B}\frac{z^2 G_h(1 - x_h)}{4c(1 + x_h)} + \mathcal{B}\frac{z x_h}{c(1 + x_h)}. \end{aligned} \tag{S.4.35}$$

The polynomial condition is implied by the stronger bound

$$c > c_{\mathrm{quad}} := \frac{r_z^2}{4h(r_z + q_z)}. \tag{S.4.36}$$

Corollary S.4.14 (Simple midpoint certificate) Under the positivity conditions above, suppose

$$c \geq c_{\mathrm{sep}} := \frac{r_z}{2h}, \qquad a_V c \frac{\mathcal{E}_{\mathrm{eff}}}{\mathcal{B}} > \kappa(1/2). \tag{S.4.37}$$

Then the polynomial and boundary conditions of Theorem S.4.13 hold, and hence (4.3) holds for every finite $u \geq 0$.

The gap is strictly positive at each fixed critical strength, but not uniformly positive as the strength grows without bound:

$$F_{\mathrm{end}}(u) - F_{\mathrm{mix}}(u) \longrightarrow 0 \qquad (u \to \infty). \tag{S.4.38}$$

Proposition S.4.15 (Local unrestricted-oracle envelope) For $s \in \{\mathrm{end}, \mathrm{mix}\}$, let

$$R_s(\gamma; u) = R_s^{(0)}(\gamma) + uH_s(\gamma), \qquad H_s(\gamma) = \nu \ell_{s,\gamma} \geq 0, \qquad \gamma \in D_s.$$

First fix any nonempty compact set $K_s \subset D_s$ on which $R_s^{(0)}$ and H_s are finite and continuous, and define

$$F_{s,K_s}(u) = \min_{\gamma \in K_s} R_s(\gamma; u), \qquad \mathcal{A}_{s,K_s}(u) = \operatorname*{argmin}_{\gamma \in K_s} R_s(\gamma; u).$$

Then F_{s,K_s} is finite, nondecreasing, concave, and continuous for $u \geq 0$, and $\mathcal{A}_{s,K_s}(u)$ is nonempty and compact. Its one-sided derivatives are

$$\frac{d^+}{du} F_{s,K_s}(u) = \min_{\gamma \in \mathcal{A}_{s,K_s}(u)} H_s(\gamma),$$

and, for $u > 0$,

$$\frac{d^-}{du} F_{s,K_s}(u) = \max_{\gamma \in \mathcal{A}_{s,K_s}(u)} H_s(\gamma).$$

Only the right derivative is asserted at $u = 0$.

Now let $N \in \mathbb{R}$ be the common finite null-predictor level and fix a finite $u_0 \geq 0$. Assume that $R_s^{(0)}$ and H_s are finite and continuous on compact subsets of D_s, that

$$\lim_{\Gamma \to \infty} \sup_{\substack{\gamma \in D_s \\ \gamma \geq \Gamma}} |R_s(\gamma; u) - N| = 0$$

for every finite u, and that $F_s(u_0) < N$. Whenever $D_s = (0, \infty)$, additionally assume that there is $\epsilon_0 > 0$ such that, with

$$u_-^{(0)} = \max\{0, u_0 - \epsilon_0\},$$

one has

$$\lim_{\gamma \downarrow 0} R_s(\gamma; u_-^{(0)}) = +\infty.$$

There are $\epsilon > 0$ and method-specific compact sets $K_s \subset D_s$, independent of u, such that, for

$$I = \begin{cases} [u_0 - \epsilon, u_0 + \epsilon], & u_0 > 0, \\ [0, \epsilon], & u_0 = 0, \end{cases}$$

one has

$$F_s(u) = \min_{\gamma \in K_s} R_s(\gamma; u), \qquad \varnothing \neq \operatorname*{argmin}_{\gamma \in D_s} R_s(\gamma; u) \subset K_s \quad (u \in I).$$

For a stable domain $D_s = [0, \infty)$, one may take $K_s = [0, \Gamma_s]$. For an open domain $D_s = (0, \infty)$, one may take $K_s = [\gamma_-, \Gamma_s]$ with $\gamma_- > 0$. Consequently, on I, the unrestricted oracle F_s has all the compact properties and one-sided derivative formulas above, with $\mathcal{A}_{s,K_s}(u)$ replaced by $\operatorname{argmin}_{\gamma \in D_s} R_s(\gamma; u)$.

If both unrestricted argmin sets at u_0 are singletons, denote their unique elements by $\gamma_s^*(u_0)$. Then the unrestricted oracle gap is differentiable at $u_0 > 0$, and has a right derivative at $u_0 = 0$, with

$$\frac{d}{du}\{F_{\mathrm{end}}(u) - F_{\mathrm{mix}}(u)\}\Big|_{u=u_0} = \nu\{\ell_{\mathrm{end},\gamma^*_{\mathrm{end}}(u_0)} - \ell_{\mathrm{mix},\gamma^*_{\mathrm{mix}}(u_0)}\}. \tag{S.4.39}$$

Theorem S.4.16 (Finite-sample oracle and argmin-set convergence) Retain the centered odd critical-observation assumptions, the deterministic index in (S.4.1), and the separate energy limits (S.4.2). Fix a method s with mixing weight λ_s, and use one of the following branches:

$$\begin{aligned} &\text{(i)} && \lambda_s \in (0,1),\ \sigma \text{ nonaffine},\ D_s = [0,\infty),\\ &\text{(ii)} && \lambda_s = 1,\ c \neq 1,\ \nu > 0,\ D_s = [0,\infty),\\ &\text{(iii)} && \lambda_s = 1,\ c = 1,\ \nu > 0,\ u > 0,\ D_s = (0,\infty). \end{aligned}$$

In branches (i)–(ii), $R_s(0;u)$ denotes its finite right limit. In branch (iii), one may equivalently adjoin $R_1(0;u) = +\infty$. With $u_n \to u$ as in the model and $N = r_a^2 + \rho_{\perp,\infty}^{2,(0)}$, assume

$$F_s(u) := \inf_{\gamma \in D_s} R_s(\gamma;u) < N.$$

Then

$$\inf_{\gamma \geq 0} \overline{\mathrm{Risk}}_{s,n}(\gamma;u_n) \xrightarrow{\mathbb{P}} F_s(u). \tag{S.4.40}$$

The finite-sample infimum includes the Moore–Penrose value at zero in every branch. The deterministic set

$$\mathcal{A}_s(u) = \operatorname*{argmin}_{\gamma \in D_s} R_s(\gamma;u)$$

is nonempty and compact. Every measurable $o_{\mathbb{P}}(1)$-approximate minimizer $\widehat{\gamma}_{s,n}$, taken over $\gamma \geq 0$, obeys

$$\operatorname{dist}\{\widehat{\gamma}_{s,n}, \mathcal{A}_s(u)\} \xrightarrow{\mathbb{P}} 0. \tag{S.4.41}$$

If $\mathcal{A}_s(u) = \{\gamma_s^*(u)\}$, then $\widehat{\gamma}_{s,n} \xrightarrow{\mathbb{P}} \gamma_s^*(u)$.

Proof of Theorem S.4.16. In branches (i)–(ii), strict null-risk improvement and right continuity at zero provide a fixed $\gamma_c > 0$ such that $R_s(\gamma_c;u) < N$. In branch (iii), $D_s = (0,\infty)$ and $F_s(u) < N$; by the definition of the infimum there is therefore a fixed positive $\gamma_c \in D_s$ satisfying $R_s(\gamma_c;u) < N$. Put

$$\mathfrak{g} = \frac{N - R_s(\gamma_c;u)}{5} > 0.$$

The deterministic large-ridge tail gives $\Gamma_0 > \gamma_c$ such that

$$\inf_{\gamma \geq \Gamma_0} R_s(\gamma;u) > N - \mathfrak{g} = R_s(\gamma_c;u) + 4\mathfrak{g}.$$

In branches (i)–(ii), set $\gamma_0 = 0$. In branch (iii), (S.4.20) and $u > 0$ give $R_1(\gamma;u) \to +\infty$ as $\gamma \downarrow 0$; choose $\gamma_0 \in (0,\gamma_c)$ such that

$$\inf_{0 < \gamma \leq \gamma_0} R_1(\gamma;u) > R_1(\gamma_c;u) + 4\mathfrak{g}.$$

Continuity on $[\gamma_0,\Gamma_0]$ now proves that $\mathcal{A}_s(u)$ is nonempty and compact.

Fix $\rho > 0$. The random large-ridge estimate (S.4.22) allows us to choose a deterministic $\Gamma \geq \Gamma_0$ such that

$$\limsup_n \mathbb{P}\left\{\sup_{\gamma \geq \Gamma} |\overline{\mathrm{Risk}}_{s,n}(\gamma;u_n) - N_n| > \mathfrak{g}\right\} < \rho.$$

Since $N_n \to N$ and $\overline{\mathrm{Risk}}_{s,n}(\gamma_c;u_n) \to R_s(\gamma_c;u)$, outside an event whose limiting probability is at most ρ,

$$\inf_{\gamma \geq \Gamma} \overline{\mathrm{Risk}}_{s,n}(\gamma;u_n) > \overline{\mathrm{Risk}}_{s,n}(\gamma_c;u_n) + \mathfrak{g}.$$

In branches (i)–(ii), take $K = [0,\Gamma]$. Lemma S.4.6 gives uniform convergence on this compact set. In branch (iii), the hard-edge containment (S.4.21) gives $0 < \widetilde{\gamma}_- < \gamma_c$ such that, with probability tending to one,

$$\inf_{0 \leq \gamma \leq \widetilde{\gamma}_-} \overline{\mathrm{Risk}}_{1,n}(\gamma;u_n) > \overline{\mathrm{Risk}}_{1,n}(\gamma_c;u_n) + 2\eta$$

for some $\eta > 0$. Set

$$\gamma_- = \min\{\widetilde{\gamma}_-, \gamma_0\}, \qquad K = [\gamma_-, \Gamma].$$

The compact positive-ridge equivalent applies on K, and the choice $\gamma_- \le \gamma_0$ preserves the deterministic lower-end exclusion. Thus, in every branch,

$$\Delta_n := \sup_{\gamma \in K} |\overline{\mathrm{Risk}}_{s,n}(\gamma; u_n) - R_s(\gamma; u)| \xrightarrow{\mathbb{P}} 0,$$

the deterministic minimum lies in K, and the finite-sample minimum over $\gamma \ge 0$, including its Moore–Penrose value, lies in K with probability at least $1 - \rho - o(1)$. On this event,

$$\left| \inf_{\gamma \ge 0} \overline{\mathrm{Risk}}_{s,n}(\gamma; u_n) - F_s(u) \right| \le \Delta_n.$$

Letting first $n \to \infty$ and then $\rho \downarrow 0$ proves (S.4.40).

