Block-Wishart Formulas for Three Spectra in Multi-Index Phase Retrieval

Abstract

We derive deterministic formulas for three spectral objects along a frozen multi-index phase-retrieval trajectory. The Hessian and the empirical gradient covariance are exactly block-Wishart matrices after removing the finite teacher–student subspace. The predictable covariance of the Muon-filtered weight noise has the same form. Their limiting densities solve one matrix fixed-point equation, while their support edges and logarithmic potentials are given by the variational formulas of Montanari and Saeed. Coherent gradients, learned teacher directions, and the finite projection removed by the cavity decomposition are finite-rank deformations; their visible eigenvalues are determined by a Schur equation using the same bulk resolvent. The theorem applies directly to the frozen or fresh-sample problem with fixed student width. It does not identify the singular-value law of the averaged gradient, the raw weight spectrum without a martingale closure, or a same-sample trajectory without leave-one-out transport. Numerical experiments compare the matrix fixed point, both variational edges, and the logarithmic potential with finite block-Wishart spectra at three stages of learning.

1 The question and the three matrices

Let $x \in \mathbb{R}^d$ be centered and isotropic. The teacher and student are

$$f_\star(x) = \sum_{j=1}^{k} \mu_j (\theta_j^\top x)^2, \qquad f_W(x) = \frac{1}{P}\sum_{a=1}^{P} (w_a^\top x)^2, \tag{1.1}$$

where P and k remain fixed as $n, d \to \infty$, and $n/d \to \alpha$. We use the square loss $\ell_W(x) = \frac{1}{2}(f_W(x) - f_\star(x))^2$. At time t, write

$$h = W_t^\top x, \qquad y = \Theta^\top x, \qquad s_t(h, y) = \frac{1}{P}\|h\|^2 - y^\top \Lambda y. \tag{1.2}$$

Three different matrices are often called "the spectrum during training." They must be separated.

The first is the empirical Hessian $H_{n,t}$ acting on dP parameter coordinates. The second is the empirical covariance of the individual sample gradients,

$$\mathcal{F}_{n,t} = \frac{1}{n}\sum_{\ell=1}^{n} \operatorname{vec}(\nabla_W \ell_{W_t}(x_\ell)) \operatorname{vec}(\nabla_W \ell_{W_t}(x_\ell))^\top. \tag{1.3}$$

It is not the singular spectrum of the averaged gradient $G_{n,t} = n^{-1}\sum_\ell \nabla_W \ell_{W_t}(x_\ell)$. The third is the predictable covariance of the stochastic component of the weight update. It controls the weight bulk after coherent drift has been removed. The raw matrix $W_t W_t^\top/P$ also contains the integrated drift and its cross terms.

The result below gives a common equation for the bulks of these three objects. Their outliers remain channel dependent.

2 The block-Wishart variational theorem

Let $Z \in \mathbb{C}^{n\times d}$ have independent, centered, isotropic rows and satisfy a Hanson–Wright concentration inequality. Let $T_1, \dots, T_n \in \mathbb{C}^{P\times P}$ be self-adjoint blocks, deterministic or independent of Z, and define

$$\mathcal{W}_{n,P} = \frac{1}{n}(Z^* \otimes I_P)\,\mathrm{diag}(T_1, \dots, T_n)(Z \otimes I_P). \tag{2.1}$$

Assume that $n/d \to \alpha \in (0,\infty)$ and that the empirical law of the blocks converges to a compactly supported probability measure ν on the self-adjoint $P \times P$ matrices.

