THE STATISTICAL COST OF REDUNDANT WIDTH IN BAYESIAN QUADRATIC NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

How much prediction error do redundant directions contribute when a quadratic network learns a low-rank target, and when does regularization suppress that contribution? We study this question through the singular geometry of the factor map $W \mapsto WW^T$. For a global Gaussian factor prior, fixed dimensions, and width $m \geq d+1$, we prove a noisy boundary law for expected Gibbs and posterior-mean prediction risks. At matched likelihood temperature, its unregularized mean risk is $\sigma^2[p+\alpha(d-r)]/n+o(n^{-1})$, where $p = d(d+1)/2$ and $\alpha = (m-d-1)/2$: each additional neuron costs $\sigma^2(d-r)/(2n)+o(n^{-1})$ within the same PSD function class. An Euler–Stein identity establishes beneficial vanishing regularization; a PSD-estimator theorem calibrates the risk gap. The same boundary law resolves teacher modes of order $n^{-1/2}$. At zero added decay, every finite nonzero PSD local signal lowers mean risk below the zero-signal endpoint. At $m = d+1$, the zero- and strong-signal endpoints coincide, yet the risk is strictly lower between them. The population quartic-to-quadratic crossover as a regularization gap opens explains the geometry; the empirical law retains the surviving random tilt. Finite-data experiments and limiting-law quadrature examine these distinct predictions. The broader local-prior empirical crossover remains conjectural. The proved law links the width cost to regularization and a weak-signal transition that fixed-rank coefficients alone cannot describe.

1 INTRODUCTION

Once a quadratic network $W \in \mathbb{R}^{d\times m}$ can represent every positive semidefinite (PSD) matrix $A = WW^\top$, further width does not expand its function class. It still changes a Gaussian factor prior's mass near a low-rank truth: the induced matrix density is proportional to $(\det A)^\alpha e^{-\gamma \operatorname{tr} A/2}$, where $\alpha = (m-d-1)/2$. We ask how this boundary mass affects prediction when noise, regularization, and weak teacher modes act on comparable scales. A noisy boundary law for the global learner determines their combined effect on expected risk.

The simplest consequence is an exact leading cost of width. For a rank-r teacher, fixed dimensions, matched likelihood temperature, and zero added decay, the Gaussian-prior result for $m \geq d+1$ gives

$$\mathbb{E}R_{\text{mean}} = \frac{\sigma^2}{n}\{p+\alpha(d-r)\} + o(n^{-1}), \qquad p = \frac{d(d+1)}{2}. \tag{1}$$

The excess over the $\alpha = 0$ matrix-prior reference is $\alpha(d-r)\sigma^2/n$: each additional neuron costs $\sigma^2(d-r)/(2n) + o(n^{-1})$. Holding the represented-matrix prior fixed instead gives identical posteriors across saturated widths. The effect concerns the specified factor-prior family.

Why an endpoint coefficient is not enough. The coefficient in Equation (1) equals $2\sigma^2\lambda$, with λ the Gram learning coefficient. This endpoint does not describe risk under statistical-scale regularization or for teachers approaching rank loss. Both questions require the posterior law with its random training score retained.

The boundary law and its consequences. Redundant factor directions enter $WW^\top$ quadratically and fluctuate on an $n^{-1/4}$ scale, so their represented matrix fluctuates on the same $n^{-1/2}$ scale as

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.

the score. Our central theorem turns this law into expected Gibbs and posterior-mean risks for a learner on the entire factor space (Theorem 3.1). At matched temperature, regularization shifts a determinant-weighted Gaussian law on the PSD cone. An Euler–Stein identity gives the risk curve under $\eta_n = 2u/\sqrt{n}$: suppressing inactive predictions yields a linear initial gain against quadratic active shrinkage bias (Corollary 3.2). A PSD estimator on the same data calibrates the cost, separating the gap already present at $\alpha = 0$ from the additional width contribution (Theorem 4.1 and Corollary 4.2).

A transition invisible to the endpoint coefficient. Weak teacher eigenvalues of order $n^{-1/2}$ shift the same inactive law, with risk now centered at the shifted truth (Proposition 5.1). At zero added decay, every finite nonzero PSD local signal strictly lowers mean risk relative to the zero-signal endpoint (Corollary 5.2). At critical width $m = d + 1$, the zero- and strong-signal endpoints both equal $p\sigma^2$, yet the risk lies strictly below them for every finite nonzero local signal. Coincident endpoints therefore miss the intermediate behavior. At larger widths, only improvement over the zero endpoint is universal. Finite-data matrix experiments test width and regularization effects; noncentral quadrature separately evaluates the weak-mode limiting law.

What the analysis adds to prior work. Singular learning theory (SLT) relates local volume exponents to Bayesian errors (Watanabe, 2009; 2010b;a). Our Gram coefficient has close precedents in probabilistic PCA and factor analysis (Drton et al., 2026; 2025); Appendix J.7 compares the models and dimension ranges explicitly. The compact-prior endpoint in Proposition 3.3 makes this connection explicit. The Gaussian crossover additionally determines the regularization and local-signal dependence absent from that coefficient.

Prior-dependent boundary limits are also established. Bochkina & Green (2014, Sec. 4) obtain polynomially tilted truncated Gaussian components under an orthant condition. Our inactive PSD block requires a different boundary analysis. Schur-complement coordinates identify its determinant-weighted law; normalized moment bounds, including the critical exponent $\alpha = 0$, and exceptional-design control justify *expected squared risk* for a global learner. The bounds support compact-parameter uniformity in regularization and local signal; the resulting risk identities and weak-signal inequality go beyond identifying the posterior limit. Prior dependence itself is not a novelty claim.

Wishart factor volume and finite temperature appear in extensive-width quadratic learning (Maillard et al., 2024; Martin et al., 2026). The width-invariant regularized ERM of Erba et al. (2025) concerns a joint limit with n of order d^2 and widths of order d. This is compatible with our fixed-dimensional posterior-mean result: a width comparison must specify its learner and limit. Spectral work such as Defilippis et al. (2026) motivates examining weak modes rather than only exact rank. Here the global Gaussian law supplies that transition; population results explain the geometry, while the broader local-prior empirical crossover remains conjectural.

2 Model and Prediction Risk

For $x \sim \mathcal{N}(0, I_d)$, use centered quadratic features and observe

$$X(x) = \frac{xx^\top - I_d}{\sqrt{2}}, \quad f_W(x) = \langle X(x), WW^\top \rangle, \quad y_i = \langle X_i, A_* \rangle + \epsilon_i, \tag{2}$$

where $W \in \mathbb{R}^{d\times m}$, $X_i = X(x_i)$, and the independent noises are $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The teacher is PSD with rank $r < \min(d, m)$ and fixed positive active eigenvalues. Dimensions are fixed unless explicitly stated otherwise; this separates the prior-induced width effect from a growing-dimensional change in estimation difficulty. Matched-temperature results require $\sigma > 0$; separate noiseless reference statements allow $\sigma = 0$. For symmetric H, $\mathbb{E}_x \langle X(x), H\rangle^2 = \|H\|_F^2$, so excess prediction risk is exactly

$$R(W) = \mathbb{E}_x(f_W(x) - f_*(x))^2 = K(W) := \|WW^\top - A_*\|_F^2. \tag{3}$$

Let $L_n(W) = n^{-1}\sum_i (y_i - f_W(x_i))^2$. The global learner uses Gaussian prior precision $\gamma > 0$, likelihood coefficient $\beta > 0$, and added weight decay $\eta_n = 2u/\sqrt{n}$; its posterior is given in Equation (11). Likelihood matching means $\beta = 1/(2\sigma^2)$. The teacher is fixed, not drawn from the

prior, and no prior-averaged optimality is assumed. The learner uses neither the true rank nor a local cutoff. With $A = WW^\top$ and $\bar{A}_n = \mathbb{E}_{\pi_n} A$, define

$$R_G = \mathbb{E}_{\pi_n}\|A - A_*\|_F^2, \qquad R_{\mathrm{mean}} = \|\bar{A}_n - A_*\|_F^2, \tag{4}$$

$$R_G - R_{\mathrm{mean}} = \mathbb{E}_{\pi_n}\|A - \bar{A}_n\|_F^2. \tag{5}$$

The statistical targets average these conditional risks over the training data $\mathcal{D}_n$. In particular, the mean predictor uses $\mathbb{E}[WW^\top]$, not $\mathbb{E}[W]\mathbb{E}[W]^\top$.

In a teacher eigenbasis, $A_* = \mathrm{diag}(A_+, 0_h)$. We use $h = d - r$, $l = m - r$, $k = hl$, $q = dr - r(r-1)/2$, and $p = d(d+1)/2$. The transverse factor block $D \in \mathbb{R}^{h\times l}$ has $Q(D) = \|DD^\top\|_F^2$ and $T(D) = \|D\|_F^2$. For represented blocks,

$$\|A - A_*\|_F^2 = \|A_{11} - A_+\|_F^2 + 2\|A_{12}\|_F^2 + \mathrm{tr}(A_{22}^2). \tag{6}$$

The count $q = r(r+1)/2 + rh$ consists of the symmetric active block and the active–inactive cross block. These directions can fluctuate to either side of the truth locally. The remaining $h(h+1)/2$ represented coordinates form the inactive PSD block, whose true value lies at the tip of the cone. They cannot fluctuate as an unconstrained centered Gaussian matrix. In factor coordinates there are instead $k = hl$ transverse variables, and their matrix product generates this block. Increasing m changes k and the volume of factor configurations representing a given matrix, while leaving the represented dimension $p = q + h(h+1)/2$ fixed. These two coordinate counts are the geometric source of the distinction between capacity and prior-induced statistical cost.

The inactive contribution is thus distinct from a change in total risk between widths. Write $\alpha = (m-d-1)/2$, $g = \eta/2$, and $\tau = n\beta$; τ is an independent concentration parameter for the population reference. The notation $\mathrm{LT}_+(r)$ denotes lower-triangular matrices with positive diagonal; $W_*\,\mathrm{O}(m)$ is the factor orbit representing A_*.

3 A NOISY BOUNDARY LAW FOR PREDICTION RISK

We first identify the law that determines the risk curve. Its random score and regularization shift remain of the same order in the inactive block; the zero-decay coefficient follows as a special case. Let Z be the centered Gaussian on $\mathrm{Sym}(\mathbb{R}^d)$ with covariance $\sigma^2 I$ in Frobenius coordinates; its diagonal and off-diagonal variances are σ^2 and $\sigma^2/2$. For $Z_\perp$ its inactive block, define

$$p_{u,Z_\perp}(D) \propto \exp[-\beta\{Q(D) + 2\,\mathrm{tr}((uI_h - Z_\perp)DD^\top)\}], \qquad D \in \mathbb{R}^{h\times l}. \tag{7}$$

The quartic term makes this law integrable even for negative eigenvalues of $uI_h - Z_\perp$. Write $M = DD^\top$, $\langle\cdot\rangle_{u,Z_\perp}$ for conditional expectation, and

$$C_G(u) = \mathbb{E}_Z\langle\|M\|_F^2\rangle, \qquad C_{\mathrm{mean}}(u) = \mathbb{E}_Z\|\langle M\rangle\|_F^2, \tag{8}$$

$$\mathcal{T}(u) = \mathbb{E}_Z\langle\mathrm{tr}\,M\rangle, \qquad \mathcal{V}(u) = C_G(u) - C_{\mathrm{mean}}(u). \tag{9}$$

A matrix view of the boundary law. For $m \geq d+1$ at matched temperature, the same conditional law can be written in represented-matrix coordinates:

$$p(M \mid Z_\perp, u) \propto (\det M)^\alpha \exp\left[-\frac{\|M - (Z_\perp - uI_h)\|_F^2}{2\sigma^2}\right], \qquad M \succ 0. \tag{10}$$

This is a Gaussian score observation restricted to the PSD cone and weighted by the factor-induced determinant power. At critical width $m = d+1$, $\alpha = 0$, so the restriction alone remains. Additional width changes the weight inside the same cone. Even at critical width the conditional mean is not the unconstrained score: averaging over the cone leaves a positive inactive component. Regularization shifts the center by $-uI_h$, reducing that component while also shrinking the active teacher block. The theorem below makes this description precise for both posterior sampling and posterior-mean prediction.

Theorem 3.1 (global Gaussian-prior empirical crossover)**.** *For the data model of Section 2, fix d, m, r with $0 \le r < d$ and $m \ge d+1$, and fix $\beta, \sigma, \gamma > 0$. On all of $\mathbb{R}^{d\times m}$, use the posterior*

$$\pi_n^{\mathrm{Gauss}}(\,\mathrm{d}W \mid \mathcal{D}_n) \propto e^{-\gamma\|W\|_F^2/2} e^{-n\beta\{L_n(W)+(2u/\sqrt{n})\|W\|_F^2\}}\,\mathrm{d}W. \tag{11}$$

Define its risks as in Equation (4). Uniformly for u in compact subsets of $[0,\infty)$,

$$n\mathbb{E}_{\mathcal{D}_n} R_G \longrightarrow q\sigma^2 + \frac{q}{2\beta} + ru^2 + C_G(u), \tag{12}$$

$$n\mathbb{E}_{\mathcal{D}_n} R_{\mathrm{mean}} \longrightarrow q\sigma^2 + ru^2 + C_{\mathrm{mean}}(u), \tag{13}$$

$$n\mathbb{E}_{\mathcal{D}_n}\mathbb{E}_{\pi_n^{\mathrm{Gauss}}} \operatorname{tr}(A_{22}^2) \longrightarrow C_G(u). \tag{14}$$

Corollary 3.2 (beneficial vanishing regularization)**.** *Under Theorem 3.1, let $r \ge 1$ and $\beta = 1/(2\sigma^2)$. The limiting mean-risk coefficient satisfies*

$$F(u) := q\sigma^2 + ru^2 + C_{\mathrm{mean}}(u) = \sigma^2[p+\alpha h] + ru^2 - u\mathcal{T}(u), \qquad F'(0+) = -\mathcal{T}(0) < 0. \tag{15}$$

It attains a global minimum at a positive finite u_*; every minimizer obeys $u_* < \sigma\sqrt{k/(2r)}$ and $2ru_* = \mathcal{T}(u_*) - 2\beta u_* \mathbb{E}_Z \operatorname{Var}_{u_*,Z}(\operatorname{tr} M)$. For every sufficiently small fixed $u > 0$, weight decay $\eta_n = 2u/\sqrt{n}$ strictly reduces the data-averaged mean-predictor risk relative to $u = 0$ for all sufficiently large n.

From the law to the risk curve. Euler's identity and Gaussian integration by parts give the balance (Proposition A.2)

$$C_G(u) + u\mathcal{T}(u) = \frac{k}{4\beta} + 2\beta\sigma^2\mathcal{V}(u), \qquad C_{\mathrm{mean}}(u) = C_G(u) - \mathcal{V}(u).$$

Likelihood matching makes $2\beta\sigma^2 = 1$, so the variance terms cancel in the mean risk, yielding Equation (15). The gain begins linearly in u, whereas active shrinkage bias is quadratic. At $u = 0$, one additional neuron adds $\sigma^2 h/(2n) + o(n^{-1})$ to mean risk. The coefficient does not depend on fixed prior precision or fixed positive teacher eigenvalues (Corollary A.3); ordinary fixed-width variance normalization therefore preserves this effect. A fixed represented-matrix prior, likelihood, and trace penalty instead yield identical posteriors at all saturated widths. The result is specific to the factor-prior family.

Mechanism and proof. Near the teacher, q regular coordinates enter the represented matrix linearly, while k inactive factor coordinates enter through $DD^\top$. The Gram learning coefficient is $\lambda = q/2 + k/4$ (Proposition G.1): regular and inactive factor scales are $n^{-1/2}$ and $n^{-1/4}$, respectively. Both represented blocks thus fluctuate on the statistical $n^{-1/2}$ scale. The population crossover interpolates the inactive *energy coefficient* from $k/4$ to $k/2$ on the scale $s = g\sqrt{\tau}$, with its dimension-dependent window $s \asymp \sqrt{\max(h,l)}$ (Appendices H and I); this coefficient is distinct from prediction risk.

The distinction between energy and prediction is essential here. In the population reference, each inactive factor coordinate has quartic confinement at zero gap and quadratic confinement once the gap dominates its fluctuation scale. This changes its contribution to the rescaled *penalized energy* from one quarter to one half. A larger energy coefficient does not imply larger test error: the coefficient also counts the penalty, and the underlying measure has no realized training score. For noisy prediction one must first form the posterior conditional on that score, then average its prediction error over data. Setting the score to its zero mean before taking posterior expectations would change the answer, because normalization and the PSD constraint make those expectations nonlinear in the score.

To see how training noise enters, put $B = \sqrt{n}(A - A_*)$, $Z_n = n^{-1/2}\sum_i \epsilon_i X_i$, and $\mathcal{I}_n H = n^{-1}\sum_i \langle X_i, H\rangle X_i$. The loss is exactly quadratic in the represented matrix:

$$n\{L_n(A_* + B/\sqrt{n}) - L_n(A_*)\} = \langle B, \mathcal{I}_n B\rangle - 2\langle Z_n, B\rangle, \qquad \mathcal{I}_n \to I, \quad Z_n \Rightarrow Z.$$

Thus the score contributes $-2\langle Z_\perp, DD^\top\rangle$ at leading order, and cannot be discarded as a smaller perturbation of the quartic loss. The statistical scale also matches the penalty: $n(2u/\sqrt{n})\operatorname{tr} A$ contributes $2u \operatorname{tr} B$ after removing its constant part.

The Gaussian factor prior pushes forward to $(\det A)^{\alpha} e^{-\gamma \operatorname{tr} A/2}$. In a Schur-complement chart, its determinant splits into a regular positive-block factor and the inactive determinant weight $(\det M)^{\alpha}$. The regular factor and the fixed γ disappear from the normalized leading law, but α survives. Equivalently, pushing forward the flat factor volume $\mathrm{d}D$ in Equation (7) gives precisely this inactive weight. The limit separates a q-dimensional tangent Gaussian with mean $Z_T - uP_r$ and covariance $(2\beta)^{-1} I_q$ from the tilted inactive law. Here P_r is the projector onto the active space, and Z_T retains the active and cross blocks of Z. Its three risk contributions are sampling variance $q\sigma^2$, Gibbs variance $q/(2\beta)$, and shrinkage bias ru^2.

For $r > 0$, the Schur complement makes the separation concrete. Since A_+ is positive definite, A_{11} remains invertible near the teacher, and

$$S = A_{22} - A_{12}^T A_{11}^{-1} A_{12} \succeq 0, \qquad \det A = \det A_{11} \det S. \tag{16}$$

For $r = 0$, simply take $S = A$. When $r > 0$, the cross block is of order $n^{-1/2}$, so its quadratic correction to A_{22} is of order n^{-1}. Thus $\sqrt{n}S$ and $\sqrt{n}A_{22}$ have the same leading limit M, while the chart retains the exact finite-n PSD constraint. The factor $\det A_{11}$ tends to a positive constant and cancels under normalization. Rescaling S extracts a power of n from $(\det S)^{\alpha}$ but leaves $(\det M)^{\alpha}$ in the limiting density. This is why the teacher's fixed positive spectrum drops out of the leading coefficient whereas the width-dependent exponent does not.

From posterior limits to expected risk. Convergence of the local density alone does not control squared error averaged over training sets. On designs whose sensing operator is close to the identity, the proof bounds normalized posterior moments by polynomials in the Gaussian score, retaining the PSD boundary contribution at $\alpha = 0$. On the remaining designs, a rough global bound from the fixed Gaussian factor prior combines with their small probability to make their expected-risk contribution vanish (Appendix C). Thus γ disappears from the leading coefficient but still helps control the finite-sample posterior. These bounds justify the global theorem and its compact-u uniformity for $m \geq d + 1$; the singular determinant weight at $m = d$ requires further work.

The endpoint and its wider prior scope. At zero decay, SLT recovers the same mean-risk coefficient for smooth positive compact factor priors, including widths outside the Gaussian crossover theorem. This identifies the common endpoint, while leaving the full regularization and weak-signal law for those priors unresolved. The global compact-ball alternative and the local version of Equation (33) are stated together below.

Proposition 3.3 (matched Bayesian prediction risk)**.** *For the fixed-dimensional data model of Section 2, $r < \min(d, m)$, and the local prior of Equation (33), let $\sigma > 0$, $\eta_n = 0$, and $\beta = 1/(2\sigma^2)$. Then*

$$\mathbb{E}_{\mathcal{D}_n} R_{\mathrm{mean}} = \frac{\sigma^2}{n}\left(q + \frac{k}{2}\right) + o(n^{-1}), \qquad n\mathbb{E}_{\mathcal{D}_n} R_G \longrightarrow 2\sigma^2\left(\frac{q}{2} + \frac{k}{4} + \nu_{\mathrm{SLT}}\right). \tag{17}$$

Here $q + k/2 = 2\lambda$, and ν_{SLT} is the singular fluctuation at likelihood temperature one. The same conclusions hold for a global posterior on a fixed compact ball containing the true orbit in its interior, with a smooth strictly positive prior density on the ball.

4 Calibrating the boundary cost

The boundary law specifies the Bayesian learner's risk. To assess its size, we compare with a PSD estimator on the same data and at the same rescaled trace penalty. This distinguishes the additional width cost from the gap already present at $\alpha = 0$; the smaller c/n term controls poorly conditioned designs.

Theorem 4.1 (PSD estimator and its expected risk)**.** *Fix d, a rank-$r < d$ PSD teacher with fixed positive active eigenvalues, $\sigma > 0$, and $c > 0$. For $n \geq d(d+1)/2$, let*

$$\widehat{A}_n(u) = \operatorname*{argmin}_{A \succeq 0}\left\{ L_n(A) + \left(\frac{2u}{\sqrt{n}} + \frac{c}{n}\right) \operatorname{tr} A \right\}. \tag{18}$$

The minimizer exists and is almost surely unique. Uniformly for u in compact subsets of $[0, \infty)$,

$$n\mathbb{E}\|\widehat{A}_n(u) - A_*\|_F^2 \longrightarrow F_{\mathrm{PSD}}(u) := q\sigma^2 + ru^2 + \mathbb{E}\|(Z_\perp - uI_h)_+\|_F^2. \tag{19}$$

In particular, $F_{\mathrm{PSD}}(0) = \sigma^2[q + h(h+1)/4]$.

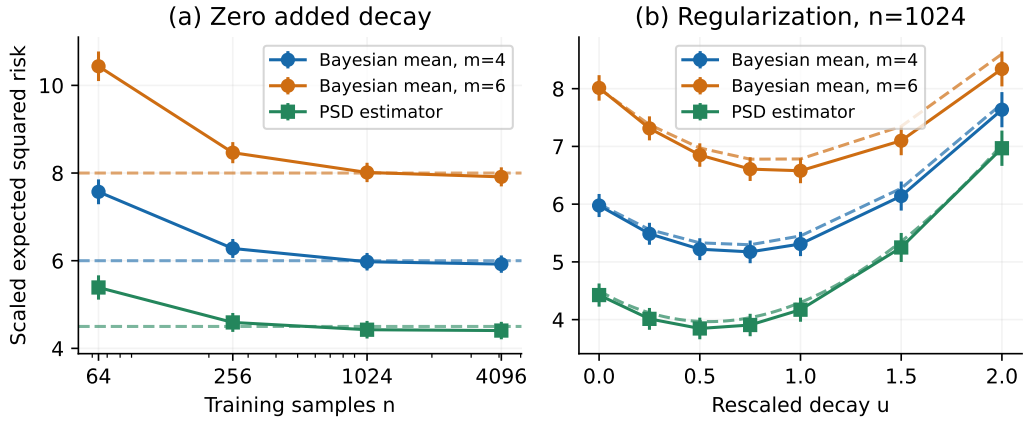

Figure 1: Bayesian means and the PSD estimator on identical data with $d = 3$, $A_* = \mathrm{diag}(1, 0, 0)$, $\sigma = \gamma = 1$, and $\beta = 1/2$. (a) Zero added decay. (b) Regularization at $n = 1024$. Dashed lines are the limiting coefficients, including the full $F_{\mathrm{PSD}}(u)$ curve. The PSD estimator includes the fixed $1/n$ trace term. Bars are twice the standard error over 1024 data replicates; Bayesian risks include posterior Monte Carlo noise.

Corollary 4.2 (two contributions to the zero-decay risk gap)**.** *Under Theorem 3.1 at matched temperature and $u = 0$, for the comparator in Equation (18),*

$$n\left\{\mathbb{E}R_{\mathrm{mean}} - \mathbb{E}\|\widehat{A}_n(0) - A_*\|_F^2\right\} \longrightarrow \sigma^2\left\{\frac{h(h+1)}{4} + \alpha h\right\}. \tag{20}$$

The first gap term is present at $\alpha = 0$; the second is the additional boundary-exponent cost. For $d = 3, r = 1, \sigma = 1$, the zero-decay coefficients are 4.5 for the comparator and 6 for the $\alpha = 0$ Bayesian mean. Beneficial regularization refers to improvement over the learner's own unregularized risk; no all-u ordering or optimality is asserted. The proof uses tangent-cone projection and exceptional-design control (Appendix D). At finite n, eigenvalue clipping alone generally does not solve Equation (18). The weak-mode extension is Corollary A.5.

For this example $h = 2$, $q = 3$, and $p = 6$. Widths $m = 4, 5, 6$ correspond to $\alpha = 0, 1/2, 1$, giving Bayesian mean-risk coefficients $6, 7, 8$. The PSD coefficient stays at 4.5, so the predicted gaps are $1.5, 2.5, 3.5$. These are coefficients of $1/n$, rather than nonvanishing error floors. They also clarify the regularization tradeoff: because $r = 1$, the active bias contributes u^2, while the inactive contribution falls initially. The existence of a useful positive u is therefore compatible with risk rising again when shrinkage becomes too strong. Both features are visible in the finite-data comparisons below.

The contrast is especially transparent in the inactive block. In the limit, the PSD estimator minimizes a quadratic objective over the cone and returns $(Z_\perp - uI_h)_+$. The Bayesian predictor instead averages all positive definite matrices under Equation (10). Projection and conditional averaging therefore differ even when $\alpha = 0$. For example, in one inactive dimension at $\alpha = 0$, writing $y = Z_\perp - u$, the posterior mean is $y + \sigma\phi(y/\sigma)/\Phi(y/\sigma)$, whereas projection is $\max(y, 0)$; here ϕ, Φ are the standard normal density and distribution function. This illustrates the boundary contribution to the risk gap without attributing the whole gap to excess width. It also explains why the comparison must specify the estimator, rather than only its PSD function class.

5 Beyond Fixed Rank: Weak Teacher Modes

At critical width, Equation (1) has the same coefficient for every fixed rank. Nevertheless, the risk between rank endpoints is not constant. To resolve this transition, we keep the same learner and move the teacher on the statistical scale. Take a rank-r base teacher $A_0 = \mathrm{diag}(A_+, 0_h)$ and

generate labels under

$$A_{*,n} = A_0 + n^{-1/2}\Theta, \qquad \Theta = \mathrm{diag}(0_r, \Theta_\perp), \qquad \Theta_\perp \succeq 0. \tag{21}$$

Here A_+ and $\Theta_\perp$ are fixed; h, l, q, k refer to the base rank r. The learner uses the same global prior and trace penalty as before and does not require knowledge of this decomposition. A weak eigenvalue $a_n = t/\sqrt{n}$ has signal-to-noise ratio t/σ in the limiting experiment. In this section, Z denotes the $h \times h$ inactive Gaussian block.

Proposition 5.1 (local alternatives at a rank boundary)**.** *Under the assumptions of Theorem 3.1, with the teacher replaced by Equation (21), put $\mu_\Theta(u, Z) = \langle M \rangle_{u, Z+\Theta_\perp}$, using Equation (7). Then*

$$n\mathbb{E}_{\mathcal{D}_n} R_{\mathrm{mean}} \longrightarrow F_\Theta(u) := q\sigma^2 + ru^2 + \mathbb{E}_Z \|\mu_\Theta(u, Z) - \Theta_\perp\|_F^2, \tag{22}$$

$$n\mathbb{E}_{\mathcal{D}_n} R_G \longrightarrow q\sigma^2 + \frac{q}{2\beta} + ru^2 + \mathbb{E}_Z \langle \|M - \Theta_\perp\|_F^2 \rangle_{u, Z+\Theta_\perp}. \tag{23}$$

Expectations use data generated under $A_{*,n}$, and both risks are measured against $A_{*,n}$. Convergence is uniform when u ranges over a compact subset of $[0, \infty)$ and $\Theta_\perp$ over a compact subset of the PSD cone.

The exact loss expansion adds $\mathcal{I}_n\Theta$ to the score when centering at A_0. This shift depends on the design and converges to Θ; its limiting tangent projection is zero. The tangent Gaussian is therefore unchanged, while the inactive law retains the weak signal. The proof in Appendix E controls the centered moments and exceptional designs for this sequence of teachers.

Corollary 5.2 (weak-mode risk balance and strict improvement)**.** *Under Proposition 5.1 at $\beta = 1/(2\sigma^2)$, write $\bar{\mu}_\Theta(u) = \mathbb{E}_Z \mu_\Theta(u, Z)$ and $\mathcal{T}_\Theta(u) = \mathrm{tr}\, \bar{\mu}_\Theta(u)$. Then*

$$F_\Theta(u) = q\sigma^2 + ru^2 + \frac{k\sigma^2}{2} - u\mathcal{T}_\Theta(u) + \|\Theta_\perp\|_F^2 - \langle \Theta_\perp, \bar{\mu}_\Theta(u) \rangle. \tag{24}$$

For $\Theta_\perp = tH$, with $H \succeq 0$ nonzero, denote this coefficient by $F_{tH}(u)$. At zero added decay,

$$\frac{\mathrm{d}}{\mathrm{d}t} F_{tH}(0)\Big|_{t=0+} = -\frac{\mathcal{T}_0(0)}{h}\, \mathrm{tr}\, H < 0. \tag{25}$$

For every width $m \geq d + 1$ and every nonzero finite $\Theta_\perp \succeq 0$,

$$F_\Theta(0) < F_0(0) = \sigma^2\{p + \alpha h\}, \qquad p = d(d+1)/2. \tag{26}$$

The subtraction of $\Theta_\perp$ in Equation (22) matters as much as the score shift. A positive weak signal raises the inactive posterior mean, but prediction is now judged against that same positive signal. Consequently a larger predicted inactive matrix need not mean a larger error. The balance Equation (24) separates these effects: $\|\Theta_\perp\|_F^2$ is offset by its inner product with the averaged posterior mean. At the zero-signal endpoint the mean already points into the cone, so moving the truth a small distance in a PSD direction decreases risk to first order. The following argument extends this initial decrease to a comparison with the zero endpoint for every finite nonzero PSD signal; it does not assert that risk decreases along the entire path.