For the argmin assertion, fix $\epsilon > 0$ and retain the arbitrary $\rho > 0$ used to choose K. Define

$$B_\epsilon = \{\gamma \in K : \mathrm{dist}\{\gamma, \mathcal{A}_s(u)\} \ge \epsilon\}.$$

If B_ϵ is nonempty, continuity and compactness give

$$h_\epsilon = \min_{\gamma \in B_\epsilon} R_s(\gamma; u) - F_s(u) > 0.$$

Let $\zeta_n = o_{\mathbb{P}}(1)$ be the approximation error of $\widehat{\gamma}_{s,n}$. On the event that excludes exterior minimizers, whose probability is at least $1 - \rho - o(1)$, all values outside K exceed the comparison value at γ_c by a fixed positive margin $m > 0$. Whenever $\zeta_n < m$, approximate optimality therefore forces $\widehat{\gamma}_{s,n} \in K$. On that same event,

$$\begin{aligned} R_s(\widehat{\gamma}_{s,n}; u) &\le \overline{\mathrm{Risk}}_{s,n}(\widehat{\gamma}_{s,n}; u_n) + \Delta_n \\ &\le \inf_{\gamma \in K} \overline{\mathrm{Risk}}_{s,n}(\gamma; u_n) + \zeta_n + \Delta_n \le F_s(u) + \zeta_n + 2\Delta_n. \end{aligned}$$

If $B_\epsilon = \varnothing$, containment in K already gives $\mathrm{dist}\{\widehat{\gamma}_{s,n}, \mathcal{A}_s(u)\} < \epsilon$. Otherwise, $\widehat{\gamma}_{s,n} \notin B_\epsilon$ whenever $\zeta_n + 2\Delta_n < h_\epsilon$. In either case, since $\zeta_n, \Delta_n \to 0$ in probability,

$$\limsup_{n \to \infty} \mathbb{P}[\mathrm{dist}\{\widehat{\gamma}_{s,n}, \mathcal{A}_s(u)\} \ge \epsilon] \le \rho.$$

Sending $\rho \downarrow 0$ proves (S.4.41); the singleton conclusion follows immediately.

□

Proof of Corollary S.4.1. Applying Theorem S.4.16 to the fixed endpoint and midpoint methods gives

$$O_{\mathrm{end},n} \xrightarrow{\mathbb{P}} F_1(u), \qquad O_{\mathrm{mid},n} \xrightarrow{\mathbb{P}} F_{1/2}(u).$$

Subtraction proves convergence of the fixed-method gap.

For every realization, the joint feasible set contains the entire midpoint ridge slice, so

$$O_{\mathrm{joint},n} \le O_{\mathrm{mid},n}.$$

On the event

$$O_{\mathrm{end},n} - O_{\mathrm{mid},n} > \frac{\Delta(u)}{2},$$

whose probability tends to one, this feasibility inequality gives (S.4.5).

For measurability, fix a countable dense subset $D_\Lambda \subset [1/2, 1]$. At positive ridge the finite risk is jointly continuous in (λ, γ), and its Moore–Penrose value is the right limit as $\gamma \downarrow 0$ on the score range. Thus

$$O_{\mathrm{joint},n} = \inf_{\substack{\lambda \in D_\Lambda \\ \gamma \in \mathbb{Q}_{>0}}} \overline{\mathrm{Risk}}_{\lambda,n}(\gamma; u_n).$$

The right-hand side is a countable infimum of measurable random variables. □

Proof of Proposition S.4.8. The variance matching and (S.3.1) give $\rho^2_{\perp,\infty} = \rho^{2,(0)}_{\perp,\infty}$. At the endpoint, normalize the feature coordinates by $\sqrt{\nu}$. The population coefficient and residual energies in this unit-variance model are $\mathcal{B}$ and $\mathcal{E}_{\text{eff}}$, respectively. Put $\delta = \gamma/\nu$ and $s_\sigma = \mu_1^2/\nu$. After this rescaling, the Marchenko–Pastur Stieltjes fixed-point identity (Silverstein & Bai, 1995, Theorem 1.1 and equation (1.1)) gives

$$\nu t_* = c m_c(\delta).$$

Direct substitution in (S.1.6), together with $J_{2,c} = -m_c'$, gives

$$\mathfrak{L} = 1 + c\{m_c(\delta) - \delta J_{2,c}(\delta)\}, \qquad \mathfrak{K} = (1 - s_\sigma)\mathfrak{L} + s_\sigma \delta^2 J_{2,c}(\delta).$$

Since the odd endpoint risk is $\mathfrak{K} r_a^2 + \mathfrak{L}\rho^{2,(0)}_{\perp,\infty}$, it follows that

$$R^{(0)}_{\text{end}}(\nu\delta) = \mathcal{E}_{\text{eff}} + \mathcal{B}\delta^2 J_{2,c}(\delta) + c\mathcal{E}_{\text{eff}} \int \frac{x}{(x+\delta)^2}\, dH_c(x).$$

Since

$$\int \frac{x}{(x+\delta)^2}\, dH_c(x) = m_c(\delta) - \delta J_{2,c}(\delta),$$

this is (S.4.24).

Because $J_{2,c}(\delta) = -m_c'(\delta)$, differentiation under the integral sign yields

$$\frac{d}{d\delta} R^{(0)}_{\text{end}}(\nu\delta) = 2(\mathcal{B}\delta - c\mathcal{E}_{\text{eff}}) \int \frac{x}{(x+\delta)^3}\, dH_c(x).$$

The integral is strictly positive for $\delta > 0$. Thus the derivative changes sign exactly once, at $\delta_* = c\mathcal{E}_{\text{eff}}/\mathcal{B}$. At this value, $\mathcal{B}\delta_*^2 = c\mathcal{E}_{\text{eff}}\delta_*$, so the two $J_{2,c}$-terms cancel. This proves (S.4.25).

For $c \neq 1$, the endpoint coefficient limits established in Theorem S.3.3, together with (S.3.9), give the finite continuous right limit at zero ridge. Since the derivative above is negative before $\delta_* > 0$ and positive afterward, adjoining this boundary value does not change the optimizer. When $c = 1$ and $\mathcal{E}_{\text{eff}} > 0$, the endpoint root satisfies $\delta(\nu t_*)(1 + \nu t_*) = 1$, so $\nu t_* \to \infty$ as $\delta \downarrow 0$. The residual-fitting term of the displayed risk consequently diverges. Thus the limiting domain is positive ridge. Equivalently, the zero-ridge value may be set to $+\infty$. □

Proof of Proposition S.4.9. In the odd branch, the background residual contribution to the fixed-ridge risk limit is $(1 + \nu\ell_{s,\gamma})\rho^{2,(0)}_{\perp,\infty}$. By (S.4.23), the critical observation adds $u\nu\ell_{s,\gamma}$. Adding the common constant u gives exactly the risk change from increasing both the homoscedastic training variance and the test variance to $\sigma_\varepsilon^2 + u$.

For a bounded heterogeneous background, Theorem 3.1 applies to the baseline sample. Its proof gives L^2 convergence of the noise-averaged baseline risk to its equivalent, and the assumed energy limits identify the fixed limit. Conditional on (X, W), increasing only the i_n-th training variance by nu_n adds exactly $u_n L_{i_n,n}$ to the noise-averaged risk, by (S.1.2) and (S.1.3). Lemma S.4.3 concerns only the design and gives $L_{i_n,n} \to \nu\ell$. The limiting risk is therefore the background formula at mean variance σ^2_{bg}, plus $u\nu\ell$. □

Proof of Proposition S.4.10. By Proposition S.4.9, $R_{\text{end}}(\gamma; u)+u$ is the homoscedastic endpoint curve with $\mathcal{E}_{\text{eff}}$ replaced by $\mathcal{E}_{\text{eff}} + u$ and $\mathcal{B}$ unchanged. Making this substitution in (S.4.24) and subtracting u proves (S.4.26). If $\mathcal{B} > 0$ and $\mathcal{E}_{\text{eff}} + u > 0$, the oracle formula of Proposition S.4.8, applied to this auxiliary homoscedastic curve, gives the unique optimizer (S.4.27) and the value (S.4.28). The constant shift does not change the optimizer.

For $c \neq 1$, the critical profile satisfies (S.3.1), so Theorem S.3.3 and (S.3.9) give the finite continuous right limit of the endpoint coefficients and risk at zero, including when $\mathcal{E}_{\text{eff}} + u = 0$. Under $\mathcal{B} > 0$ and $\mathcal{E}_{\text{eff}} + u > 0$, the same boundary argument for the auxiliary homoscedastic curve shows that adjoining zero ridge does not change the optimizer. For the remaining boundary $c = 1$, put $x = \nu t_*$ and $\delta = \gamma/\nu$. The endpoint root equation is $\delta x(1 + x) = 1$, so $x \to \infty$ as $\delta \downarrow 0$, and direct substitution gives

$$\nu\ell_{\text{end},\gamma} = \frac{x^2}{2x+1} \longrightarrow +\infty \qquad (\delta \downarrow 0).$$

If $\mathcal{E}_{\mathrm{eff}} + u > 0$, the contribution proportional to $\mathcal{E}_{\mathrm{eff}} + u$ diverges, so the limiting domain is $(0, \infty)$, or equivalently one may assign $R_{\mathrm{end}}(0; u) = +\infty$. If $\mathcal{E}_{\mathrm{eff}} + u = 0$, then $\mathcal{E}_{\mathrm{eff}} = u = 0$, and differentiating $\delta x(1 + x) = 1$ gives

$$\delta^2 J_{2,1}(\delta) = \frac{1}{1 + 2x} \longrightarrow 0.$$

Hence the endpoint curve instead has the finite right limit zero. □

Proof of Proposition S.4.11. The endpoint formula gives $F_{\mathrm{end}}(u) = R_{\mathrm{end}}(\gamma_e(u); u)$, while the midpoint oracle may use the same ridge, so $F_{\mathrm{mix}}(u) \le R_{\mathrm{mix}}(\gamma_e(u); u)$. Subtracting these two inequalities and using (S.4.23) gives the displayed witness. □

Proof of Proposition S.4.12. Implicit differentiation of (S.4.29) gives

$$x_c'(\delta) = -\frac{x_c(\delta)}{\delta + \{1 + x_c(\delta)\}^{-2}} < 0.$$

Eliminating δ from (S.4.30) gives, with $x = x_c(\delta)$,

$$G_c(\delta) = \frac{x^2}{c + 2cx + (c - 1)x^2}.$$

This expression has derivative $2cx(1+x)/\{c+2cx+(c-1)x^2\}^2 > 0$ with respect to x, so G_c is strictly decreasing in δ.

The susceptibility formulas follow directly from the fixed-point coefficients. In the balanced branch, let $x = 2m_{\mathrm{bal}} t_*$. Since $V = m_{\mathrm{bal}} \mathbf{1}_2 \mathbf{1}_2^{\mathsf{T}}$ has eigenvalue $2m_{\mathrm{bal}}$ in the direction of $w = \mathbf{1}_2/2$, the fixed-point equation becomes

$$c = x\left\{\frac{\gamma + \kappa(1/2)}{2m_{\mathrm{bal}}} + \frac{1}{1 + x}\right\}, \qquad w^{\mathsf{T}} V C_*^2 w = \frac{m_{\mathrm{bal}}}{(1 + x)^2}.$$

Substitution into the definitions of τ_{fp} and ℓ in (S.1.6) yields

$$\ell_{\mathrm{mix},\gamma} = \frac{1}{4m_{\mathrm{bal}}} G_c\left(\frac{\gamma + \kappa(1/2)}{2m_{\mathrm{bal}}}\right).$$

At the endpoint, the same calculation with $x = \nu t_*$ gives

$$c = x\left\{\frac{\gamma}{\nu} + \frac{1}{1 + x}\right\}, \qquad \ell_{\mathrm{end},\gamma} = \frac{1}{\nu} G_c\left(\frac{\gamma}{\nu}\right).$$

This proves (S.4.32). Nonaffinity gives $\kappa(1/2) > 0$. The first inequality in (S.4.31) yields $(\gamma + \kappa(1/2))/(2m_{\mathrm{bal}}) > \gamma/\nu$, and the second yields $1/(4m_{\mathrm{bal}}) \le 1/\nu$. Apply the strict monotonicity to (S.4.32). □

Proof of Theorem S.4.13. Strict Gaussian Cauchy–Schwarz and nonaffinity give $\mu_1^2 < \nu$, hence $\mathcal{B} > 0$ and $\mathcal{E}_{\mathrm{eff}} > 0$. Also $r_z = h + q_z > 0$. Writing $y = ch$,

$$P_c(x) = y + (2y - r_z)x + (y + q_z)x^2.$$

If $2y - r_z \ge 0$, this quadratic is increasing on the positive half-line. Otherwise its positive vertex has value

$$\frac{4y(r_z + q_z) - r_z^2}{4(y + q_z)}.$$

Consequently

$$P_c(x) > 0 \text{ for every } x > 0 \quad \Longleftrightarrow \quad c > c_{\mathrm{quad}}. \tag{S.4.42}$$

Indeed, at $c = c_{\mathrm{quad}}$ the positive double root is $r_z/(r_z + 2q_z)$, so equality cannot be included. If $c < c_{\mathrm{quad}} < c_{\mathrm{sep}} = r_z/(2h)$, the vertex lies on the positive half-line and the displayed vertex value is negative, proving the reverse implication as well. More generally, convexity shows that its minimum on $[0, x_{\max}]$ is attained at x_*; since $P_c(0) = ch > 0$, condition (S.4.33) is equivalent to positivity on the attained interval $0 < x \le x_{\max}$.