For an invertible $S \in \mathbb{C}^{P\times P}$, define the matrix K-transform

$$\mathcal{K}_{\alpha,\nu}(S) = \int T(I_P + ST)^{-1}\nu(\mathrm{d}T) - \frac{1}{\alpha}S^{-1}. \tag{2.2}$$

Its derivative in a self-adjoint direction Δ is

$$\begin{aligned} D\mathcal{K}_{\alpha,\nu}(S)[\Delta] &= \frac{1}{\alpha}S^{-1}\Delta S^{-1} \\ &\quad - \int T(I_P + ST)^{-1}\Delta(I_P + TS)^{-1}T\,\nu(\mathrm{d}T). \end{aligned} \tag{2.3}$$

Theorem 2.1 (Block-Wishart law and matrix fixed point)**.** *Under the assumptions above, the empirical spectral distribution of $\mathcal{W}_{n,P}$ converges almost surely to a deterministic measure $\mu_{\alpha,\nu}$. For $z \in \mathbb{C}_+$, there is a unique $S(z) \in \mathbb{C}^{P\times P}$ in the matrix upper half-plane such that*

$$\mathcal{K}_{\alpha,\nu}(S(z)) = zI_P. \tag{2.4}$$

The Stieltjes transform of the limiting measure is

$$m_{\alpha,\nu}(z) = \frac{\alpha}{P}\,\mathrm{Tr}\,S(z), \qquad \rho_{\alpha,\nu}(E) = \frac{\alpha}{\pi P}\lim_{\eta\downarrow 0}\Im\,\mathrm{Tr}\,S(E + \mathrm{i}\eta). \tag{2.5}$$

This is Proposition 1 of Montanari and Saeed [1] in the normalization $n/d \to \alpha$. The fixed-point equation can equivalently be written

$$S(z) = \left\{\alpha\left[\int T(I_P + S(z)T)^{-1}\nu(\mathrm{d}T) - zI_P\right]\right\}^{-1}. \tag{2.6}$$

Define

$$\mathcal{P}_-(\alpha,\nu) = \left\{\; S \succ 0 \;\middle|\; \begin{array}{l} S^{-1} + T \succ 0 \quad \nu\text{-a.s.}, \\ D\mathcal{K}_{\alpha,\nu}(S) \succ 0 \text{ on } \mathrm{Sym}_P \end{array} \right\}. \tag{2.7}$$

Theorem 2.2 (Variational edges and logarithmic potential)**.** *Let $E_- = \inf\mathrm{supp}(\mu_{\alpha,\nu})$. Then*

$$E_- = sup_{S\in\mathcal{P}_-(\alpha,\nu)}\lambda_{\min}(\mathcal{K}_{\alpha,\nu}(S)). \tag{2.8}$$

The right edge follows by reflection:

$$E_+(\alpha,\nu) = -E_-(\alpha,\check{\nu}), \qquad \check{\nu} = \mathrm{Law}(-T), \quad T \sim \nu. \tag{2.9}$$

For $u < E_-$, the logarithmic potential is

$$\begin{aligned} \Phi_{\alpha,\nu}(u) &= \int \log|u - E|\,\mu_{\alpha,\nu}(\mathrm{d}E) \\ &= \frac{1}{P}\inf_{S\in\mathcal{P}_-(\alpha,\nu)}\Big\{-\alpha u\,\mathrm{Tr}\,S + \alpha\int \log\det(I_P + TS)\nu(\mathrm{d}T) \\ &\qquad\qquad - \log|\det S| - P(\log\alpha + 1)\Big\}. \end{aligned} \tag{2.10}$$

Equations (2.8) and (2.10) are Theorems 1 and 2 of [1]. The important point for non-convex losses is that the blocks T need not be positive semidefinite.

Remark 2.3 (Regular edges)**.** *At a regular extremal edge, the maximizing matrix S_e satisfies*

$$\mathcal{K}_{\alpha,\nu}(S_e) = E_eI_P, \qquad \lambda_{\min}(D\mathcal{K}_{\alpha,\nu}(S_e)) = 0. \tag{2.11}$$

The first identity supplies $P(P+1)/2 - 1$ scalar equations and the second supplies the final equation. This is the system used in the numerical experiments below; no empirical edge is fitted.