Why the improvement is strict. At $u = 0$ and matched temperature, write $Y = Z + \Theta_\perp$ and let Ξ be an independent isotropic Gaussian on $\mathrm{Sym}(\mathbb{R}^h)$ with covariance $\sigma^2 I$. The Gaussian convolution of the determinant weight is

$$\Psi_\alpha(Y) = \mathbb{E}_\Xi[(\det(Y + \Xi))^\alpha \mathbf{1}_{\{Y+\Xi \succ 0\}}], \qquad \mu_\Theta(0, Z) = Y + \sigma^2 \nabla \log \Psi_\alpha(Y).$$

For every $\alpha \geq 0$, Ψ_α has strictly positive derivatives along every nonzero PSD direction. A Schur-complement factorization proves this including $\alpha = 0$, where the derivative comes from the moving boundary (Appendix E). Substitution into Equation (24) gives

$$F_\Theta(0) = F_0(0) - \sigma^2 \mathbb{E}_Z \langle \Theta_\perp, \nabla \log \Psi_\alpha(Z + \Theta_\perp) \rangle < F_0(0).$$

The strict inequality holds for $\Theta_\perp \succeq 0$ nonzero: the improvement follows from the positive definite shift of the posterior mean from the score.

Reading the local signal scale. The parameter t measures a weak eigenvalue relative to statistical fluctuations, rather than on the scale of the fixed active eigenvalues. For a teacher eigenvalue $a_n = t/\sqrt{n}$, increasing sample size along the local-alternative sequence leaves this ratio fixed. By contrast, holding a positive eigenvalue a fixed makes $\sqrt{n}\,a$ grow: eventually that mode belongs to the regular active block. The local experiment bridges these descriptions by allowing t to vary only after taking the sample-size limit. It describes why a small eigenvalue can behave differently from both an exact zero and a well-resolved positive eigenvalue. Uniformity over bounded t gives control within this local regime, but does not justify taking t to infinity together with n. The strong endpoint below is therefore an iterated limit, not a rate statement for fixed finite data.

The strong-signal endpoint and nonuniformity. If H has rank $s \geq 1$, recentering the inactive law at tH gives the rank-$(r+s)$ law as $t \to \infty$ (Lemma E.1). In particular, at matched temperature,

$$\lim_{t\to\infty} F_{tH}(0) = \sigma^2\{p + \alpha(h-s)\}. \tag{27}$$

For $\alpha = 0$ the zero- and strong-signal endpoints coincide, and every finite positive t lies strictly below both. The equal fixed-rank coefficients therefore hide a nonconstant local risk curve (Figure 2). For $\alpha > 0$, the strong-signal endpoint is lower than the zero-signal endpoint by $\alpha s\sigma^2$. The strict inequality above does not compare the curve with this lower endpoint; such a comparison can fail (Remark E.2). No monotonicity in t is asserted. Already the sequence $a_n = t/\sqrt{n}$ proves that the fixed-positive-spectrum coefficient is not uniform near rank loss: at $\alpha = 0$ every finite positive t has a different coefficient, while at $\alpha > 0$ continuity near $t = 0$ separates it from the lower strong endpoint. The limit in Equation (27) is taken after the local-alternative limit; substituting $t = \sqrt{n}\,a$ for fixed $a > 0$ supplies no finite-sample error bound.

6 Numerical validation

Width, regularization, and the PSD benchmark. We test the global Gaussian learner with $d = 3$, $A_* = \mathrm{diag}(1, 0, 0)$, $\sigma = \gamma = 1$, and $\beta = 1/2$. The 1024 independent design/score replicates are paired across widths and regularization values. Four-chain Hamiltonian Monte Carlo targets an exact reparameterization of the global posterior, retaining the empirical sensing operator and fixed prior. The PSD estimator uses the same data and $c = 1$ (Figure 1). Methods and diagnostics, including posterior Monte Carlo error, are in Appendices O and P; Appendix N retains all settings, scalar calibrations, and figures.

At $n = 4096$, the paired scaled mean-risk increments for $m = 4 \to 5$ and $5 \to 6$ are 0.987 ± 0.019 and 1.003 ± 0.018, against the predicted unit cost. Here $\pm$ is twice the across-data standard error, including residual posterior Monte Carlo noise. At $n = 64$, these increments are 1.412 ± 0.044 and 1.452 ± 0.042: finite-sample corrections are substantial. At $n = 1024$, $u = 0.5$ reduces scaled risk by 0.757 ± 0.075 and 1.163 ± 0.076 for widths $4, 6$. Subsequent increases in risk reflect active shrinkage bias; the grid does not identify a continuous optimizer. At $n = 4096, u = 0$, Bayesian-minus-PSD gaps are 1.516 ± 0.020 and 3.507 ± 0.047, versus the predictions 1.5 and 3.5.

These comparisons test distinct parts of the theory. Paired width increments isolate the predicted boundary-exponent cost while sharing the same data fluctuations; they do not by themselves validate the absolute risk level. The Bayesian-minus-PSD gap tests the additional cost of conditional averaging at the boundary as well as the width contribution. The regularization sweep then tests the balance between suppressing inactive predictions and biasing active ones. The substantial corrections at $n = 64$ show why the limiting coefficient should not be read as an accurate finite-sample formula at every sample size. Error bars describe sampling uncertainty in the reported estimates; they do not bound the asymptotic remainder or certify absence of posterior-sampling bias.

Weak modes and finite positive spectra. Noncentral matrix quadrature evaluates Equation (22) for $d = 3$, base rank $r = 1$, $\Theta_\perp = t\,\mathrm{diag}(1, 0)$, unit noise, matched temperature, and $u = 0$ (Figure 2). At $\alpha = 0$, the risk coefficient at $t = 1.5$ is 5.358, below the common endpoint 6. At $\alpha = 1$, the value 6.677 at $t = 2$ also numerically falls below the strong endpoint 7; only improvement over the zero endpoint 8 is general at positive α. Order refinement, Cartesian-score integration, and the shifted risk identity check the quadrature (Appendix Q).

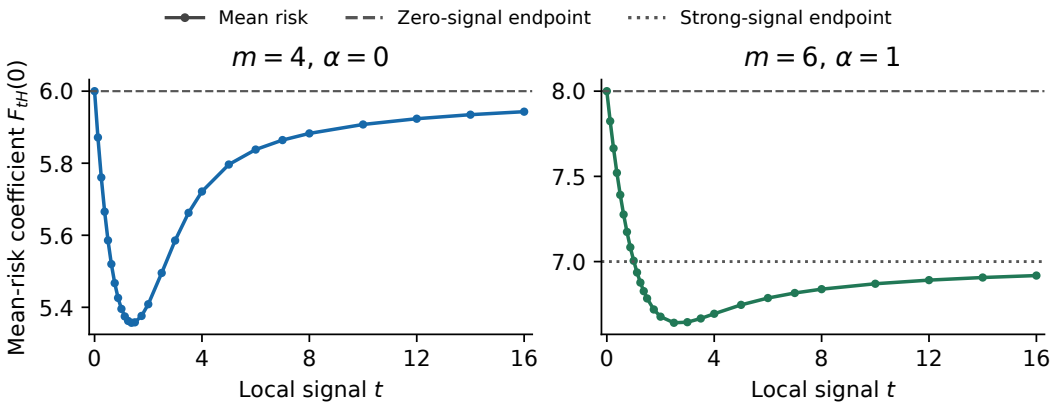

Figure 2: Weak-mode risk for $d = 3$, base rank $r = 1$, $\sigma = 1$, $\beta = 1/2$, and $u = 0$, with $a_n = t/\sqrt{n}$ in one inactive direction. At $m = 4$ ($\alpha = 0$), every finite $t > 0$ lies below both endpoints analytically. At $m = 6$ ($\alpha = 1$), the theorem proves improvement over the zero-signal endpoint; improvement over the lower strong-signal endpoint is illustrated numerically here and is not universal at positive α (see Remark E.2). Points are deterministic quadrature; lines join values. These are limiting coefficients, not finite-sample risks.

Separate rank-two experiments use $d = 4$, $A_* = \mathrm{diag}(1, a, 0, 0)$ with $a = 0.3, 1$. At $n = 4096, u = 0, m = 5$, changing only a gives paired scaled-risk difference -0.076 ± 0.018, although the fixed-spectrum limits agree. This motivates resolving the weak-mode transition; the two-by-two quadrature does not evaluate this experiment's three-by-three local prediction or bound its remainder.

The two figures consequently have different evidential roles. Figure 1 uses finite training samples and the actual global learner. Figure 2 numerically integrates the limiting experiment after the sample-size limit has been taken. Agreement between quadrature resolutions checks that integration, but does not demonstrate how quickly finite-data weak-mode risks approach it. The rank-two contrast above records precisely the kind of residual spectrum dependence that a finite-sample analysis would need to explain; its different inactive-block dimension prevents treating the displayed two-dimensional weak-mode curves as a quantitative fit.

7 Discussion and Limitations

The noisy boundary law connects the width cost, beneficial vanishing regularization, and weak-mode transition. At critical width, rank-independent endpoint coefficients conceal strictly lower risk at finite nonzero local signals. The PSD benchmark separately shows why equal function classes need not give equal risk for different prediction rules.

The scope is fixed-dimensional quadratic regression with isotropic Gaussian inputs and noise. The global Gaussian crossover is proved for $m \geq d + 1$; the compact-prior endpoint has wider width scope, while Gaussian widths $m \leq d$ and the general local-prior crossover remain unresolved. Weak-mode improvement is relative to the zero-signal endpoint; a dip below the strong endpoint is not universal at positive α. The results provide neither finite-sample remainder bounds, an optimal data-dependent tuning rule, nor a joint growing-dimension limit.

Connecting local and fixed-spectrum regimes at finite sample size remains open. Bounds away from rank loss miss the transition; compact-local-signal convergence does not control approach to a fixed positive spectrum. The positive-spectrum discrepancies and limiting-law quadrature motivate this analysis without supplying remainder estimates. The boundary law reveals how prior mass becomes prediction error beyond what a single learning coefficient describes.

AI USE STATEMENT

Generative AI tools, including Anthropic's Claude and OpenAI's Codex, assisted with literature search, mathematical formulation, proof development, manuscript editing, and numerical code. The author is responsible for the final text, mathematical claims, citations, and reported results.

REPRODUCIBILITY STATEMENT

An anonymized code archive is provided as supplementary material. All script paths below are relative to its root; the accompanying README documents dependencies, reproduction commands, and the correspondence between scripts and reported results. The population results, matched Bayesian risk endpoint, Gaussian-prior empirical crossover, PSD-comparator expected-risk law, weak-mode extension, and exact reduced-model identities are accompanied by proofs in Appendices B to F and J to L. The general local-prior crossover remains Conjecture A.4; the square-block equilibrium prediction retains its moment-concentration assumption. The scalar experiments use synthetic Gaussian inputs and label noise. Their script, random seed, quadrature settings, raw numerical summaries, and generated table are included in `scripts/scalar_risk/`. Deterministic matrix quadratures in `scripts/empirical_proof_checks/` check the empirical proof's boundary inequality and conditional limits at a fixed score and sensing operator; they do not measure risk averaged over random training sets. The separate positive-rank matrix study in `scripts/matrix_risk/` includes random-data sufficient statistics, per-dataset risk and diagnostic summaries, exact-posterior samplers, independent limiting quadrature, and figure generation. Its sampling checks and uncertainty conventions are stated in Appendix O. The PSD comparison and rank-two study in `scripts/v005_validation/` add a convex solver, exact global samplers in two coordinate systems, paired sufficient statistics and risk summaries, and diagnostic records (Appendix P). Noncentral two-by-two quadrature in `scripts/weak_modes/` integrates the shifted score and its orientation, checks the risk identity against direct squared error, and records order and cutoff refinement plus a separate Cartesian-score integration (Appendix Q). It evaluates a limiting law, not finite-sample risk. The earlier population spectral figure is generated by `scripts/crossover_figure/`.

ETHICS STATEMENT

This work studies mathematical models and synthetic data. It involves no human subjects, personal data, or deployed systems.

REFERENCES

Dennis Amelunxen, Martin Lotz, Michael B. McCoy, and Joel A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014. arXiv:1303.6672.

Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*. Cambridge University Press, 2010.

Miki Aoyagi and Sumio Watanabe. Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Networks*, 18(7):924–933, 2005.

Natalia A. Bochkina and Peter J. Green. The Bernstein–von Mises theorem and nonregular models. *The Annals of Statistics*, 42(5):1850–1878, 2014. doi: 10.1214/14-AOS1239. arXiv:1211.3434.

Zhongtian Chen and Daniel Murfet. Modes of sequence models and learning coefficients. *arXiv preprint arXiv:2504.18048*, 2025.

Ben Cullen, Sergio Estan-Ruiz, Riya Danait, and Jiayi Li. A basin-selection perspective on grokking via singular learning theory. *arXiv preprint arXiv:2603.01192*, 2026.

Leonardo Defilippis, Yizhou Xu, Julius Girardin, Emanuele Troiani, Vittorio Erba, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Scaling laws and spectra of shallow neural networks in the feature learning regime. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.24882.

Mathias Drton, Elizabeth Gross, Dimitra Kosta, Anton Leykin, Andrew McCormack, Seth Sullivant, and Daniel Windisch. Singular learning theory for factor analysis. *arXiv preprint arXiv:2511.15419*, 2025.

Mathias Drton, Andrew McCormack, and Daniel Windisch. Asymptotics for model selection in probabilistic principal component analysis. *arXiv preprint arXiv:2608.23513*, 2026.

Vittorio Erba, Emanuele Troiani, Lenka Zdeborová, and Florent Krzakala. The nuclear route: Sharp asymptotics of ERM in overparameterized quadratic networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.17958.

Peter J. Forrester. *Log-Gases and Random Matrices*. London Mathematical Society Monographs. Princeton University Press, 2010.

Antoine Maillard, Emanuele Troiani, Simon Martin, Lenka Zdeborová, and Florent Krzakala. Bayes-optimal learning of an extensive-width neural network from quadratically many samples. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Simon Martin, Giulio Biroli, and Francis Bach. High-dimensional analysis of gradient flow for extensive-width quadratic neural networks. *Journal of Machine Learning Research*, 27(136):1–182, 2026. arXiv:2601.10483.

Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley, 1982.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC. *Bayesian Analysis*, 2021. doi: 10.1214/20-BA1221. arXiv:1903.08008.

Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*, volume 25 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, 2009.

Sumio Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010a.

Sumio Watanabe. Equations of states in singular statistical estimation. *Neural Networks*, 23(1): 20–34, 2010b. doi: 10.1016/j.neunet.2009.08.002. arXiv:0712.0653.

A ADDITIONAL EMPIRICAL IDENTITIES AND SCOPE

Proposition A.1 (empirical loss expansion)**.** *For symmetric H, define*

$$\mathcal{A}_n(H)_i = n^{-1/2} \langle X_i,\, H \rangle, \qquad \mathcal{I}_n = \mathcal{A}_n^* \mathcal{A}_n, \qquad Z_n = n^{-1/2} \sum_{i=1}^{n} \epsilon_i X_i. \tag{28}$$

Regarding L_n as a function of the represented matrix, one has exactly

$$L_n(A_* + H) - L_n(A_*) = \|\mathcal{A}_n(H)\|_F^2 - \frac{2}{\sqrt{n}} \langle Z_n,\, H \rangle. \tag{29}$$

At fixed d, $\mathcal{I}_n \to I$ almost surely in operator norm and $Z_n \Rightarrow Z$, where Z is a centered Gaussian on $\mathrm{Sym}(\mathbb{R}^d)$ with $\mathbb{E}\langle Z,\, H\rangle \langle Z,\, J\rangle = \sigma^2 \langle H,\, J\rangle$ for symmetric H, J.

Proposition A.2 (risk balance in the tilted ensemble)**.** *For all fixed $h, l \geq 1$, $u \geq 0$, $\beta > 0$, and $\sigma \geq 0$, the quantities in Equations (8) and (9) are finite and satisfy*

$$C_G(u) = \frac{k}{4\beta} - u\mathcal{T}(u) + 2\beta\sigma^2 \mathcal{V}(u). \tag{30}$$

Consequently, at $\sigma > 0$ and $\beta = 1/(2\sigma^2)$,

$$C_{\mathrm{mean}}(u) = \frac{k}{4\beta} - u\mathcal{T}(u), \qquad \frac{C_G(0)}{2\sigma^2} = \frac{k}{4} + \beta\mathcal{V}(0). \tag{31}$$

The Euler–Stein proof is in Appendix B. The conditional variance term distinguishes Gibbs and mean risk.

Noise activation and width. For a realized $Z_\perp$, minimizing the energy in Equation (7) over $M \succeq 0$, $\operatorname{rank} M \leq l$, retains the largest positive eigenvalues of $Z_\perp - uI_h$, up to rank l. This rank constraint becomes inactive when $l \geq h$. The finite-temperature distribution still depends on l: already for $h = 1$, the represented scalar $a = \|D\|_F^2$ has density

$$p_l(a \mid z) \propto a^{l/2-1} e^{-\beta\{a^2 + 2(u-z)a\}}, \qquad a > 0. \tag{32}$$

Width saturation of the feasible set therefore does not imply width-independent posterior prediction risk. For the Gaussian factor prior with $m \geq d$, the represented density is proportional to $(\det A)^\alpha e^{-\gamma \operatorname{tr} A/2}$ on $A \succ 0$, with $\alpha = (m - d - 1)/2$. At matched temperature and $u = 0$, the coefficient proved for $m \geq d + 1$ can be written $q + k/2 = d(d+1)/2 + \alpha(d - r)$. At $m = d + 1$ this equals the ambient symmetric-matrix dimension $d(d+1)/2$, regardless of teacher rank; every further neuron adds $(d - r)/2$. Width changes the prior's behavior near the rank boundary even after the PSD function class has saturated. With a fixed prior on represented matrices, the same likelihood and the same trace penalty define exactly the same posterior across saturated widths, at every sample size. The width cost here is therefore specific to the factor-prior family. At matched Bayes temperature the rescaled noise amplitude $\sigma\sqrt{\beta} = 1/\sqrt{2}$ remains nonzero. The population ensemble is a reference calculation, not a replacement for this near-threshold empirical fluctuation law.

Corollary A.3 (fixed-width normalization and teacher spectrum)**.** *For each fixed width $m \geq d + 1$, the limits in Theorem 3.1 are unchanged if the fixed prior precision is any $\gamma_m > 0$. They also agree for all fixed teachers with the same rank, even when their fixed positive active eigenvalues differ, under the stated isotropic design. In particular, at matched temperature and zero added decay, two fixed adjacent widths with arbitrary fixed precisions have leading risk difference $\sigma^2(d - r)/(2n)$.*

These are direct consequences of the theorem: its coefficients contain neither γ nor the positive teacher eigenvalues. For example, the normalization $A = WW^\top/m$ with a fixed-precision Gaussian prior on W is equivalent to a Gaussian prior of precision $m\gamma$ on the factors representing A. It retains the same boundary exponent. This statement concerns zero added decay, or a trace penalty matched in represented coordinates; it does not assert invariance of finite-sample risks or allow n-dependent precision. The local-alternative result in Proposition 5.1 resolves the lack of uniformity as a positive teacher eigenvalue approaches zero.

Compact-prior alternatives. For the local endpoint and the separate local conjecture, replace the Gaussian prior in Equation (11) by $\chi(W)\varphi(W)$, where φ is a fixed smooth positive density and χ is smooth, nonnegative, compactly supported in a sufficiently small neighbourhood of the true factor orbit, and positive on that orbit:

$$\pi_n^{\mathrm{loc}}(\,\mathrm{d}W \mid \mathcal{D}_n) \propto \chi(W)\varphi(W)e^{-n\beta\{L_n(W)+\eta_n\|W\|_F^2\}}\,\mathrm{d}W. \tag{33}$$

This is an oracle local learner. The zero-decay matched endpoint also covers the fixed global compact prior specified in Proposition 3.3. Its scope is distinct from that of the global Gaussian-prior crossover.

Conjecture A.4 (general local empirical risk limit)**.** *For the local learner of Equation (33), at every fixed width with $r < \min(d, m)$, $\beta > 0$, $\sigma \geq 0$, and $\eta_n = 2u/\sqrt{n}$ with $u \geq 0$, the limits Equations (12) to (14) hold with its own posterior in place of π_n^{Gauss}, uniformly on compact u-intervals.*

The global Gaussian theorem does not prove this general local-prior conjecture or its Gaussian analogue at $m \leq d$.

Corollary A.5 (PSD risk for weak teacher modes)**.** *Under the data model and fixed-parameter assumptions of Theorem 4.1, replace the teacher by Equation (21). The same estimator satisfies*

$$n\mathbb{E}\|\widehat{A}_n(u) - A_{*,n}\|_F^2 \longrightarrow q\sigma^2 + ru^2 + \mathbb{E}\|(Z_\perp + \Theta_\perp - uI_h)_+ - \Theta_\perp\|_F^2. \tag{34}$$

The convergence is uniform on compact nonnegative u intervals and compact PSD sets of $\Theta_\perp$. Its proof is in Appendix E.

When $c = \gamma/(2\beta)$, the PSD estimator is the mode of the $\alpha = 0$ posterior in represented-matrix coordinates; it need not be the factor-coordinate mode.

Additional observable conventions. Test error against a fresh noisy label is $\sigma^2 + R$. For uncentered $x^\top Ax$, risk is $2\|A - A_*\|_F^2 + (\mathrm{tr}(A - A_*))^2$, so its risk formulas differ. Bayesian predictive KL loss is a separate observable; Lemma F.2 establishes the matched-endpoint conversion.

Factor prior	Width range	Matched zero-decay mean risk	Empirical crossover
Smooth local compact	$r < \min(d, m)$	Proved	Conjectured
Smooth global compact	$r < \min(d, m)$	Proved	Not established here
Global isotropic Gaussian	$m \geq d + 1$	Proved	Proved
Global isotropic Gaussian	$r < m \leq d$	Not established here	Not established here

Table 1: Statistical scope at fixed dimensions and $\sigma > 0$, with a rank-$r < d$ teacher. Compact priors satisfy the positivity and support conditions in Equation (33) and Proposition 3.3. The crossover uses $\eta_n = 2u/\sqrt{n}$ and is uniform over bounded nonnegative u-intervals; Proposition 5.1 also allows compact sets of PSD local signals for the global Gaussian learner. Population results and reduced-ensemble identities have their separately stated scopes.

B Exact empirical identities and the tilted ensemble

B.1 Proof of Proposition A.1

For $A = A_* + H$, the residual is $\epsilon_i - \langle X_i, H\rangle$. Expanding its square and subtracting $n^{-1}\sum_i \epsilon_i^2$ gives Equation (29). Since $\mathbb{E}\,\langle X, H\rangle\,\langle X, J\rangle = \langle H, J\rangle$, the strong law applied to the entries of the empirical covariance operator gives $\mathcal{I}_n \to I$ in operator norm in the fixed-dimensional space $\mathrm{Sym}(\mathbb{R}^d)$. Independence of input and noise gives

$$\mathbb{E}\big[\,\langle \epsilon X, H\rangle\,\langle \epsilon X, J\rangle\,\big] = \sigma^2\,\langle H, J\rangle\,.$$

The multivariate central limit theorem gives the asserted limit of Z_n. The two convergences hold jointly by Slutsky's theorem; no independence between Z_n and $\mathcal{I}_n$ at finite n is asserted. In a Frobenius-orthonormal basis, each Gaussian coordinate has variance σ^2, so an off-diagonal matrix entry has variance $\sigma^2/2$. □

B.2 Conditional moments and Gaussian integration by parts

Fix $Z \in \mathrm{Sym}(\mathbb{R}^h)$ and abbreviate $\langle\cdot\rangle = \langle\cdot\rangle_{u,Z}$. Since $Q(D) \geq \|D\|_F^4/\min(h,l)$, the density Equation (7) and all its polynomial moments are integrable for each Z. Integration of $\mathrm{div}_D(De^{-\beta\{Q+2\,\mathrm{tr}((uI-Z)DD^\top)\}})$ gives

$$\frac{k}{4\beta} = \langle Q\rangle + u\langle T\rangle - \langle Z,\, \langle M\rangle\rangle\,. \tag{35}$$

These moments grow at most polynomially in Z at the orders needed here. Indeed, with $b = \|Z - uI_h\|_F$ and $x = \sqrt{\langle Q\rangle}$, Cauchy–Schwarz and Equation (35) imply $x^2 \leq k/(4\beta) + bx$, and hence

$$\langle Q\rangle \leq b^2 + \frac{k}{2\beta}. \tag{36}$$

Also $\|\langle M\rangle\|_F \leq x$ and $\langle T\rangle \leq \sqrt{h}\,x$.

Let (E_a) be a Frobenius-orthonormal basis of $\mathrm{Sym}(\mathbb{R}^h)$, and write $Z_a = \langle Z,\, E_a\rangle$, $M_a = \langle M,\, E_a\rangle$. Differentiation of the normalized integral gives

$$\partial_{Z_a}\langle M_b\rangle = 2\beta\big(\langle M_a M_b\rangle - \langle M_a\rangle\langle M_b\rangle\big).$$

On compact sets of Z, quartic decay justifies differentiation under the integral. The polynomial bound Equation (36) bounds both the means and the covariance trace, so Gaussian integration by parts is legitimate for $Z_a \sim \mathcal{N}(0,\sigma^2)$:

$$\mathbb{E}_Z\,\langle Z,\,\langle M\rangle\rangle = \sigma^2\mathbb{E}_Z\sum_a \partial_{Z_a}\langle M_a\rangle = 2\beta\sigma^2\mathbb{E}_Z\langle\|M - \langle M\rangle\|_F^2\rangle. \tag{37}$$

Taking expectations in Equation (35) proves Equation (30). The variance identity $C_G - C_{\mathrm{mean}} = \mathcal{V}$ gives Equation (31). For $\sigma = 0$ the Gaussian is a point mass at zero and the same result follows directly from Equation (35). These are statements about Equation (7); they do not prove Conjecture A.4. □

B.3 Optimization and factor volume

The reduced energy, as a function of M, is $\|M - (Z - uI_h)\|_F^2 - \|Z - uI_h\|_F^2$. Frobenius projection onto the positive semidefinite matrices of rank at most l keeps the largest $\min(l,h)$ positive eigenvalues, omitting any beyond that rank. Every such M can be represented by an $h \times l$ factor. For $h = 1$, polar coordinates $D = \rho\omega$ give $\mathrm{d}D = c_l\rho^{l-1}\,\mathrm{d}\rho$; setting $a = \rho^2$ yields Equation (32). The dependence on l persists even though the represented parameter space is already $[0,\infty)$ for every $l \geq 1$.

C The empirical crossover for a global Gaussian factor prior

This appendix proves Theorem 3.1 and Corollary 3.2. Dimensions and the teacher are fixed throughout. The proof uses the represented matrix to control normalized posterior moments, including at the critical width $m = d + 1$.

C.1 The represented prior and conditional score

Put $p = d(d+1)/2$ and $\alpha = (m-d-1)/2 \geq 0$. The rectangular Gram change of variables (Muirhead, 1982, Thm. 2.1.14) pushes the Gaussian density $\exp(-\gamma\|W\|_F^2/2)\,\mathrm{d}W$ to

$$c_{d,m,\gamma}(\det A)^\alpha e^{-\gamma\,\mathrm{tr}\,A/2}\,\mathrm{d}A, \qquad A \succ 0, \tag{38}$$

where $\mathrm{d}A$ is Lebesgue measure on the symmetric matrices. Constants from the choice of Frobenius-orthonormal coordinates cancel in posterior ratios. More generally, this change of variables holds for every integrable function of $WW^\top$, so it applies to the empirical likelihood without an invariance assumption on the input design.

Conditional on the inputs, the score in Equation (28) is exactly Gaussian:

$$Z_n \mid X_{1:n} \sim \mathcal{N}(0, \sigma^2\mathcal{I}_n). \tag{39}$$

Consequently, for data-averaged posterior observables we may use the coupling $Z_n = \sigma \mathcal{I}_n^{1/2} G$, where G is an independent standard Gaussian in $\mathrm{Sym}(\mathbb{R}^d)$. This couples the marginal distributions needed for expectations; it does not assert independence of Z_n and $\mathcal{I}_n$ in the original data.

Set $B = \sqrt{n}(A - A_*)$ and

$$v_n(u) = Z_n - \left(u + \frac{\gamma}{4\beta\sqrt{n}}\right) I_d, \qquad Q_n(B) = \langle B,\, \mathcal{I}_n B \rangle\,. \tag{40}$$

The exact density of B, up to its normalizer, is

$$[\det(A_* + n^{-1/2}B)]^{\alpha} e^{-\beta\{Q_n(B) - 2\langle v_n(u),\, B\rangle\}} \mathbf{1}_{\{A_* + n^{-1/2}B \succ 0\}}\, \mathrm{d}B. \tag{41}$$

C.2 A MOMENT BOUND WITH THE PSD BOUNDARY INCLUDED

Lemma C.1 (normalized matrix moments)**.** *Let $A_0 \succeq 0$, $a > 0$, $\alpha \geq 0$, $\beta > 0$, and $v \in \mathrm{Sym}(\mathbb{R}^d)$. Let $\mathcal{J} \succeq cI$ be a self-adjoint operator on $\mathrm{Sym}(\mathbb{R}^d)$, with $c > 0$. Under the normalized density*

$$[\det(A_0 + aB)]^{\alpha} e^{-\beta\{Q(B) - 2\langle v,\, B\rangle\}} \mathbf{1}_{\{A_0 + aB \succ 0\}}\, \mathrm{d}B, \qquad Q(B) = \langle B,\, \mathcal{J}B \rangle\,,$$

every integer $j \geq 1$ satisfies

$$\langle Q^j \rangle \leq C_j(1 + \|v\|_F^{2j}), \tag{42}$$

where C_j depends only on j, d, α, β, c, not on A_0 or a.