Put $t = \mathcal{E}_{\text{eff}} + u$, $\delta_e(t) = ct/\mathcal{B}$, and $x = x_e(t)$. Proposition S.4.10 gives

$$F_{\text{end}}(u) = \mathcal{E}_{\text{eff}} + tx. \tag{S.4.43}$$

Consider the matching ridge

$$\gamma_w(t) := a_V \delta_e(t) - \kappa(1/2).$$

It is nonnegative exactly when $t \geq t_f$, and it makes the balanced fixed-point coordinate equal to the endpoint coordinate x. With $G = G_c(\delta_e(t))$, direct one-dimensional diagonalization of $V = m_{\text{bal}} \mathbf{1}_2 \mathbf{1}_2^{\mathsf{T}}$ gives

$$\theta = \frac{d_\alpha}{2(1+x)}, \qquad q = \frac{x}{a_V c}, \qquad \tau_{\text{fp}} = \frac{G(1+x)^2}{a_V^2},$$
$$\ell = \frac{G}{2a_V}, \qquad K = \frac{G}{2a_V} + \frac{d_\alpha^2 G(1-x)}{4a_V^2 c(1+x)}.$$

Substitution in the odd-risk formula, using $r_a^2 = \mathcal{B}\nu/\mu_1^2$, yields

$$R_{\text{mix}}(\gamma_w(t); t - \mathcal{E}_{\text{eff}}) = \mathcal{B} + \mathcal{E}_{\text{eff}} + \alpha_0(\mathcal{B} + t)G + \mathcal{B}\frac{z^2 G(1-x)}{4c(1+x)} - \mathcal{B}\frac{zx}{c(1+x)}. \tag{S.4.44}$$

The scalar root equation gives

$$\delta_e(t) = \frac{c + (c-1)x}{x(1+x)}, \qquad G = \frac{x^2}{c(1+x)^2 - x^2}.$$

Subtracting (S.4.44) from (S.4.43) and simplifying gives the exact witness gap

$$F_{\text{end}}(t - \mathcal{E}_{\text{eff}}) - R_{\text{mix}}(\gamma_w(t); t - \mathcal{E}_{\text{eff}}) = \mathcal{B}\frac{xP_c(x)}{c(1+x)\{c(1+x)^2 - x^2\}}. \tag{S.4.45}$$

Its denominator is positive because it equals $c(1+x)x\{1 + \delta_e(t)(1+x)^2\}$. Thus the matching witness is strictly better throughout $t \geq \max\{\mathcal{E}_{\text{eff}}, t_f\}$.

The possible interval $\mathcal{E}_{\text{eff}} \leq t < t_f$, where $\gamma_w(t) < 0$, requires a separate argument. At zero ridge the balanced coordinate and response are x_h, G_h, and the same substitution, with D_{bd} defined in (S.4.35), gives exactly $D_{\text{bd}}(t)$. Differentiating the scalar equation gives

$$D'_{\text{bd}}(t) = G_c(ct/\mathcal{B}) - \alpha_0 G_h, \qquad D''_{\text{bd}}(t) = \frac{c}{\mathcal{B}} G'_c(ct/\mathcal{B}) < 0.$$

At $t = t_f$, formula (S.4.45) applies with $x = x_h$, so the junction value is positive under (S.4.33). When $\mathcal{E}_{\text{eff}} < t_f$, condition (S.4.34) gives positivity at the other endpoint. Strict concavity places D_{bd} above their positive chord, so the zero-ridge witness covers the whole remaining interval. Taking the mixup infimum proves (4.3). Equation (S.4.42) proves the simpler condition (S.4.36). □

Proof of (S.4.38). Put $t = \mathcal{E}_{\text{eff}} + u$. The relation $x_e(t) \sim \mathcal{B}/t$ implies that the gap for the chosen witness is of order t^{-1}. To control the optimized gap, write the balanced risk relative to the zero-predictor limit $N = \mathcal{B} + \mathcal{E}_{\text{eff}}$. For a candidate ridge $\gamma \geq 0$, set

$$\delta = \frac{\gamma + \kappa(1/2)}{a_V}, \qquad x = x_c(\delta), \qquad G = G_c(\delta).$$

Then

$$R_{\text{mix}}(\gamma; u) - N = \alpha_0(\mathcal{B} + t)G + \mathcal{B}\left\{\frac{z^2 G(1-x)}{4c(1+x)} - \frac{zx}{c(1+x)}\right\}.$$

Any near-minimizing ridge sequence as $t \to \infty$ must have $x \to 0$. Otherwise the first nonnegative term diverges while the zero predictor gives an upper bound N. The braced term then vanishes, so the mixup oracle tends to N. The endpoint formula gives the same limit. This proves (S.4.38). □

Proof of Corollary S.4.14. The thresholds satisfy

$$\frac{c_{\text{quad}}}{c_{\text{sep}}} = \frac{r_z}{2(r_z + q_z)} \le \frac{1}{2},$$

so $c \ge c_{\text{sep}}$ implies $c > c_{\text{quad}}$. The second condition in (S.4.37) is equivalent to $\mathcal{E}_{\text{eff}} > t_f$, so no negative-ridge boundary interval remains. Hence (S.4.42) implies (S.4.33). Theorem S.4.13 now applies. □

Proof of Theorem 4.3. The certificate depends algebraically on the two background energies. To apply it to the deterministic equivalent in the main text, use

$$r_a^2 = \|W^{\mathrm{T}}\beta^*\|^2, \qquad \rho_{\perp,\infty}^{2,(0)} = \|\beta_\perp^*\|^2 + \sigma_\varepsilon^2, \qquad u = \tau^2.$$

Here these symbols denote the curve's input values. The theorem's two quantitative assumptions are exactly the simple certificate (S.4.37). Corollary S.4.14 therefore gives (4.3). □

Proof of Proposition S.4.15. We first prove the assertions on compact sets. By continuity, the minimum on K_s is finite and attained, and its argmin set is a nonempty closed subset of K_s. Since $H_s \ge 0$, pointwise comparison of the affine curves shows that F_{s,K_s} is nondecreasing. For $0 \le t \le 1$,

$$\begin{aligned} R_s(\gamma; (1-t)u + tv) &= (1-t)R_s(\gamma; u) + tR_s(\gamma; v) \\ &\ge (1-t)F_{s,K_s}(u) + tF_{s,K_s}(v). \end{aligned}$$

Taking the minimum over $\gamma \in K_s$ proves concavity. Moreover,

$$|F_{s,K_s}(u) - F_{s,K_s}(v)| \le |u - v| \max_{\gamma \in K_s} H_s(\gamma),$$

so the compact oracle is continuous.

Fix $u \ge 0$, let $h > 0$, and choose $\gamma_h \in \mathcal{A}_{s,K_s}(u+h)$. For every $\gamma \in \mathcal{A}_{s,K_s}(u)$, optimality at the two strengths gives

$$H_s(\gamma_h) \le \frac{F_{s,K_s}(u+h) - F_{s,K_s}(u)}{h} \le H_s(\gamma).$$

Every limit point of γ_h as $h \downarrow 0$ belongs to $\mathcal{A}_{s,K_s}(u)$: this follows by taking a convergent subsequence in K_s and passing to the limit in its defining optimality inequality. Continuity of H_s therefore gives the stated right derivative and the minimum active slope. The same calculation for $h < 0$ reverses the inequalities after division and gives the left derivative as the maximum active slope. At $u = 0$, only the right-hand calculation is available.

We next show that the unrestricted minimizers lie locally in fixed compact sets. Fix a method s and choose $\gamma_{c,s} \in D_s$ such that $R_s(\gamma_{c,s}; u_0) < N$, and put

$$d_s = N - R_s(\gamma_{c,s}; u_0) > 0.$$

Because $H_s(\gamma_{c,s})$ is finite, choose $\epsilon_s > 0$, with $\epsilon_s < u_0$ when $u_0 > 0$, such that

$$R_s(\gamma_{c,s}; u) \le N - \frac{3d_s}{4}$$

throughout

$$I_s = \begin{cases} [u_0 - \epsilon_s, u_0 + \epsilon_s], & u_0 > 0, \\ [0, \epsilon_s], & u_0 = 0. \end{cases}$$

For an open domain also require $\epsilon_s \le \epsilon_0$. Let a_s, b_s be the endpoints of I_s. The large-ridge limit at these two fixed strengths gives $\Gamma_s > \gamma_{c,s}$ such that

$$\sup_{\substack{\gamma \in D_s \\ \gamma \ge \Gamma_s}} |R_s(\gamma; a_s) - N| < \frac{d_s}{4}, \qquad \sup_{\substack{\gamma \in D_s \\ \gamma \ge \Gamma_s}} |R_s(\gamma; b_s) - N| < \frac{d_s}{4}.$$

If $u = (1-t)a_s + tb_s$, affineness in u gives

$$R_s(\gamma; u) - N = (1-t)\{R_s(\gamma; a_s) - N\} + t\{R_s(\gamma; b_s) - N\}.$$

Hence the two endpoint bounds hold uniformly for $u \in I_s$. In particular, every $\gamma \geq \Gamma_s$ is strictly worse than $\gamma_{c,s}$.

For $D_s = [0, \infty)$, continuity now shows that the unrestricted minimum is attained on $K_s = [0, \Gamma_s]$. For $D_s = (0, \infty)$, choose $\gamma_- > 0$ so that

$$R_s(\gamma; u_-^{(0)}) > N \qquad (0 < \gamma \leq \gamma_-).$$

Every $u \in I_s$ satisfies $u \geq u_-^{(0)}$. Since $H_s \geq 0$,

$$R_s(\gamma; u) = R_s(\gamma; u_-^{(0)}) + (u - u_-^{(0)}) H_s(\gamma) > N$$

for $0 < \gamma \leq \gamma_-$. Thus the unrestricted minimum is attained on $K_s = [\gamma_-, \Gamma_s]$. Taking half the smaller of the two ϵ_s's produces the common interval I, while the compact sets remain method-specific. A minimizer on K_s has value $F_s(u)$ and is therefore an unrestricted minimizer. Conversely, the strict exterior bounds place every unrestricted minimizer in K_s. Thus the compact and unrestricted argmin sets coincide on each I_s, hence on a neighborhood relative to $[0, \infty)$ of every point of I. Applying the compact part of the proposition gives the asserted monotonicity, concavity, continuity, and one-sided derivative formulas for the unrestricted oracle, including when minimizers are not unique. When both argmin sets at u_0 are singletons, the minimum and maximum active slopes coincide for each method. Subtracting those two method-specific derivatives proves (S.4.39), with only the right derivative used when $u_0 = 0$. □

S.4.5 Additional theorem-matched diagnostics

This subsection gives chordwise and ridge-path views of the random feature experiment in Section 5.1. The calculations use the same fixed Gaussian design and analytically averaged population risks as Figure 2(b).