3 Exact phase-retrieval blocks

Lemma 3.1 (Local derivatives). *For the loss in* (1.1)*, define*

$$u_t(h, y) = \nabla_h \frac{1}{2} s_t(h, y)^2 = \frac{2}{P} s_t(h, y)h. \tag{3.1}$$

The local Hessian in the student coordinates is

$$T_t^H(h, y) = \nabla_h^2 \frac{1}{2} s_t(h, y)^2 = \frac{4}{P^2} hh^\top + \frac{2}{P} s_t(h, y)I_P. \tag{3.2}$$

For one sample,

$$\nabla_W \ell_{W_t}(x) = xu_t(h, y)^\top, \qquad \nabla_W^2 \ell_{W_t}(x) = xx^\top \otimes T_t^H(h, y), \tag{3.3}$$

up to the fixed permutation associated with the vectorization convention.

Proof. Since $\nabla_h s_t = 2h/P$ and $\nabla_h^2 s_t = 2I_P/P$,

$$\nabla_h \frac{1}{2} s_t^2 = s_t \nabla_h s_t, \qquad \nabla_h^2 \frac{1}{2} s_t^2 = (\nabla_h s_t)(\nabla_h s_t)^\top + s_t \nabla_h^2 s_t.$$

The map $W \mapsto W^\top x$ is linear with Jacobian $x^\top \otimes I_P$. The chain rule gives (3.3). □

Let $\mathcal{S}_t = \operatorname{span}(W_t, \Theta)$ and let $r_t = \dim \mathcal{S}_t$. Choose orthonormal matrices $U_t \in \mathbb{R}^{d\times r_t}$ and $U_t^\perp \in \mathbb{R}^{d\times(d-r_t)}$. For Gaussian design,

$$x = U_t q + (U_t^\perp)z, \qquad q \perp z, qquadq \sim N(0, I_{r_t}), \quad z \sim N(0, I_{d-r_t}). \tag{3.4}$$

Both h and y are functions of q only. Hence every local block in Lemma 3.1 is independent of the cavity design z.

Lemma 3.2 (Finite-rank restoration of the projected coordinates). *Let $\mathcal{W}_{n,P}^\perp$ be the block-Wishart matrix constructed with the cavity coordinates z_ℓ and the same blocks T_ℓ. Let $\mathcal{W}_{n,P}^{\text{full}}$ be constructed with x_ℓ. After embedding the cavity matrix in the full parameter space,*

$$\operatorname{rank}(\mathcal{W}_{n,P}^{\text{full}} - \mathcal{W}_{n,P}^\perp) \le 2r_t P. \tag{3.5}$$

Consequently, both matrices have the same limiting empirical spectral law. The difference can nevertheless create finitely many BBP outliers.

Proof. Insert $X = QU_t^\top + Z(U_t^\perp)^\top$ on both sides of the block-diagonal matrix. In the decomposition into the signal parameter space of dimension $r_t P$ and its orthogonal complement, the difference has block form $\left(\begin{smallmatrix} A & B \\ B^* & 0 \end{smallmatrix}\right)$. Its kernel contains the kernel of B inside the orthogonal complement, hence its rank is at most $2r_t P$. The rank inequality for empirical spectral distributions then proves the second statement. □

4 The Hessian spectrum

Set

$$\nu_t^H = \operatorname{Law}(T_t^H(h_t, y)), \qquad (h_t, y) \sim N(0, \mathcal{G}_t), \tag{4.1}$$

where $\mathcal{G}_t$ is the finite Gram summary of (W_t, Θ). For a bounded loss, or after a smooth truncation, this law is compactly supported.

Theorem 4.1 (Frozen Hessian bulk)**.** *Suppose W_t is fixed independently of the Hessian sample, P, k are fixed, $n/d \to \alpha_H$, and the local loss has bounded second derivatives. Then the Hessian bulk converges almost surely to μ_{α_H,ν_t^H}. Its density, edges, and logarithmic potential are obtained from Equations* (2.5), (2.8), (2.9), *and* (2.10) *with*

$$\boxed{\mathcal{K}_t^H(S) = \mathbb{E}\left[T_t^H(I_P + ST_t^H)^{-1}\right] - \frac{1}{\alpha_H}S^{-1}.} \tag{4.2}$$

The full Hessian differs from its cavity bulk by a finite-rank deformation.