Proof. All moments exist: the determinant factor has at most polynomial growth and the quadratic form is coercive. The normalizer is positive because the domain is nonempty and open. Write $B = t\theta$ in polar coordinates of the p-dimensional symmetric space. By convexity of the PSD cone, the intersection of a ray with $A_0 + aB \succ 0$ is either empty or an interval $0 < t < R_\theta$, with $R_\theta \leq \infty$. Indeed, if $A_0 + at\theta \succ 0$, every strict convex combination of this matrix with A_0 is positive definite.

On a nonempty ray put $g(t) = [\det(A_0 + at\theta)]^{\alpha}$, $b_\theta = \langle \theta,\, \mathcal{J}\theta \rangle \geq c$, and $z_\theta = \langle v,\, \theta \rangle$. If $\alpha > 0$,

$$t g'(t) = \alpha\{d - \mathrm{tr}[(A_0 + at\theta)^{-1} A_0]\} g(t) \leq \alpha d\, g(t).$$

For $\alpha = 0$, $g = 1$ and the same inequality holds. Integrate the derivative of

$$t^{p+2j-2} g(t) e^{-\beta b_\theta t^2 + 2\beta z_\theta t}$$

over the ray. Its lower endpoint is zero. At an infinite upper endpoint the Gaussian decay makes the limit zero. At a finite upper endpoint the limit is zero for $\alpha > 0$ and nonnegative for $\alpha = 0$. The determinant along a ray is a polynomial positive in the interior, so its nonnegative power has the required one-sided limits; when $\alpha > 0$, its derivative is locally integrable at any finite root. One may first integrate on a closed interior interval and then take these limits.

Multiply by b_θ^{j-1} and integrate over directions. The nonnegative upper boundary term and the preceding derivative inequality give

$$2\beta\langle Q^j \rangle \leq (p + 2j - 2 + \alpha d)\langle Q^{j-1} \rangle + 2\beta\langle \langle v,\, B \rangle\, Q^{j-1} \rangle. \tag{43}$$

There is thus no omitted zero-boundary assumption at $\alpha = 0$. Since $|\langle v,\, B \rangle| \leq c^{-1/2}\|v\|_F \sqrt{Q}$, Young's inequality bounds the two terms on the right by $\beta\langle Q^j \rangle + C_j'(1 + \|v\|_F^{2j})$. Absorbing the first term proves Equation (42). □

For $0 \leq u \leq U_0$ and

$$E_n = \{\|\mathcal{I}_n - I\|_{\mathrm{op}} \leq 1/2\}, \tag{44}$$

the lemma implies, for every integer $j \geq 1$,

$$\sup_{u \in [0, U_0]} \langle \|B\|_F^{2j} \rangle \leq C_{j,U_0}(1 + \|Z_n\|_F^{2j}). \tag{45}$$

Under the coupling Equation (39), the right side on E_n is bounded by $C'_{j,U_0}(1 + \|G\|_F^{2j})$, which has finite expectation at every fixed order.

C.3 Local convergence of normalized moments

We first consider deterministic sequences $\mathcal{J}_n \to I$, $z_n \to z$, and $u_n \to u \geq 0$, replacing $\mathcal{I}_n, Z_n, u$ in Equation (41). Work in a teacher eigenbasis $A_* = \mathrm{diag}(A_+, 0_h)$. For $r > 0$, choose a fixed neighborhood $\|P - A_+\| < \delta$ on which $P \succ 0$ and P^{-1} is uniformly bounded. The Schur-complement coordinates are

$$A = \begin{pmatrix} P & C \\ C^\top & S + C^\top P^{-1} C \end{pmatrix}, \qquad P \succ 0, \quad S \succ 0.$$

Their Jacobian is constant (equal to one in entry coordinates), and $\det A = \det P \det S$. Rescale

$$P = A_+ + n^{-1/2} U, \qquad C = n^{-1/2} V, \qquad S = n^{-1/2} M.$$

In the $\mathrm{d}B$ density of Equation (41), this transformation has a constant Jacobian independent of n: $\mathrm{d}B$ is a constant multiple of $\mathrm{d}U\ \mathrm{d}V\ \mathrm{d}M$. The determinant contributes $n^{-\alpha h/2}(\det P \det M)^\alpha$. Equivalently, starting from $\mathrm{d}A$ contributes the common power $n^{-(p+\alpha h)/2}$, since $\mathrm{d}A = n^{-p/2}\,\mathrm{d}B$. Either common factor cancels in normalized posterior moments. The matrix in the exponent is

$$B_n(U, V, M) = \begin{pmatrix} U & V \\ V^\top & M + n^{-1/2} V^\top P^{-1} V \end{pmatrix}. \tag{46}$$

For $r = 0$, omit P, U, V, put $M = B$, and use $\det P = 1$; the same argument applies.

On this neighborhood, positivity gives

$$\|B_n\|_F^2 \geq \|U\|_F^2 + 2\|V\|_F^2 + \|M\|_F^2.$$

For bounded scores and regularization, the exponential in Equation (41) is bounded by a constant times $\exp[-c_0 \|B_n\|_F^2]$, with $c_0 > 0$. The factor $(\det P)^\alpha$ is bounded, and $(\det M)^\alpha$ has polynomial growth and no boundary divergence because $\alpha \geq 0$. Polynomial observables of B_n are bounded by polynomials in U, V, M on this neighborhood. Dominated convergence therefore applies to both normalizing and moment integrals after removing the common power of n. The limiting normalizer is finite and strictly positive.

The limiting tangent matrix $B_T = \left[\begin{smallmatrix} U & V \\ V^\top & 0 \end{smallmatrix}\right]$ and the inactive matrix M are conditionally independent. Their laws are respectively

$$B_T \sim \mathcal{N}\big(z_T - u P_r, (2\beta)^{-1} I_q\big), \qquad P_r = \mathrm{diag}(I_r, 0_h), \tag{47}$$

in the tangent Frobenius space, and

$$\frac{1}{Z_{u,z}} (\det M)^\alpha e^{-\beta\{\|M\|_F^2 + 2\,\mathrm{tr}((u I_h - z_\perp) M)\}} \mathbf{1}_{\{M \succ 0\}}\, \mathrm{d}M. \tag{48}$$

The latter is the pushforward of Equation (7), because $(l - h - 1)/2 = (m - d - 1)/2 = \alpha$.

The preceding ratios give moments of the posterior conditioned on the chosen P-neighborhood. They also give the full posterior limits. Indeed, outside the chosen P-neighborhood, $\|B\|_F \geq \delta\sqrt{n}$. By Equation (45), uniformly for the bounded deterministic sequences above, its posterior mass is $O(n^{-j})$ and its contribution to a second moment is $O(n^{-(j-1)})$ for any integer $j > 1$. First moments are controlled in the same way. The conditioning event has probability tending to one, so decomposing each full moment into its inside and outside contributions proves convergence. Thus both the mean of B and its second moments converge to those of Equations (47) and (48). The corresponding statement for B_{22} follows from the same moment argument.

The argument holds for every convergent sequence u_n in a fixed compact interval, and the limiting moments are continuous in u by dominated convergence. Sequential compactness consequently gives uniformity over that interval. In particular, it covers the square of the posterior mean as well as posterior second moments.

C.4 Exceptional designs and averaging over training sets

For completeness, the global Gaussian prior also controls moments when the empirical sensing operator is poorly conditioned. Put $t = \|W\|_F^2$, $s_i(W) = \langle X_i, WW^\top \rangle$, and $c_n = \gamma/2 + 2\beta u\sqrt{n} \geq$

$\gamma/2$. The factor posterior is proportional to $\exp[-\beta\sum_i(y_i - s_i)^2 - c_n t]$. Integrating the divergence of Wt^{j-1} against this density gives

$$2c_n\langle t^j\rangle = (dm + 2j - 2)\langle t^{j-1}\rangle - 4\beta\left\langle t^{j-1}\sum_i(s_i^2 - y_i s_i)\right\rangle.$$

There is no boundary term: the likelihood is at most one and the fixed Gaussian prior controls polynomial factors and their derivatives. Since $\sum_i(s_i^2 - y_i s_i) = \sum_i(s_i - y_i/2)^2 - \|y\|^2/4$, induction yields

$$\sup_{u\geq 0}\langle t^j\rangle \leq \gamma^{-j}\prod_{a=1}^{j}(dm + 2a - 2 + \beta\|y\|^2) \leq C_j(1 + \|y\|^2)^j. \tag{49}$$

In particular, since $\|WW^\top\|_F \leq t$,

$$\sup_{u\geq 0} nR_G \leq Cn(1 + \|y\|^4). \tag{50}$$

The same bound holds for nR_{mean} and the scaled inactive second moment.

Every entry of $\mathcal{I}_n - I$ is an average of centered independent variables with moments of all orders. Expansion of an even centered-sum moment shows $\mathbb{E}\|\mathcal{I}_n - I\|_{\mathrm{op}}^{2J} = O(n^{-J})$ at fixed d: terms containing a singleton sample index vanish, leaving at most J distinct indices. Hence $\mathbb{P}(E_n^c) = O(n^{-J})$ for every fixed integer J. Also y_i has all moments and $\mathbb{E}\|y\|^8 = O(n^4)$. Cauchy–Schwarz applied to Equation (50) gives

$$\mathbb{E}\left[\mathbf{1}_{E_n^c}\sup_{u\geq 0} nR_G\right] \leq Cn^3\mathbb{P}(E_n^c)^{1/2} = o(1) \tag{51}$$

by taking $J > 6$. This step uses no independence between labels and the design.

Now couple the scores as in Equation (39). Almost surely, $\mathcal{I}_n \to I$ and $Z_n \to \sigma G$, so the local moment limits hold uniformly on each compact u-interval. On E_n, Equation (45) dominates all three scaled risk observables by $C(1 + \|G\|_F^2)$, uniformly in u in that interval. The limiting observables obey the same type of bound, either by taking limits in the moment inequality or by the conditional Euler bound of Appendix B. Dominated convergence therefore passes the uniform limits through the data average on E_n; the complement vanishes by Equation (51).

For an isotropic Gaussian score $Z = \sigma G$, the orthogonal tangent and inactive components are independent. The tangent Gaussian in Equation (47) has mean-square contribution $\mathbb{E}\|Z_T - uP_r\|_F^2 = q\sigma^2 + ru^2$, and its posterior variance contributes $q/(2\beta)$ to Gibbs risk. The inactive moments in Equation (48) are exactly $C_{\mathrm{mean}}(u)$ and $C_G(u)$. This proves all three uniform expected-risk limits in Theorem 3.1. □

C.5 Proof of the regularization corollary

At matched temperature, Proposition A.2 gives

$$F(u) = q\sigma^2 + \frac{k\sigma^2}{2} + ru^2 - u\mathcal{T}(u).$$

Differentiation of the normalized tilted integral gives

$$\mathcal{T}'(u) = -2\beta\mathbb{E}_Z\,\mathrm{Var}_{u,Z}(\mathrm{tr}\,M).$$

This differentiation also holds after the Gaussian average: on a compact u-interval, $\mathrm{Var}(\mathrm{tr}\,M) \leq \langle(\mathrm{tr}\,M)^2\rangle \leq h\langle\|M\|_F^2\rangle$, and Equation (36) supplies an integrable polynomial bound in Z. Since $k > 0$, $\mathcal{T}(0) > 0$, so $F'(0+) = -\mathcal{T}(0) < 0$. For $r > 0$, $F(u) \geq q\sigma^2 + ru^2 \to \infty$. Continuity therefore supplies a strictly positive finite global minimizer. Since its value is below $F(0)$, every minimizer satisfies $ru_*^2 < k\sigma^2/2$. Its stationarity equation follows by differentiation. Choose any sufficiently small fixed $u > 0$ with $F(u) < F(0)$. The empirical theorem gives $n\{\mathbb{E}R_{\mathrm{mean}}(u) - \mathbb{E}R_{\mathrm{mean}}(0)\} \to F(u) - F(0) < 0$, proving the claimed eventual strict improvement. This does not identify a global finite-n optimizer or a data-dependent tuning procedure. □

D EXPECTED RISK OF THE PSD COMPARATOR

We prove Theorem 4.1. All dimensions, the teacher, and $c > 0$ are fixed. The coordinate polynomials of $X(x)$ are linearly independent, since their covariance is the identity. Hence there are $p = d(d+1)/2$ input points whose feature vectors are linearly independent. The determinant of their evaluation matrix is a nonzero polynomial, so Gaussian samples make it nonzero almost surely. Thus $\mathcal{I}_n \succ 0$ almost surely for $n \geq p$. The objective is strictly convex on the closed PSD cone. The positive trace penalty makes it coercive there for every design, so its minimizer exists and is almost surely unique for these sample sizes.

A deterministic boundary limit. Put $\widehat{B}_n = \sqrt{n}(\widehat{A}_n - A_*)$ and $v_n = Z_n - (u + c/(2\sqrt{n}))I_d$. The exact loss expansion gives

$$\widehat{B}_n = \operatorname*{argmin}_{B \in \mathcal{C}_n} \{\langle B, \mathcal{I}_n B\rangle - 2\langle v_n, B\rangle\}, \qquad \mathcal{C}_n = \{B : A_* + n^{-1/2}B \succeq 0\}.$$

Since $0 \in \mathcal{C}_n$, on $\mathcal{I}_n \succeq I/2$ the minimizing inequality yields $\frac{1}{2}\|\widehat{B}_n\|_F^2 \leq 2\|v_n\|_F\|\widehat{B}_n\|_F$ and hence

$$\|\widehat{B}_n\|_F \leq 4\|v_n\|_F. \tag{52}$$

Consider deterministic sequences $\mathcal{I}_n \to I$, $z_n \to z$, $u_n \to u$. In a teacher eigenbasis the limiting cone is

$$\mathcal{C} = \left\{ \begin{pmatrix} U & V \\ V^\top & M \end{pmatrix} : U \in \mathrm{Sym}(\mathbb{R}^r),\ V \in \mathbb{R}^{r\times h},\ M \succeq 0 \right\}.$$

A limit of feasible B_n belongs to $\mathcal{C}$, because their inactive principal blocks are PSD. Conversely, given $B = [U, V; V^\top, M] \in \mathcal{C}$ and $r > 0$, let $P_n = A_+ + n^{-1/2}U \succ 0$ for sufficiently large n, and set

$$\widetilde{B}_n = \begin{pmatrix} U & V \\ V^\top & M + n^{-1/2}V^\top P_n^{-1} V \end{pmatrix}.$$

The Schur complement shows that $\widetilde{B}_n \in \mathcal{C}_n$, and $\widetilde{B}_n \to B$. If $r = 0$, $\mathcal{C}_n = \mathcal{C}$ and no correction is needed. Bound Equation (52) makes the minimizers bounded. For any convergent subsequence, compare its objective with every such recovery sequence and pass to the limit. Its limit must minimize $\|B\|_F^2 - 2\langle z - uI, B\rangle$ over $\mathcal{C}$. This minimizer is unique and equals

$$\widehat{B}(z,u) = z_T - uP_r + (z_\perp - uI_h)_+, \tag{53}$$

where the final block is embedded in the inactive sector and $(\cdot)_+$ denotes the positive spectral part. Every subsequence has this limit, proving convergence. The same argument for arbitrary convergent u_n, together with continuity of PSD projection, proves uniformity on compact nonnegative u-intervals.

Expected risk and exceptional designs. Comparison of the original objective with $A = 0$ gives, for all $u \geq 0$, $c \operatorname{tr} \widehat{A}_n \leq \|y\|^2$. Since $\|\widehat{A}_n\|_F \leq \operatorname{tr} \widehat{A}_n$,

$$\sup_{u \geq 0} n\|\widehat{A}_n - A_*\|_F^2 \leq Cn(1 + \|y\|^4).$$

For E_n in Equation (44), the all-orders moment estimate for $\mathcal{I}_n - I$ and $\mathbb{E}\|y\|^8 = O(n^4)$, proved in Appendix C, therefore give

$$\mathbb{E}\left[\mathbf{1}_{E_n^c} \sup_{u \geq 0} n\|\widehat{A}_n - A_*\|_F^2\right] \leq Cn^3 \mathbb{P}(E_n^c)^{1/2} = o(1).$$

On E_n, the exact coupling $Z_n = \sigma\mathcal{I}_n^{1/2}G$ and Equation (52) dominate the squared error by $C(1 + \|G\|_F^2)$ uniformly on each compact u-interval. The deterministic limit and dominated convergence now pass through the data average, uniformly on that interval. Orthogonality of the tangent and inactive blocks in Equation (53) gives $\mathbb{E}\|Z_T - uP_r\|_F^2 = q\sigma^2 + ru^2$, proving the risk formula.

At $u = 0$, symmetry under $Z_\perp \mapsto -Z_\perp$ and the orthogonal positive/negative spectral decomposition give

$$\mathbb{E}\|(Z_\perp)_+\|_F^2 = \tfrac{1}{2}\mathbb{E}\|Z_\perp\|_F^2 = \frac{\sigma^2 h(h+1)}{4}.$$

This is the self-dual PSD-cone statistical-dimension calculation (Amelunxen et al., 2014, Sec. 3.2). Subtracting the already established Bayesian endpoint proves Corollary 4.2. □

E LOCAL ALTERNATIVES AND WEAK TEACHER MODES

E.1 PROOF OF PROPOSITION 5.1

Use the model Equation (21), with fixed A_+, and let $B = \sqrt{n}(A - A_0)$. The residual at the true teacher remains ϵ_i, so Equation (39) is unchanged. The exact loss identity is

$$\begin{aligned} n\{L_n(A) - L_n(A_{*,n})\} &= Q_n(B - \Theta) - 2\langle Z_n,\, B - \Theta\rangle \\ &= Q_n(B) - 2\langle Z_n + \mathcal{I}_n\Theta,\, B\rangle + Q_n(\Theta) + 2\langle Z_n,\, \Theta\rangle\,. \end{aligned} \tag{54}$$

The posterior density in Equation (41) therefore has base A_0 and score

$$v_n^{\Theta}(u) = Z_n + \mathcal{I}_n\Theta - \left(u + \frac{\gamma}{4\beta\sqrt{n}}\right) I_d.$$

For deterministic sequences $\mathcal{J}_n \to I$, $z_n \to z$, $u_n \to u$, and $\Theta_n \to \Theta$, the score converges to $z + \Theta - uI_d$. The Schur coordinates Equation (46) are still based at A_0. The tangent law is Equation (47), since $\Theta_T = 0$; the inactive law is Equation (48) with $z_\perp$ replaced by $z_\perp + \Theta_\perp$. At finite n, $(\mathcal{I}_n\Theta)_T$ need not vanish.

Let $u \le U$ and $\Theta_\perp$ belong to a fixed compact set K. On E_n, Lemma C.1 and the conditional Gaussian coupling give

$$\sup_{u\le U,\,\Theta_\perp\in K} \langle \|B\|_F^{2j}\rangle \le C_{j,U,K}(1 + \|G\|_F^{2j}).$$

This controls normalized moments and removes the complement of the fixed P-neighborhood exactly as in Appendix C. In Schur coordinates the inactive block of B is $M + n^{-1/2}V^\top P^{-1}V$, which converges to M; the risk is centered at $\Theta_\perp$, not at zero.

The rough factor-moment inequality Equation (49) holds conditional on the labels for every teacher. Here $\sup_{n,\Theta_\perp\in K}\|A_{*,n}\|_F < \infty$, so the Gaussian design and noise give $\sup_{\Theta_\perp\in K}\mathbb{E}\,\|y\|^8 = O(n^4)$. The rough centered risk bound remains $Cn(1 + \|y\|^4)$ with a uniform constant. The exceptional-design expectation is consequently $o(1)$ by Equation (51).

On good designs, the moment bound dominates the risks centered at Θ by $C_{U,K}(1 + \|G\|_F^2)$. Local convergence holds along every convergent sequence (u_n, Θ_n); compactness and continuity of the limiting normalized integrals therefore give uniform convergence on $[0, U] \times K$. Dominated convergence passes this uniform limit through the Gaussian and design averages. Orthogonality of the tangent and inactive blocks gives Equations (22) and (23). □

E.2 SHIFTED EULER–STEIN IDENTITY

Abbreviate $\vartheta = \Theta_\perp$, $\mu = \mu_\Theta(u, Z)$, and define

$$\mathcal{V}_\Theta(u) = \mathbb{E}_Z\langle \|M - \mu\|_F^2\rangle_{u,Z+\vartheta}.$$

The conditional Euler identity and Gaussian integration by parts give

$$\frac{k}{4\beta} = \langle \|M\|_F^2\rangle + u\langle \operatorname{tr} M\rangle - \langle Z + \vartheta,\, \mu\rangle\,, \qquad \mathbb{E}_Z\,\langle Z,\, \mu\rangle = 2\beta\sigma^2\mathcal{V}_\Theta(u).$$

The moment bound in Equation (36), with Z replaced by $Z + \vartheta$, justifies integration by parts after the Gaussian average. Subtracting the conditional variance and centering at ϑ gives

$$\mathbb{E}_Z\|\mu - \vartheta\|_F^2 = \frac{k}{4\beta} - u\mathcal{T}_\Theta(u) + (2\beta\sigma^2 - 1)\mathcal{V}_\Theta(u) - \langle \vartheta,\, \bar\mu_\Theta(u)\rangle + \|\vartheta\|_F^2. \tag{55}$$

At matched temperature this proves Equation (24). For $\vartheta = tH$, continuity at $t = 0$ suffices to differentiate its last two terms at zero. The needed continuity follows from quartic decay and polynomial moment control on compact t intervals. Rotational invariance implies $\bar\mu_0(0) = \mathcal{T}_0(0)I_h/h$, and $\mathcal{T}_0(0) > 0$, proving Equation (25).

E.3 Strict improvement over the zero-signal endpoint

For $\alpha \geq 0$, $u = 0$, and matched temperature, the represented inactive density is

$$p_\alpha(M \mid Y) \propto (\det M)^\alpha e^{-\|M-Y\|_F^2/(2\sigma^2)} \mathbf{1}_{\{M \succ 0\}}, \qquad Y = Z + \vartheta.$$

Let Ξ be an independent Gaussian in $\mathrm{Sym}(\mathbb{R}^h)$ with covariance $\sigma^2 I$ in Frobenius coordinates. Define

$$\Psi_\alpha(Y) = \mathbb{E}_\Xi[(\det(Y+\Xi))^\alpha \mathbf{1}_{\{Y+\Xi \succ 0\}}].$$

This positive smooth function is the Gaussian convolution of a locally integrable weight of polynomial growth; it is not a normalized marginal probability density. Differentiating the Gaussian kernel gives

$$b_\alpha(Y) := \mu_\alpha(Y) - Y = \sigma^2 \nabla \log \Psi_\alpha(Y). \tag{56}$$

We claim $b_\alpha(Y) \succ 0$ for every finite Y. It suffices to show $D_{vv^\top}\Psi_\alpha(Y) > 0$ for every unit vector v. Rotate v to the first coordinate and write $Y + \Xi = [a, w^\top; w, C]$. The first diagonal Gaussian coordinate a is independent of (w, C) and has variance σ^2. Except on the zero-probability event that C is singular, PSD feasibility is equivalent to $C \succ 0$ and $a \geq \tau := w^\top C^{-1} w$, and $\det(Y+\Xi) = \det C\,(a-\tau)$. Let ϕ_σ be the centered Gaussian density of variance σ^2. For $\alpha > 0$, integration by parts gives

$$\frac{\partial}{\partial y_{11}} \int_\tau^\infty (a-\tau)^\alpha \phi_\sigma(a - y_{11})\,\mathrm{d}a = \alpha \int_\tau^\infty (a-\tau)^{\alpha-1} \phi_\sigma(a-y_{11})\,\mathrm{d}a > 0.$$

The boundary term vanishes and the integrable endpoint exponent is $\alpha - 1 > -1$. For $\alpha = 0$, the derivative instead equals $\phi_\sigma(\tau - y_{11}) > 0$. Averaging with the nonnegative factor $\mathbf{1}_{\{C \succ 0\}}(\det C)^\alpha$ proves the claim, since $\mathbb{P}(C \succ 0) > 0$. Fubini applies to the differentiated Gaussian integral by polynomial growth; the nonnegative expression after integration by parts can also be averaged by Tonelli. For $h = 1$, omit w, C and set $\tau = 0$. Since $\mathbb{E}Y = \vartheta$, Equation (24) becomes

$$F_\Theta(0) = \sigma^2\{p + \alpha h\} - \mathbb{E}_Z \left\langle \vartheta,\, b_\alpha(Z+\vartheta) \right\rangle.$$

For $\vartheta \succeq 0$ nonzero the subtracted expectation is strictly positive and finite. This proves Equation (26) and completes the proof of Corollary 5.2. □

E.4 Strong-signal endpoint

Lemma E.1 (activating a fixed PSD direction)**.** *Fix the parameters of Proposition 5.1 and $H \succeq 0$ of rank $s \geq 1$ in $\mathrm{Sym}(\mathbb{R}^h)$. Put $h' = h - s$, $l' = l - s$, and $q' = q + hs - s(s-1)/2$. As $t \to \infty$,*

$$F_{tH}(u) \longrightarrow q'\sigma^2 + (r+s)u^2 + C_{\mathrm{mean}}^{(h',l')}(u), \tag{57}$$

where the superscript specifies the dimensions of the original inactive law. For Gibbs risk the corresponding limit is $q'\sigma^2 + q'/(2\beta) + (r+s)u^2 + C_G^{(h',l')}(u)$. An inactive contribution with $h' = 0$ is defined to be zero. Both convergences are uniform on compact nonnegative u intervals.

Proof. Recenter the represented inactive matrix as $C = M - tH$. Its density, after discarding a constant, is

$$[\det(tH + C)]^\alpha e^{-\beta\{\|C\|_F^2 - 2\langle Z - uI_h,\, C\rangle\}} \mathbf{1}_{\{tH + C \succ 0\}}\,\mathrm{d}C.$$

Apply Lemma C.1 with base matrix tH, scale $a = 1$, operator I, and score $Z - uI_h$. Its constants are independent of tH, so all centered moments obey bounds polynomial in $\|Z\|_F + u$, uniformly in t.

In an eigenbasis of H, its positive block is $H_+ \succ 0$. Write

$$M = \begin{pmatrix} P & V \\ V^\top & S + V^\top P^{-1} V \end{pmatrix}, \qquad P = tH_+ + U.$$

On $\|U\|_F \leq \delta t$, with δ smaller than half the least eigenvalue of H_+, $P^{-1} = O(t^{-1})$ and $t^{-\alpha s}(\det P)^\alpha$ is bounded. At each fixed U, V, S this factor converges to $(\det H_+)^\alpha$, while $V^\top P^{-1} V \to 0$. The determinant identity and constant Schur Jacobian leave the weight $(\det S)^\alpha$ in the remaining block. The centered Gaussian exponential dominates polynomial observables by

an integrable Gaussian times a polynomial in U, V, S, as in Appendix C. The uniform centered moment bound removes the complement $\|U\|_F > \delta t$ and upgrades convergence to first and second normalized moments.

The new tangent space has dimension $hs - s(s-1)/2$, Gaussian mean $Z_{T,s} - uP_s$, and covariance $(2\beta)^{-1}I$. The remaining inactive law has dimensions h', l', because $(l' - h' - 1)/2 = \alpha$. If $s = h$, the entire limiting space is tangent and the inactive block is absent. Averaging over Z is justified by the same uniform centered moment bounds; compact-u uniformity follows by the convergent-sequence argument. Adding the original tangent risk proves the lemma. At matched temperature and $u = 0$, the identity $q' + h'l'/2 = p + \alpha(h - s)$ proves Equation (27). □

Remark E.2 (a lower strong-signal endpoint need not be undercut). Take $h = 1$, $\alpha = 2$ (thus $m = d + 5$), and $\Theta_\perp = t > 0$. The zero-signal coefficient is $(q + 3)\sigma^2$ and the strong-signal coefficient is $(q+1)\sigma^2$. Nevertheless $F_t(0) > (q+1)\sigma^2$ for every finite t. To see this, let $X = y+\Xi$, $\Psi(y) = \mathbb{E}X_+^2$, and $b(y) = \sigma^2\Psi'(y)/\Psi(y)$, so that $\mu(y) = y + b(y)$. Then $\Psi' = 2\mathbb{E}X_+$ and $\Psi'' = 2\mathbb{P}(X > 0)$, and strict Cauchy–Schwarz gives

$$2\Psi\Psi'' - (\Psi')^2 = 4\{\mathbb{P}(X > 0)\mathbb{E}X_+^2 - (\mathbb{E}X_+)^2\} > 0.$$

Gaussian integration by parts for $Y = t + Z$ therefore yields

$$\mathbb{E}(\mu(Y) - t)^2 = \sigma^2 + \mathbb{E}\{b(Y)^2 + 2\sigma^2 b'(Y)\} = \sigma^2 + \sigma^4\mathbb{E}\left[\frac{2\Psi\Psi'' - (\Psi')^2}{\Psi^2}(Y)\right] > \sigma^2.$$

The truncated Gaussian moments ensure polynomial growth of b, b' and justify integration by parts. Adding the q tangent coordinates proves the claim. Thus the general zero-signal improvement must not be interpreted as a universal dip below the strong-signal endpoint.

E.5 PSD COMPARATOR UNDER LOCAL ALTERNATIVES

For Corollary A.5, center the optimization at A_0. The feasible sets in Appendix D are unchanged and the score becomes $Z_n + \mathcal{I}_n\Theta - (u + c/(2\sqrt{n}))I_d$. Its limiting cone projection is

$$Z_T - uP_r + (Z_\perp + \Theta_\perp - uI_h)_+.$$

The good-design minimizing inequality controls its norm by four times the shifted score norm. On compact $u, \Theta_\perp$ sets this is bounded by $C(1 + \|G\|_F)$. The labels have the uniform moment bounds already proved above, so the rough comparison with $A = 0$ again removes exceptional designs. Centering the risk at Θ and repeating the compactness argument proves the stated uniform expected-risk limit. □

F THE MATCHED BAYESIAN ENDPOINT

This appendix proves Proposition 3.3 for the compactly supported local prior in Equation (33), and for the global compact-prior alternative in the proposition. All dimensions and the teacher are fixed, $\sigma > 0$, $\eta_n = 0$, and $\beta = 1/(2\sigma^2)$. We first control localization in training-set expectation, then verify the hypotheses of singular learning theory (SLT) and relate its predictive KL observable to squared prediction risk.