Panel (c) of Figure S.4.4 numerically verifies the pointwise prediction-influence limit $n\mathcal{I}_{u,i_*} \to \nu\ell_u$ used in Corollary 4.2. The finite-design curves follow the deterministic equivalents throughout the displayed ridge range. At zero ridge, balanced mixup lowers the susceptibility from approximately 4.000 for endpoint interpolation to 1.553, providing a theorem-supported analogue of the smoothing in the toy model. Panel (d) shows that the population risk advantage remains when each method tunes ridge separately.

S.5 Auxiliary Lemmas and Standard Results

Sections S.2 and S.4 share the tools collected below. Unlike the supporting lemmas that accompany those proofs, these results do not depend on the paper-specific fixed-point coefficients V, t_*, θ, ℓ, or K.

S.5.1 Deterministic matrix and resolvent identities

For invertible matrices of compatible dimensions, we repeatedly use the resolvent identity, the Sherman–Morrison rank-one formula, and the Woodbury finite-rank formula. In the notation $R_A(z) = (A + zI)^{-1}$,

$$R_A(z) - R_B(z) = R_A(z)(B - A)R_B(z),$$

and, whenever the displayed inverse exists,

$$(A + UCV^\top)^{-1} = A^{-1} - A^{-1}U(C^{-1} + V^\top A^{-1}U)^{-1}V^\top A^{-1}.$$

We also use cyclicity of the trace, the trace–rank inequality, and Schur complements. In each application, we state the relevant spectral gap or positive-real-part condition.

For $\lambda \in [0, 1]$, let $\mathsf{M} \in \mathbb{R}^{n^2 \times n}$ be the full-pair label-mixing matrix whose (i, j)th row is $\lambda e_i^\top + (1 - \lambda)e_j^\top$. Then

$$\begin{aligned} \mathsf{M}^\top\mathsf{M} &= n\{\lambda^2 + (1 - \lambda)^2\}I_n + 2\lambda(1 - \lambda)\mathbf{1}_n\mathbf{1}_n^\top, \\ \|\mathsf{M}\|_{\mathrm{op}}^2 &= n, \qquad \mathsf{M}\mathsf{M}^\top \preceq nI_{n^2}. \end{aligned} \tag{S.5.1}$$

Summing over the ordered pairs gives the Gram matrix in the first line. Its eigenvalue is $n\{\lambda^2 + (1 - \lambda)^2\}$ on $\mathbf{1}_n^\perp$ and n on $\mathrm{span}\{\mathbf{1}_n\}$, which proves the second line.

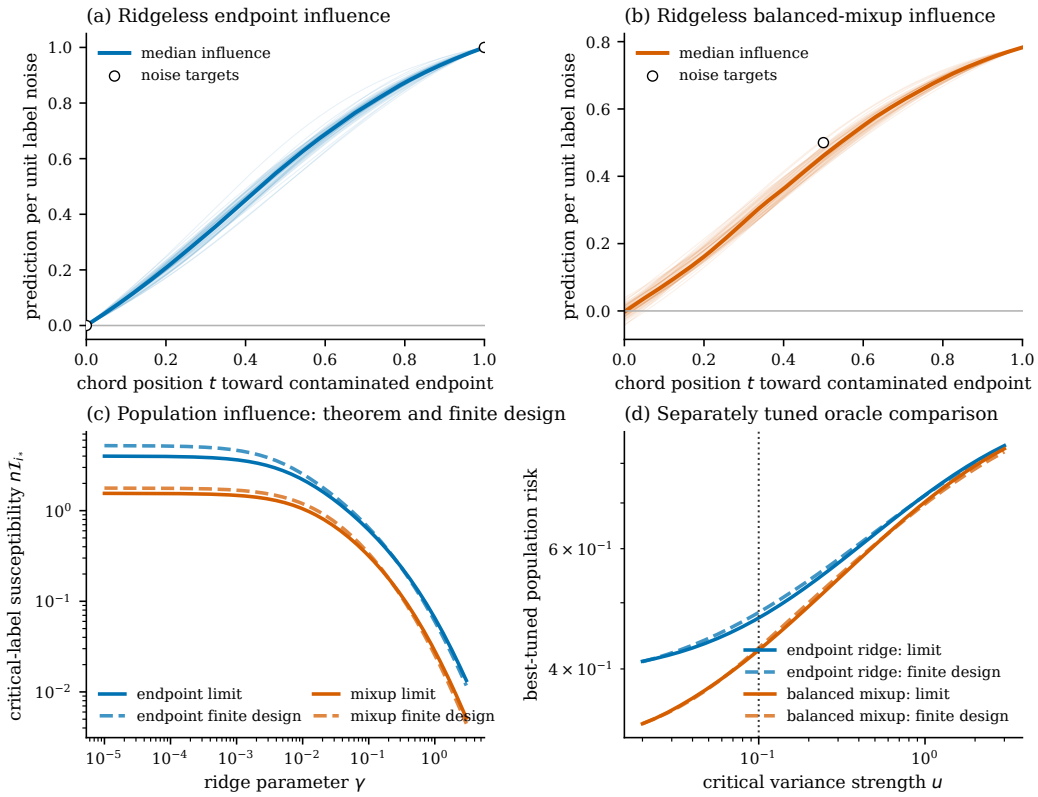

Figure S.4.4: Chordwise influence and population risk diagnostics in the same theorem-matched model. (a)–(b) Ridgeless prediction per unit perturbation of y_{i_*} along 28 prespecified chords, with $t = 0$ at a clean endpoint and $t = 1$ at the contaminated endpoint. Thin curves are individual chords, the thick curve is their median, and the band spans the 10%–90% quantiles. Endpoint training constrains the two chord endpoints, whereas balanced mixup places the contaminated–clean target at $(t, y) = (1/2, 1/2)$. (c) Critical-label susceptibility $J = n\mathcal{I}_{i_*}$ versus ridge: deterministic limits are solid and exact values for the fixed finite design are dashed. At the ridgeless boundary the limiting susceptibilities are approximately 4.000 for endpoint interpolation and 1.553 for balanced mixup. (d) Separately tuned population risks as the critical variance strength varies; the vertical dotted line marks $u = 0.1$. All risks average the training noise and integrate the Gaussian test covariate without sampled responses or test points.

S.5.2 Moment and concentration inequalities

The scalar and vector estimates also use Gaussian Poincaré inequalities. For standard independent-entry matrix norms, Gaussian Lipschitz concentration, and Hanson–Wright, we use only the forms in (Vershynin, 2018, Theorems 4.4.5, 5.2.2, and 6.2.1).

For independent centered scalar or real-Hilbert-valued summands, we use Rosenthal's inequality throughout the appendix as stated in (Pinelis, 1994, Theorem 5.3, equation (5.8)), in the form

$$\left\| \sum_j X_j \right\|_{L^q} \le C \left\{ q \left(\sum_j \mathbb{E}\|X_j\|^q \right)^{1/q} + \sqrt{q} \left(\sum_j \mathbb{E}\|X_j\|^2 \right)^{1/2} \right\}.$$

S.5.3 Matrix concentration and dependent sums

Operator-norm bounds use matrix Bernstein in the form of Tropp (2012). For order-two dependent sums, the proofs use the forward canonical Banach-valued decoupling of (Minsker & Wei, 2019, Lemma 5.2); no symmetry is required for this forward direction. After conditioning on

one endpoint family, the rectangular matrices are independent and centered, and (Dirksen et al., 2022, Theorem 10) gives the rectangular L^q moment estimate. The comparisons in (de la Peña & Montgomery-Smith, 1995, Theorem 1) and (Junge & Xu, 2008, Theorem 4.6) provide background and alternative support; general treatments appear in (de la Peña & Giné, 1999; Junge & Xu, 2003).

S.5.4 Standard random-matrix limits

The endpoint transform calculations use the Marchenko–Pastur law and its Stieltjes transform (Silverstein & Bai, 1995). Exact first and second inverse real-Wishart moments, with existence thresholds $m > d + 1$ and $m > d + 3$, respectively, are taken from (Matsumoto, 2012, arXiv:1004.4717v3, Theorem 2 and Appendix A). Away from $c = 1$, the endpoint pseudoinverse argument uses only the fixed positive smallest-singular-value floor supplied by (Rudelson & Vershynin, 2009, Theorem 1.1), after variance normalization and transposition when needed. No exact edge convergence, inverse-moment uniform integrability, or aspect-one pseudoinverse control is attributed to that result.

S.5.5 Haar projection facts

Lemma S.5.1 Let $W \sim \mathrm{Haar}\{\mathrm{St}(d, p)\}$, let $b \in \mathbb{R}^d$ be deterministic, and put

$$e = \frac{\mathbf{1}}{\sqrt{p}}, \qquad P_\perp = I_p - ee^{\mathsf{T}}, \qquad T_\perp = \|P_\perp W^{\mathsf{T}} b\|^2, \qquad T_0 = \|W^{\mathsf{T}} b\|^2.$$

For $b \neq 0$,

$$\frac{T_\perp}{\|b\|^2} \sim \mathrm{Beta}\left(\frac{p-1}{2}, \frac{d-p+1}{2}\right), \qquad \frac{T_0}{\|b\|^2} \sim \mathrm{Beta}\left(\frac{p}{2}, \frac{d-p}{2}\right), \tag{S.5.2}$$

with the natural point-mass interpretation when the projection rank is zero or d. Their exact moments are

$$\begin{aligned} \mathbb{E}_W T_\perp &= \frac{p-1}{d}\|b\|^2, & \mathrm{Var}_W(T_\perp) &= \frac{2(p-1)(d-p+1)}{d^2(d+2)}\|b\|^4, \\ \mathbb{E}_W T_0 &= \frac{p}{d}\|b\|^2, & \mathrm{Var}_W(T_0) &= \frac{2p(d-p)}{d^2(d+2)}\|b\|^4. \end{aligned} \tag{S.5.3}$$

Moreover, for $e = \mathbf{1}/\sqrt{p}$,

$$\mathbb{E}_W (e^{\mathsf{T}} W^{\mathsf{T}} b)^2 = \frac{\|b\|^2}{d}. \tag{S.5.4}$$

Proof. The right invariance of the Haar law allows us to rotate e to a coordinate vector. Consequently, $WP_\perp W^{\mathsf{T}}$ and WW^{T} are Haar-uniform orthogonal projectors of ranks $p - 1$ and p, respectively. More generally, if P_k is a Haar-uniform rank-k projector and $u = b/\|b\|$, rotational invariance gives

$$u^{\mathsf{T}} P_k u \overset{d}{=} \frac{Z_1^2 + \cdots + Z_k^2}{Z_1^2 + \cdots + Z_d^2}, \qquad Z_1, \ldots, Z_d \overset{\text{iid}}{\sim} \mathcal{N}(0, 1).$$

The numerator and the remaining part of the denominator are independent chi-square variables. Their ratio is $\mathrm{Beta}(k/2, (d-k)/2)$, whose mean and variance are

$$\frac{k}{d}, \qquad \frac{2k(d-k)}{d^2(d+2)}.$$

Taking $k = p - 1$ and $k = p$ proves (S.5.2)– (S.5.3). Finally, We is uniform on the unit sphere in $\mathbb{R}^d$; hence $\mathbb{E}_W \langle We, b\rangle^2 = \|b\|^2/d$, which proves (S.5.4). □

S.5.6 Analytic tools

Once the Jacobian gap is established, the implicit-function or contraction argument gives uniform analytic continuation of the deterministic fixed-point branch. We then apply Cauchy's integral formula on a fixed inner circle to differentiate the deterministic equivalent at $\omega = 0$. These steps justify the passage from first-resolvent limits to the squared-resolvent quantities appearing in the risk.