Proof. Lemma 3.1 gives the block-Wishart factorization. The cavity split (3.4) makes the design independent of the blocks. Theorem 2.1 and Theorem 2.2 apply. Finally use Lemma 3.2. □

The sign of the lower edge gives a bulk stability criterion:

$$E_-^H(t) > 0 \quad \Longleftrightarrow \quad \text{the limiting Hessian bulk is strictly positive.} \tag{4.3}$$

This does not exclude a negative finite-rank outlier. Such an outlier is a separate Schur root, described in Section 7.

5 The gradient spectrum: what is and is not covered

For the sample gradient define

$$T_t^F(h,y) = u_t(h,y)u_t(h,y)^\top, \qquad \nu_t^F = \mathrm{Law}(T_t^F(h_t,y)). \tag{5.1}$$

Then (1.3) has the exact block representation

$$\mathcal{F}_{n,t} = \frac{1}{n}(X^\top \otimes I_P)\,\mathrm{diag}(T_{t,1}^F,\dots,T_{t,n}^F)(X \otimes I_P). \tag{5.2}$$

Theorem 5.1 (Gradient-covariance bulk)**.** *Under the frozen assumptions of Theorem 4.1, the cavity bulk of the empirical gradient covariance converges to μ_{α_F,ν_t^F}. Its matrix K-transform is*

$$\boxed{\mathcal{K}_t^F(S) = \mathbb{E}\left[u_tu_t^\top(I_P + Su_tu_t^\top)^{-1}\right] - \frac{1}{\alpha_F}S^{-1}.} \tag{5.3}$$

All blocks are positive semidefinite, so the limiting bulk is supported on $[0,\infty)$.

Proof. The vectorized sample gradient is $u_t \otimes x$. Its outer product is $xx^\top \otimes u_tu_t^\top$, which proves (5.2). The remaining argument is identical to the proof of Theorem 4.1. □

Proposition 5.2 (The averaged gradient is a different object)**.** *The singular values of*

$$G_{n,t} = \frac{1}{n}\sum_{\ell=1}^{n} x_\ell u_{t,\ell}^\top \tag{5.4}$$

are not determined by μ_{α_F,ν_t^F}. Indeed,

$$G_{n,t}G_{n,t}^\top = \frac{1}{n^2}\sum_{\ell,m} x_\ell(u_{t,\ell}^\top u_{t,m})x_m^\top \tag{5.5}$$

contains all cross-sample terms, whereas (5.2) *contains only the diagonal terms $\ell = m$. For fixed P, $G_{n,t}$ has at most P nonzero singular values.*

This proposition prevents an incorrect use of the block-Wishart theorem. Muon acts on the singular values of $G_{n,t}$, not directly on the spectrum of $\mathcal{F}_{n,t}$. The latter controls the covariance of fresh gradient noise and therefore enters the state evolution and the linearized denoising problem. A proportional-width singular-value law requires a separate rectangular random-matrix theorem.

6 The weight spectrum after linearized Muon

For $\lambda > 0$ and $a \in \mathbb{R}$, define the regularized spectral map

$$\Psi_{a,\lambda}(G) = G(G^\top G + \lambda^2 I_P)^{(a-1)/2}. \tag{6.1}$$

Let $\bar{G}_t$ be the conditional mean gradient and set

$$R_{a,t} = (\bar{G}_t^\top \bar{G}_t + \lambda_t^2 I_P)^{(a-1)/2}. \tag{6.2}$$

Lemma 6.1 (Bulk part of the Muon linearization). *For a perturbation $E \in \mathbb{R}^{d\times P}$,*

$$D\Psi_{a,\lambda}(\bar{G}_t)[E] = ER_{a,t} + \bar{G}_t\,\mathcal{L}_{a,t}[E], \tag{6.3}$$

where $\mathcal{L}_{a,t}[E]$ is a $P \times P$ matrix. The range of the second term is contained in $\mathrm{col}(\bar{G}_t)$ and has ambient rank at most P. The extensive noise bulk is therefore generated by $ER_{a,t}$; the second term belongs to the finite deformation sector.