F.1 LOCALIZATION WITH AN EXPECTATION BOUND

Let $\rho(W)\,\mathrm{d}W$ denote either prior, and let π_n be its posterior. Its support is compact, and its density is smooth and strictly positive on a neighborhood of the entire true orbit. The Gram loss $K(W) = \|WW^\top - A_*\|_F^2$ is coercive and its zero set is that orbit. Consequently, for a sufficiently small regular value $a > 0$, the compact semialgebraic set $\Omega_a = \{W : K(W) \le a\}$ lies in this positive-density neighborhood. Write π_n^a for the posterior obtained by restricting the same prior to Ω_a.

Lemma F.1 (compact-prior localization). *For every fixed $p > 0$ there is $C_p < \infty$ such that*

$$\mathbb{E}_{\mathcal{D}_n}\pi_n(K > a) \le C_p n^{-p}. \tag{58}$$

In particular, the differences between the expected Gibbs risks, and between the expected mean-predictor risks, for π_n and π_n^a are $o(n^{-1})$.

Proof. Put $S_n = n^{-1}\sum_i \epsilon_i X_i = Z_n/\sqrt{n}$. Each coordinate of $\mathcal{I}_n - I$ and S_n is a mean of independent centered random variables having all finite moments. Expanding their even moments and applying Markov's inequality gives, for every integer $j \geq 1$ and fixed $\delta > 0$,

$$\mathbb{P}\big(\|S_n\|_F > \delta \text{ or } \|\mathcal{I}_n - I\|_{\mathrm{op}} > \delta\big) \leq C_{j,\delta} n^{-j}.$$

On the complementary event, the exact loss expansion gives, uniformly in W,

$$(1-\delta)K - 2\delta\sqrt{K} \leq L_n(W) - L_n(A_*) \leq (1+\delta)K + 2\delta\sqrt{K}.$$

Choose $\delta \leq \min\{1/4, \sqrt{a}/8\}$. The lower bound is at least $a/2$ for $K > a$. Choose a fixed $b \in (0, a)$ so small that $(1+\delta)b + 2\delta\sqrt{b} \leq a/4$. The prior mass $c_b = \int_{\Omega_b} \rho(W)\,\mathrm{d}W$ is positive. Comparing the posterior numerator outside Ω_a with its denominator restricted to Ω_b yields $\pi_n(K > a) \leq c_b^{-1} e^{-n\beta a/4}$ on this event, for a normalized prior. On its complement the probability is at most one. Taking $j > p$ proves Equation (58).

The represented matrices and their losses are bounded on the prior support. Since π_n^a is the conditional law of π_n on Ω_a, the differences of their Gibbs risks and of their posterior matrix means are bounded by a constant times $\pi_n(K > a)$. Boundedness of both means gives the same bound for their squared risks. □

F.2 SLT ASSUMPTIONS ON THE RESTRICTED PARAMETER SET

We apply Watanabe (2010b, Sec. 2, assumptions (A.1)–(A.3)) on Ω_a. The parameter set is compact with analytic defining inequality $a - K \geq 0$ and, by the choice of a regular value, is the closure of its interior. The restricted prior is a smooth strictly positive density; the analytic factor in (A.1) can be taken to be one. This restriction avoids imposing analyticity on the original cutoff or its boundary.

Include the input density in the statistical model: $p_W(x,y) = \phi_d(x)\phi_\sigma(y - f_W(x))$, with true distribution p_*. For $H = WW^\top - A_*$, put $\delta_W(x) = \langle X(x),\, H\rangle$ and write $y = f_*(x) + \epsilon$. The log density ratio and its expectation are

$$\ell_W(x,\epsilon) = \log\frac{p_*}{p_W} = \frac{\delta_W(x)^2}{2\sigma^2} - \frac{\epsilon\delta_W(x)}{\sigma^2}, \qquad \kappa(W) := \mathbb{E}_*\ell_W = \frac{K(W)}{2\sigma^2}. \tag{59}$$

The polynomial dependence on W, with polynomial coefficients in Gaussian x, ϵ, is analytic into $L^s(p_*)$ for every finite s. On a bounded complex neighborhood of Ω_a, its supremum has all finite moments. For the remaining envelope condition, compactness gives $|f_W(x)| \leq B(x) := C(1 + \|x\|^2)$. The supremum of $p_W(x,y)$ is bounded by

$$\frac{\phi_d(x)}{\sqrt{2\pi}\sigma}\exp\left[-\frac{\mathrm{dist}(y,[-B(x),B(x)])^2}{2\sigma^2}\right].$$

The squared log-ratio envelope grows polynomially in (x,y). Integration in y against this bound gives polynomial growth in x, which is integrable against ϕ_d. Thus (A.3) holds. In the equivalent relative-variance formulation, Gaussian fourth moments also give

$$\mathbb{E}_*\ell_W^2 = \frac{K(W)}{\sigma^2} + \frac{\mathbb{E}_x\delta_W(x)^4}{4\sigma^4} \leq C\{K(W) + K(W)^2\} \leq C'\kappa(W)$$

on this compact set.

Multiplication of the loss by $1/(2\sigma^2)$ leaves its learning coefficient unchanged. By Proposition G.1, and positivity of the density on the whole zero orbit, $\lambda = q/2 + k/4$ with multiplicity one. Theorem 1 and Corollaries 2–3 of Watanabe (2010b), at likelihood temperature one, therefore give

$$n\mathbb{E}B_g \longrightarrow \lambda, \qquad n\mathbb{E}G_g \longrightarrow \lambda + \nu_{\mathrm{SLT}}, \tag{60}$$

for the restricted posterior. Here $B_g = \mathbb{E}_* \log(p_*/\langle p_W\rangle)$ and $G_g = \langle\kappa\rangle$, where brackets denote π_n^a expectation.

F.3 From predictive KL to squared prediction risk

Lemma F.2 (Gaussian predictive KL and the posterior mean)**.** *For the restricted posterior above,*

$$\mathbb{E}_{\mathcal{D}_n}\left|B_g - \frac{R_{\mathrm{mean}}}{2\sigma^2}\right| = o(n^{-1}), \qquad G_g = \frac{R_G}{2\sigma^2}. \tag{61}$$

Proof. The identity for G_g follows from Equation (59). The second-order predictive-log expansion, with its remainder controlled as in Watanabe (2010b, Sec. 4.5, equations (30)–(33)), is

$$B_g = \langle \kappa \rangle - \frac{1}{2}\mathbb{E}_*\langle \ell_W^2 \rangle + \frac{1}{2}\mathbb{E}_*\langle \ell_W \rangle^2 + o_{\mathbb{P}}(n^{-1}).$$

The posterior moment estimate in Lemma 7 of that paper, with power two, gives $\langle K^2 \rangle = O_{\mathbb{P}}(n^{-2})$. The Gaussian noise is independent of a fresh input, so direct substitution of Equation (59) yields

$$\mathbb{E}_*\langle \ell_W^2 \rangle = \sigma^{-2}\mathbb{E}_x\langle \delta_W^2 \rangle + (4\sigma^4)^{-1}\mathbb{E}_x\langle \delta_W^4 \rangle,$$
$$\mathbb{E}_*\langle \ell_W \rangle^2 = \sigma^{-2}\mathbb{E}_x\langle \delta_W \rangle^2 + (4\sigma^4)^{-1}\mathbb{E}_x\langle \delta_W^2 \rangle^2.$$

The first term in the first line cancels $2\langle \kappa \rangle$. Moreover, $\mathbb{E}_x\langle \delta_W^4 \rangle \le C\langle K^2 \rangle$, and Jensen's inequality bounds $\mathbb{E}_x\langle \delta_W^2 \rangle^2$ by the same quantity. The surviving leading term is $\mathbb{E}_x\langle \delta_W \rangle^2/(2\sigma^2) = R_{\mathrm{mean}}/(2\sigma^2)$. Thus $n(B_g - R_{\mathrm{mean}}/(2\sigma^2)) \to 0$ in probability.

To pass to training-set expectation, both B_g and $R_{\mathrm{mean}}/(2\sigma^2)$ lie in $[0, G_g]$, by Jensen's inequality. Lemma 3 of Watanabe (2010b) gives uniform integrability asymptotically for nG_g, so the convergence is in L^1 as asserted. We make no $O(n^{-2})$ claim for the expected remainder. □

F.4 Completion of Proposition 3.3

Combine Equations (60) and (61) and $\lambda = q/2 + k/4$ for π_n^a. Then Lemma F.1 transfers the two expected squared-risk limits to the original prior. This proves both the local and global compact-prior versions without invoking Conjecture A.4. The width increment follows by subtracting the two fixed-width mean-risk expansions; it does not identify an individual block contribution or a simultaneous width–sample-size limit. □

G Singular geometry of redundant directions

The null space in Equation (6) is reached nonlinearly in factor coordinates. Near a rank-r factor, the redundant block changes $WW^\top$ to second order and the risk to fourth order. This mixed scaling determines the local volume and the population reference law.

Local learning coefficients. For a local integral $Z_\chi(\tau) = \int \chi\varphi e^{-\tau K}\,\mathrm{d}W$, the learning coefficient λ and multiplicity μ are defined by $Z_\chi(\tau) = C\tau^{-\lambda}(\log\tau)^{\mu-1}(1 + o(1))$ as $\tau \to \infty$ (Watanabe, 2009). A quadratic block of dimension q contributes $q/2$; a positive homogeneous quartic block of dimension k contributes $k/4$. These integral exponents enter statistical asymptotics under additional assumptions, verified for the Bayesian endpoint in Appendix F. They are not by themselves Gibbs squared-risk coefficients. Definitions and the equivalent-loss argument are in Appendix J.

For the Gram loss, the regular and redundant dimensions are

$$q_r = dr - \binom{r}{2} = \dim\{A \succeq 0 : \operatorname{rank} A = r\}, \qquad k_r = (d-r)(m-r). \tag{62}$$

An orbit chart puts the factor in the form $[L, 0; C, D]$, with $L \in \mathrm{LT}_+(r)$, and gives regular coordinates $(U, V) = (LL^\top - A_+, LC^\top)$ of total dimension q. On this slice,

$$K = \|U\|_F^2 + 2\|V\|_F^2 + \|CC^\top + DD^\top\|_F^2, \qquad C = O(\|V\|_F). \tag{63}$$

Positivity gives $\|CC^\top + DD^\top\|_F^2 \ge Q(D)$, while $\min(h,l)^{-1}\|D\|_F^4 \le Q(D) \le \|D\|_F^4$. Thus the loss is locally equivalent to a quadratic regular block plus a coupled quartic redundant block.

Proposition G.1 (Gram benchmark). *For $0 \le r \le \min(d,m)$, the local learning coefficient on $W_*\,\mathrm{O}(m)$ is*

$$\lambda_{d,m,r} = \frac{q_r}{2} + \frac{k_r}{4} = \frac{dm + r(d-m+1)}{4}, \qquad \mu = 1. \tag{64}$$

Appendix J proves the orbit reduction and the normal form. The count is a benchmark with precedents in probabilistic PCA and factor analysis (Drton et al., 2026; 2025); the comparison and its limitations are given in Appendix J.7. The coefficient $k/4$ describes the whole homogeneous block, not k independent scalar quartics. The mixed scales are $U, V = O(\tau^{-1/2})$ and $D = O(\tau^{-1/4})$, so the represented inactive matrix $DD^\top$ and the regular matrix perturbation both occur at order $\tau^{-1/2}$.

H Weight decay and the local crossover theorem

The gap is weight decay. For the fixed-readout model with ℓ_2 regularization, write $A = WW^\top$ and $R_\eta(W) = \|A - A_*\|_F^2 + \eta\|W\|_F^2$. Since $\|W\|_F^2 = \operatorname{tr} A$, completing the square gives the exact identity

$$R_\eta(W) = \big\|WW^\top - \big(A_* - \tfrac{\eta}{2} I_d\big)\big\|_F^2 + \eta \operatorname{tr} A_* - \tfrac{d\eta^2}{4}. \tag{65}$$

Weight decay therefore shifts every target eigenvalue by $-\eta/2$: modes with $\lambda_i(A_*) > \eta/2$ survive with reduced amplitude $a_i = \lambda_i - \eta/2$, and modes with $\lambda_i(A_*) < \eta/2$ are thresholded and acquire a *gap* $g_i = \eta/2 - \lambda_i > 0$, provided at most m eigenvalues survive (the equal-gap model below has $r < \min(d,m)$ and is safe). This is the trace-penalization mechanism behind the spectral thresholds of Erba et al. (2025), read here at finite temperature for complete observations; with finite data the empirical sensing operator and score also enter the optimality conditions (Theorem 4.1 and Proposition A.1).

We study the equal-gap model. For $0 \le r < \min(d,m)$, positive $a_1, \dots, a_r$ and $g \ge 0$ put

$$h = d - r, \quad l = m - r, \quad k = hl, \quad q = q_r, \qquad B_g = \operatorname{diag}(a_1, \dots, a_r, -gI_h), \tag{66}$$

$$E_g(W) = \big\|WW^\top - B_g\big\|_F^2 - hg^2. \tag{67}$$

The positive-semidefinite projection of B_g is $\operatorname{diag}(a, 0)$ for every $g \ge 0$, so $E_g \ge 0$ vanishes exactly on the orbit $W_*\,\mathrm{O}(m)$ with $W_*W_*^\top = \operatorname{diag}(a,0)$; at $g = 0$ it is the Gram loss of Appendix G, and for $g > 0$ the minimizer keeps its rank while the redundant block acquires a quadratic term. The reduced integral that governs it is

$$\Psi_{h,l}(s) = \int_{\mathbb{R}^{h\times l}} \exp\big(-\|DD^\top\|_F^2 - 2s\|D\|_F^2\big)\,\mathrm{d}D, \qquad s \ge 0, \tag{68}$$

and $\mathbb{E}_s$ denotes expectation under its normalized integrand.

Theorem H.1 (local crossover, equal-gap model). *Fix d, m, $0 \le r < \min(d,m)$ and $a_1, \dots, a_r > 0$, and let E_g be the equal-gap energy Equation (67). Let $\chi \ge 0$ be smooth, supported in a sufficiently small neighbourhood of the orbit $W_*\,\mathrm{O}(m)$ and positive on a nonempty open subset of it, let φ be a smooth positive density there, and let $Z_\chi(\tau, g) = \int \chi(W)\varphi(W)e^{-\tau E_g(W)}\,\mathrm{d}W$. Then for every $S < \infty$, uniformly in $s \in [0, S]$, as $\tau \to \infty$,*

$$Z_\chi\big(\tau,\, s/\sqrt{\tau}\big) = C\,\tau^{-q/2-k/4}\,\Psi_{h,l}(s)\,\big(1 + o(1)\big), \tag{69}$$

where $C > 0$ depends on χ, φ and the positive block but not on s. Moreover, with $\mathbb{E}_{\tau,g}$ the expectation under the localized density $\propto \chi\varphi e^{-\tau E_g}$,

$$\tau\,\mathbb{E}_{\tau, s/\sqrt{\tau}}\big[E_{s/\sqrt{\tau}}(W)\big] \;\longrightarrow\; \frac{q}{2} + \frac{k}{4} + s\,\mathbb{E}_s\|D\|_F^2, \tag{70}$$

again uniformly for $s \in [0, S]$.

Proof overview. The proof (Appendix K) meets three obstacles. *Symmetries versus fluctuations:* $W \mapsto WW^\top$ is not a change of coordinates, so the orbit is quotiented by a factor-space chart with a smooth positive induced density (Lemma J.2). *Mixed orders:* on the slice the energy is exactly $E_g = \|U\|_F^2 + 2\|V\|_F^2 + \|C_1C_1^\top + DD^\top\|_F^2 + 2g(\|C_1\|_F^2 + \|D\|_F^2)$ with U, V regular coordinates

and $C_1 = O(\|V\|)$, so U, V are rescaled by $\tau^{-1/2}$ and D by $\tau^{-1/4}$. *Uniform control:* positivity, $\|C_1 C_1^\top + DD^\top\|_F^2 \geq \|DD^\top\|_F^2$, supplies one s-independent dominating function for the partition function and for its energy moment, and dominated convergence gives both limits uniformly in s; Equation (70) then follows from the Euler relation $k = 4\mathbb{E}_s\|DD^\top\|_F^2 + 4s\mathbb{E}_s\|D\|_F^2$, an integration by parts in the reduced variables.

Corollary H.2 (endpoints, and why the limits do not commute)**.** *Define the reduced effective coefficient*

$$\Lambda_{h,l}(s) = \frac{k}{4} + s\,\mathbb{E}_s\|D\|_F^2 = \frac{k}{2} - \mathbb{E}_s\|DD^\top\|_F^2. \tag{71}$$

Then $\Lambda_{h,l}(0) = k/4$, so at zero gap Equation (70) recovers the Gram coefficient $q/2 + k/4$, and $\Lambda_{h,l}(s) \to k/2$ as $s \to \infty$ (Proposition I.1), the regular count $q/2 + k/2$ of the ordinary Laplace expansion at fixed $g > 0$; the two iterated limits of the normalized energy differ by $k/4$, and $g = s/\sqrt{\tau}$ is the scale that connects them.

A small gap can thus be invisible in two ways: at fixed block size when $g\sqrt{\tau} \ll 1$, and, when the block grows, through the second scale that its coupled fluctuations introduce (Appendix I).

Corollary H.3 (population risk for a fixed teacher)**.** *Fix $A_* = \mathrm{diag}(\alpha_1, \ldots, \alpha_r, 0_h)$ with $\alpha_i > 0$, and fix a smooth positive density and a cutoff supported near its factor orbit and positive on that orbit. Let $\eta = 2g = 2s/\sqrt{\tau}$, and let $\mathbb{E}_{\tau,g}$ now denote the local population Gibbs law $\propto \chi\varphi e^{-\tau(K + 2g\|W\|_F^2)}$. Its minimizer represents $A_g = \mathrm{diag}(\alpha_1 - g, \ldots, \alpha_r - g, 0_h)$ for all sufficiently large τ. Uniformly for s in any fixed bounded interval,*

$$\tau\,\mathbb{E}_{\tau, s/\sqrt{\tau}} K \longrightarrow rs^2 + \frac{q}{2} + \mathbb{E}_s Q, \tag{72}$$

$$\tau\|\mathbb{E}_{\tau, s/\sqrt{\tau}}[WW^\top] - A_*\|_F^2 \longrightarrow rs^2 + \frac{(\mathbb{E}_s T)^2}{h}. \tag{73}$$

The proof extends the local rescaling uniformly along the moving positive block $a_i = \alpha_i - g$, which stays away from zero (Appendix K.5). The term rs^2 is active-mode shrinkage bias. In the reduced model, the Gibbs inactive-risk coefficient is $\mathbb{E}_s Q = k/2 - \Lambda_{h,l}(s)$, which equals $k/4$ at zero gap and tends to zero as $s \to \infty$; the mean-predictor coefficient is $(\mathbb{E}_s T)^2/h = \|\mathbb{E}_s[DD^\top]\|_F^2$. At $s = 0$ the latter is strictly smaller than $k/4$ by the nonzero conditional variance of $DD^\top$. Neither coefficient is the regularised-energy count $\Lambda_{h,l}(s)$.

Corollary H.4 (minimizing the limiting population-risk coefficient)**.** *Assume $r \geq 1$. The functions $\mathcal{C}_G(s) = rs^2 + q/2 + \mathbb{E}_s Q$ and $\mathcal{C}_{\mathrm{mean}}(s) = rs^2 + (\mathbb{E}_s T)^2/h$ attain their minima at positive finite s. Every minimizer satisfies, respectively,*

$$2rs = \mathbb{E}_s T - 2s\,\mathrm{Var}_s\, T, \qquad rs = \frac{2\mathbb{E}_s T\,\mathrm{Var}_s\, T}{h}. \tag{74}$$

This optimizes a fixed-dimensional leading coefficient along $\eta = 2s/\sqrt{\tau}$. It does not identify the global finite-τ optimum or the optimum of the empirical learner. For $r = 0$ there is no active-mode bias and the reduced risk has infimum zero as $s \to \infty$; the interior-minimizer argument does not apply. The width bounds in Appendix I concern the reduced ensemble and require a separate uniform approximation before they can be used in a joint width–sample-size limit.

A fixed Gaussian prior contributes a scaled gap that vanishes as $\tau \to \infty$, whereas fixed weight decay inside the energy gives $g\sqrt{\tau} \to \infty$. These conventions follow different low-temperature paths; see Remark K.2.

I Width and regularization in the population reference model

The reduced matrix integral gives bounds valid at every block size. Its eigenvalue representation and square-block equilibrium are developed in Appendix M.

Proposition I.1 (bounds across dimensions)**.** *Put $k = hl$, $\nu = \min(h, l)$, $M = \max(h, l)$, and let $\Lambda_{h,l}(s)$ be as in Equation (71). For all $h, l \geq 1$ and $s \geq 0$,*

$$0 \;\leq\; \frac{\Lambda_{h,l}(s)}{k} - \frac{1}{4} \;\leq\; \min\Big\{\frac{1}{4},\, \frac{s}{2\sqrt{M}}\Big\}. \tag{75}$$

For $s > 0$,

$$0 \le \frac{1}{2} - \frac{\Lambda_{h,l}(s)}{k} \le \min\Big\{\frac{1}{4}, \frac{h+l+1}{16s^2}\Big\}, \qquad 0 \le \log\frac{(\pi/2s)^{k/2}}{\Psi_{h,l}(s)} \le \frac{k(h+l+1)}{16s^2}. \quad (76)$$

The first bound uses $\|D\|_F^4 \le \nu\|DD^\top\|_F^2$ and the Euler relation; the second regards the reduced density as the Gaussian $e^{-2s\|D\|_F^2}$ tilted by $e^{-\|DD^\top\|_F^2}$, a tilt that decreases the mean of $\|DD^\top\|_F^2$, together with the Gaussian fourth-moment identity $\mathbb{E}_G\|DD^\top\|_F^2 = hl(h+l+1)/(16s^2)$; the free-energy bound is Jensen's inequality (Appendix L). All three inequalities are exact statements about the reduced model, and they were checked numerically for $(h,l) \in \{(2,2),(3,1),(1,3),(3,2)\}$.

Reading the bounds. The coefficient per redundant parameter tends to $1/4$ if $s/\sqrt{M} \to 0$ and to $1/2$ if $s^2/(h+l) \to \infty$; since $h+l+1 \le 3M$, both can be read on one scale: for $s = c\sqrt{M}$,

$$\frac{\Lambda_{h,l}(c\sqrt{M})}{k} \le \frac{1}{4} + \frac{c}{2} \qquad \text{and} \qquad \frac{\Lambda_{h,l}(c\sqrt{M})}{k} \ge \frac{1}{2} - \frac{3}{16c^2}, \quad (77)$$

so the crossover of every block, whatever its shape, is confined to the window $s \asymp \sqrt{\max(h,l)}$. An absolute $o(1)$ error in the Gaussian approximation of the log partition function needs the stronger, sufficient condition $s^2/[k(h+l)] \to \infty$; the distinction between *intensive* (per parameter) and *absolute* (free-energy) accuracy is material for basin comparisons, where free energies are subtracted.

Remark I.2 (required inverse temperature). For a gap $g > 0$, block (h,l) and tolerance $0 < \varepsilon < \frac{1}{4}$, the bounds say that the block may be treated as quartic to within ε per parameter whenever $\tau \le 4\varepsilon^2 M/g^2$ and as quadratic to within ε per parameter whenever $\tau \ge (h+l+1)/(16\varepsilon g^2)$; an absolute error at most ε in the Gaussian log-partition function needs the k-fold larger $\tau \ge k(h+l+1)/(16\varepsilon g^2)$. For the reduced Gibbs risk $\mathcal{R}_\perp^{\mathrm{red}} = \tau^{-1}\mathbb{E}_s Q$, these bounds imply

$$0 \le \mathcal{R}_\perp^{\mathrm{red}} \le \min\left\{\frac{k}{4\tau}, \frac{k(h+l+1)}{16g^2\tau^2}\right\}. \quad (78)$$

Thus $g^2\tau \ge (h+l+1)/(16\varepsilon)$ gives $\mathcal{R}_\perp^{\mathrm{red}} \le k\varepsilon/\tau$, a statement per redundant coordinate. For an absolute risk tolerance $\delta > 0$, a sufficient condition is instead $g^2\tau^2 \ge k(h+l+1)/(16\delta)$. Neither condition is a finite-sample guarantee for Equation (11).

For a square block $h = l = \nu \gg r$, balancing the equilibrium inactive risk with shrinkage bias predicts $g_* \asymp \nu^{3/4} r^{-1/4} \tau^{-1/2}$ (Appendix M.1). This is a population Gibbs-risk approximation; its use for growing blocks or noisy samples requires additional uniform control.

J Proofs for Appendix G

Throughout, $A_* \succeq 0$ has rank r, $0 \le r \le \min(d,m)$, and $W_* W_*^\top = A_*$.

J.1 Local integral and equivalent losses

For a nonnegative analytic loss K and a smooth positive density near a compact minimizing orbit, use a smooth cutoff supported in a neighbourhood isolating that orbit and put

$$Z_\chi(\tau) = \int \chi(W)\varphi(W)e^{-\tau K(W)}\,\mathrm{d}W. \quad (79)$$

The learning coefficient and multiplicity specify its leading power and logarithmic order.

Lemma J.1 (equivalent losses)**.** *If $cK' \le K \le CK'$ near a compact piece of the common zero set, with $0 < c \le C$, the local integral orders, and hence the learning coefficients and multiplicities, agree.*

Proof. $Z_\chi^{K'}(C\tau) \le Z_\chi^K(\tau) \le Z_\chi^{K'}(c\tau)$. □

J.2 Reduction to block form

Choose $U \in \mathrm{O}(d)$ with $U^\top A_* U = \mathrm{diag}(\Lambda_r, 0)$, $\Lambda_r \succ 0$. The map $W \mapsto U^\top W$ is a linear isometry of $\mathbb{R}^{d\times m}$ that carries K to the Gram loss with target $\mathrm{diag}(\Lambda_r, 0)$, so we assume $A_* = \mathrm{diag}(\Lambda_r, 0)$ from now on. Every W with $WW^\top = A_*$ is of the form $W_* Q$ with $Q \in \mathrm{O}(m)$, and we fix the representative

$$W_* = \begin{bmatrix} L_0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad L_0 \in \mathrm{LT}_+(r), \quad L_0 L_0^\top = \Lambda_r \tag{80}$$

(Cholesky). Let $h = d - r$, $l = m - r$, and define the *slice*

$$\mathcal{S} = \Big\{ S(L, C, D) = \begin{bmatrix} L & 0 \\ C & D \end{bmatrix} : \; L \in \mathrm{LT}_+(r),\; C \in \mathbb{R}^{h\times r},\; D \in \mathbb{R}^{h\times l} \Big\}, \tag{81}$$

of dimension $\dim \mathcal{S} = \frac{r(r+1)}{2} + hr + hl = q_r + k_r$.

J.3 The orbit chart

Lemma J.2 (orbit chart). *Let $\Phi : \mathcal{S} \times \mathrm{O}(m) \to \mathbb{R}^{d\times m}$, $\Phi(S, Q) = SQ$. There is a neighbourhood $\mathcal{S}_0$ of W_* in $\mathcal{S}$ such that:*

(a) Φ is a submersion on $\mathcal{S}_0 \times \mathrm{O}(m)$, and $V := \Phi(\mathcal{S}_0 \times \mathrm{O}(m))$ is an open neighbourhood of the orbit $W_* \mathrm{O}(m)$;

(b) the fibres of Φ are exactly the orbits of the free right action of $G = \mathrm{O}(l)$ given by $(S, Q) \cdot R = \big(S \,\mathrm{diag}(I_r, R),\; \mathrm{diag}(I_r, R^\top) Q\big)$;

(c) for every smooth positive φ on V and every integrable f on V that is $\mathrm{O}(m)$-invariant, $f(SQ) = f(S)$,

$$\int_V f(W)\varphi(W)\,\mathrm{d}W = \int_{\mathcal{S}_0} f(S)\,\tilde{\varphi}(S)\,\mathrm{d}S \tag{82}$$

with $\tilde{\varphi}$ smooth and positive on $\mathcal{S}_0$. If φ is replaced by $\chi\varphi$ with $\chi \geq 0$ smooth and positive on a nonempty open subset of the orbit, then $\tilde{\varphi}$ is smooth, nonnegative, and positive at W_*.

Proof. (a) Tangent vectors to $\mathrm{O}(m)$ at Q are ΩQ with $\Omega \in \mathfrak{so}(m)$, so $d\Phi_{(S,Q)}(\dot{S}, \Omega Q) = (\dot{S} + S\Omega)Q$, and surjectivity at (S, Q) is equivalent to surjectivity of $(\dot{S}, \Omega) \mapsto \dot{S} + S\Omega$. At $S = W_*$ write $\Omega = \left[\begin{smallmatrix} \Omega_{11} & \Omega_{12} \\ -\Omega_{12}^\top & \Omega_{22} \end{smallmatrix}\right]$ and $\dot{S} = \left[\begin{smallmatrix} \dot{L} & 0 \\ \dot{C} & \dot{D} \end{smallmatrix}\right]$; then

$$\dot{S} + W_*\Omega = \begin{bmatrix} \dot{L} + L_0\Omega_{11} & L_0\Omega_{12} \\ \dot{C} & \dot{D} \end{bmatrix}.$$

The lower blocks are free, the upper-right block is onto because L_0 is invertible, and the upper-left block is onto $\mathbb{R}^{r\times r}$ because the lower-triangular matrices $\{\dot{L}\}$ and $L_0\,\mathfrak{so}(r)$ have complementary dimensions $\binom{r+1}{2} + \binom{r}{2} = r^2$ and intersect trivially: $\dot{L} = -L_0\Omega_{11}$ forces $L_0^{-1}\dot{L}$ to be simultaneously lower triangular and antisymmetric, hence zero. So $d\Phi$ is onto at (W_*, I), hence at (W_*, Q) for every Q by equivariance, hence on $\mathcal{S}_0 \times \mathrm{O}(m)$ for a neighbourhood $\mathcal{S}_0$ by openness and compactness of $\mathrm{O}(m)$. We take $\mathcal{S}_0$ of the product form $\{(L, C) \in N\} \times \{\|D\|_F < \varrho\}$, which is invariant under $D \mapsto DR$, $R \in \mathrm{O}(l)$, so that the fibres in (b) lie entirely in $\mathcal{S}_0 \times \mathrm{O}(m)$. A submersion is open, so V is open, and it contains the orbit.