Table S.6.1: Search grids for the random-feature risk comparisons. Mixing-weight grids are equally spaced on $[1/2, 1]$, and ridge grids are logarithmically spaced. Local refinement, where used, minimizes risk between the two neighboring log-ridge grid points when the grid minimum is interior. Each method is tuned separately.

Comparison	Mixing weights	Ridge grid	Local refinement
Finite-sample and limit	$1/2, 1$	501 points in $[10^{-7}, 10^{1.5}]$	Yes
Joint limiting-risk search	61 points	401 points in $[10^{-8}, 30]$	No
Noise and covariate stress tests	13 points	241 points in $[10^{-10}, 10^{2}]$	Yes

S.6 Additional random feature numerical studies

These studies expand the risk comparisons in Section 5.1. We first compare finite-sample and limiting risks, then examine joint tuning, noise allocation, covariate anisotropy, and weak regularization. Throughout, $\mathcal{E}(\lambda, \gamma)$ and $\chi(\lambda, \gamma)$ denote the signal reconstruction error coefficient and residual fitting error coefficient from Theorem 3.1, and τ^2 denotes noise spike strength. Absolute risk reduction is the scalar-ridge risk minus the mixup risk; percentage reduction divides this difference by the scalar-ridge risk and multiplies by 100.

S.6.1 Experimental setup and finite-sample comparison

All experiments use tanh features and average prediction risk over training noise and a fresh test observation. Finite-sample comparisons condition on the sampled training covariates and the feature frame. Training noise is integrated analytically; Gaussian test moments are computed by deterministic Gauss–Hermite quadrature. The fitted random-feature coefficients depend linearly on the labels, so an observation's noise contribution equals its variance times its squared prediction influence on a fresh covariate. Under the single-noisy-observation model, this contribution converges to $\tau^2\chi(\lambda, \gamma)$ for fixed (λ, γ) with $\gamma > 0$, as established in Section S.4.3. The finite-sample sensitivity can differ from its limit χ.

Figure 2(b) uses one Gaussian training design with $(n, p) = (128, 160)$, an independent orthonormal feature frame, and a unit-norm teacher in the feature span. One predetermined label has noise variance $n\tau^2$; all other training and test labels are noiseless. We evaluate 30 logarithmically spaced strengths $\tau^2 \in [0.02, 3]$, minimizing finite-sample and limiting risks separately for each method and strength. The finite-sample curve closely follows the limiting percentage reduction.

Table S.6.1 summarizes the tuning protocols. The stress tests use $(n, p) = (48, 60)$ and $(64, 80)$, with $d = p$, a fixed orthogonal feature frame, and four independently drawn Gaussian training designs per setting. The same four random seeds are reused across settings, and each design is shared by the compared methods. Points report means; bars span the four observed values and describe design variation. A solution at the lower ridge boundary is interpreted as near-ridgeless; the displayed stress-test optima do not reach the upper boundary.

S.6.2 Joint tuning of mixing weight and ridge strength

For a unit-norm teacher in the feature span, no background noise, and $c = 1.25$, Corollary 4.2 gives

$$\overline{\mathrm{Risk}}_{\mathrm{het}}(\lambda, \gamma; \tau^2) = \mathcal{E}(\lambda, \gamma) + \tau^2\chi(\lambda, \gamma).$$

We approximate the improvement over optimally regularized scalar ridge,

$$\overline{\mathrm{Risk}}^{*}_{\mathrm{het}}(1; \tau^2) - \inf_{\lambda \in [1/2, 1]} \overline{\mathrm{Risk}}^{*}_{\mathrm{het}}(\lambda; \tau^2),$$

using the joint grid in Table S.6.1 at 81 equally spaced strengths $\tau^2 \in [0, 2]$. These are numerical approximations to the continuous-parameter optima in (4.2). At every evaluated strength, the grid minimum occurs at $\lambda = 1/2$, so the jointly tuned and midpoint curves coincide on this grid.

Gaussian quadrature supports the sufficient conditions of Corollary S.4.14: the feature-ratio threshold is approximately $0.671 < 1.25$, and the two sides of its second inequality are approximately

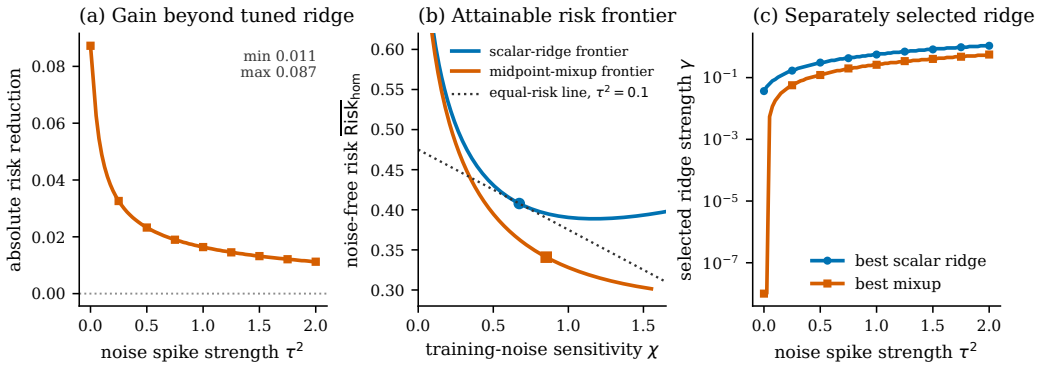

Figure S.6.5: Grid approximations to optimally tuned risks for one noisy observation, with $c = 1.25$, a unit-norm teacher in the feature span, zero background noise, and tanh. (a) Absolute risk reduction from jointly tuning mixup. (b) Noise-free risk $\mathcal{E} = \overline{\text{Risk}}_{\text{hom}}$ versus sensitivity χ, with the equal-risk line for the best grid-selected scalar ridge at $\tau^2 = 0.1$. (c) Selected ridge strengths. The joint search selects $\lambda = 1/2$ at all 81 evaluated noise strengths. All panels evaluate deterministic risk formulas without sampling training or test observations.

$0.02485 > 0.00765$. The required moment positivity conditions also hold numerically. Theorem 4.3 therefore supplies the midpoint-versus-ridge comparison under these conditions; the search quantifies the gain and examines the additional choice of mixing weight.

The absolute gap in Figure S.6.5(a) ranges from 0.0112 to 0.0873 and narrows as noise strength increases. In panel (b), the dotted line consists of pairs satisfying $\mathcal{E} + 0.1\chi$ equal to the best scalar-ridge risk on the grid. The selected midpoint lies below this line, illustrating the improved tradeoff between signal reconstruction error and residual fitting error. Near $\tau^2 = 0$, mixup selects the lower ridge boundary; all plotted solutions use positive ridge.

S.6.3 Heterogeneous training-noise profiles

To examine whether the gain persists when noise affects several labels, we distribute a fixed average variance across a set I of m observations:

$$\sigma_{n,i}^2 = \frac{n\tau^2}{m}\mathbf{1}\{i \in I\}, \qquad \overline{\sigma}_{\varepsilon,n}^2 = \tau^2.$$

We use isotropic Gaussian covariates, a unit-norm linear teacher, noiseless test labels, and $\tau^2 = 0.05$. The profiles have $m = 1, 7, 12$ for $n = 48$, and $m = 1, 8, 16$ for $n = 64$. The exogenous profiles place noise at the first m observations, independently of their covariates. Theorem 3.1 predicts the same limiting risk for these profiles at any fixed (λ, γ) with $\gamma > 0$. We also place noise at the m largest teacher projections $|x_i^\top \beta^*|$ to examine a design-dependent allocation beyond this assumption. Scalar ridge tunes γ; mixup tunes both λ and γ.

At $(n, p) = (48, 60)$, Figure S.6.6(a) gives mean scalar-ridge risks of 0.477–0.479 and mixup risks of 0.381–0.385; every paired design gives a positive reduction. In panel (b), the mean gap across the displayed profiles is 0.0926 for design-dependent locations and 0.0943 for exogenous locations. At $(n, p) = (64, 80)$, the mean gaps are 0.088–0.089. Thus the benefit persists as the same noise budget is spread across more observations. Noise allocation can affect finite-sample risk because individual labels have different prediction influences; for the exogenous profiles, its limiting contribution at fixed parameters is $\tau^2\chi(\lambda, \gamma)$. The similar observed risks are consistent with this prediction.

S.6.4 Anisotropic covariates

To examine sensitivity to the isotropic-covariate assumption, we use

$$\Sigma_s = \frac{p}{s + p - 1}\text{diag}(s, 1, \ldots, 1), \qquad s \in \{1, 3, 10, 30, 100\}.$$

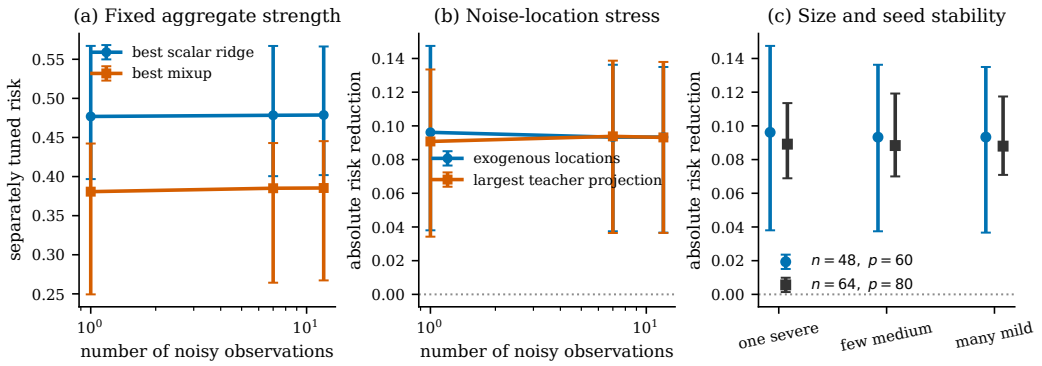

Figure S.6.6: Noise allocation at fixed average variance $\tau^2 = 0.05$. (a) Separately tuned risks when noise affects one, seven, or twelve observations at $(n, p) = (48, 60)$. (b) Absolute risk reductions for exogenous locations and locations with the largest teacher projections, at the same sample size. (c) Exogenous profiles at $(n, p) = (48, 60)$ and $(64, 80)$. Points are means and bars span four independently drawn Gaussian training designs per setting. All panels use tanh features and average training noise analytically and test covariates by quadrature.