Proof. Apply the product rule to $G \mapsto Gf(G^\top G)$ with $f(A) = (A+\lambda^2 I_P)^{(a-1)/2}$. The derivative of the first factor gives $ER_{a,t}$. The derivative of the matrix function gives a $P \times P$ matrix $\mathcal{L}_{a,t}[E]$ multiplied on the left by $\bar{G}_t$. □

Define the filtered local score and block

$$v_{a,t} = R_{a,t}u_t, \qquad T_t^W = v_{a,t}v_{a,t}^\top, \qquad \nu_t^W = \mathrm{Law}(T_t^W). \tag{6.4}$$

Since $R_{a,t}$ is symmetric, row-vector conventions replace $v_{a,t}$ by $u_t^\top R_{a,t}$ without changing the block.

Theorem 6.2 (Predictable weight-noise spectrum). *Consider fresh mini-batches and linearize the Muon update around the frozen state $\bar{G}_t$. The predictable covariance of its extensive stochastic part has cavity representation*

$$\mathcal{Q}_{n,t}^W = \frac{1}{n}(X^\top \otimes I_P)\,\mathrm{diag}(T_{t,1}^W, \dots, T_{t,n}^W)(X \otimes I_P). \tag{6.5}$$

Its limiting bulk is μ_{α_W,ν_t^W}, with

$$\boxed{\mathcal{K}_t^W(S) = \mathbb{E}\left[v_{a,t}v_{a,t}^\top(I_P + Sv_{a,t}v_{a,t}^\top)^{-1}\right] - \frac{1}{\alpha_W}S^{-1}.} \tag{6.6}$$

For a time-inhomogeneous fresh stream, the same formula holds with ν_t^W replaced by the weighted empirical mixture of the propagated blocks from all earlier times.

Proof. The centered sample-gradient perturbation has separable part $E_\ell = x_\ell u_{t,\ell}^\top - \bar{G}_t$. By Lemma 6.1, its extensive filtered part is $x_\ell v_{a,t,\ell}^\top$; subtracting its conditional mean changes only the finite coherent sector. Its covariance block is $xx^\top \otimes v_{a,t}v_{a,t}^\top$, giving (6.5). A triangular array of fresh samples at different times is covered whenever its weighted block empirical measure converges. □

Remark 6.3 (Raw weights). *Theorem 6.2 is a theorem about the predictable covariance of propagated weight noise. To identify the raw spectrum of $W_tW_t^\top/P$, one must additionally prove that the martingale part realizes this covariance in the required anisotropic topology. The integrated drift, initialization, and derivative term $\bar{G}_t\mathcal{L}_{a,t}$ are then finite or coherent deformations. This dynamic closure is not a consequence of the static block-Wishart theorem.*

7 One bulk equation, three BBP equations

Let $B^c_{n,t}$ denote any of the three cavity bulks and suppose the restored finite sector has a factorization

$$M^c_{n,t} = B^c_{n,t} + U^c_t A^c_t (U^c_t)^\top, \qquad \mathrm{rank}(U^c_t) = r_c = O(1). \tag{7.1}$$

For z outside the bulk, the determinant lemma gives

$$\det(M^c_{n,t} - zI) = \det(B^c_{n,t} - zI)\det\left(I_{r_c} + A^c_t (U^c_t)^\top (B^c_{n,t} - zI)^{-1} U^c_t\right). \tag{7.2}$$

Thus the limiting outliers solve

$$\boxed{\det\left(I_{r_c} + A^c_t \mathcal{R}^c_t(z)\right) = 0, \qquad \mathcal{R}^c_t(z) = \lim (U^c_t)^\top (B^c_{n,t} - zI)^{-1} U^c_t.} \tag{7.3}$$

The outlier residue follows by differentiating the same finite determinant.

The three K-transforms determine the bulk resolvents entering $\mathcal{R}^c_t$. They do not make the finite matrices A^c_t identical. For the gradient covariance, A^F_t contains the coherent mean-gradient sector. For the weights, A^W_t contains learned teacher directions and integrated drift. For the Hessian, A^H_t contains residual and aligned curvature directions. This is the precise sense in which the three BBP clocks share one random-matrix mechanism without sharing one transition time.