(b) $\Phi(S, Q) = \Phi(S', Q')$ iff $S' = SR$ with $R = QQ'^{-1} \in \mathrm{O}(m)$. Comparing the first r rows, $[L, 0]R = [L', 0]$ gives $R_{1:r,\cdot} = [L^{-1}L', 0]$; the rows of R are orthonormal, so $L^{-1}L' \in \mathrm{O}(r)$ is lower triangular with positive diagonal, hence $L^{-1}L' = I$, and orthogonality of R then forces $R = \mathrm{diag}(I_r, R')$ with $R' \in \mathrm{O}(l)$. Thus $S' = S\,\mathrm{diag}(I, R')$ and $Q' = \mathrm{diag}(I, R'^\top)Q$, which is the stated action; it is free because $Q \mapsto \mathrm{diag}(I, R'^\top)Q$ is.

(c) Let $\mathrm{NJ}\Phi = \sqrt{\det(d\Phi\, d\Phi^\top)}$, smooth and positive on $\mathcal{S}_0 \times \mathrm{O}(m)$ by (a). The coarea formula gives, for measurable $F \geq 0$, $\int F\,\mathrm{NJ}\Phi\,\mathrm{d}S\,\mathrm{d}Q = \int_V \big(\int_{\Phi^{-1}(W)} F\,\mathrm{d}\mathcal{H}\big)\,\mathrm{d}W$, where $\mathrm{d}Q$ is Haar measure and $\mathcal{H}$ the Hausdorff measure of dimension $\dim G$ on the compact fibres. Put $\omega(W) =$

Directions near $W_*\,\mathrm{O}(m)$	Number	Contribution at $g=0$
exact $\mathrm{O}(m)$ symmetry (moves W, not $WW^\top$)	$r(2m-r-1)/2$	orbit volume
identifiable (first order; quadratic loss)	q_r	$q_r/2$
redundant block D (second order; quartic loss $\|DD^\top\|_F^2$)	$k_r=(d-r)(m-r)$	$k_r/4$ (gap: Appendix H)

Table 2: The three kinds of direction at a rank-r minimizing orbit of the Gram loss.

$\int_{\Phi^{-1}(W)}(\mathrm{NJ}\Phi)^{-1}\,\mathrm{d}\mathcal{H}$, smooth and positive on V, and apply the formula to $F=(f\varphi/\omega)\circ\Phi\cdot(\mathrm{NJ}\Phi)^{-1}$:

$$\int_{\mathcal{S}_0\times\mathrm{O}(m)} f(SQ)\,\frac{\varphi(SQ)}{\omega(SQ)}\,\mathrm{d}S\,\mathrm{d}Q=\int_V f\varphi\,\mathrm{d}W.$$

Since $f(SQ)=f(S)$, the left side equals $\int_{\mathcal{S}_0} f(S)\tilde\varphi(S)\,\mathrm{d}S$ with $\tilde\varphi(S)=\int_{\mathrm{O}(m)}\varphi(SQ)/\omega(SQ)\,\mathrm{d}Q$, smooth and positive. With $\chi\varphi$ in place of φ the same formula gives $\tilde\varphi\ge 0$, and $\tilde\varphi(W_*)=\int_{\mathrm{O}(m)}\chi(W_*Q)\varphi(W_*Q)/\omega(W_*Q)\,\mathrm{d}Q>0$ when $\chi>0$ on an open subset of the orbit. □

Since $K(SQ)=K(S)$, Equation (82) shows that the learning coefficient and multiplicity of K on the orbit, with respect to $\chi\varphi$, equal those of $K|_{\mathcal{S}}$ at the point W_*, with respect to the weight $\tilde\varphi$. Note that $K|_{\mathcal{S}}$ vanishes only at W_* in $\mathcal{S}_0$: $SS^\top=A_*$ forces $LL^\top=\Lambda_r$, $LC^\top=0$ (so $C=0$) and $CC^\top+DD^\top=0$ (so $D=0$), and $L=L_0$ by uniqueness of the Cholesky factor.

J.4 Normal form on the slice

For $S=S(L,C,D)$,

$$K(S)=\left\|LL^\top-\Lambda_r\right\|_F^2+2\left\|LC^\top\right\|_F^2+\left\|CC^\top+DD^\top\right\|_F^2. \tag{83}$$

Lemma J.3 (normal form)**.** *There are local coordinates (u,D) on $\mathcal{S}$ near W_*, with $u=(LL^\top-\Lambda_r,\,LC^\top)\in\mathrm{Sym}(\mathbb{R}^r)\times\mathbb{R}^{r\times h}\cong\mathbb{R}^{q_r}$ and D unchanged, such that $(L,C)\mapsto u$ is a diffeomorphism near $(L_0,0)$ with $C=C(u)$, $\|C(u)\|_F\le c\,\|u\|$, and constants $0<c_1\le c_2$ with*

$$c_1\big(\|u\|^2+\|D\|_F^4\big)\;\le\;K\;\le\;c_2\big(\|u\|^2+\|D\|_F^4\big) \tag{84}$$

on a neighbourhood of $(u,D)=(0,0)$.

Proof. If $k_r=0$ (i.e. $r=d$ or $r=m$) there is no D block, the third term of Equation (83) is absent or equals $\|CC^\top\|_F^2=O(\|u\|^4)$, and $K\asymp\|u\|^2$ is regular on the slice; so assume $h,l\ge 1$. The Cholesky map $\mathrm{LT}_+(r)\to\{\Lambda\succ 0\}$, $L\mapsto LL^\top$, is a diffeomorphism, and for fixed invertible L the map $C\mapsto LC^\top$ is a linear isomorphism, so $(L,C)\mapsto u$ is a local diffeomorphism at $(L_0,0)$ with $C=(L^{-1}u_2)^\top=O(\|u_2\|)$. Lower bound: $CC^\top$ and $DD^\top$ are positive semidefinite, so $\langle CC^\top,\,DD^\top\rangle=\mathrm{tr}(CC^\top DD^\top)=\|D^\top C\|_F^2\ge 0$ and $\|CC^\top+DD^\top\|_F^2\ge\|DD^\top\|_F^2$; hence $K\ge\|u_1\|^2+2\|u_2\|^2+\|DD^\top\|_F^2\ge\min(1,\nu^{-1})(\|u\|^2+\|D\|_F^4)$, using $\|DD^\top\|_F^2=\sum_i\sigma_i(D)^4\ge\nu^{-1}(\sum_i\sigma_i(D)^2)^2=\nu^{-1}\|D\|_F^4$ with $\nu=\min(h,l)$. Upper bound: $\|CC^\top+DD^\top\|_F^2\le 2\|C\|_F^4+2\|DD^\top\|_F^2\le 2c^4\|u\|^4+2\|D\|_F^4$, and $\|u\|^4\le\|u\|^2$ near 0. □

J.5 Direction counts and rank dependence

Corollary J.4 (rank is not a proxy)**.** *At fixed d,m, $\lambda_{d,m,r+1}-\lambda_{d,m,r}=(d-m+1)/4$. Hence the coefficient is independent of the target rank when $m=d+1$ and* decreases *with the target rank when $m>d+1$. For $d=2$, $m=4$ the values at $r=0,1,2$ are $2,\ 7/4,\ 3/2$.*

The corollary compares different targets, so it is not by itself a posterior comparison between two basins for one data set; it says that a universal "lower rank ⇒ lower learning coefficient" premise fails for the fixed-readout network as soon as $m>d+1$.

J.6 Proof of Proposition G.1 and Corollary J.4

By Lemma J.2 and the discussion following it, the coefficient of K on the orbit equals that of $K|_{\mathcal{S}}$ at W_* with a smooth positive weight; by Lemma J.3 and Lemma J.1 it equals that of $\|u\|^2 + \|D\|_F^4$ at the origin of $\mathbb{R}^{q_r} \times \mathbb{R}^{k_r}$ with a smooth positive weight ψ. Rescaling $u \mapsto \tau^{-1/2}u$, $D \mapsto \tau^{-1/4}D$ and applying dominated convergence,

$$\int e^{-\tau(\|u\|^2+\|D\|_F^4)}\psi(u,D)\,\mathrm{d}u\,\mathrm{d}D = \tau^{-q_r/2-k_r/4}\Big(\psi(0,0)\int e^{-\|u\|^2}\,\mathrm{d}u\int e^{-\|D\|_F^4}\,\mathrm{d}D + o(1)\Big),$$

so $\lambda = q_r/2 + k_r/4$ with multiplicity 1. The closed form follows from $2q_r + k_r = 2dr - r(r-1) + dm - dr - mr + r^2 = dm + r(d-m+1)$, and Corollary J.4 is the difference of two such expressions. □

Numerical sanity check. The volume exponent was estimated by Monte Carlo as the slope of $\log \mathbb{P}(K < t)$ against $\log t$ for W uniform in the cube of half-width 0.4 around W_* (2×10^7 samples, t between 10^{-3} and $10^{-1.5}$): at $(d,m,r) = (2,2,1),(2,4,0),(2,4,1),(2,4,2),(3,2,1)$ the slopes were $1.25, 2.02, 1.76, 1.50, 1.99$ against the predicted $5/4, 2, 7/4, 3/2, 2$.

J.7 Precedents in detail

The Gram map is the symmetric cousin of reduced rank regression, whose learning coefficients were computed by Aoyagi & Watanabe (2005); the answers differ (for $d = m = 1$, $r = 0$ the asymmetric loss $(ab)^2$ has coefficient $1/2$ with multiplicity 2, the symmetric loss w^4 has $1/4$ with multiplicity 1). For probabilistic PCA, $\Sigma = \Lambda\Lambda^\top + \sigma^2 I$ with one isotropic noise parameter, Drton et al. (2026, Thm. 4.1) obtain $[pk + r(p-k+1) + 2]/4$ for $0 \le r \le k \le p-1$, which is Equation (64) plus $\frac{1}{2}$ after relabelling $(d,m) \to (p,k)$. For factor analysis, $\Sigma = \mathrm{diag}(\psi) + \Lambda\Lambda^\top$ with p free diagonal variances, Drton et al. (2025) obtain $[p(k+2) + r(p-k+1)]/4$ at generic non-saturated parameters, numerically Equation (64) plus $p/2$. In that result the p diagonal nuisance parameters are eliminated and the residual ideal is the off-diagonal Gram ideal, with a positive-dimensional fibre, so the numerical agreement with Equation (64) plus $p/2$ is not a formal implication in either direction, and the formula cannot extend to saturated cases: for $p = 2$, $k = r = 1$ the one-factor covariance map has a rank-3 Jacobian and coefficient $3/2$. The probabilistic-PCA range $k \le p-1$ excludes the overcomplete regime $m > d$ that Proposition G.1 includes.

K Proofs for Appendix H

K.1 Completing the square

With $A = WW^\top$ and $\|W\|_F^2 = \mathrm{tr}\,A$, $\|A - A_*\|_F^2 + \eta\,\mathrm{tr}\,A = \|A\|_F^2 - 2\langle A,\, A_* - \frac{\eta}{2}I\rangle + \|A_*\|_F^2 = \|A - (A_* - \frac{\eta}{2}I)\|_F^2 + \|A_*\|_F^2 - \|A_* - \frac{\eta}{2}I\|_F^2$, and $\|A_*\|_F^2 - \|A_* - \frac{\eta}{2}I\|_F^2 = \eta\,\mathrm{tr}\,A_* - d\eta^2/4$. This is Equation (65). The Frobenius projection of $A_* - \frac{\eta}{2}I$ onto the positive semidefinite cone clips the negative eigenvalues to zero, which gives the threshold $\eta/2$ and the gaps $g_i = \eta/2 - \lambda_i(A_*)$.

K.2 The energy on the slice

Put $\Lambda_+ = \mathrm{diag}(a_1,\dots,a_r)$ and use the slice of Appendix J with target $\mathrm{diag}(\Lambda_+, 0)$ and $W_* = \mathrm{diag}(L_0, 0)$, $L_0L_0^\top = \Lambda_+$. For $S = S(L, C_1, D)$, $SS^\top - B_g = \left[\begin{smallmatrix} LL^\top - \Lambda_+ & LC_1^\top \\ C_1L^\top & C_1C_1^\top + DD^\top + gI_h \end{smallmatrix}\right]$, and with $U = LL^\top - \Lambda_+$, $V = LC_1^\top$,

$$E_g(S) = \|U\|_F^2 + 2\|V\|_F^2 + \|C_1C_1^\top + DD^\top\|_F^2 + 2g\big(\|C_1\|_F^2 + \|D\|_F^2\big), \tag{85}$$

because $\|X + gI_h\|_F^2 = \|X\|_F^2 + 2g\,\mathrm{tr}\,X + hg^2$ and $\mathrm{tr}(C_1C_1^\top + DD^\top) = \|C_1\|_F^2 + \|D\|_F^2$. Since E_g is $\mathrm{O}(m)$-invariant, Lemma J.2 applies verbatim: $Z_\chi(\tau, g) = \int_{\mathcal{S}_0} e^{-\tau E_g(S)}\tilde{\varphi}(S)\,\mathrm{d}S$ with $\tilde{\varphi}$ smooth, nonnegative, positive at W_*, and $E_g|_{\mathcal{S}_0}$ vanishes only at W_* (the same argument as for K, since the positive semidefinite projection of B_g is $\mathrm{diag}(\Lambda_+, 0)$). Pass to the coordinates (U, V, D) of Lemma J.3, whose Jacobian $J(U,V)$ is smooth, positive and independent of D, and write $C_1 = C_1(U,V)$ with $\|C_1\|_F \le c\|V\|_F$.

K.3 Proof of Theorem H.1

Lower bound (domination). For $g \geq 0$ the last term of Equation (85) is nonnegative, and $\|C_1C_1^\top + DD^\top\|_F^2 \geq \|DD^\top\|_F^2$ by positivity (Lemma J.3), so

$$E_g(S) \;\geq\; \|U\|_F^2 + 2\|V\|_F^2 + \|DD^\top\|_F^2 \qquad \text{for all } g \geq 0. \tag{86}$$

Rescaling. Substitute $U = \tau^{-1/2}U'$, $V = \tau^{-1/2}V'$, $D = \tau^{-1/4}D'$ and $g = s/\sqrt{\tau}$. Then $\mathrm{d}U\,\mathrm{d}V\,\mathrm{d}D = \tau^{-q/2-k/4}\,\mathrm{d}U'\,\mathrm{d}V'\,\mathrm{d}D'$, and since $\|C_1\|_F \leq c\tau^{-1/2}\|V'\|_F$,

$$\begin{aligned} \tau E_g &= \|U'\|_F^2 + 2\|V'\|_F^2 + \|D'D'^\top\|_F^2 + 2s\|D'\|_F^2 + \mathcal{R}, \\ \mathcal{R} &:= \tau\|C_1C_1^\top\|_F^2 + 2\tau\left\langle C_1C_1^\top,\, DD^\top\right\rangle + 2s\sqrt{\tau}\|C_1\|_F^2, \end{aligned} \tag{87}$$

with $0 \leq \mathcal{R} \leq c^4\tau^{-1}\|V'\|_F^4 + 2c^2\tau^{-1/2}\|V'\|_F^2\|D'\|_F^2 + 2sc^2\tau^{-1/2}\|V'\|_F^2$. Hence, for $\tau \geq 1$ and $s \leq S$, $\mathcal{R} \leq \tau^{-1/2}(1+S)\,P(U', V', D')$ for a fixed polynomial P, so $\mathcal{R} \to 0$ pointwise, uniformly in $s \in [0, S]$. *Convergence.* The integrand of $\tau^{q/2+k/4}Z_\chi(\tau, s/\sqrt{\tau})$ in the primed variables is $\tilde{\varphi}J\,e^{-\tau E_g}$ evaluated at the rescaled point; it converges pointwise to $\tilde{\varphi}(W_*)J(0,0)\exp(-\|U'\|_F^2 - 2\|V'\|_F^2 - \|D'D'^\top\|_F^2 - 2s\|D'\|_F^2)$, and by Equation (86) it is dominated by $\sup(\tilde{\varphi}J)\exp(-\|U'\|_F^2 - 2\|V'\|_F^2 - \|D'D'^\top\|_F^2)$, which is integrable because $\|D'D'^\top\|_F^2 \geq \nu^{-1}\|D'\|_F^4$, and does not depend on s. The cutoff region in (U', V', D') exhausts $\mathbb{R}^{q+k}$ as $\tau \to \infty$, and the contribution of $\mathcal{S}_0$ outside a fixed ball around W_* is $O(e^{-c\tau})$ because E_g is bounded below there uniformly in $g \geq 0$. Uniformity in s comes from the fact that the difference between the rescaled integrand and its limit is bounded by an s-independent function: writing $F_0 = \|U'\|_F^2 + 2\|V'\|_F^2 + \|D'D'^\top\|_F^2$, κ_τ for the value of $\tilde{\varphi}J$ at the rescaled point, $\kappa_0 = \tilde{\varphi}(W_*)J(0,0)$, and using $\mathcal{R} \leq \tau^{-1/2}(1+S)P$,

$$\left|\kappa_\tau e^{-\tau E_g} - \kappa_0 e^{-F_s}\right| \leq |\kappa_\tau - \kappa_0|\,e^{-F_0} + \kappa_0\,\min\{1, \tau^{-1/2}(1+S)P\}\,e^{-F_0},$$

and the right side tends to zero pointwise, is dominated by $3\sup(\tilde{\varphi}J)e^{-F_0}$, and does not depend on s; hence $\sup_{s\leq S}$ of the integrated difference tends to zero by dominated convergence, which gives Equation (69) with $C = \tilde{\varphi}(W_*)J(0,0)\int e^{-\|U'\|_F^2 - 2\|V'\|_F^2}\,\mathrm{d}U'\,\mathrm{d}V' > 0$.

Energy moment. Insert τE_g in the integrand. By Equation (87), $\tau E_g \leq F + \tau^{-1/2}(1+S)P$ with $F = \|U'\|_F^2 + 2\|V'\|_F^2 + \|D'D'^\top\|_F^2 + 2s\|D'\|_F^2$, and $\tau E_g \geq F - 2s\|D'\|_F^2 \geq 0$, so $\tau E_g\,e^{-\tau E_g}$ is dominated by $(F + (1+S)P)\,e^{-\|U'\|_F^2 - 2\|V'\|_F^2 - \|D'D'^\top\|_F^2}$, integrable. The same dominated convergence gives

$$\tau\,\mathbb{E}_{\tau, s/\sqrt{\tau}}[E_g] \;\to\; \frac{\int F\,e^{-F}}{\int e^{-F}} = \mathbb{E}\big[\|U'\|_F^2\big] + \mathbb{E}\big[2\|V'\|_F^2\big] + \mathbb{E}_s\big[\|DD^\top\|_F^2 + 2s\|D\|_F^2\big],$$

where the limit measure factorises into a Gaussian in (U', V') with density $\propto e^{-\|U'\|_F^2 - 2\|V'\|_F^2}$ and the reduced density $\propto e^{-\|DD^\top\|_F^2 - 2s\|D\|_F^2}$ in D. The Gaussian contributes $\frac{1}{2}\dim U' + \frac{1}{2}\dim V' = q/2$, and the reduced part contributes $\mathbb{E}_s[\|DD^\top\|_F^2 + 2s\|D\|_F^2] = k/4 + s\mathbb{E}_s\|D\|_F^2$ by Lemma K.1 below. □

Lemma K.1 (Euler identity). *For $s \geq 0$ let $\mathbb{E}_s$ be the expectation with density $\propto e^{-F_s(D)}$, $F_s(D) = \|DD^\top\|_F^2 + 2s\|D\|_F^2$ on $\mathbb{R}^{h\times l}$. Then $k = 4\mathbb{E}_s\|DD^\top\|_F^2 + 4s\mathbb{E}_s\|D\|_F^2$.*

Proof. F_s is a sum of homogeneous terms of degrees 4 and 2, so $D\cdot\nabla F_s = 4\|DD^\top\|_F^2 + 4s\|D\|_F^2$. Integrating $\mathrm{div}(D\,e^{-F_s}) = (k - D\cdot\nabla F_s)e^{-F_s}$ over $\mathbb{R}^{h\times l}$ gives zero, because $e^{-F_s} \leq e^{-\|D\|_F^4/\nu}$ kills the boundary term. □

K.4 Proof of Corollary H.2

$\Lambda_{h,l}(0) = k/4$ by Lemma K.1 at $s = 0$. The limit $s \to \infty$ is Equation (76). For fixed $g > 0$, E_g restricted to the slice has, at W_*, the positive definite quadratic part $\|U\|_F^2 + 2\|V\|_F^2 + 2g\|C_1(U,V)\|_F^2 + 2g\|D\|_F^2$ in the $q+k$ variables (U, V, D) (the term $2g\|C_1\|_F^2$ is quadratic in V to leading order and only changes the Gaussian form in V), so the classical Laplace expansion gives $\tau\mathbb{E}_{\tau,g}[E_g] \to (q+k)/2$; the two iterated limits differ by $k/4$. □

K.5 Proof of Corollary H.3

Fix the teacher $A_* = \mathrm{diag}(\alpha, 0_h)$ and let $g = s/\sqrt{\tau}$, $0 \le s \le S$. For all sufficiently large τ, the positive block $\Lambda_g = \mathrm{diag}(\alpha_1 - g, \dots, \alpha_r - g)$ has its spectrum in a fixed compact subset of $(0, \infty)$. Work in the same slice as before, now with $U = LL^\top - \Lambda_g$ and $V = LC_1^\top$. The Cholesky map and its inverse depend smoothly on (g, U), and $C_1 = (L^{-1}V)^\top$ satisfies $\|C_1\|_F \le c\|V\|_F$ with one constant c for this range of g. The orbit-integrated density and the Jacobian are uniformly bounded on a common small neighbourhood. At the rescaled point they converge uniformly in $s \le S$ to their positive value at the true orbit, since $g \to 0$ uniformly.

After subtracting the minimum, the regularised energy has exactly the form Equation (85). Its lower bound Equation (86) and the polynomial remainder estimate Equation (87) therefore hold with constants independent of s and τ. The limiting density in (U', V', D') is the same product of a centered Gaussian and the reduced law as in Theorem H.1. The fixed cutoff contains a neighbourhood of every moving minimizer for sufficiently large τ. Outside a common small coordinate ball, the energy is uniformly positive by compactness and continuity at $g = 0$. These observations give uniform dominated convergence for the partition function and for all the polynomial moments used below.

On the slice,

$$WW^\top - A_* = \begin{bmatrix} U - gI_r & V \\ V^\top & C_1C_1^\top + DD^\top \end{bmatrix}.$$

After $U, V = \tau^{-1/2}(U', V')$ and $D = \tau^{-1/4}D'$, this implies

$$\tau K = \|U' - sI_r\|_F^2 + 2\|V'\|_F^2 + \|D'D'^\top + \sqrt{\tau}C_1C_1^\top\|_F^2.$$

The last correction vanishes in the required moments because $\sqrt{\tau}\|C_1\|_F^2 \le c^2\tau^{-1/2}\|V'\|_F^2$. The limiting Gaussian is centered and contributes $q/2$, the deterministic shift contributes rs^2, and the inactive block contributes $\mathbb{E}_s Q$. This proves Equation (72). For the mean predictor, dominated convergence gives $\mathbb{E}U' \to 0$, $\mathbb{E}V' \to 0$, and $\sqrt{\tau}\,\mathbb{E}[C_1C_1^\top + DD^\top] \to \mathbb{E}_s[DD^\top]$ uniformly in $s \le S$. The reduced law is invariant under left multiplication by $\mathrm{O}(h)$, so $\mathbb{E}_s[DD^\top] = (\mathbb{E}_s T/h)I_h$. Squaring the limiting mean matrix proves Equation (73). Only convergence of the first moments is needed; no finite-τ rate for their means is asserted. □

K.6 Proof of Corollary H.4

Differentiation of the reduced density gives $\partial_s \mathbb{E}_s T = -2\,\mathrm{Var}_s T$. The Euler identity $\mathbb{E}_s Q = k/4 - s\mathbb{E}_s T$ then yields $\partial_s \mathbb{E}_s Q = -\mathbb{E}_s T + 2s\,\mathrm{Var}_s T$. Consequently $\mathcal{C}_G'(0) = -\mathbb{E}_0 T < 0$ and $\mathcal{C}_{\mathrm{mean}}'(0) = -4\mathbb{E}_0 T\,\mathrm{Var}_0 T/h < 0$. Both coefficients are continuous and tend to infinity as $s \to \infty$ when $r \ge 1$, because they are at least rs^2. They therefore have positive finite minimizers; their stationary equations are Equation (74). At $r = 0$ the reduced Gibbs and mean risks are positive at finite s and tend to zero by Equation (76) and Jensen's inequality. No uniqueness, global finite-temperature optimality, or empirical-risk optimality is claimed. □

Remark K.2 (fixed prior versus fixed weight decay). A Gaussian factor $e^{-\gamma\|D\|^2/2}$ placed outside τ rescales to an effective scaled gap of order $\gamma/\sqrt{\tau}$ (in the convention of Equation (68), $s = \gamma/(4\sqrt{\tau})$), which tends to zero: a fixed prior or localizer becomes irrelevant at low temperature and the coefficient tends to the singular value, whereas fixed weight decay inside the energy gives $s = g\sqrt{\tau} \to \infty$ and the regular value. The two conventions follow opposite low-temperature paths, and an estimator must specify which it uses (Appendix R).

K.7 Scalar calibration

The case $h = l = 1$ of Equation (68) is $\Psi_{1,1}(s) = \int_{\mathbb{R}} e^{-v^4 - 2sv^2}\,\mathrm{d}v$; let $\mathbb{E}_s$ denote expectation under the density $\propto e^{-v^4-2sv^2}$. With $u = \tau^{-1/4}v$, $I(\tau, g) = \int_{\mathbb{R}} e^{-\tau(u^4+2gu^2)}\,\mathrm{d}u = \tau^{-1/4}\Psi_{1,1}(s)$, $s = g\sqrt{\tau}$, and $\partial s/\partial \log\tau = s/2$ at fixed g, so

$$\Lambda_{1,1}(s) = -\frac{\partial \log I}{\partial \log \tau} = \frac{1}{4} - \frac{s}{2}\frac{\Psi_{1,1}'(s)}{\Psi_{1,1}(s)} = \frac{1}{4} + s\,\mathbb{E}_s[v^2] = \mathbb{E}_s[v^4 + 2sv^2], \tag{88}$$

where the third expression uses $-\Psi_{1,1}'(s)/\Psi_{1,1}(s) = 2\mathbb{E}_s[v^2]$ and the last follows from integrating $v\,\partial_v e^{-v^4-2sv^2}$ by parts, which gives $4\mathbb{E}_s v^4 + 4s\mathbb{E}_s v^2 = 1$ (the case $k = 1$ of Lemma K.1).

$\Lambda_{1,1}(0) = \frac{1}{4}$ and $\Lambda_{1,1}(s) \to \frac{1}{2}$ as $s \to \infty$. The values quoted in Appendix H, $\Lambda_{1,1}(0.05) = 0.266240$, $\Lambda_{1,1}(0.5) = 0.366980$, $\Lambda_{1,1}(5) = 0.493036$, $\Lambda_{1,1}(50) = 0.499925$, were computed by adaptive quadrature of $\Psi_{1,1}$ and $\Psi'_{1,1}$ and checked against the last identity in Equation (88) (agreement to 10^{-12}). The factor coordinate u and the represented coordinate $A = u^2$ scale differently: $\operatorname{Var} u \asymp \tau^{-1/2}$ and $\operatorname{Var} A \asymp \tau^{-1}$ at $g = 0$, whereas $\operatorname{Var} u \sim 1/(4g\tau)$ and $\operatorname{Var} A \asymp \tau^{-2}$ at fixed $g > 0$, while a regular matrix coordinate at an active optimum has variance $\asymp \tau^{-1}$; a prediction about "posterior variance at the spectral threshold" must therefore say whether it concerns factors or predictions.

L Proofs for Appendix I

L.1 Proof of Proposition M.1

Assume $h \geq l$ (otherwise replace D by $D^\top$, which leaves $\|DD^\top\|_F^2 = \|D^\top D\|_F^2$ and $\|D\|_F^2$ unchanged), so $\nu = l$ and $a = h - l$. The integrand of Equation (68) is a function of $X = D^\top D \in \operatorname{Sym}(\mathbb{R}^l)$ alone, namely $\exp(-\operatorname{tr} X^2 - 2s \operatorname{tr} X)$. For a density on $\mathbb{R}^{h\times l}$ that depends on D only through X, the induced density of X on the positive definite cone is proportional to $(\det X)^{(h-l-1)/2}$ times that function (Muirhead, 1982, Thm. 2.1.14), and the induced density of the eigenvalues $x_1, \dots, x_l$ of X carries the additional Jacobian $\prod_{i<j} |x_i - x_j|$ of the spectral decomposition, the angular integral over $\mathrm{O}(l)$ being a constant because the density is orthogonally invariant (Muirhead, 1982, Thm. 3.2.17); see also Forrester (2010, Ch. 3). Since $\operatorname{tr} X^2 = \sum_i x_i^2$ and $\operatorname{tr} X = \sum_i x_i$, this is Equation (89). □

L.2 Proof of Proposition I.1

Write $Q = \|DD^\top\|_F^2 = \sum_i \sigma_i^4$ and $T = \|D\|_F^2 = \sum_i \sigma_i^2$ in terms of the ν nonzero singular values. Cauchy–Schwarz gives $T^2 \leq \nu Q$, and Lemma K.1 gives $\mathbb{E}_s Q = k/4 - s\mathbb{E}_s T \leq k/4$, so that $\Lambda_{h,l}(s)/k - \frac{1}{4} = s\mathbb{E}_s T/k = \frac{1}{4} - \mathbb{E}_s Q/k \in [0, \frac{1}{4}]$. Moreover $\mathbb{E}_s T \leq (\mathbb{E}_s T^2)^{1/2} \leq (\nu \mathbb{E}_s Q)^{1/2} \leq \frac{1}{2}\sqrt{\nu k}$, hence $s\mathbb{E}_s T/k \leq s\sqrt{\nu k}/(2k) = s/(2\sqrt{M})$ because $k/\nu = M$. This is Equation (75).