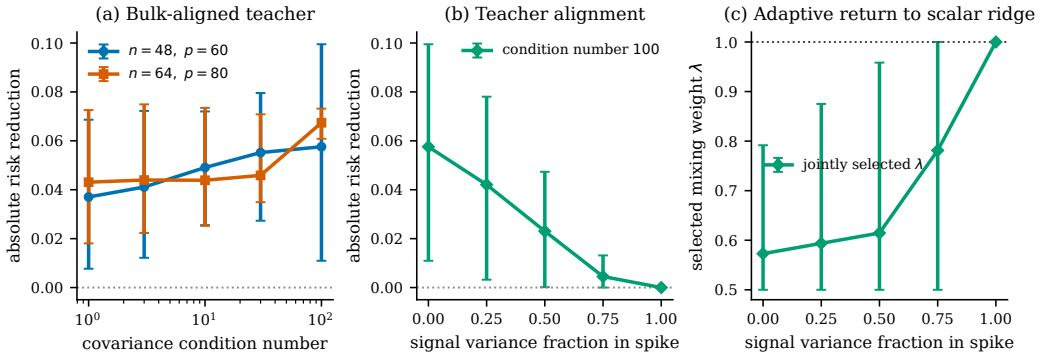

Figure S.6.7: Covariate anisotropy and teacher alignment. (a) Absolute risk reduction with the teacher in a bulk covariance direction, at $(n, p) = (48, 60)$ and $(64, 80)$. (b) Risk reduction as the fraction of signal variance in the leading covariance direction increases, at condition number 100. (c) Jointly selected mixing weight; $\lambda = 1$ recovers scalar ridge. Panels (b)–(c) use $(n, p) = (48, 60)$. Points are means and bars span four independently drawn Gaussian training designs. Risks average training noise analytically and Gaussian test moments by quadrature.

The covariance condition number is s, and its average eigenvalue is one. We normalize the teacher by $\beta^{*\mathsf{T}}\Sigma_s\beta^* = 1$ and set both training and test noise variances to 0.1. The feature frame is fixed independently of the training designs. Each method uses the stress-test tuning grid in Table S.6.1.

For bulk-aligned signals, Figure S.6.7(a) shows a positive gap at every displayed condition number and design. At condition number 100, the mean reductions are 0.0576 and 0.0674 for the two sample sizes. As signal variance concentrates in the leading covariance direction, panels (b)–(c) show the gap closing and the selected weight returning to $\lambda = 1$. Thus gains persist under strong anisotropy for bulk-aligned signals, while alignment changes the useful amount of mixing. These finite-sample results motivate extending the risk theory to general covariance matrices.

S.6.5 Weak regularization near interpolation

To isolate the interpolation peak, we use a unit-norm teacher in the feature span and training and test noise variance 4. The limiting risk is

$$\overline{\mathrm{Risk}}_{\mathrm{hom}}(\lambda, \gamma) = \mathcal{E}(\lambda, \gamma) + 4\chi(\lambda, \gamma) + 4.$$

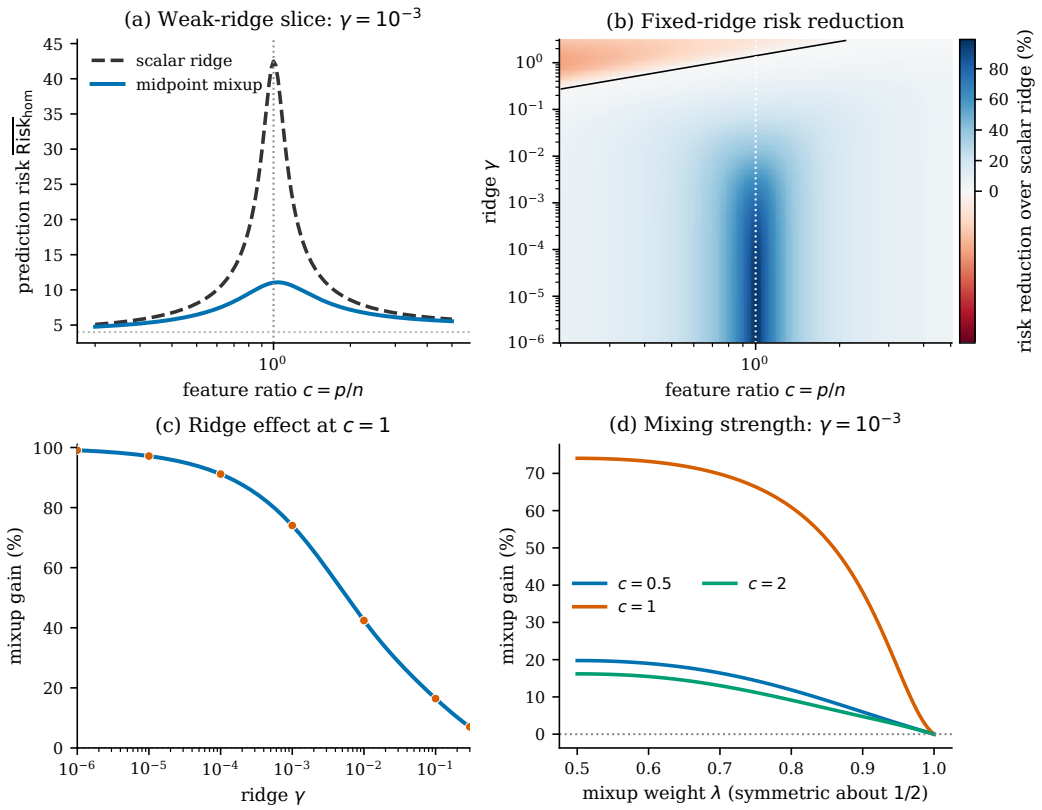

Figure S.6.8: Weak regularization with tanh features, unit signal energy, and training and test noise variance 4. (a) Limiting risks at $\gamma = 10^{-3}$. (b) Percentage reduction at the same ridge strength for each method. (c) Percentage reduction at $c = 1$, evaluated at 241 logarithmically spaced ridge values in $[10^{-6}, 0.3]$. (d) Mixing-weight dependence at $\gamma = 10^{-3}$, using 101 equally spaced weights in $[1/2, 1]$ and $c \in \{0.5, 1, 2\}$. All panels evaluate Theorem 3.1 deterministically.

We first compare endpoint ridge and midpoint mixup at the common strength $\gamma = 10^{-3}$. The main grid uses 161 logarithmically spaced ratios $c \in [0.2, 5]$ and 151 logarithmically spaced ridge values $\gamma \in [10^{-6}, 3]$.

Figure S.6.8(a) shows that midpoint mixup suppresses the interpolation peak near $c = 1$, consistent with its positive interaction variance $\kappa(1/2)$. Panel (c) traces this effect as ridge decreases, while panels (b) and (d) show its dependence on the feature ratio and mixing weight. This common-ridge comparison complements the separately tuned comparisons above.

S.7 Learned-feature neural-network experiments

These experiments examine whether concentrating mixup on observations with large fitting-progress scores improves prediction with learned neural features. We compare five procedures on three regression datasets, varying the strength and prevalence of controlled training-label perturbations. The study extends the instance-level perspective empirically: feature learning and adaptive pairing lie beyond the fixed-feature, full-pair theory.

S.7.1 Data and controlled perturbations

Table S.7.2 summarizes the datasets (Yeh, 1998; Tsanas & Xifara, 2012; Brooks et al., 1989). For each of ten seeds $\{2026091701, \ldots, 2026091710\}$, a NumPy default generator permutes the N row

Table S.7.2: Regression datasets and split sizes. All three datasets are available from the UCI repository under CC BY 4.0.

Dataset	Response	N	Covariates	Train	Validation	Test
Concrete	Compressive strength	1,030	8	618	206	206
Energy	Heating load	768	8	460	153	155
Airfoil	Sound pressure level	1,503	5	901	300	302

indices. The first $\lfloor 0.6N \rfloor$ rows form the training set, the next $\lfloor 0.2N \rfloor$ the validation set, and the remainder the test set. The same split and perturbations are used for every method.

Let n_{tr} be the training size, y_i^0 the original recorded response, and s_0 its training standard deviation with divisor n_{tr}. A fixed random permutation defines nested sets C_ρ of size round(ρn_{tr}). Independent random signs $\xi_i \in \{-1, 1\}$, with equal probabilities, are shared across perturbation levels. The observed training labels are

$$y_i = y_i^0 + \delta_i, \qquad \delta_i = a s_0 \xi_i \mathbf{1}\{i \in C_\rho\}.$$

Thus a controls perturbation magnitude and ρ its prevalence. We vary $a \in \{0, 1, 2, 4, 6, 8\}$ at $\rho = 0.01$, and $\rho \in \{0, 0.01, 0.05, 0.10, 0.20\}$ at $a = 4$. The two scans share the unperturbed fit and $(a, \rho) = (4, 0.01)$, giving nine distinct conditions. Nominal prevalence 1% corresponds to six, five, and nine perturbed labels on Concrete, Energy, and Airfoil. Validation and test labels remain unchanged.

Covariates are standardized using training means and standard deviations; responses are standardized using the observed training labels. The transformations are shared with both holdouts, and predictions are returned to original units for test root mean squared error (RMSE). The original responses and perturbation masks are used to construct the experiment and compute diagnostics; in perturbed conditions, fitting and selection receive only the observed labels.

Splits contain disjoint source rows, although repeated covariates can cross splits. Concrete has 996 distinct covariate vectors, with 5–11 exact covariate groups shared by training and test sets across the ten splits. Energy has 12 building shapes, defined by its first five covariates; all occur in both training and test sets, although no complete covariate vector crosses these sets. Airfoil also has no such exact matches. The experiment evaluates held-out rows from the same data source; unseen-building-shape prediction would require a different split.

S.7.2 Selection rule and comparison methods

A preliminary empirical-risk-minimization (ERM) network ranks training observations by their reduction in squared fitting error. With standardized inputs, standardized observed labels $\widetilde{y}_i$, and pilot prediction f_t after t updates, define

$$s_i = \left[(f_{75}(x_i) - \widetilde{y}_i)^2 - (f_{600}(x_i) - \widetilde{y}_i)^2\right]_+, \qquad [z]_+ = \max\{z, 0\}.$$

The largest $\lceil q n_{\text{tr}} \rceil$ scores form S_q, with ties resolved by training-row order. The selection fraction q is validation-tuned, while ρ specifies the experimental perturbation prevalence. This fitting-progress score can capture both signal learning and noise fitting; it does not estimate the theoretical sensitivity χ. The pilot uses training observations only. Label-dependent pairing also appears in C-Mixup and RC-Mixup (Yao et al., 2022; Hwang et al., 2024); here selection uses fitting progress.

Table S.7.3 summarizes the five procedures. All candidate fits start from a shared fresh initialization and retain only the pilot ranking. Each minibatch draws its first endpoints uniformly with replacement. For selected mixup, an endpoint $i \in S_q$ receives a uniform partner $j \in S_q^c$, and the training pair is

$$(x^{\text{mix}}, y^{\text{mix}}) = (\lambda x_i + (1 - \lambda) x_j, \lambda y_i + (1 - \lambda) y_j), \qquad \lambda \sim \text{Beta}(0.4, 0.4).$$

An endpoint outside S_q remains unchanged. Random-set mixup applies the same rule to a fixed random ranking, with nested sets as q varies. Uniform mixup mixes every first endpoint with an independently uniform partner from the full training set. The selective rule also reduces selected labels' total expected coefficients in the training targets because only unselected labels serve as partners.

Table S.7.3: The five neural-network procedures. Every procedure minimizes squared loss and selects a saved model by validation error. Additional tuning parameters are listed below; all grids are given in Section S.7.3.