8 Numerical verification

The experiment is designed so that the imported theorem applies without a same-sample approximation. The input has independent Rademacher coordinates, which satisfy Hanson–Wright concentration. Teacher and student vectors are supported on a fixed eight-dimensional coordinate subspace. The local blocks therefore depend only on these eight coordinates and are independent of the remaining $d = 120$ cavity coordinates. Their law has exactly 2^8 atoms and is compactly supported.

For each captured mass $r \in \{0.12, 0.52, 0.88\}$, we construct the three block laws in Equations (3.2), (5.1), and (6.4). We use $P = 2$, $n/d = 4$, $a = 0.35$, and solve (2.4) without fitting the empirical histogram. The two support edges are obtained from (2.11). The logarithmic potential is evaluated from (2.10) at a point to the left of the support.

Across the nine panels, the median relative error of the two empirical extremes is 0.106 and the median Kolmogorov distance is 0.151. The median relative error of the logarithmic potential is 2.83×10^{-3}, with a maximum of 1.21×10^{-2}. The largest edge error is 0.284 and occurs in a rank-one positive channel. This ordering is expected: weak convergence and the log potential are global observables, while convergence to a support edge is controlled by the rarest blocks and is slower at finite dimension. The numerical stationary residual in (2.11) is below 3×10^{-10} in every panel. The full-minus-cavity rank never exceeds $32 = 2 \times 8 \times 2$, exactly the deterministic coordinate-cavity bound in Lemma 3.2.

The experiment verifies the three deterministic bulk formulas. It does not verify the dynamic martingale closure required to replace the predictable weight covariance by the raw weight spectrum, nor the leave-one-out theorem needed when the same observations generate W_t and the spectral probe.

9 Scope of the result

The block-Wishart theorem closes the frozen bulk calculation more strongly than a scalar Marchenko–Pastur approximation. It retains noncommuting $P \times P$ local Hessian blocks, permits