For Equation (76), let $\mathbb{E}_G$ denote expectation under the Gaussian density $\propto e^{-2s\|D\|_F^2}$, under which the entries of D are i.i.d. $\mathcal{N}(0, 1/4s)$. The reduced density is this Gaussian tilted by e^{-Q}: $\mathbb{E}_s Q = \mathbb{E}_G[Qe^{-Q}]/\mathbb{E}_G[e^{-Q}]$. Since $x \mapsto x$ is nondecreasing and $x \mapsto e^{-x}$ nonincreasing, Chebyshev's association inequality $\mathbb{E}_G[Qe^{-Q}] \leq \mathbb{E}_G[Q]\,\mathbb{E}_G[e^{-Q}]$ (proved by expanding $\mathbb{E}_G[(Q - Q')(e^{-Q} - e^{-Q'})] \leq 0$ for an independent copy Q') yields $\mathbb{E}_s Q \leq \mathbb{E}_G Q$. With $\sigma^2 = 1/4s$,

$$\mathbb{E}_G Q = \sum_i \mathbb{E}\Big(\sum_k D_{ik}^2\Big)^2 + \sum_{i\neq j} \mathbb{E}\Big(\sum_k D_{ik}D_{jk}\Big)^2 = \big[h(l^2+2l) + h(h-1)l\big]\sigma^4 = \frac{hl(h+l+1)}{16s^2}.$$

Therefore $\frac{1}{2} - \Lambda_{h,l}(s)/k = \mathbb{E}_s Q/k \leq (h+l+1)/(16s^2)$, and $\leq \frac{1}{4}$ by the first part. Finally $\Psi_{h,l}(s) = \int e^{-2s\|D\|_F^2}\,\mathrm{d}D \cdot \mathbb{E}_G[e^{-Q}] = (\pi/2s)^{k/2}\mathbb{E}_G[e^{-Q}]$, and Jensen's inequality $\mathbb{E}_G[e^{-Q}] \geq e^{-\mathbb{E}_G Q}$ together with $e^{-Q} \leq 1$ gives $0 \leq -\log \mathbb{E}_G[e^{-Q}] \leq \mathbb{E}_G Q$, which is the free-energy bound. □

L.3 The scaled functional

In Equation (89) substitute $x_i = \sqrt{\nu}\, y_i$ and $s = \sqrt{\nu}\, t$. The exponent becomes $-\nu \sum_i (y_i^2 + 2ty_i)$, the Vandermonde becomes $\nu^{\nu(\nu-1)/4} \prod_{i<j} |y_i - y_j|$, and the wall factor becomes $\nu^{\nu(a-1)/4} \prod_i y_i^{(a-1)/2}$. With ρ_ν the empirical measure of the y_i and $a = \delta\nu + o(\nu)$,

$$\begin{aligned}
&\nu \sum_i (y_i^2 + 2ty_i) - \sum_{i<j} \log|y_i - y_j| - \frac{a-1}{2} \sum_i \log y_i \\
&\qquad = \nu^2 \Big[\int (y^2 + 2ty)\,\mathrm{d}\rho_\nu - \frac{1}{2} \iint_{y\neq z} \log|y - z|\;\mathrm{d}\rho_\nu\,\mathrm{d}\rho_\nu - \frac{\delta}{2} \int \log y\,\mathrm{d}\rho_\nu \Big] + o(\nu^2),
\end{aligned}$$

up to terms independent of t and ρ_ν. The bracket is $\mathcal{E}_t^{\neq}(\rho_\nu)$; the diagonal must be excluded because $\iint \log|y - z|\;\mathrm{d}\rho_\nu\,\mathrm{d}\rho_\nu$ is $-\infty$ for an atomic measure, and the passage to $\mathcal{E}_t$ on absolutely continuous

measures is the usual large-deviation statement for log-gases (Anderson et al., 2010). The scales quoted in Appendix I follow: at bounded t the squared singular values of the scaled variable are $O(\sqrt{\nu})$, so in the original block $D = \tau^{-1/4} D'$ the operator norm is $O((\sqrt{\nu}/\sqrt{\tau})^{1/2})$ and the squared Frobenius norm is $O(\nu^{3/2}\tau^{-1/2})$.

L.4 Proof of Proposition M.2 and Corollary M.3

Let $\delta = 0$, $t \geq 0$, $b = b_t$, and let ρ_t be given by Equation (91). Define $G(z) = 2z + 2t - (2z + 2t + b)\sqrt{(z-b)/z}$ with the branch of the square root that tends to 1 as $z \to \infty$; G is analytic off $[0, b]$. For $y \in (0, b)$, $(y-b)/y < 0$, so $G(y \pm i0) = 2y + 2t \mp i(2y + 2t + b)\sqrt{(b-y)/y}$: the real part is $2y + 2t$ and $-\frac{1}{\pi}\operatorname{Im} G(y + i0) = \rho_t(y)$. Expanding at infinity with $\sqrt{1 - b/z} = 1 - \frac{b}{2z} - \frac{b^2}{8z^2} - \frac{b^3}{16z^3} - \frac{5b^4}{128z^4} - \dots$,

$$G(z) = \frac{tb + \frac{3}{4}b^2}{z} + \frac{b^2(t+b)/4}{z^2} + \frac{(9b^4 + 8tb^3)/64}{z^3} + O(z^{-4}).$$

Since G is analytic off the support and decays at infinity, it is the Stieltjes transform $\int \rho_t(y)\,\mathrm{d}y/(z - y)$ of the measure with density $-\frac{1}{\pi}\operatorname{Im} G(y + i0)$, and the $1/z$ coefficient is its total mass; so ρ_t is a probability measure exactly when $tb + \frac{3}{4}b^2 = 1$, i.e. $b = \frac{2}{3}(\sqrt{t^2 + 3} - t)$. The next coefficients give $m_1 = \frac{1}{4}b^2(t + b)$ and $m_2 = (9b^4 + 8tb^3)/64$, and $m_2 + tm_1 = b^2(3b + 4t)^2/64 = \frac{1}{4}$ by the normalization identity $b(3b + 4t) = 4$.

On the support, $\operatorname{Re} G(y \pm i0) = \mathrm{PV}\int \rho_t(z)\,\mathrm{d}z/(y - z) = 2y + 2t = V'(y)$ with $V(y) = y^2 + 2ty$, which is the differentiated Euler–Lagrange equation $V(y) - \int \log|y - z|\,\mathrm{d}\rho_t(z) = \text{const}$ on $\operatorname{supp}\rho_t$. Off the support, for $y > b$, the derivative of the effective potential is $V'(y) - G(y) = (2y + 2t + b)\sqrt{(y-b)/y} > 0$, so the effective potential is strictly larger than its constant value on the support. These two conditions characterise the minimizer of the strictly convex functional Equation (90) on probability measures supported in $[0, \infty)$, which is unique (Anderson et al., 2010, Sec. 2.6); hence ρ_t is the equilibrium measure. For Corollary M.3, $\Lambda_{\nu,\nu}(s) = \nu^2/4 + s\mathbb{E}_s\|D\|_F^2$ with $s = \sqrt{\nu}t$ and $\|D\|_F^2 = \sum_i x_i = \nu^{3/2}\int y\,\mathrm{d}\rho_\nu$, so that $\Lambda_{\nu,\nu}(\sqrt{\nu}t)/\nu^2 = \frac{1}{4} + t\,\mathbb{E}\int y\,\mathrm{d}\rho_\nu \to \frac{1}{4} + t\,m_1(t)$ under the stated convergence hypothesis; the large-t form follows from $b_t = t^{-1} - \frac{3}{4}t^{-3} + O(t^{-5})$. □

L.5 Numerical record

The normalization, the moments, the identity $m_2 + tm_1 = \frac{1}{4}$ and the closed form of G against direct quadrature of the Stieltjes integral were checked at $t \in \{0, \frac{1}{2}, 1, 3, 10\}$ (agreement to 10^{-14}). The convergence hypothesis was tested at $t = 1$ by single-site Metropolis sampling of the scaled eigenvalue density $\propto \exp(-\nu\sum_i(y_i^2 + 2y_i))\prod_{i<j}|y_i - y_j|\prod_i y_i^{-1/2}$ on $(0, \infty)^\nu$ with Gaussian proposals of width $0.45\,b_1/\sqrt{\nu}$ (acceptance 0.33–0.43), 12,000 sweeps of which the first 2,000 were discarded. The quoted uncertainties are standard errors of the mean obtained by successive blocking (integrated autocorrelation times of 6–9 sweeps). The estimates of $\Lambda_{\nu,\nu}(\sqrt{\nu})/\nu^2 = \frac{1}{4} + \mathbb{E}\int y\,\mathrm{d}\rho_\nu$ were 0.4298(23), 0.4325(11), 0.4348(6), 0.4346(2), 0.4348(1) for $\nu = 4, 8, 16, 32, 64$, and a weighted fit of the form $F + c/\nu$ gave $F = 0.4351(2)$, $c = -0.017(6)$ with χ^2 per degree of freedom 0.63, against $F(1) = 47/108 = 0.43519$.

Figure 3 was produced by the same single-site Metropolis sampler applied to the unscaled eigenvalue density Equation (89) with $a = 0$, at $t = s/\sqrt{\nu} \in \{0, 0.1, 0.2, 0.35, 0.5, 0.75, 1, 1.5, 2, 3, 4\}$ and $\nu \in \{2, 4, 8, 16, 32, 64\}$: 256 ($\nu \leq 8$), 128 ($\nu = 16, 32$) or 64 ($\nu = 64$) independent chains per point, initialized uniformly on the expected support, Gaussian proposals of width $0.6\,b_t/\sqrt{\nu}$ (acceptance 0.4–0.65), 3,000 sweeps of which the first 1,000 were discarded (2,500 and 800 for $\nu \geq 16$; the points with $t \geq 1.5$ and $\nu \geq 16$ were rerun with 7,000 sweeps, 3,000 discarded, and proposals of width $b_t/\sqrt{\nu}$). The plotted quantity is $\frac{1}{4} + s\,\mathbb{E}[\sum_i x_i]/\nu^2$; its uncertainty is the standard error over chains of the chain means, so no autocorrelation correction is needed, and the largest error bar is 1.1×10^{-3}. The $\nu = 1$ curve is $\Lambda_{1,1}(s)$ by quadrature (Appendix K.7), and the sampler was checked against it at $s \in \{0.05, 0.5, 1, 5\}$.

M SPECTRAL DESCRIPTION OF THE POPULATION REDUCED ENSEMBLE

Proposition M.1 (rectangular eigenvalue representation). *Let $\nu = \min(h,l)$, $a = |h-l|$, and let $x_1, \dots, x_\nu$ be the nonzero squared singular values of D. Then, for a constant $C_{h,l} > 0$ independent of s,*

$$\Psi_{h,l}(s) = C_{h,l} \int_{(0,\infty)^\nu} \prod_{i<j} |x_i - x_j| \prod_{i=1}^{\nu} x_i^{(a-1)/2}\, e^{-x_i^2 - 2s x_i}\, \mathrm{d}x. \tag{89}$$

This is a real Laguerre-type ensemble, with Vandermonde exponent one, a quadratic potential in the squared singular values, a linear source $2s$, and a hard wall at zero. For $l > h$, the $l - h$ deterministic zero eigenvalues of $D^\top D$ carry no density. The factor-volume exponent $hl/4$ comes from quartic homogeneity, not the hard-edge exponent $(a-1)/2$. The change-of-variables proof is in Appendix L.

The collective scale. For an extensive block, $h, l \asymp \nu$ with $a/\nu \to \delta \geq 0$, the window Equation (77) is $s \asymp \sqrt{\nu}$, and the log-gas makes it explicit. Set $x = \sqrt{\nu}\, y$ and $s = \sqrt{\nu}\, t$ in Equation (89). Writing $\rho_\nu = \nu^{-1} \sum_i \delta_{y_i}$ for the empirical measure and excluding the diagonal from the logarithmic sum, the exponent of the joint density is $-\nu^2 \mathcal{E}_t^{\neq}(\rho_\nu) + o(\nu^2)$ with

$$\mathcal{E}_t(\rho) = \int (y^2 + 2ty)\, \mathrm{d}\rho(y) - \frac{1}{2} \iint \log|y - z|\, \mathrm{d}\rho(y)\, \mathrm{d}\rho(z) - \frac{\delta}{2} \int \log y\, \mathrm{d}\rho(y), \tag{90}$$

where $\mathcal{E}_t^{\neq}$ denotes the same expression with the diagonal $y = z$ removed (Appendix L). In the original variables the extensive block therefore crosses over when $g\sqrt{\tau/\nu} = O(1)$, not when $g\sqrt{\tau} = O(1)$. Squared singular values of order $\sqrt{\nu}$ in the scaled variables also mean that the localized expansion can describe the block only when $\tau \gg \nu$ (operator norm) or $\tau \gg \nu^3$ (Frobenius norm), a scaling deduction rather than a localization guarantee; unequal gaps leave a Harish-Chandra–Itzykson–Zuber integral, which we do not treat.

The square block: an explicit crossover. For $\delta = 0$ the variational problem Equation (90) is a log-gas with the convex potential $y^2 + 2ty$ and a hard wall at 0, and its minimizer is explicit.

Proposition M.2 (square-block equilibrium measure). *Let $\delta = 0$ and $t \geq 0$, and put $b_t = \frac{2}{3}\big(\sqrt{t^2+3} - t\big)$. The unique minimizer of Equation (90) over probability measures on $[0,\infty)$ is*

$$\rho_t(y) = \frac{2y + 2t + b_t}{\pi} \sqrt{\frac{b_t - y}{y}}\, \mathbf{1}_{(0,b_t)}(y), \tag{91}$$

with Stieltjes transform $G_t(z) = 2z + 2t - (2z + 2t + b_t)\sqrt{(z - b_t)/z}$, first moment $m_1(t) = \frac{1}{4} b_t^2 (t + b_t)$ and $m_2(t) + t\, m_1(t) = \frac{1}{4}$.

Corollary M.3 (equilibrium prediction for the crossover). *If the scaled first moment of the square ensemble Equation (89) with $h = l = \nu$ converges in expectation, $\nu^{-3/2}\, \mathbb{E}_{\sqrt{\nu}t}\big[\sum_i x_i\big] \to m_1(t)$ as $\nu \to \infty$, then the intensive coefficient of a square block crosses over as*

$$F(t) := \lim_{\nu\to\infty} \frac{\Lambda_{\nu,\nu}(\sqrt{\nu}\, t)}{\nu^2} = \frac{1}{4} + \frac{t\, b_t^2 (t + b_t)}{4}, \qquad F(0) = \frac{1}{4}, \quad F(1) = \frac{47}{108}, \quad F(t) \to \frac{1}{2}, \tag{92}$$

with $F(t) = \frac{1}{2} - 1/(8t^2) + O(t^{-4})$ as $t \to \infty$.

The density has the hard-edge singularity $y^{-1/2}$ at the wall for every t, and the support shrinks as $b_t \sim 1/t$: the gap pushes the spectrum of $D^\top D$ toward zero while the Euler identity $m_2 + t m_1 = \frac{1}{4}$ is preserved. The proof (Appendix L) is a standard resolvent computation: the Euler–Lagrange equation on the support and the exterior inequality off it follow from the Stieltjes transform, whose expansion at infinity gives the moments and the normalization $t b_t + \frac{3}{4} b_t^2 = 1$. The hypothesis of Corollary M.3 is a concentration statement for the finite-ν log-gas (a large deviation principle for the empirical measure with a tail bound for the largest eigenvalue); we do not prove it, so $F(t)$ is an *equilibrium prediction*. The rectangular case $\delta > 0$ is not computed here.

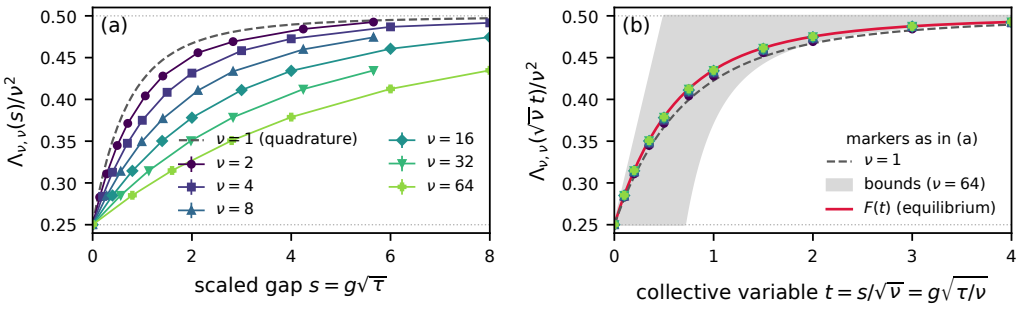

Figure 3: Crossover of the reduced ensemble. Left: the intensive coefficient $\Lambda_{\nu,\nu}(s)/\nu^2$ of a square $\nu \times \nu$ block against the scaled gap $s = g\sqrt{\tau}$ for $\nu = 1$ (quadrature) and $\nu = 2, 4, 8, 16, 32, 64$ (Metropolis, 64–256 independent chains per point; error bars, standard errors over chains, are smaller than the markers). Right: the same data against the collective variable $t = s/\sqrt{\nu}$, with the equilibrium prediction $F(t)$ of Corollary M.3 (line) and the envelope of the bounds of Proposition I.1 for $\nu = 64$ (shaded).

M.1 A WIDTH-DEPENDENT POPULATION REGULARIZATION PREDICTION

For a square redundant block $h = l = \nu$ and $r \geq 1$, put $s = \sqrt{\nu}\, t$. Combining the population Gibbs-risk coefficient in Corollary H.3 with the equilibrium approximation replaces $(\mathcal{C}_G(s) - q/2)/\nu^2$ by

$$J_\nu(t) = \frac{r}{\nu}t^2 + \frac{1}{2} - F(t) = \frac{r}{\nu}t^2 + \frac{1}{8t^2} + O(t^{-4}). \tag{93}$$

As $r/\nu \to 0$, the minimizer of this proxy satisfies

$$t_* \sim \left(\frac{\nu}{8r}\right)^{1/4}, \qquad g_*^{\text{proxy}} = \frac{\sqrt{\nu}\, t_*}{\sqrt{\tau}} \sim \frac{\nu^{3/4}}{(8r)^{1/4}\sqrt{\tau}}. \tag{94}$$

Indeed, $\frac{1}{2} - F(t) > 0$ for bounded t, whereas choosing $t \asymp (\nu/r)^{1/4}$ makes the proxy tend to zero. Its minimizer therefore tends to infinity, where the two leading terms in Equation (93) balance. The remaining inactive-risk fraction, relative to its unregularized value $\nu^2/4$, is asymptotic to $\sqrt{2r/\nu}$. The shrinkage bias contributes the same leading amount: the bias-plus-inactive-risk fraction is $2\sqrt{2r/\nu}$. Both comparisons exclude the unchanged regular term $q/2$. For $r = 2$ and $\nu = 64, 256, 1024$, numerical minimization of the full proxy gives $t_* = 1.0692, 1.7380, 2.6365$; its large-$t$ approximation gives $1.4142, 2, 2.8284$.

These are predictions of the equilibrium proxy, not consequences of the fixed-t convergence assumption alone. Transferring its optimizer to the finite-ν reduced ensemble requires uniform moment estimates along $t = t_\nu \to \infty$ and control of minimizers. Transferring it to the original network additionally requires a uniform population approximation as both dimensions and s grow. Neither step supplies the optimal regularization of the noisy empirical learner or of a different prediction rule.

Numerical illustration. Figure 3 shows $\Lambda_{\nu,\nu}(s)/\nu^2$ for $\nu = 2, \ldots, 64$ by Metropolis sampling of Equation (89) (Appendix L.5): against s the curves separate with ν, as the window $s \asymp \sqrt{\nu}$ requires; against $t = s/\sqrt{\nu}$ they collapse, approach $F(t)$ from below with a correction of order $1/\nu$ (at $t = 1$ the longer runs extrapolate to $0.4351(2)$ against $47/108 = 0.43519$) and stay inside the bounds of Proposition I.1. This checks the reduced theory, not the full network (Section 7).

N COMPLETE EXPERIMENTAL RESULTS AND CALIBRATION

A positive-rank matrix learner. We test the global Gaussian-prior posterior with $d = 3$, $A_* = \mathrm{diag}(1, 0, 0)$, $\sigma = \gamma = 1$, and $\beta = 1/2$. The inactive block is genuinely matrix-valued: $h = 2$, $l = m - 1$. We use 1024 independent design/score replicates, paired across widths and regularization

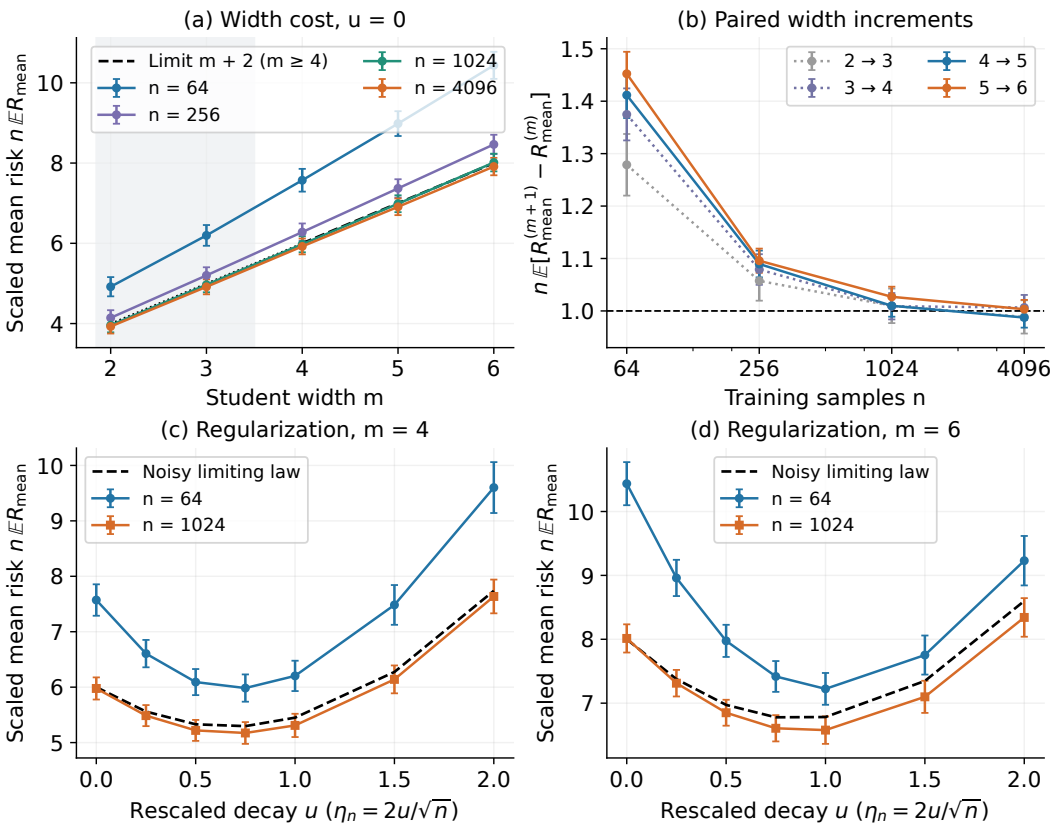

Figure 4: Positive-rank matrix prediction with a global Gaussian factor prior. (a) Total mean-predictor risk and (b) paired extra-neuron cost at zero added decay. The shaded widths in (a) and dotted width comparisons in (b) include cases outside Theorem 3.1; the dotted continuation of $m+2$ is exploratory. (c,d) Total risk under $\eta_n = 2u/\sqrt{n}$, including active shrinkage bias. Points use 1024 independent design/score replicates; bars are twice their standard errors. Dashed risk curves are independent quadratures of the noisy limiting law. Lines connecting the regularization grid do not identify a continuous optimizer.

values. Conditional Gaussian score sampling is exact for each input design; finite-sample posteriors retain the empirical sensing operator and fixed prior. Four-chain Hamiltonian Monte Carlo targets an exact reparameterization of the global factor posterior, without an oracle cutoff. Methods, numerical checks, and the separate assessment of posterior Monte Carlo error appear in Appendix O.

The cost of additional neurons. At zero added decay, Figure 4(a,b) compares widths $2, 3, 4, 5, 6$ at $n = 64, 256, 1024, 4096$. For $m \geq 4$, Theorem 3.1 predicts $n\mathbb{E}R_{\text{mean}} \to m + 2$ and a unit increment per additional neuron. At $n = 4096$, the paired increments for $4 \to 5$ and $5 \to 6$ are 0.987 ± 0.019 and 1.003 ± 0.018, respectively. Here and below, $\pm$ denotes twice the standard error across independent data replicates, including residual posterior Monte Carlo noise. At $n = 64$, the corresponding increments are 1.412 ± 0.044 and 1.452 ± 0.042: the asymptotic coefficient understates the width cost at this sample size. Widths $2, 3$ provide exploratory evidence outside the Gaussian-prior theorem, rather than an extension of its proof.

Regularization and prior dependence. For widths 4 and 6, we use the seven-point grid $u = 0, 0.25, 0.5, 0.75, 1, 1.5, 2$ at $n = 64, 1024$. At the common grid value $u = 0.5$ and $n = 1024$, paired reductions in scaled total mean-predictor risk are 0.757 ± 0.075 and 1.163 ± 0.076, respectively. The curves subsequently turn upward as active shrinkage becomes costly (Figure 4(c,d)). These observations support the mechanism in Corollary 3.2; they do not supply a tuning guarantee or a unique optimizer. The supporting experiment separates regular and inactive contributions and

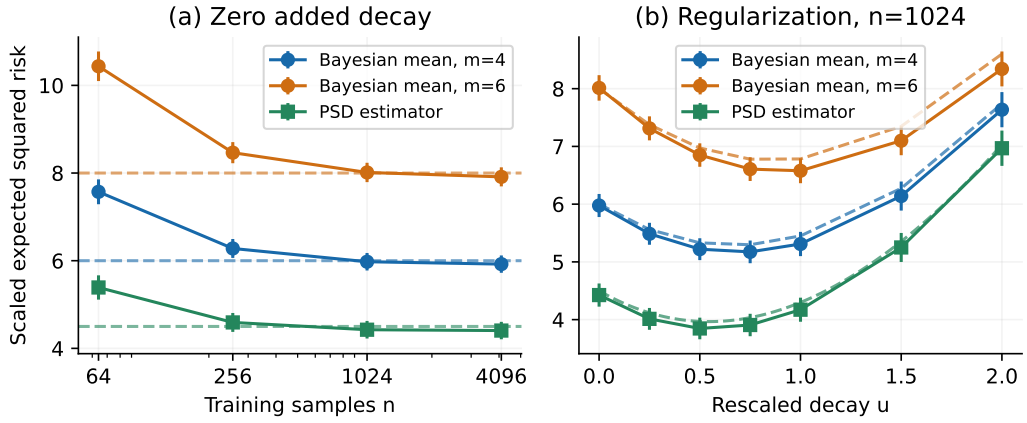

Figure 5: Bayesian posterior means and the PSD estimator on identical rank-one data replicates. (a) Zero added decay, with dashed asymptotic coefficients. (b) Total risk at $n = 1024$, with dashed limiting curves. The PSD estimator includes the fixed $1/n$ trace term. Bars are twice the standard error over 1024 independent replicates; Bayesian risks retain posterior Monte Carlo uncertainty.

compares Gibbs risk with the noisy and population references (Figure 9). Holding the represented-matrix prior fixed instead gives exactly the same posterior at each saturated width. Our control reuses this common posterior, explicitly separating the factor-prior width effect from representational capacity. Finite-sample corrections remain measurable at some $n = 1024$ settings; the limiting curve is not a finite-sample equality.

A PSD comparator on the same data. We compute Equation (18) with $c = 1$ on all 1024 existing data replicates, retaining the empirical sensing operator. At $u = 0$, its predicted scaled risk is 4.5, compared with 6 and 8 for the Bayesian means at widths 4 and 6. At $n = 4096$, the paired Bayesian-minus-PSD gaps are 1.516 ± 0.020 and 3.507 ± 0.047, against the predictions 1.5 and 3.5 (Figure 5). Thus the Bayesian reference with $\alpha = 0$ already has a nonzero comparison gap, and the additional determinant exponent accounts for the further width cost. Regularization improves both learners on this grid; the zero-decay theorem does not assert an ordering at every positive u. Appendix P gives the convex optimization checks and limiting quadrature, separate from posterior sampling error.

Matched inactive geometry at rank two. We next use $d = 4$, $r = 2$, $A_* = \mathrm{diag}(1, a, 0, 0)$ with $a = 1$ or 0.3, and widths 5, 7, retaining $\sigma = \gamma = 1$ and $\beta = 1/2$. There are 512 independent data replicates, coupled to the first 512 rank-one replicates. The pairs $(d, r, m) = (3, 1, 4)$ and $(4, 2, 5)$ share $h = 2, l = 3$; the widths 6, 7 likewise share $h = 2, l = 5$. The inactive limiting law is therefore the same in each pair. Since q increases from 3 to 7 and r from 1 to 2, the leading total mean-risk curves differ by $4 + u^2$, independently of fixed $a > 0$. We test $n = 64, 1024$ at $u = 0, 0.5, 1, 2$, and $n = 4096$ at $u = 0$.

Figure 6(a) compares the shared limiting curve with two finite-sample curves at $n = 1024$. To isolate finite-sample effects in the rank comparison, panel (b) subtracts the conditional limiting shift on the shared Gaussian score; its expectation is $4 + u^2$ (Equation (101)). At $n = 4096, u = 0, m = 5$, the remaining residual is -0.008 ± 0.029 for $a = 1$, but -0.084 ± 0.033 for $a = 0.3$. Changing only a gives a paired scaled-risk difference of -1.133 ± 0.228 at $n = 64$ and -0.076 ± 0.018 at $n = 4096$, at $u = 0, m = 5$. Finite positive-spectrum sensitivity is thus still detectable at our largest sample size. This supports distinguishing the fixed-spectrum asymptotic statement from a finite-sample equality or a limit uniform near loss of teacher rank. All settings and replicates are retained; Appendix P reports diagnostics, an alternative-coordinate sampler check, and the raw paired shifts.