Method	Mixed first endpoints	Partner pool	Additional tuning
ERM	None	—	None; zero decay
Tuned weight decay	None	—	Weight decay
Uniform mixup	All	All observations	Weight decay
Selected mixup	Top fitting-progress scores	Unselected set	Weight decay, q
Random-set mixup	Fixed random subset	Unselected set	Weight decay, q

S.7.3 Training and validation

The network has three hidden layers of width 128 and a scalar linear output; all parameters are trained. For tanh, hidden weights use Xavier-normal initialization with gain one. Output weights are independent normal variables with standard deviation $128^{-1/2}$, and all biases start at zero. AdamW uses learning rate 10^{-3}, minibatch size 96, and 1,500 updates per candidate. Weight decay supplies the neural-network regularization baseline. The pilot uses its own initialization, zero decay, and 600 updates, saving predictions at updates 75 and 600 for the score. Candidate methods share initialization and first-endpoint draws within each activation and split. Separate random streams supply partners and mixing weights, preserving the paired comparison as q varies.

Each method selects a checkpoint, meaning saved model parameters, from updates $\{100, 250, 500, 750, 1000, 1500\}$. All methods except zero-decay ERM search weight-decay coefficients $\{0, 10^{-4}, 10^{-2}, 10^{-1}, 1, 10, 100\}$. Selected and random-set mixup also search $q \in \{0.05, 0.10, 0.20\}$. Minimum validation mean squared error selects each method's settings; ties prefer an earlier checkpoint, smaller decay, then smaller q. ERM reuses the zero-decay endpoint trajectory with the same checkpoint rule. Thus selected and random-set mixup each use 21 trajectories per condition, while endpoint fitting and uniform mixup each use seven. The fixed-q comparisons below match selection-set size and retain separate validation choices of decay and checkpoint.

The three datasets, ten splits, and nine conditions require 15,120 candidate trajectories and 270 pilot fits per activation. Computation uses twelve single-thread CPU workers, PyTorch 2.3.1, and NumPy 2.4.6. The update budget is fixed for every trajectory. All selections are frozen before test scoring. Independent NumPy forward calculations checked retained-model predictions, and saved records were checked for validation selection and diagnostic consistency. Detailed run records accompany the reproducibility materials. This study follows earlier exploratory comparisons on these datasets; both activation protocols were fixed before their test evaluation. We report all three datasets. Means and standard deviations summarize ten paired repetitions of splitting, initialization, and training; the standard deviations describe this variation and are not confidence intervals.

S.7.4 Perturbation strength and prevalence on Concrete

Figure S.7.9(a) shows how targeting high-score observations changes prediction under increasingly strong perturbations. Without added noise, mean RMSE is 4.98 for selected mixup and 5.09 for tuned weight decay. At $a = 4$, the means are 5.63 ± 0.41, 6.43 ± 0.60, and 6.31 ± 0.54 for selected mixup, tuned weight decay, and random-set mixup, respectively; selected mixup has lower RMSE than both comparators in all ten paired repetitions. At $a = 8$, the corresponding means are 6.26, 6.96, and 7.10. The advantage thus persists as sparse label perturbations become stronger.

In the prevalence scan, Figure S.7.9(b), selected mixup has the lowest mean RMSE among the five methods at every tested positive prevalence. At $\rho = 20\%$, its mean is 8.40, compared with 9.14 for tuned weight decay. The next section examines whether the advantage over random sets persists when their sizes are matched.

S.7.5 Other datasets and fixed-q comparisons

Figure S.7.10 extends the two scans to Energy and Airfoil. At $a = 4$ and $\rho = 1\%$, selected mixup and tuned weight decay have mean RMSE 1.12 versus 1.88 on Energy, and 2.20 versus 2.21 on Airfoil.

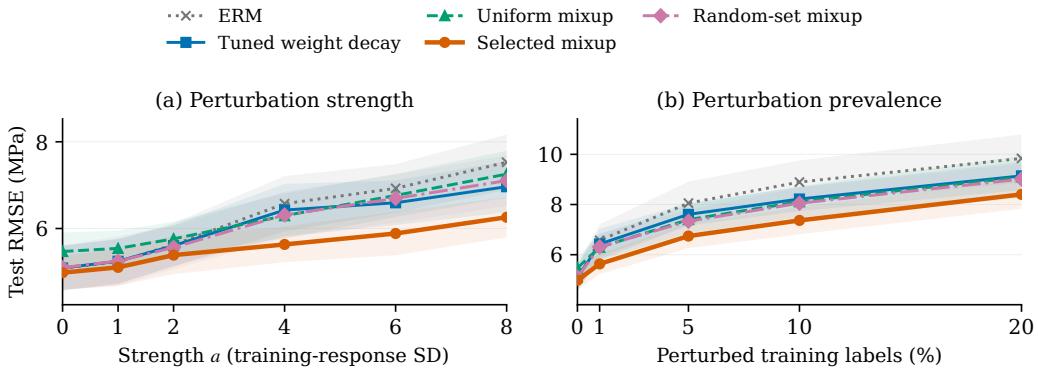

Figure S.7.9: Selected mixup with a learned tanh network on Concrete. (a) Perturbation magnitude a at prevalence $\rho = 1\%$. (b) Prevalence ρ at magnitude $a = 4$. Lines show mean test RMSE in megapascals (MPa), and bands show one standard deviation over ten paired splits. Every marker is an evaluated condition; the two zero-perturbation points share the same fit for each method and split. All settings and checkpoints are validation-selected.

Airfoil's separation grows at $a = 8$, to 2.63 versus 3.10. Without added perturbations, selected mixup has slightly higher means on Energy (0.67 versus 0.58) and Airfoil (1.91 versus 1.79). The strongest gains therefore occur under substantial label perturbations, with their onset varying by dataset.

The selected and random-set curves tune q independently. To compare their rankings at the same mixing frequency, Figure S.7.11 fixes a common q at $(a, \rho) = (4, 0.01)$, allowing each method to select its own decay and checkpoint. Concrete's selected-set means range from 5.73 to 5.85, versus 6.36 to 6.38 for random sets; Energy has the same ordering at all three set sizes. Airfoil is sensitive to set size: at $q = 20\%$, selected mixup gives 2.49, versus 2.38 for random-set mixup. Both the ranking and the selection fraction matter.

S.7.6 Selection and noise-fitting diagnostics

Known perturbation locations let us measure selection precision $|S_q \cap C_\rho|/|S_q|$ and recall $|S_q \cap C_\rho|/|C_\rho|$ after training. To describe how a fitted predictor absorbs the injected perturbations, we also compute the training noise-fitting projection

$$A(\widehat{f}; \delta) = \frac{\sum_i \delta_i \{\widehat{f}(x_i) - y_i^0\}}{\sum_i \delta_i^2}.$$

Predictions and responses are in original units, making A dimensionless. Fitting the perturbed labels exactly at affected observations gives $A = 1$; fitting their original responses gives $A = 0$. The projection can lie outside $[0, 1]$ and is undefined without added perturbations. It describes training residuals relative to recorded responses y_i^0, which may themselves contain noise. It does not estimate χ or decompose test risk.

On Concrete at $a = 4$, $\rho = 1\%$, and $q = 5\%$, selection recall averages 85%, versus 5% for random sets, with precision 16.5%. At prevalence 10% and 20%, recall falls to 50% and 25%, while precision reaches 100%. This illustrates the coverage limit imposed by a fixed selection budget. At $a = 4$, the noise-fitting projection averages 0.165 for validation-tuned selected mixup and 0.291 for tuned weight decay. At $a = 6$ and 8, weight decay has the smaller projection but larger test RMSE, showing why prediction performance must be assessed on the held-out observations as well.

S.7.7 ReLU extension

We repeat the nine perturbation conditions with rectified linear units (ReLU), preserving the splits, perturbations, pilot rule, optimizer, training budget, tuning grids, and validation procedure. Hidden

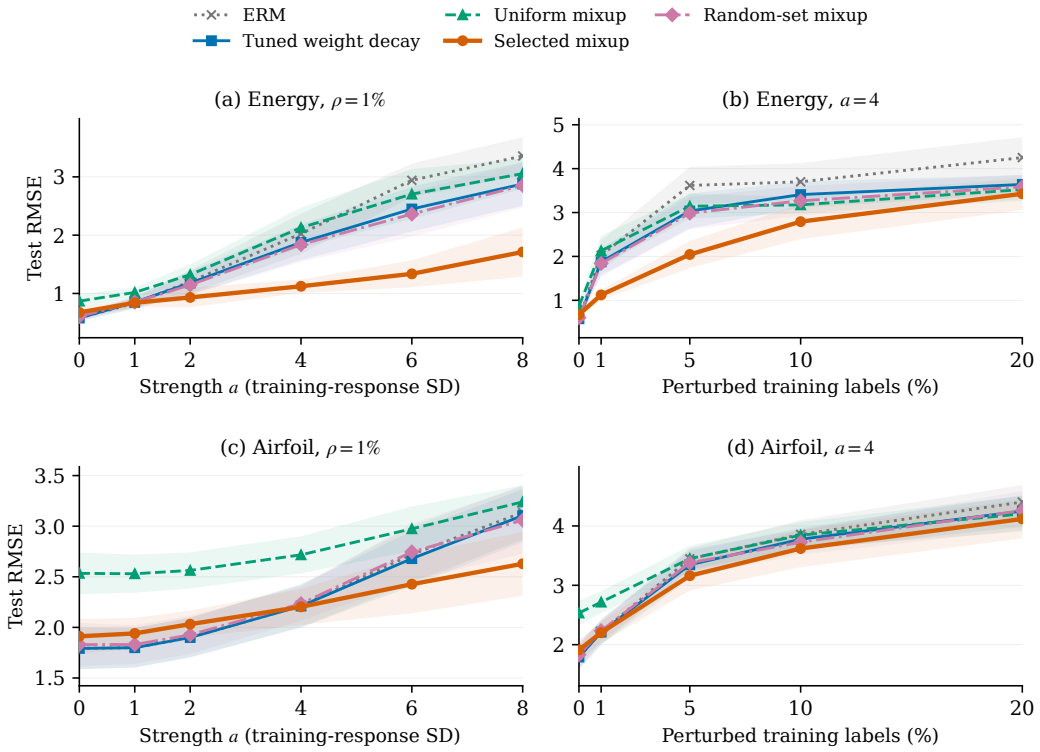

Figure S.7.10: Learned tanh networks on Energy and Airfoil under the same strength and prevalence scans as Figure S.7.9. Lines and bands show mean test RMSE and one standard deviation over ten paired splits. Each condition uses separately validation-selected settings; the two zero points share the same fit for each method and split.

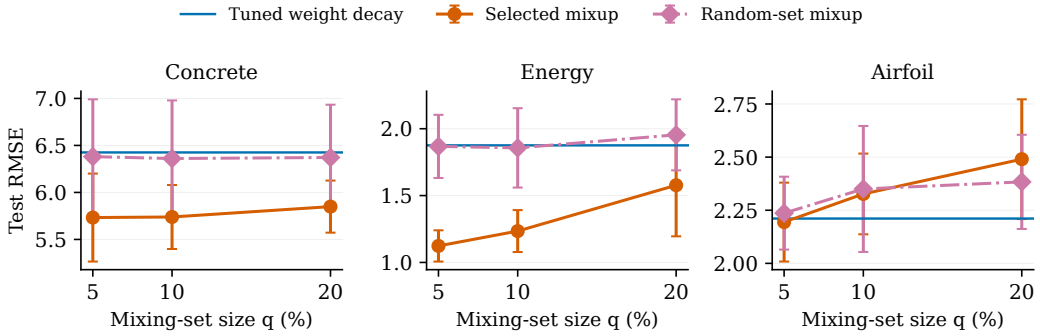

Figure S.7.11: Fixed-q comparison for learned tanh networks at $a = 4$ and $\rho = 1\%$. Selected and random-set mixup use the same fraction q, with decay and checkpoint separately validation-selected. Points and bars show mean test RMSE and one standard deviation over ten paired splits. The horizontal reference is tuned weight decay's mean.

weights use Kaiming-normal initialization; the output initialization and zero biases are unchanged. Both activation protocols were fixed before test evaluation. This comparison assesses the procedure with each activation's specified initialization.