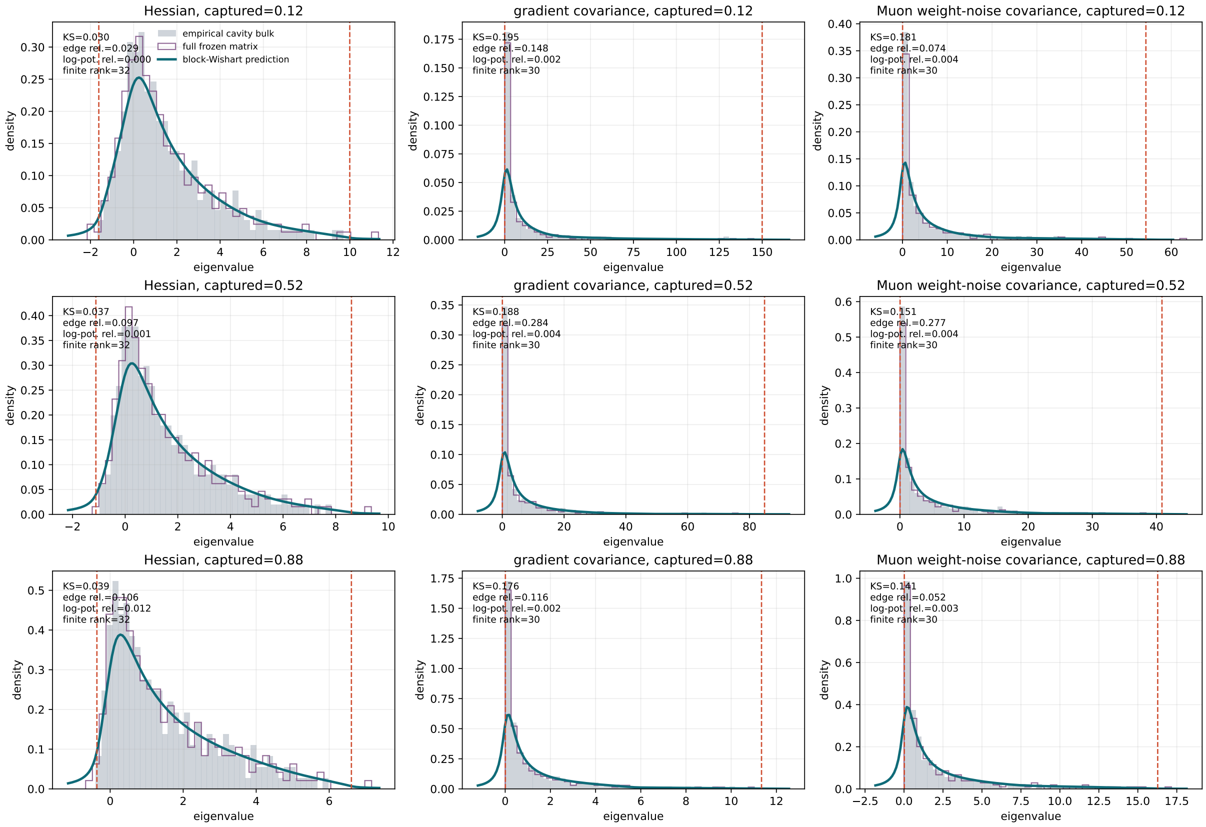

Figure 1: Empirical cavity spectra and deterministic block-Wishart densities. Rows correspond to increasing captured teacher mass. Columns show the Hessian, the gradient covariance, and the predictable covariance of Muon-filtered weight noise. Red dashed lines are the two variational edges, not empirical fits. The Hessian density is already accurate at this size. The rank-one positive blocks in the last two columns have a rare-event right edge and display visibly slower convergence of the extreme eigenvalue than of the global density and logarithmic potential.

an indefinite Hessian, computes both edges, and gives the log determinant needed in landscape calculations.

Four restrictions remain explicit.

First, the imported theorem keeps P fixed. A regime in which both P and d diverge requires an operator-valued or two-aspect extension. Using $P = 512$ as a finite approximation is numerical evidence, not a consequence of Theorem 2.1.

Second, the Hessian and gradient-covariance blocks depend on the finite teacher–student projections. Gaussian cavity splitting removes this dependence from the extensive design exactly at a frozen state. Along a same-sample trajectory, W_t also depends on the cavity coordinates; a leave-one-out comparison is still required.

Third, the singular spectrum to which Muon is applied is the spectrum of the averaged gradient. Proposition 5.2 shows why it is not equal to the block-Wishart Fisher spectrum. The latter supplies the noise covariance used by the denoising or state-evolution calculation.

Fourth, the raw weight spectrum equals a deterministic drift plus propagated noise only after a dynamic martingale or Volterra closure. The block-Wishart formula computes the covariance of that noise exactly once the closure has identified its local blocks.

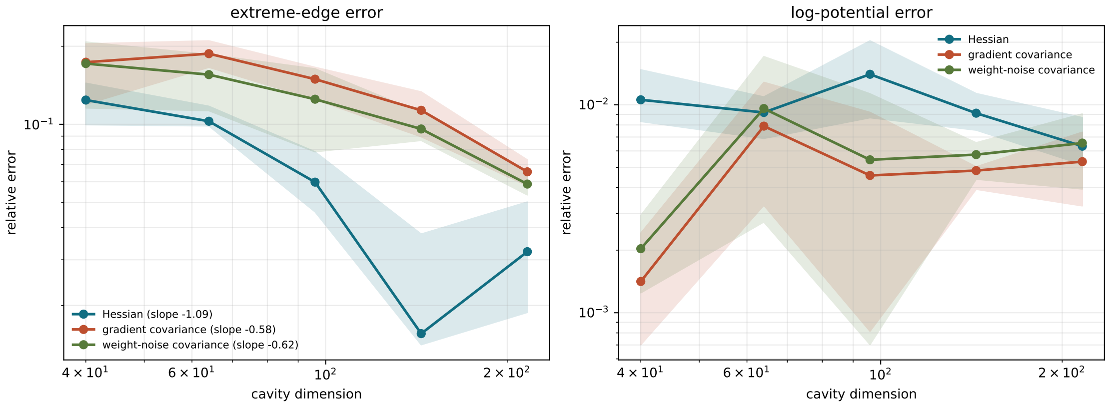

Figure 2: Dimension scaling at captured mass $r = 0.52$, with four independent designs per dimension. Lines show medians and bands show interquartile ranges. The edge error decreases as the cavity dimension grows; fitted log–log slopes are -1.09 for the Hessian, -0.58 for the gradient covariance, and -0.62 for the weight-noise covariance. The logarithmic-potential error is already at the percent level and fluctuates around its finite Monte Carlo floor.