Resolving a weak positive teacher mode. We evaluate the local law for $d = 3$, base rank $r = 1$, $m = 4$, $\sigma = 1$, $\beta = 1/2$, $u = 0$, and $\Theta_\perp = t\,\mathrm{diag}(1, 0)$. Both the zero- and strong-signal endpoints are 6. Noncentral two-by-two matrix quadrature gives $F_{tH}(0) = 5.396, 5.358, 5.721, 5.943$ at $t =$

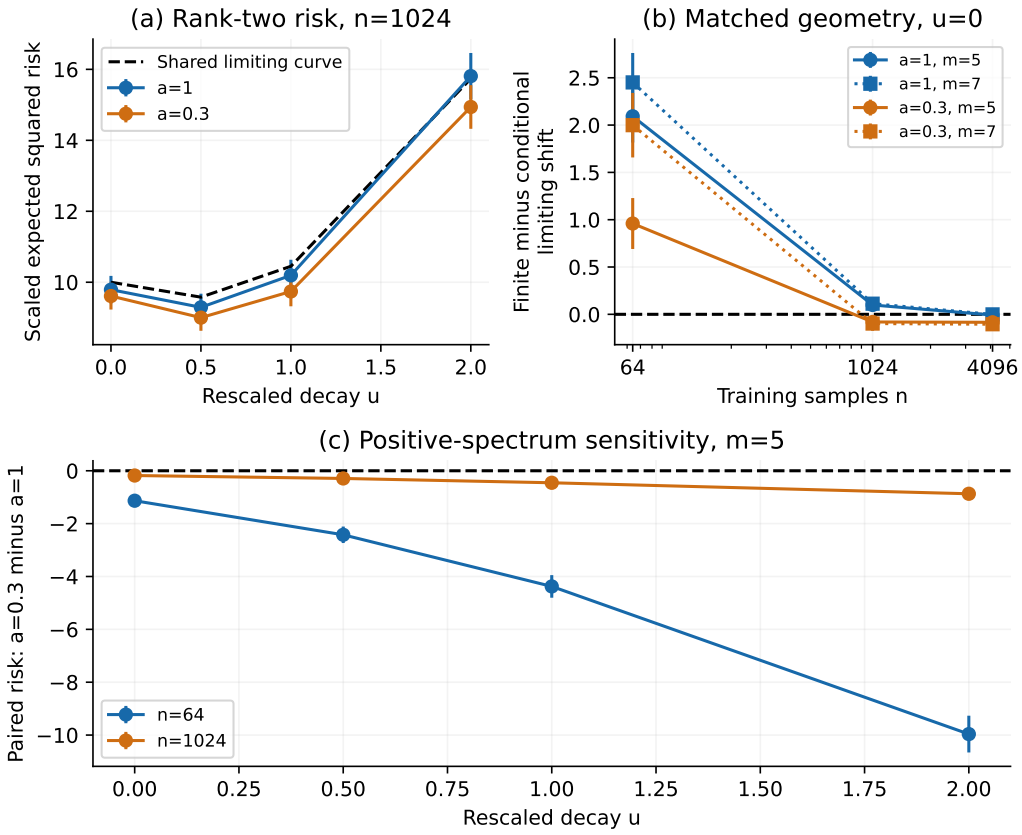

Figure 6: Rank-two robustness and finite-sample corrections, with 512 independent data replicates. (a) Mean-predictor risk at width 5, for two fixed positive teacher spectra sharing a limiting curve. (b) Paired rank-two-minus-rank-one risk, minus its conditional limiting shift; zero is the asymptotic reference. (c) Paired risk difference between $a = 0.3$ and $a = 1$, whose limit is zero. Bars are twice the across-data standard error. Panel (b)'s variance-reducing diagnostic was added during analysis; the archive also retains the raw shifts.

$1, 1.5, 4, 16$, respectively (Figure 7); at $t = 32$ it is 5.972. These values illustrate the strict inequality proved in Corollary 5.2. They are coefficients of the local-alternative limit, not measurements of finite-sample risk. The same calculation at widths $m = 5, 6$ gives $6.016, 6.677$ at $t = 2$, respectively. Their zero-signal endpoints are $7, 8$, and their strong-signal endpoints are $6.5, 7$. These examples numerically fall below the lower endpoint, but the general theorem asserts only improvement over the zero-signal value. The calculation retains the noncentral score distribution and the orientation needed to center the squared error at the signal. Order and cutoff refinement, a separate Cartesian-score integration, and the shifted risk identity check the calculation (Appendix Q). The existing rank-two residuals remain finite-sample diagnostics. A more direct observable for a future three-by-three limiting calculation is the paired spectrum difference -0.0765 ± 0.0183 at $n = 4096, u = 0, m = 5$ (twice the standard error). Its local-law comparison would be $F_{19.2H}(0) - F_{64H}(0)$ for $d = 4$, base rank one, and $H = \mathrm{diag}(1, 0, 0)$. This contrast cancels the reference learner used in the rank-comparison residual. We have not verified that three-by-three prediction; the two-by-two quadrature here does not evaluate it.

The noisy limiting law. For $h = l = 1$, $\sigma = 1$, and $\beta = 1/2$, Equation (7) is a one-dimensional integral conditional on $Z \sim \mathcal{N}(0, 1)$. Table 3 compares its two risk coefficients with the population Gibbs coefficient, obtained by setting $Z = 0$. At $u = 0$, the noisy Gibbs coefficient is 0.845947, rather than 0.5; the noisy mean coefficient is 0.5, as required by Equation (31), and matches the Bayesian endpoint in Proposition 3.3. The noisy mean and population *Gibbs* coefficients both equal

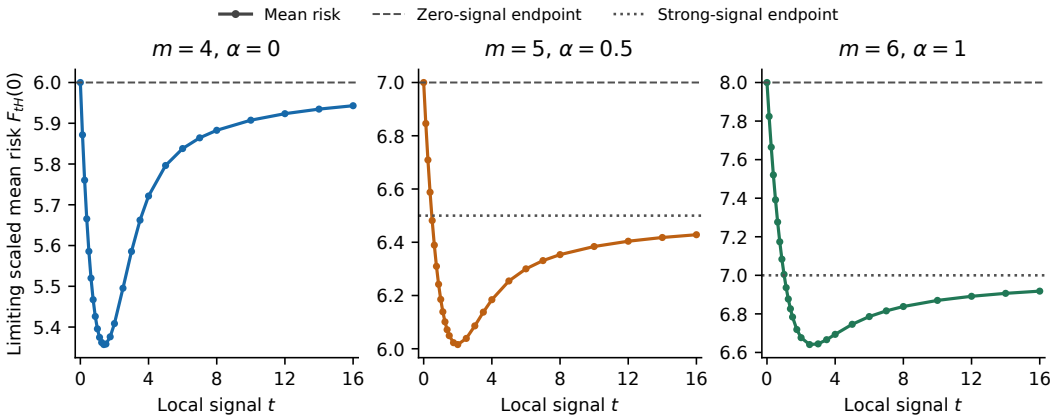

Figure 7: Weak teacher modes in the limiting experiment, with $d = 3$, base rank $r = 1$, widths $m = 4, 5, 6$, matched temperature, unit noise, and zero added decay. The weak eigenvalue is $a_n = t/\sqrt{n}$. The dashed line marks the zero-signal endpoint; the dotted line marks the strong-signal endpoint when different. At $m = 4$, every finite $t > 0$ lies below the common coefficient 6 analytically. At $m = 5, 6$, the theorem gives strict improvement over the zero endpoint; the plotted curves also numerically undercut the lower strong endpoint. Points are deterministic quadrature and lines join values. No finite-sample accuracy or general below-strong-endpoint claim is made.

u	Population Gibbs	Noisy Gibbs	Noisy mean
0	0.500000	0.845947	0.500000
0.5	0.317021	0.521794	0.276716
1	0.210398	0.327562	0.156428
2	0.105165	0.143340	0.058075
5	0.026106	0.028790	0.009986

Table 3: Risk coefficients in the scalar limiting models, in units of $1/n$, at $\sigma = 1$, $\beta = 1/2$. These are quadratures of limiting laws, not finite-training-set measurements.

$2\sigma^2(k/4)$ here; this coincidence does not equate the predictors, since the population mean coefficient is 0.228473. The Euler–Stein identity holds numerically to 10^{-13}; increasing the noise and factor quadrature orders from $(40, 192)$ to $(96, 512)$ changes the displayed coefficients by less than 10^{-9}.

A finite-sample scalar learner. We also evaluate actual finite-sample posterior integrals for $d = m = 1$, $r = 0$ and a fixed standard Gaussian factor prior. This calibration uses the global scalar Gibbs law, without an oracle cutoff. Put $z_i = (x_i^2 - 1)/\sqrt{2}$, $a_n = n^{-1}\sum_i z_i^2$, and $b_n = n^{-1}\sum_i z_i \epsilon_i$. Up to a constant independent of w, the regularised loss is exactly $a_n w^4 + 2(u/\sqrt{n} - b_n)w^2$. Conditional on the inputs, $\sqrt{n} b_n \sim \mathcal{N}(0, \sigma^2 a_n)$. Including the prior and setting $w = (n\beta a_n)^{-1/4} v$ therefore reduces the posterior to

$$p(v \mid a_n, \xi) \propto e^{-v^4 - 2 s_n v^2}, \qquad s_n = \frac{\sqrt{\beta}\{u + 1/(4\beta\sqrt{n})\}}{\sqrt{a_n}} - \sigma\sqrt{\beta}\,\xi, \quad \xi \sim \mathcal{N}(0, 1). \tag{95}$$

The scaled risks are $(\beta a_n)^{-1}\mathbb{E}[v^4 \mid a_n, \xi]$ and $(\beta a_n)^{-1}(\mathbb{E}[v^2 \mid a_n, \xi])^2$. We average over 4096 independent input designs at each n, and integrate the conditional Gaussian label noise by quadrature. Thus no Gaussian approximation to the finite-sample score is used in this calculation.

Figure 8 shows substantial finite-sample corrections at small n, particularly without weight decay. At $u = 0$, the scaled Gibbs risk decreases from 1.193 at $n = 16$ to 0.845 at $n = 1024$, close to the noisy limiting coefficient 0.846. The population coefficient 0.5 misses this effect. Corrections from empirical curvature and the fixed prior remain visible at other gaps. A resolution check on 64 of the $n = 16$ designs gives quadrature differences below 10^{-8}, smaller than the reported Monte Carlo errors. The script and numerical results are supplied in `scripts/scalar_risk/`. The Gaussian prior lies outside the compact-prior hypothesis of Proposition 3.3. The width $m = d = 1$ also lies

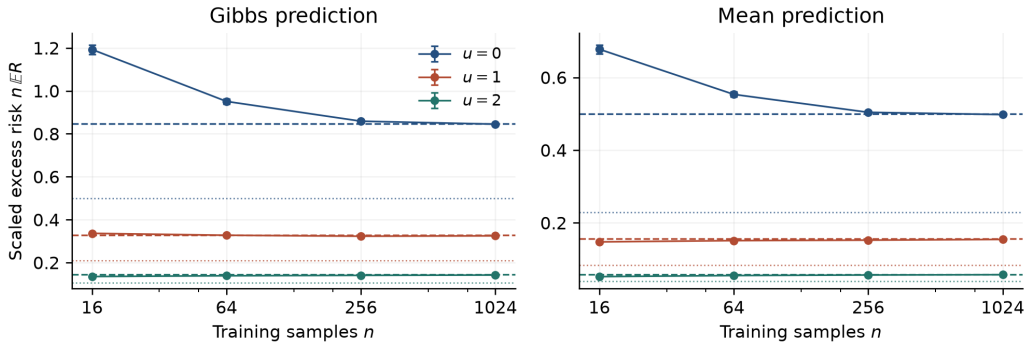

Figure 8: Finite-sample scalar risk with a standard Gaussian factor prior, $\sigma = 1$, $\beta = 1/2$, and $g_n = u/\sqrt{n}$. Points: 4096 input designs per sample size; bars: twice the Monte Carlo standard error over designs, after conditional integration of label noise. Dashed lines: noisy limiting coefficients for the corresponding prediction rule. Dotted lines: population limiting coefficients for that same rule.

outside Theorem 3.1. This scalar calibration isolates the distinction between prediction rules; the positive-rank matrix experiment above tests the setting of the Gaussian-prior theorem.

O Positive-rank matrix experiment and numerical checks

Design and exact data distribution. The experiment in Figures 4 and 9 uses the global learner in Theorem 3.1, with $d = 3$, $r = 1$, $A_* = \mathrm{diag}(1, 0, 0)$, $\sigma = \gamma = 1$, and $\beta = 1/2$. It uses 1024 independent replicates with seed 20260926, sample sizes $64, 256, 1024, 4096$, and widths $2, 3, 4, 5, 6$ at $u = 0$. The regularization grid is $0, 0.25, 0.5, 0.75, 1, 1.5, 2$ for widths $4, 6$ at $n = 64, 1024$. The grid and diagnostic criteria were fixed before the production run; a separate 16-replicate pilot was used to check the sampler.

Use Frobenius coordinates $(A_{11}, \sqrt{2}A_{12}, \sqrt{2}A_{13}, A_{22}, \sqrt{2}A_{23}, A_{33})$ and let ϕ_i be the coordinates of the sensing matrix X_i. We form $I_n = n^{-1}\sum_i \phi_i\phi_i^\top$ from actual Gaussian inputs. Conditional on these inputs, the score has exactly the distribution $Z_n = I_n^{1/2}G$, $G \sim \mathcal{N}(0, I_6)$. We sample it in this form rather than simulate each label. Widths and regularization values use identical sufficient statistics. Input prefixes and a common G pair the sample sizes; this coupling preserves each sample size's exact marginal data law but does not define a nested sequence of labels. The limiting-law calculation uses the same G, allowing paired finite-minus-limit comparisons without replacing the finite-sample sensing operator by its limit.

An exact global sampling parameterization. Let $P = A_{11} > 0$, $C = A_{1,2:3}$, and write the Schur complement as $A_{2:3,2:3} - C^\top C/P = DD^\top$, $D \in \mathbb{R}^{2\times(m-1)}$. Integrating the orientation of the first row of W gives, up to a constant, the factor volume element

$$P^{(m-4)/2}\, dP\, dC\, dD. \tag{96}$$

Indeed, the radial first-row measure contributes $P^{m/2-1}dP$, and expressing each of the two remaining rows' parallel components through C contributes $P^{-1/2}$. Set $P = e^{s/\sqrt{n}}$, $C = V/\sqrt{2n}$, and $D = n^{-1/4}F$. Then the represented scaled error is exactly

$$B = \sqrt{n}(A - A_*) = \begin{pmatrix} \sqrt{n}(P-1) & V/\sqrt{2} \\ V^\top/\sqrt{2} & FF^\top + V^\top V/(2\sqrt{n}P) \end{pmatrix}. \tag{97}$$

Using the Euclidean inner product of Frobenius coordinates, its negative log density in (s, V, F), up to an additive constant, is

$$\frac{1}{2}\langle B, I_n B\rangle - \left\langle Z_n - \left(u + \frac{1}{2\sqrt{n}}\right) I_3, B \right\rangle - \frac{m-2}{2\sqrt{n}}s. \tag{98}$$

The $1/(2\sqrt{n})$ term is the fixed Gaussian factor prior; the last term is the remaining Jacobian. The Schur correction is retained, and P ranges over all positive values. Thus this is a global finite-sample posterior, not a local asymptotic substitute. At width 2, the rectangular $F \in \mathbb{R}^{2\times 1}$ retains the rank constraint.

Posterior computation and risk estimation. Each of the 62464 finite or limiting posterior jobs uses four independently seeded Hamiltonian Monte Carlo chains, with 512 warmup and 1024 retained iterations per chain. Dual averaging targets acceptance 0.85 during warmup; production uses half the adapted step, randomized integration times, and Metropolis correction. For chain means $\widehat{b}_c$ of the six represented coordinates of B, we estimate scaled mean-predictor risk by

$$\widehat{R}^{\mathrm{scaled}}_{\mathrm{mean}} = \frac{1}{4\cdot 3}\sum_{c\neq c'}\langle \widehat{b}_c, \widehat{b}_{c'}\rangle. \tag{99}$$

For independent stationary chains this removes the positive variance bias of squaring one estimated posterior mean. It does not remove finite-run sampler bias. Gibbs risk uses posterior averages of $\|B\|_F^2$; regular and inactive contributions use the corresponding coordinate blocks. Across-data standard errors use the 1024 independent replicates. Conditional posterior Monte Carlo variance is separately estimated by a jackknife over the four chain means. The resulting posterior Monte Carlo standard error of the data-averaged mean risk is approximately 0.002–0.004 across the finite-sample cases.

Numerical review and difficult cases. We diagnose represented matrix entries, total and inactive squared norms, and inactive trace using rank-normalized folded split $\widehat{R}$ and bulk effective sample size (Vehtari et al., 2021). The criteria are $\widehat{R} \le 1.01$, effective sample size at least 400, and no retained trajectory with nonfinite energy error or an energy increase exceeding 1000. The initial run had 51 jobs failing at least one criterion. All were repeated with 2048 warmup and 8192 retained iterations per chain and one quarter of the original production step factor. No datasets were dropped. The final maximum $\widehat{R}$ is 1.00997, the minimum bulk effective sample size exceeds 765, and no retained trajectories trigger the energy-error criterion.

A preselected subset of the first 32 data replicates and eight cases was checked with longer chains, smaller steps, and two alternative coordinate systems: the original factor W, and Schur-complement Cholesky coordinates when $m \ge 3$. Their Jacobians are included. After repeating one flagged benchmark job, all benchmark diagnostics pass. Averaging paired mean-risk differences within each dataset before estimating uncertainty, the long Schur, original-factor, and Cholesky estimates minus the production estimates are, respectively, 0.00145 ± 0.01102, -0.00307 ± 0.01242, and 0.00109 ± 0.01340 (twice standard errors across 32 independent dataset clusters). Finite-difference gradient checks have maximum relative discrepancy below 4×10^{-9}; checks against the original factor target agree to 3×10^{-14}. These checks assess computation; they are not a proof of stationarity or a bound on residual numerical bias.

Independent matrix-valued limiting quadrature. For $h = 2$, $\beta = 1/2$, write the eigenvalues of $M = FF^\top$ as $\lambda \ge \mu \ge 0$, and set $\alpha = (m-4)/2$. For a standard isotropic Gaussian $Z \in \mathrm{Sym}_2$, its trace is $\sqrt{2}s$ and its eigenvalue gap is $2\sqrt{t}$, with independent $s \sim \mathcal{N}(0,1)$ and $t \sim \mathrm{Exp}(1)$. Integrating the relative eigenspace angle gives the conditional eigenvalue weight

$$(\lambda\mu)^\alpha(\lambda-\mu)e^{-(\lambda^2+\mu^2)/2+(s/\sqrt{2}-u)(\lambda+\mu)}I_0\big(\sqrt{t}(\lambda-\mu)\big)\,d\lambda\,d\mu, \tag{100}$$

where I_0 is the modified Bessel function. Each conditional integral is normalized *before* averaging over (s,t). If $\overline{T}$ is its mean trace and $\overline{D}$ its mean of $(\lambda-\mu)I_1(\sqrt{t}(\lambda-\mu))/I_0(\sqrt{t}(\lambda-\mu))$, then the conditional squared norm of the posterior matrix mean is $(\overline{T}^2+\overline{D}^2)/2$. The conditional Gibbs coefficient is the mean of $\lambda^2+\mu^2$. Adding $3+u^2$ or $6+u^2$ gives the total noisy mean or Gibbs coefficient, respectively. Setting the score to zero instead gives the population reference, whose regular mean and Gibbs contributions are u^2 and $3+u^2$.

We use Gauss–Hermite and Gauss–Laguerre quadrature for (s,t) and Gauss–Legendre quadrature for $(\lambda-\mu,\mu)$. Increasing the outer/inner orders from $(20,64)$ to $(32,96)$ changes the reported coefficients by less than 10^{-10}; the noisy Euler–Stein residual is also below 10^{-10}. The finite

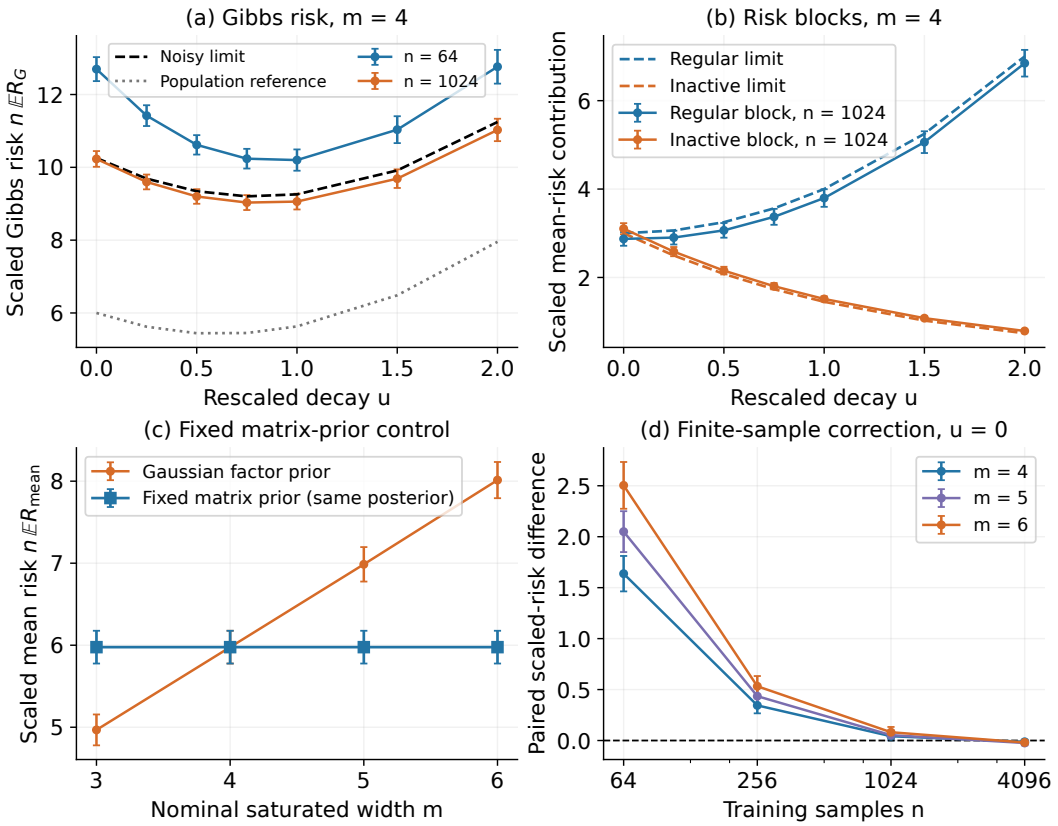

Figure 9: Supporting matrix checks. (a) Gibbs risk with noisy and population reference curves. (b) Regular and inactive contributions to mean-predictor risk; the regular limit $3+u^2$ includes sampling error and active shrinkage bias. (c) Fixing the represented matrix prior gives one common posterior, reused at each nominal saturated width; the repeated control points are not independent fits. (d) Paired finite-minus-limiting scaled mean risks share the Gaussian score variable. Bars denote twice the standard error across the 1024 independent design/score replicates.

integration range and quadrature rule are recorded in the script; these are resolution checks, not rigorous error bounds. Conditional inactive mean-risk comparisons against the sampler over the first 32 score replicates, widths $4, 6$, and the seven u values give an average difference 0.00108 ± 0.00353, with uncertainty clustered by independent score replicate.

A fixed represented-prior control. For each nominal width $m \geq 3$, set the prior on positive definite A to the one induced by the width-4 Gaussian factor model: $\rho(A) \propto e^{-\operatorname{tr} A/2}$. With identical likelihood and trace penalty, the posterior is exactly identical across these widths. Accordingly, Figure 9(c) reuses the width-4 posterior at $n = 1024, u = 0$. Width 2 cannot support this full-rank prior and is excluded. This control changes the factor-prior family across widths; it is not a counterexample to the width law under the stipulated factor priors.

Scope and remaining validation. The experiment directly tests a positive-rank teacher with matrix-valued inactive fluctuations, including total risk, width costs, regularization, prior dependence, and visible finite-sample corrections. Widths $2, 3$ remain exploratory under the Gaussian prior. The experiment does not prove the general local-prior conjecture, a growing-dimension limit, a finite-sample remainder, or a data-dependent tuning rule. A separate local-learner experiment would need to implement its stated smooth compact cutoff; the present global posterior cannot stand in for it. The earlier scalar and population spectral experiments retain their respective roles. Scripts, sufficient statistics, per-dataset summaries, and numerical checks for this matrix study are provided in `scripts/matrix_risk/`.

P PSD COMPARISON AND RANK-TWO VALIDATION

The empirical PSD estimator. The comparison in Figure 5 reuses all 1024 rank-one sufficient-statistic replicates from Appendix O. We set $c = 1 = \gamma/(2\beta)$ and solve Equation (18) at $u = 0$, $n = 64, 256, 1024, 4096$, and at the six positive regularization-grid values of the earlier experiment for $n = 64, 1024$. In Frobenius coordinates, let $b = \sqrt{n}(A - A_*)$ and $v = Z_n - (u + 1/(2\sqrt{n}))I_d$. The objective, up to scaling and an additive constant, is $\frac{1}{2}\langle b, \mathcal{I}_n b\rangle - \langle v, b\rangle$, with $b + \sqrt{n}A_* \succeq 0$. Projected gradient uses step $2/(L + \mu)$, where μ, L are the smallest and largest eigenvalues of the empirical operator. If its update map is T, then T has contraction factor $\rho = (L - \mu)/(L + \mu) < 1$. Consequently,

$$\|b - \widehat{b}\| \le e := \frac{\|b - T(b)\|}{1 - \rho}, \qquad \big|\|b\|^2 - \|\widehat{b}\|^2\big| \le e(2\|b\| + e).$$

Every solve reaches a computed scaled-risk error bound below $2 \cdot 10^{-8}$. This is a contraction bound evaluated in floating point, not interval arithmetic. On a fixed first-eight-replicate subset at four settings, independent optimization in full-factor coordinates gives objective discrepancies below $3 \cdot 10^{-15}$. An identity-curvature check on 32 random scores agrees with the exact positive-part formula to 10^{-10}. The independent factor optimizer is a numerical check; the convex solver supplies the primary error bound.

For $h = 2, \sigma = 1$, the inactive Gaussian eigenvalues are $s/\sqrt{2} \pm \sqrt{t}$, where $s \sim \mathcal{N}(0, 1)$ and t is an independent unit-rate exponential. We integrate their positive squared parts analytically over s and by one-dimensional adaptive quadrature over t, with absolute and relative tolerance 10^{-11}. Adding $3 + u^2$ gives the dashed PSD curve. Its value at zero agrees with 4.5 to 10^{-10}. Risk differences against the Bayesian means are paired within each dataset; posterior Monte Carlo error in those means remains in the across-data uncertainty.

Rank-two design and coupling. The additional experiment has $d = 4$, $r = 2$, teacher $\mathrm{diag}(1, a, 0, 0)$, $a \in \{1, 0.3\}$, widths $5, 7$, and $\sigma = \gamma = 1, \beta = 1/2$. It uses 512 independent design/score replicates and 36 settings: $n = 64, 1024$ at $u = 0, 0.5, 1, 2$, and $n = 4096$ at $u = 0$, for each spectrum and width. We reconstruct the earlier rank-one inputs from their recorded seed, insert a new independent Gaussian coordinate as the second active coordinate, and form the actual 10×10 empirical sensing operator. In row-wise upper-triangular Frobenius coordinates, the common standard Gaussian score G shares the six coordinates indexed by the first, third, and fourth input axes with the old rank-one score. Its four other coordinates are independent standard Gaussians. For each sample size, $Z_n = \mathcal{I}_n^{1/2} G$ has the exact conditional score law. Spectrum, width, and decay comparisons share these sufficient statistics. As in the earlier study, this coupling preserves the marginal data laws without claiming a nested sequence of labels.

The two rank settings have identical inactive scores and identical h, l in the limit. If g_1, g_2, g_3, g_4 are the four extra tangent Frobenius coordinates, with g_2 the new active diagonal, the conditional limiting risk difference is

$$\Delta(G, u) = g_1^2 + (g_2 - u)^2 + g_3^2 + g_4^2, \qquad \mathbb{E}\Delta(G, u) = 4 + u^2. \tag{101}$$

Subtracting Δ from each paired finite-sample risk difference cancels the leading shared-score fluctuations. This diagnostic was added during analysis, after production settings and replicates were fixed. We retain raw differences as well: at $n = 4096, u = 0, m = 5$, rank-two-minus-rank-one is 3.771 ± 0.237 for $a = 1$ and 3.695 ± 0.228 for $a = 0.3$, against the limiting value 4. At width 7, these values are 3.778 ± 0.237 and 3.674 ± 0.225. The smaller-spectrum discrepancies are not within two standard errors of the limit at this sample size. No settings or replicates are removed on that basis.

An exact global Cholesky parameterization. For the saturated widths here the represented prior has density $(\det A)^{(m-d-1)/2} e^{-\mathrm{tr}\, A/2}$. Write $A = LL^\top$ with lower-triangular L and positive diagonal. For teacher active eigenvalues $a_1 = 1, a_2 = a$, use real coordinates x_{ij} defined by

$$L_{ii} = \sqrt{a_i} \exp\left(\frac{x_{ii}}{2a_i\sqrt{n}}\right) \ (i \le r), \quad L_{ij} = \frac{x_{ij}}{\sqrt{2na_j}} \ (i > j,\ j \le r),$$

$$L_{ii} = n^{-1/4} e^{x_{ii}} \ (i > r), \qquad L_{ij} = n^{-1/4} x_{ij} \ (i > j > r).$$

The Cholesky Jacobian, determinant factor, and logarithmic diagonal change of variables together give $\prod_{i=1}^{d} L_{ii}^{m+1-i}$, up to constants. Thus the exact negative log density, with $B = \sqrt{n}(LL^\top - A_*)$, is

$$\tfrac{1}{2} \langle B, \mathcal{I}_n B \rangle - \big\langle Z_n - (u + 1/(2\sqrt{n}))I_d, \, B \big\rangle - \sum_{i \le r} \frac{(m+1-i)x_{ii}}{2a_i\sqrt{n}} - \sum_{i>r} (m+1-i)x_{ii}.$$

Every positive definite matrix is included. These coordinates use the known synthetic teacher to improve numerical conditioning; the target posterior itself requires no knowledge of the teacher.