Prediction under perturbations. Figure S.7.13 shows gains on Concrete with ReLU. At $(a, \rho) = (4, 0.01)$, selected mixup, tuned weight decay, and random-set mixup have mean RMSE 5.59, 6.15, and 6.04 MPa. At $a = 8$, their means are 6.39, 8.17, and 7.88. In the prevalence scan at $\rho = 20\%$,

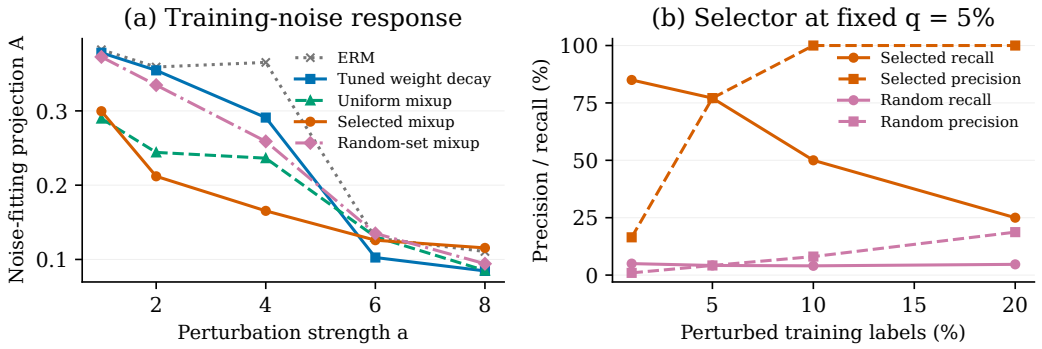

Figure S.7.12: Selection and noise-fitting diagnostics on Concrete with learned tanh features. (a) Training noise-fitting projection at $\rho = 1\%$ for validation-selected procedures. (b) Precision and recall at fixed $q = 5\%$ and $a = 4$. Points average ten paired splits. Zero-perturbation conditions are omitted because the projection and recall are undefined.

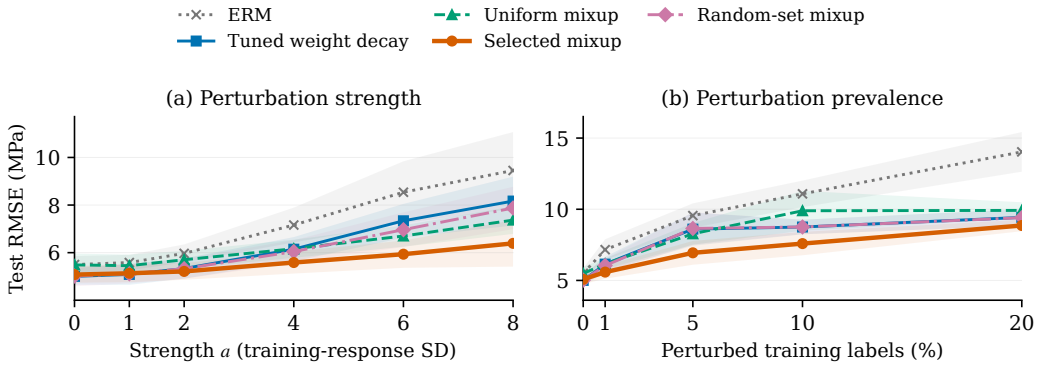

Figure S.7.13: Concrete with learned ReLU features, using the strength and prevalence scans in Figure S.7.9. Lines and bands show mean test RMSE in MPa and one standard deviation over ten paired splits. Settings and checkpoints are validation-selected; each method's two zero points share its unperturbed fit within a split.

the corresponding means are 8.86, 9.43, and 9.41. The improvement therefore persists under both stronger and more prevalent perturbations.

Figure S.7.14 extends the comparison to Energy and Airfoil. At $(a, \rho) = (4, 0.01)$, selected mixup and tuned weight decay give 1.44 versus 2.42 on Energy and 2.46 versus 2.52 on Airfoil. Airfoil's gap grows at $\rho = 10\%$, to 3.41 versus 3.99. Without added perturbations, selected mixup has slightly higher means on all three datasets: 5.08 versus 5.00 on Concrete, 0.69 versus 0.60 on Energy, and 2.25 versus 2.08 on Airfoil. Together with the tanh results, these scans locate the clearest benefits under substantial training-label perturbations for both activations.

Selection at fixed set size and fitting diagnostics. At $(a, \rho) = (4, 0.01)$, Figure S.7.15 shows lower mean RMSE for selected than random sets on Concrete and Energy at all three values of q. On Airfoil at $q = 20\%$, their means are 2.66 and 2.68, both above tuned weight decay's 2.52. For Concrete at $q = 5\%$, selection recall averages 95%, versus 5% for random sets, with precision 18.4%. The noise-fitting projection averages 0.207 for validation-tuned selected mixup and 0.381 for tuned weight decay (Figure S.7.16). These diagnostics describe selection and training behavior using the definitions in Section S.7.6.

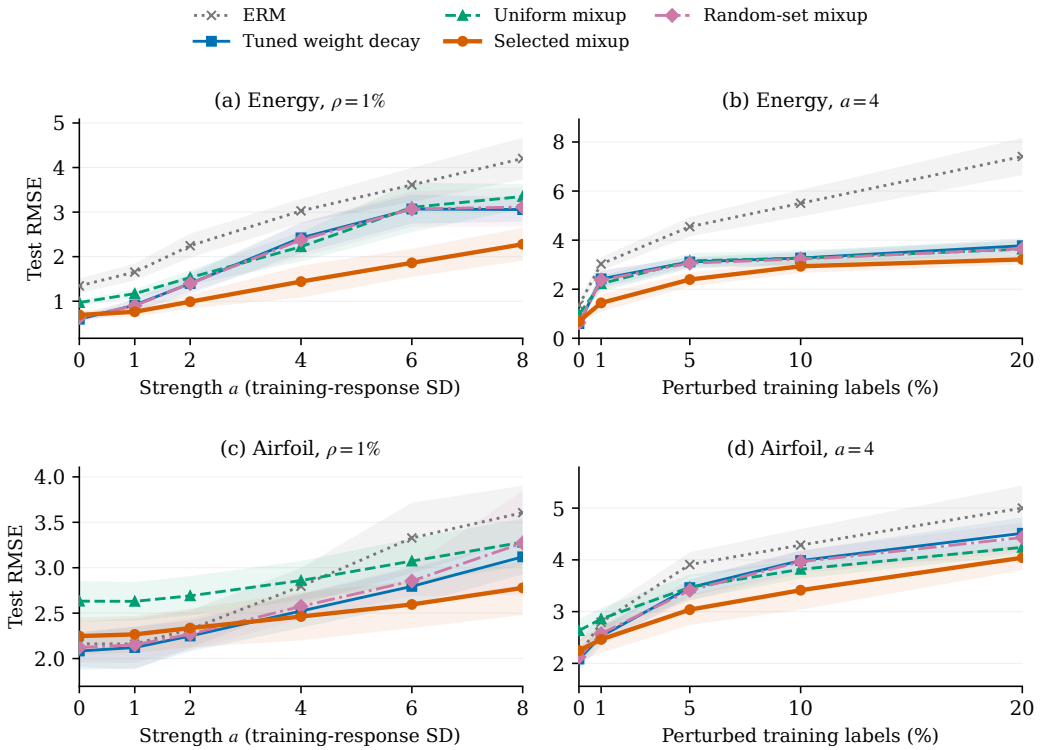

Figure S.7.14: Energy and Airfoil with learned ReLU features. Lines and bands show mean test RMSE and one standard deviation over ten paired splits for all five methods. Settings are validation-selected at each condition; the two zero points share the same fit for each method and split.

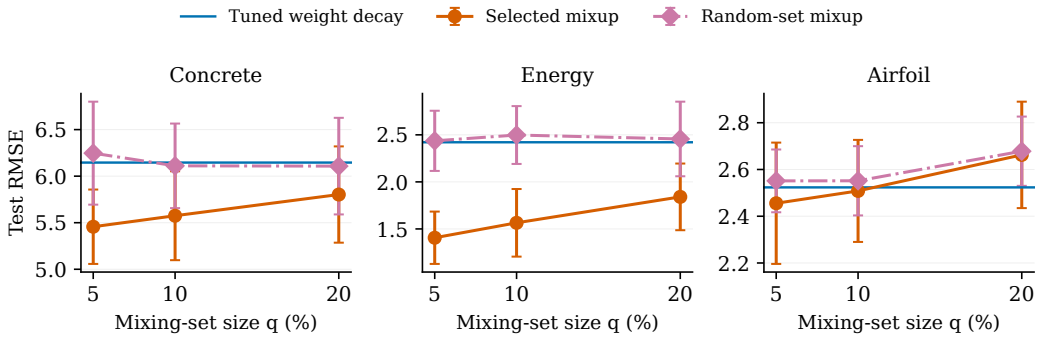

Figure S.7.15: Fixed-q comparisons with ReLU at $a = 4$ and $\rho = 1\%$. Decay and checkpoint are separately validation-selected at each common q. Points and bars show mean test RMSE and one standard deviation over ten paired splits; the horizontal reference is tuned weight decay's mean.

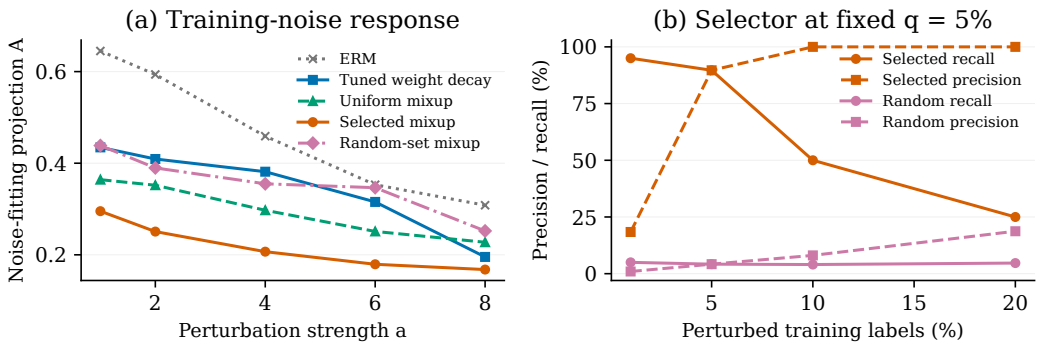

Figure S.7.16: Selection and noise-fitting diagnostics on Concrete with ReLU. (a) Training noise-fitting projection at $\rho = 1\%$ for validation-selected procedures. (b) Precision and recall at fixed $q = 5\%$ and $a = 4$. Points average ten paired splits; zero-perturbation conditions are omitted as in Figure S.7.12.