Within these boundaries, the final dictionary is

$$\boxed{\begin{aligned} T_t^H &= \frac{4}{P^2}hh^\top + \frac{2}{P}s_t I_P, & \mathcal{K}_t^H(S) &= \mathbb{E}[T_t^H(I+ST_t^H)^{-1}] - \alpha_H^{-1}S^{-1}, \\ T_t^F &= u_t u_t^\top, & \mathcal{K}_t^F(S) &= \mathbb{E}[T_t^F(I+ST_t^F)^{-1}] - \alpha_F^{-1}S^{-1}, \\ T_t^W &= (R_{a,t}u_t)(R_{a,t}u_t)^\top, & \mathcal{K}_t^W(S) &= \mathbb{E}[T_t^W(I+ST_t^W)^{-1}] - \alpha_W^{-1}S^{-1}. \end{aligned}} \tag{9.1}$$

Each line feeds the same density, edge, and logarithmic-potential formulas. Each line is dressed by its own finite Schur deformation to produce observable BBP branches.

A Scalar consistency check

Take $P = 1$ and $T = 1$ deterministically. Then

$$\mathcal{K}_\alpha(s) = \frac{1}{1+s} - \frac{1}{\alpha s}. \tag{A.1}$$

The edge equation $\mathcal{K}_\alpha'(s) = 0$ gives

$$s_- = \frac{1}{\sqrt{\alpha}-1}, \qquad s_+ = -\frac{1}{\sqrt{\alpha}+1}. \tag{A.2}$$

Substitution yields

$$E_- = \left(1 - \frac{1}{\sqrt{\alpha}}\right)^2, \qquad E_+ = \left(1 + \frac{1}{\sqrt{\alpha}}\right)^2, \tag{A.3}$$

the Marchenko–Pastur edges for $n^{-1}X^\top X$ when $n/d \to \alpha$. This fixes all normalization factors in Equations (2.2)–(2.5).

B Numerical edge equations

Let $E_1, \ldots, E_m$ be an orthonormal basis of Sym_P for the Frobenius inner product. The derivative in Equation (2.3) is represented by

$$[\mathbb{D}(S)]_{ij} = \mathrm{Tr}\left(E_i D\mathcal{K}_{\alpha,\nu}(S)[E_j]\right). \tag{B.1}$$

At a regular left edge we parameterize $S = LL^\top$; at a regular right edge we parameterize $S = -LL^\top$. We solve

$$\mathrm{offdiag}(\mathcal{K}(S)) = 0, \qquad \mathcal{K}(S)_{ii} - \mathcal{K}(S)_{PP} = 0 \quad (i < P), \qquad \lambda_{\min}(\mathbb{D}(S)) = 0. \tag{B.2}$$

For the discrete Rademacher block law, all expectations are exact finite averages over its 2^8 atoms.

References

[1] A. Montanari and B. Saeed, *Variational Formulas for the Spectrum of Block Wishart Matrices*, arXiv:2606.27774, 2026.

[2] J. Alt, L. Erdős, and T. Krüger, *Local law for random Gram matrices*, Electronic Journal of Probability 22 (2017).

[3] G. Ben Arous, R. Gheissari, J. Huang, and A. Jagannath, *Local geometry of high-dimensional mixture models: effective spectral theory and dynamical transitions*, arXiv:2502.15655, 2025.

[4] J. W. Silverstein and S.-I. Choi, *Analysis of the limiting spectral distribution of large dimensional random matrices*, Journal of Multivariate Analysis 54 (1995), 295–309.