Sampling protocol and checks. The production settings and thresholds were fixed before the run. A separate 16-replicate pilot checked difficult settings. Each production job uses four independent HMC chains, 512 warmup and 1024 retained iterations per chain, target acceptance 0.85, and half the adapted step size. Risk estimation and diagnostics follow Appendix O: cross-chain products for posterior-mean risk, folded rank-normalized split $\widehat{R} \le 1.01$, bulk ESS at least 400, and no retained nonfinite or greater-than-1000 energy errors. All 72 flagged jobs among 18432 were rerun with 2048 warmup, 8192 retained iterations, and one eighth of the adapted step. None remained flagged: maximum $\widehat{R} = 1.00998$, minimum ESS 402.7, and zero retained divergences. Numerical gradients were checked on 72 cases, with maximum relative discrepancy $1.8 \cdot 10^{-9}$; nine change-of-variable checks had discrepancy below $1.6 \cdot 10^{-9}$.

An alternative sampler uses the original full factors $W \in \mathbb{R}^{4\times m}$, without the represented-density Jacobian. It checks the first 32 replicates at eight preselected settings with $a = 0.3$, widths $5, 7$, and $(n, u) = (64, 0), (64, 2), (1024, 0.5), (1024, 2)$. Its chains use 1024 warmup and 4096 retained iterations, and one quarter of the adapted step. All 256 jobs pass the same diagnostics. Averaging differences within each dataset before computing uncertainty, its scaled mean-risk difference from the Cholesky sampler is -0.006 ± 0.014. This checks the target and parameterization but does not certify absence of finite-run sampling bias or establish the mathematical theorem.

The code, run settings, sufficient statistics, per-dataset chain and diagnostic summaries, pre/post-refinement records, checks, raw paired comparisons, and figure analysis are archived in `scripts/v005_validation/`. The archive records seeds and software versions, the change of the independent optimizer after its first implementation failed to converge, and the subsequent conditional-shift diagnostic. Full diagnostic traces are retained locally with hashes; they are not required to reproduce figures from the portable summaries.

Q Noncentral matrix quadrature for weak modes

The weak-mode figure evaluates Equation (22) with $d = 3$, base rank $r = 1$, $m \in \{4, 5, 6\}$, $\sigma = 1$, $\beta = 1/2$, $u = 0$, and $\Theta_\perp = t\,\mathrm{diag}(1, 0)$. It uses deterministic integration of the limiting law, independently of the finite-sample posterior samplers. Here $h = 2$, $l = m - 1$, and $\alpha = (m-4)/2$. The zero-signal coefficient is $m + 2$ and the strong-signal coefficient is $6 + (m-4)/2$; both equal 6 at $m = 4$.

Write $Y = Z + \Theta_\perp$, $s = \mathrm{tr}\, Y/\sqrt{2}$, and let ρ be half the eigenvalue gap of Y. The trace coordinate is $s \sim \mathcal{N}(t/\sqrt{2}, 1)$. The two traceless Frobenius coordinates are independent unit Gaussians with means $(t/\sqrt{2}, 0)$, independently of s. Polar integration therefore gives

$$f_t(\rho) = 2\rho e^{-\rho^2 - t^2/4} I_0(t\rho), \qquad \rho > 0, \tag{102}$$

where I_j are modified Bessel functions. Conditional on ρ, the mean cosine of the traceless orientation toward the signal is $I_1(t\rho)/I_0(t\rho)$. Scaled Bessel functions evaluate these expressions without exponential overflow.

For matrix eigenvalues $b + a, b$ with $a, b > 0$, the conditional integration weight is

$$a\{b(b+a)\}^\alpha e^{-\{(b+a)^2 + b^2\}/2 + (s/\sqrt{2} - u)(a+2b)} I_0(\rho a).$$

Let τ and δ be the trace and eigenvalue gap of the conditional posterior mean. Its squared norm is $(\tau^2 + \delta^2)/2$, and its first diagonal entry, averaged over the score orientation at fixed s, ρ, is

$$\frac{1}{2}\left\{\tau + \delta \frac{I_1(t\rho)}{I_0(t\rho)}\right\}.$$

Thus the squared error includes the cross term with the signal, in addition to the squared norm and t^2. Each conditional matrix integral is normalized before the outer score average. Reusing the central score-gap density would omit the noncentral signal and give the wrong risk.

For the additional widths, substituting $a = x^2, b = y^2$ in the conditional matrix integral, with Jacobian $4xy$, removes the square-root boundary factor at $\alpha = 1/2$. Direct Gauss–Legendre integration in a, b converges more slowly at this exponent. The change of variables preserves the normalized law; we also check the linear coordinates at $m = 6, t = 2$.

The main quadrature uses 40 Gauss–Hermite trace nodes, 72 Gauss–Legendre radial nodes, and 96 nodes in each positive eigenvalue coordinate. We retain a coarser rule and refine six signal values, including $t = 32$, to orders 56, 96, and 128, while increasing the tail cutoffs. For the original $m = 4$ calculation, the maximal fine-to-refined risk difference is below 2×10^{-10}. A separate integration over the three Cartesian Gaussian score coordinates agrees within 10^{-9} at $t = 1, 4$; it shares the conditional matrix integral but does not use the noncentral polar reduction. The direct squared-error calculation and Equation (24) differ by less than 2×10^{-12} on the plotted rule, and its score-mass error is below 10^{-12}. Additional checks use $m = 6, u = 0.5$ and the scalar truncated-normal posterior mean. For $m = 5, 6$, we repeat the order/cutoff and Cartesian-score checks with the square coordinates; the numerical discrepancies are recorded in `scripts/weak_modes/results/v007/verification.json`. The fine-to-refined differences are below 2×10^{-10}, and Cartesian checks agree within 10^{-11}. The deliberately coarse rule loses about 1.4×10^{-3} at $t = 32$; this value is retained as a failed resolution check, and the accepted fine rule agrees with its refinement at that signal. On the plotted interval through $t = 16$, coarse-to-fine differences are below 2×10^{-9}. The positive-width risk-identity residuals are below 4×10^{-12}. Scalar $\alpha = 2$ risk identities and selected matrix mean shifts provide additional numerical self-checks of the analytical conclusions. These are floating-point convergence checks, not certified integration-error bounds. The calculation has no sampling confidence intervals. All rules, results, checks, and figure generation are in `scripts/weak_modes/`.

The rank-two residual in Figure 6(b) is a different observable. Write $\widehat{C}_n(a)$ for its per-dataset scaled rank-two risk, $\widehat{B}_n$ for the coupled reference risk, and $\Delta(G, 0)$ for the conditional limiting shift of expectation 4. The residual is $D_n(a) = \widehat{C}_n(a) - \widehat{B}_n - \Delta(G, 0)$, so

$$10 + \mathbb{E}D_n(a) = \mathbb{E}\widehat{C}_n(a) + (6 - \mathbb{E}\widehat{B}_n).$$

Thus adding 10 to the residual replaces the reference learner's finite-sample expectation by its limit. In contrast, $D_n(0.3) - D_n(1) = \widehat{C}_n(0.3) - \widehat{C}_n(1)$ exactly for each paired dataset. The archived mean of this contrast is -0.076473, with standard error 0.009143 across 512 replicates. A reproducible three-by-three calculation could compare it with $F_{19.2H}(0) - F_{64H}(0)$ at $d = 4, m = 5$, base rank one, $H = \mathrm{diag}(1, 0, 0)$, matched temperature and zero decay. We report no unverified prediction for that contrast. Even after such a calculation, substituting $t = \sqrt{n}\, a$ for fixed a would supply an approximation without a finite-sample remainder. The present figure uses a two-by-two inactive block.

R Consequences for finite-temperature complexity measurements

An estimated local learning coefficient is an expectation under a localized posterior at one finite τ, with a Gaussian localizer in place of a cutoff and, at a trained checkpoint, a loss that need not be stationary there; this section separates the crossover it sees from that of Theorem H.1.

Lemma R.1 (quadratic soft count)**.** *Let $L(\delta) = \frac{1}{2}\delta^\top H \delta$ with $H \succeq 0$ having eigenvalues c_i, let $\gamma > 0$, and let the localized posterior be $\propto \exp(-\tau L(\delta) - \gamma \|\delta\|^2/2)$. Then*

$$\hat{\lambda}_{\mathrm{quad}} := \tau\, \mathbb{E}[L] = \frac{1}{2} \sum_i \frac{\tau c_i}{\tau c_i + \gamma}. \tag{103}$$

If instead $L(\delta) = b^\top \delta + \frac{1}{2}\delta^\top H \delta$ with a nonzero gradient b at the centre of localization, then with $A_\tau = \tau H + \gamma I$,

$$\hat{\lambda}_{\mathrm{quad}} = \frac{\tau}{2} \mathrm{tr}\left(H A_\tau^{-1}\right) - \frac{\tau^2}{2}\, b^\top A_\tau^{-1}\big(I + \gamma A_\tau^{-1}\big) b. \tag{104}$$

A direction is invisible while $\tau c_i \ll \gamma$ and counts $\frac{1}{2}$ once $\tau c_i \gg \gamma$. This *localization crossover* has endpoints 0 and $\frac{1}{2}$. It is not the crossover of Theorem H.1, whose endpoints are $\frac{1}{4}$ and $\frac{1}{2}$ and whose control parameter is a gap inside the energy. A quartic direction with a fixed Gaussian localizer has an effective scaled gap of order $\gamma/\sqrt{\tau} \to 0$ (Remark K.2), and its data-loss count approaches $\frac{1}{4}$ from below as τ grows. That an estimate at one temperature can miss directions that a larger inverse temperature resolves is not new in itself: Chen & Murfet (2025) already treat inverse temperature as a measurement resolution for the modes of a sequence model; what Theorem H.1 and Corollary H.3 add is the specific form of the dependence for a redundant factor block, and the fact that the regularised energy and the data loss cross over in opposite directions under the same sampling law.

Three cautions govern the use of Equation (103) at trained checkpoints. First, a checkpoint that is stationary for a regularised objective has a nonzero gradient b of the data loss, and the second term of Equation (104) then alters the count materially: it is not monotone in τ and can be negative (for $H = 1$, $\gamma = 1$ it equals $\tau/(2\tau+2) - \tau^2 b^2(\tau+2)/(2(\tau+1)^2)$) because the posterior centre moves off the checkpoint, so "both crossovers move upward with τ" holds only at a stationary centre; for squared loss the Hessian also carries a residual-weighted term beyond the Jacobian Gram matrix. Second, an algebraic curvature after eliminating active parameters is not the covariance of a posterior with a finite isotropic localizer. Third, a Gaussian localizer does not confine the low-temperature posterior to one basin, whereas Theorem H.1 uses a hard cutoff. Since a zero Jacobian does not imply a zero coefficient (u^4), the informative diagnostic is a curvature spectrum with a temperature sweep.

One-hot modular inputs. In the modular-arithmetic design of Cullen et al. (2026) the observable geometry differs from the ambient one: each neuron carries at most $3p-2$ first-order directions, the ambient cap of the secant count must be replaced by the observable cap $p(N-1)/2$ on the Jacobian-rank count, and an isolated inactive neuron, the trainable-readout analogue of D, has coefficient $(2p-1)/4$, $p/2$ or $p-1$ by stratum (Appendix T); this corrects a count but is not an application of Theorem H.1.

S Proofs for Appendix R

S.1 Proof of Lemma R.1

With $\gamma > 0$ the localized posterior is Gaussian with precision $A_\tau = \tau H + \gamma I \succ 0$ and mean $\mu = -\tau A_\tau^{-1} b$. Hence $\mathbb{E}[L] = b^\top \mu + \frac{1}{2}\operatorname{tr}(H A_\tau^{-1}) + \frac{1}{2}\mu^\top H \mu$, and with $H = (A_\tau - \gamma I)/\tau$, $\mu^\top H \mu = \tau\, b^\top A_\tau^{-1} b - \tau\gamma\, b^\top A_\tau^{-2} b$, so $\tau\mathbb{E}[L] = \frac{\tau}{2}\operatorname{tr}(H A_\tau^{-1}) - \frac{\tau^2}{2} b^\top A_\tau^{-1}(I + \gamma A_\tau^{-1}) b$. For $b = 0$ this is $\frac{1}{2}\sum_i \tau c_i/(\tau c_i + \gamma)$. □

T Case study: observable geometry of one-hot modular inputs

This appendix collects the statements summarised in Appendix R about the modular-arithmetic design of Cullen et al. (2026). It is an independent extension: it changes the input distribution, the number of outputs, the trainability of the readout and the observable space of the loss, and it does not produce a prediction of Theorem H.1 for a trained network.

T.1 Shift symmetry and the observable rank

Modular addition in $\mathbb{Z}_p$ uses one-hot pairs $x_{ab} = (e_a, e_b) \in \mathbb{R}^{2p}$, so $d = 2p$, with p outputs and a training set of $N \le p^2$ pairs. Let $P_{\mathrm{c}} = I_N - \mathbf{1}\mathbf{1}^\top/N$ centre each output over the training examples, and let $\Psi_D(\theta) = (P_{\mathrm{c}} f_{\theta,k}(x_i))_{k,i}$ be the centered prediction map. Two facts about the support drive everything below.

Lemma T.1 (shift symmetry and rank bound)**.** *Let $z = (\mathbf{1}_p, -\mathbf{1}_p)$. Every input satisfies $z^\top x_{ab} = 0$, so $w_j \mapsto w_j + c_j z$ leaves every prediction unchanged, and the span of the training inputs has dimension at most $2p-1$, with equality on the full grid. Together with the scale symmetry $(w_j, v_j) \mapsto (cw_j, c^{-2}v_j)$ of every nonzero neuron,*

$$\operatorname{rank} D\Psi_D(\theta) \;\le\; \min\{\, m(3p-2),\; p(N-1)\,\}. \tag{105}$$

On the full grid, unrestricted quadratic forms produce every $p \times p$ table through their cross block, so evaluation has rank p^2 ($p^2 - 1$ after centring) on $\mathrm{Sym}(\mathbb{R}^{2p})$, whose dimension is $p(2p+1)$.

The generic-position count of Cullen et al. (2026, Cor. 4.2), $\lambda = m(d+p-1)/2$ capped at $p\,d(d+1)/4$, is a statement about the ambient coefficient space $\mathrm{Sym}(\mathbb{R}^{2p})^p$. On the one-hot support the per-neuron count is at most $3p-2$, not $3p-1$, and the ambient cap must be replaced by the observable cap $p(N-1)/2 \le p(p^2-1)/2$, roughly half the ambient cap on the full grid and far below it on a training subset. Both caps bound the *Jacobian-rank* count: at an interpolator near which the Jacobian has constant rank the local coefficient is half that rank (constant rank theorem), whereas at a singular interpolator higher-order residuals can contribute beyond it, and no bound in terms of the number of residual coordinates alone can hold there ($K(x) = \|x\|^4$ on $\mathbb{R}^D$ has one residual coordinate and coefficient $D/4$). Numerically, generic parameters at $(p,m) = (3,1),(3,2),(5,1),(5,2),(5,4)$ give ranks $7, 14, 13, 26, 52$ both before and after centring, i.e. exactly $m(3p-2)$, so the bound is attained at generic parameters in these cases; centring starts to bind only near saturation (at $(p,m) = (2,2)$ the uncentered rank is 8 and the centered rank 6). None of this licenses extrapolation to the structured Fourier solutions found by training; it fixes the count that any such solution is measured against.

T.2 An isolated inactive neuron

An inactive neuron is the trainable-readout analogue of the redundant block of Appendix G: its contribution to the prediction is linear in the readout and quadratic in the hidden weight, so at zero readout the hidden weight is invisible to first order, as the block D is invisible in $WW^\top$, and the shift symmetry of Lemma T.1 plays the role of the orbit. Consider the full grid and a single p-output neuron whose contribution is $v \otimes q(w)$, $q(w)_{ab} = (\alpha_a + \beta_b)^2$ for $w = (\alpha, \beta) \in \mathbb{R}^p \times \mathbb{R}^p$, with the rest of the prediction held fixed. Write $Q(w) = P_{\mathrm{c}}\, q(w)$, $H(w) = \|Q(w)\|^2$, so that the squared residual of the neuron is $K(v,w) = \|v\|^2 H(w)$.

Proposition T.2 (inactive strata)**.** *Let $p \ge 2$ and let the local density be smooth and positive. With $z = (\mathbf{1}_p, -\mathbf{1}_p)$:*

(a) at (v_0, w_0) with $v_0 \neq 0$ and $w_0 \in \mathbb{R}z$ (in particular at $(v_0, 0)$), and at $(0, w_0)$ with $w_0 \in \mathbb{R}z$ (in particular at $(0,0)$), the coefficient is $\lambda = (2p-1)/4$ with multiplicity 1;
(b) at $(0, w_0)$ with $Q(w_0) \neq 0$, $\lambda = p/2$ with multiplicity 1;
(c) at (v_0, w_0) with $v_0 \neq 0$, $Q(w_0) = 0$ and $w_0 \notin \mathbb{R}z$, $\lambda = p-1$ with multiplicity 1;
(d) at $(0, w_0)$ with $Q(w_0) = 0$ and $w_0 \notin \mathbb{R}z$, $\lambda = \min\{p/2, p-1\}$, with multiplicity 2 if $p = 2$ and 1 if $p > 2$.

The proof (below) quotients the exact shift direction z; the zero set of the quartic H then becomes a finite union of lines through the origin (Lemma T.3), along which H vanishes quadratically in the $2p-2$ transverse directions. At the origin the zeta function factorises into a radial pole at $(2p-1)/4$ and an angular pole at $p-1$; since $(2p-1)/4 < p-1$ for $p \ge 2$ the radial pole wins and is simple, which gives (a). Away from the origin the transverse quadratic vanishing gives (c), and the product with the readout quadratic gives (b) and (d). The $p = 2$ case of (a) is transparent: with $t = \frac{1}{2}(\alpha_1+\alpha_2+\beta_1+\beta_2)$, $x = \frac{1}{2}(\alpha_1-\alpha_2)$, $y = \frac{1}{2}(\beta_1-\beta_2)$ one finds $H = 16(t^2x^2 + t^2y^2 + x^2y^2)$, whose Newton polyhedron gives $3/4$ with multiplicity one and no logarithm; the logarithm in (d) at $p = 2$ lives on a different stratum.

Three consequences for the basin-selection account. First, inactive neurons are not free: deleting them changes the local model, and an isolated one carries $(2p-1)/4$, $p/2$ or $p-1$ depending on its stratum, with a logarithmic factor on one stratum when $p = 2$. Second, isolated coefficients need not add in a shared network. For a readout-inactive neuron the extra first-order rank after eliminating the active parameters is $\mathrm{rank}[\Pi_T(Q(w_j) \otimes I_p)] \le p$, where Π_T projects off the active prediction tangent space; if the active Jacobian already spans all independent residual coordinates the complete model remains a submersion with coefficient $p(N-1)/2$, and two inactive neurons with proportional $Q(w_j)$ share their first-order directions. At zero hidden weight the quartic residual must also be projected by Π_T and Proposition T.2 need not survive that projection unchanged. We therefore do not assert a full-network coefficient for a trained generalizing solution, nor that the ordering of the memorizing and generalizing basins in Cullen et al. (2026) survives these corrections; both require a decomposition statement with an explicit transversality hypothesis that we do not prove.

Third, the corrections are testable: at a generalizing checkpoint whose inactive neurons have small but nonzero weights, Equation (103) predicts that the estimated coefficient increases with τ as the readout curvatures $\propto \|Q(w_j)\|^2$ cross γ/τ, subject to the cautions of Appendix R.

T.3 Proof of Lemma T.1

$z^\top x_{ab} = 1 - 1 = 0$ for every pair, so $(w_j + c_j z)^\top x_{ab} = w_j^\top x_{ab}$ and the training inputs lie in the hyperplane $z^\perp$, of dimension $2p-1$; on the full grid they span it (differences $e_a - e_{a'}$ and $e_{p+b} - e_{p+b'}$ together with one x_{ab}), while a training subset of N pairs spans at most $\min\{N, 2p-1\}$ dimensions. The rank of $D\Psi_D$ is at most the sum over neurons of the ranks of the per-neuron blocks. For a neuron with $v_j \neq 0$ and $w_j \notin \mathbb{R}z$, the block has the two independent kernel directions $(\delta v_j, \delta w_j) = (0, z)$ and $(-2v_j, w_j)$, so its rank is at most $3p - 2$; for $v_j = 0$ the w_j-derivative vanishes and the rank is at most p; for $w_j \in \mathbb{R}z$ the whole block vanishes. The second bound is the dimension $p(N-1)$ of the centered target space. On the full grid, a quadratic form with matrix $\left[\begin{smallmatrix} H_{11} & H_{12} \\ H_{12}^\top & H_{22} \end{smallmatrix}\right]$ evaluates to $(H_{11})_{aa} + 2(H_{12})_{ab} + (H_{22})_{bb}$ at x_{ab}, and the cross block alone produces every $p \times p$ table. □

T.4 The zero set of the isolated-neuron residual

Lemma T.3 (zero set). *For $w = (\alpha, \beta) \in \mathbb{R}^p \times \mathbb{R}^p$, $Q(w) = P_c[(\alpha_a + \beta_b)^2]_{ab} = 0$ if and only if*

$$w \in \bigcup_{\sigma \in \{\pm 1\}^p} \operatorname{span}\{z, (\sigma, 0)\} \cup \bigcup_{\tau \in \{\pm 1\}^p} \operatorname{span}\{z, (0, \tau)\}.$$

At every such $w_0 \notin \mathbb{R}z$, the differential $dQ(w_0) : \mathbb{R}^{2p} \to \mathbb{R}^{p \times p}$ has kernel $\operatorname{span}\{z, w_0\}$ and rank $2p - 2$.

Proof. $Q(w) = 0$ iff the table $(\alpha_a + \beta_b)^2$ is constant, say equal to $\kappa \geq 0$. If $\kappa = 0$ then $\alpha_a + \beta_b = 0$ for all a, b, so $\alpha \equiv \alpha_1$, $\beta \equiv -\alpha_1$, and $w = \alpha_1 z$. If $\kappa > 0$, write $\alpha_a + \beta_b = \sqrt{\kappa}\,\varepsilon_{ab}$ with $\varepsilon_{ab} \in \{\pm 1\}$. For fixed a, a' the difference $\varepsilon_{ab} - \varepsilon_{a'b} = (\alpha_a - \alpha_{a'})/\sqrt{\kappa}$ does not depend on b and lies in $\{0, \pm 2\}$, so any two rows of ε are equal or opposite: $\varepsilon_{ab} = \sigma_a \tau_b$ for sign vectors σ, τ. If σ is constant, then $\alpha_a + \beta_b = \pm\sqrt{\kappa}\,\tau_b$ forces $\alpha \equiv \alpha_1$ and $\beta = \pm\sqrt{\kappa}\,\tau - \alpha_1 \mathbf{1}$, i.e. $w = \alpha_1 z \pm \sqrt{\kappa}(0, \tau)$. If σ takes both signs, pick a, a' with $\sigma_a = -\sigma_{a'}$; then $\alpha_a - \alpha_{a'} = 2\sqrt{\kappa}\,\tau_b$ for all b, so τ is constant and by symmetry $w \in \operatorname{span}\{z, (\sigma, 0)\}$. Conversely every listed w gives a constant table.

For the differential, let $\mathsf{L}\delta w = (\delta\alpha_a + \delta\beta_b)_{ab}$, a linear map with kernel $\mathbb{R}z$ and rank $2p - 1$. At w_0 with table $\sqrt{\kappa}\,\varepsilon$, $dQ(w_0)[\delta w] = 2\sqrt{\kappa}\,P_c[\varepsilon \odot \mathsf{L}\delta w]$. Entrywise multiplication by ε is invertible, and P_c annihilates exactly the constant tables, so $\delta w \in \ker dQ(w_0)$ iff $\mathsf{L}\delta w = c\,\varepsilon = (c/\sqrt{\kappa})\mathsf{L}w_0$ for some c, iff $\delta w - (c/\sqrt{\kappa})w_0 \in \mathbb{R}z$. Since $w_0 \notin \mathbb{R}z$ the kernel is $\operatorname{span}\{z, w_0\}$, of dimension 2. □

T.5 Proof of Proposition T.2

Quotient by the shift. $H(w + cz) = H(w)$ for all c. Write $w = cz + Bu$ with $u \in \mathbb{R}^s$, $s = 2p - 1$, and $B : \mathbb{R}^s \to z^\perp$ an isometry; then for a smooth positive weight ψ, $\int H(w)^{-\xi}\psi(w)\,\mathrm{d}w = \int \tilde{H}(u)^{-\xi}\tilde{\psi}(u)\,\mathrm{d}u$ with $\tilde{H}(u) = H(Bu)$ and $\tilde{\psi}(u) = \int \psi(cz + Bu)\,\mathrm{d}c$ smooth and positive. $\tilde{H}$ is a homogeneous quartic, and by Lemma T.3 its zero set is a finite union of lines $\mathbb{R}u_\ell$ through the origin. Learning coefficients do not depend on the smooth positive weight (Watanabe, 2009), so we may take $\tilde{\psi} \equiv 1$ on the unit ball.

Quadratic vanishing along the lines. Let $e = u_\ell / \|u_\ell\|$ and write $u = te + y$, $y \perp e$. With $\tilde{Q}(u) = Q(Bu)$, homogeneity of degree 2 gives $\tilde{Q}(te) = 0$ and $d\tilde{Q}(te) = t\,d\tilde{Q}(e)$, whose restriction to $e^\perp$ is injective by Lemma T.3 (the kernel of $dQ(w_0)$ meets $z^\perp$ in $\mathbb{R}w_0$). Hence for t in a compact interval away from 0 and $\|y\|$ small, $\left\|\tilde{Q}(te + y)\right\| = \left\|d\tilde{Q}(te)[y] + O(\|y\|^2)\right\| \asymp \|y\|$, i.e. $\tilde{H}(te+y) \asymp \|y\|^2$.

Polar coordinates. Write $u = \varrho\theta$, $\theta \in S^{s-1}$, so that $\tilde{H}(u) = \varrho^4 h(\theta)$ with $h = \tilde{H}|_{S^{s-1}} \geq 0$, vanishing exactly at the finitely many points $\pm e_\ell$, and $h(\theta) \asymp |\theta - e_\ell|^2$ near each of them by the previous step. Then

$$\zeta(\xi) = \int_{\|u\| \leq 1} \tilde{H}(u)^{-\xi}\,\mathrm{d}u = \int_0^1 \varrho^{\,s-1-4\xi}\,\mathrm{d}\varrho \int_{S^{s-1}} h(\theta)^{-\xi}\,\mathrm{d}\theta = \frac{1}{s - 4\xi}\,\mathcal{A}(\xi).$$

The angular factor $\mathcal{A}$ converges absolutely and is analytic for $\operatorname{Re}\xi < (s-1)/2$, because near each zero $h^{-\xi} \asymp |\phi|^{-2\xi}$ in $s-1$ local coordinates ϕ, and $\mathcal{A}(\xi) > 0$ for real ξ in this range. Since $s \geq 3$, $s/4 < (s-1)/2$, so the smallest positive pole of ζ is $\xi = s/4$, simple. Thus $\tilde{H}$ has coefficient $s/4 = (2p-1)/4$ at the origin with multiplicity 1.

The strata. Since $H(w+cz) = H(w)$, the germ of H at any $w_0 \in \mathbb{R}z$ is the germ at 0, so it suffices to treat $w_0 = 0$ in (a). At $(v_0, 0)$ with $v_0 \neq 0$: on a product neighbourhood, $\|v\|^2$ is bounded above and below, so $K = \|v\|^2 H \asymp H$ and Lemma J.1 gives $(2p-1)/4$ with multiplicity 1. At $(0,0)$: $K = \|v\|^2 \tilde{H}(u)$ is a product of functions of disjoint variables, so its zeta function is the product of $\int \|v\|^{-2\xi} \, \mathrm{d}v$, with first pole at $\xi = p/2$, and the zeta function of $\tilde{H}$, with first pole at $(2p-1)/4 < p/2$; the smaller pole wins and remains simple. This is (a). For (b), $H \asymp 1$ on a neighbourhood of w_0, so $K \asymp \|v\|^2$ and the coefficient is $p/2$, multiplicity 1. For (c) and (d), let $Q(w_0) = 0$ with $w_0 \notin \mathbb{R}z$. By Lemma T.3, w_0 lies on exactly one of the planes (two planes meet only in $\mathbb{R}z$), $dQ(w_0)$ has kernel equal to that plane, and the argument of the previous paragraph, applied in a neighbourhood of w_0 rather than along a ray, gives $H \asymp \|y\|^2$ with $y \in \mathbb{R}^{2p-2}$ the coordinates transverse to the plane. At (v_0, w_0) with $v_0 \neq 0$ this gives $K \asymp \|y\|^2$, a nondegenerate quadratic form in $2p-2$ variables: coefficient $p-1$, multiplicity 1. At $(0, w_0)$, $K \asymp \|v\|^2 \|y\|^2$ is a product of nondegenerate quadratic forms in disjoint variables with poles at $p/2$ and $p-1$; the coefficient is the smaller of the two, and the multiplicity is 2 exactly when they coincide, i.e. when $p = 2$. □

The case $p = 2$. With $t = \frac{1}{2}(\alpha_1 + \alpha_2 + \beta_1 + \beta_2)$, $x = \frac{1}{2}(\alpha_1 - \alpha_2)$, $y = \frac{1}{2}(\beta_1 - \beta_2)$ and $c = \frac{1}{2}(\alpha_1 + \alpha_2 - \beta_1 - \beta_2)$ (the z-coordinate), the four table entries are $(t \pm x \pm y)^2 = t^2 + x^2 + y^2 \pm 2tx \pm 2ty \pm 2xy$ with the signs of the cross terms determined by the two choices; centring removes $t^2 + x^2 + y^2$ and $H = 4\sum_{\pm,\pm}(\pm tx \pm ty \pm xy)^2 = 16(t^2x^2 + t^2y^2 + x^2y^2)$, independent of c. The Newton polyhedron of $t^2x^2 + t^2y^2 + x^2y^2$ is the convex hull of $(2,2,0), (2,0,2), (0,2,2)$ plus the positive orthant; the diagonal $\varrho(1,1,1)$ meets its boundary in the relative interior of the face $\{x_1 + x_2 + x_3 = 4\}$ at $\varrho = 4/3$, so the coefficient is $3/4 = (2p-1)/4$ with multiplicity one and no logarithmic factor.

Numerical sanity checks. The ranks $7, 14, 13, 26, 52$ quoted in Appendix T.1 were obtained from explicit Jacobians at Gaussian random parameters with and without centring; the rank $2p-2$ of dQ at points of the zero set was verified for $p = 3, 5$ by finite-difference Hessians of H (two eigenvalues at machine precision, the rest of order one).