Gradients, Hessians, and Variational Weight Spectra in Trainable Quadratic Multi-Index Models

Janis Heran Aiad

July 2026

Abstract

We study a quadratic multi-index teacher–student model in which the hidden directions and the second-layer coefficients are trained jointly. We derive the exact population and empirical gradients, the complete Hessian in weight and output variables, and a closed population flow for the output coefficients and two Gram matrices. Although the population risk depends only on the learned quadratic form, its Euclidean gradient flow depends on the particular signed atomic decomposition; in particular, the captured teacher subspace is no longer autonomous.

Conditionally on the finite Gaussian projections, the orthogonal weight–weight Hessian is exactly a block-Wishart matrix. We specialize the matrix Stieltjes and K-transform theory of Montanari and Saeed to obtain the complete limiting density, variational formulas for both support edges, a variational logarithmic potential, and a positivity certificate. The pure quadratic model has unbounded matrix weights; an explicit saturated quadratic link gives a theorem-ready compactly supported model, while the unbounded extension is stated separately as a proof target. The output coordinates have vanishing spectral mass but enlarge the finite signal sector. An exact Schur complement on $(\mathrm{span}(W,\Theta)\otimes\mathbb{R}^p)\oplus\mathbb{R}^p$ yields a single root equation for weight, output, and mixed outliers, together with residue formulas that attribute each branch. Under an explicit anisotropic local-law assumption, these identities give the limiting joint BBP equations. Finite-dimensional experiments verify the differential identities, the block-Wishart density, both variational edges, and the logarithmic potential across dimensions, and separately show transient output-dominated outliers followed by weight-dominated branches.

Keywords. multi-index model; phase retrieval; trainable output weights; block-Wishart matrix; Hessian spectrum; K-transform; Schur complement; BBP transition.

2020 Mathematics Subject Classification. Primary 60B20, 62H12; secondary 68T07, 90C26.

1 Introduction

An isolated Hessian eigenvalue is informative only if its eigenvector has non-vanishing projection onto a planted direction. The creation or absorption of such an eigenvalue at a bulk edge is a BBP transition [1, 3]. In a learning problem the matrix itself evolves, so the transition is dynamic: both the bulk law and the finite-rank signal sector move with the training trajectory. Effective spectral theory provides a rigorous description when the loss depends on finitely many Gaussian projections and the Hessian is formed from a fresh sample [2]. For the matrix-valued weights created by a multi-index model, the block-Wishart K-transform gives more: the limiting density, both bulk edges, and the logarithmic potential required by Kac–Rice complexity calculations [5].

Most spectral analyses of quadratic phase retrieval freeze the output weights. Here we train them. The predictor is

$$f_{a,W}(x) = \frac{1}{p}\sum_{r=1}^{p} a_r (w_r^\top x)^2.$$

This apparently modest extension changes the state space in two ways. First, the same quadratic form admits many rescaled atomic decompositions, and Euclidean gradient flow is not invariant under those rescalings. Second, the p output coordinates form a finite Hessian block that couples to the dp weight coordinates. A spectral branch may therefore live in the weights, in the outputs, or in both.

The paper makes four contributions. Section 2 gives the exact joint population flow and proves that the learned matrix alone is not a closed state. Section 3 separates the rank requirement from effective overparameterisation. Section 4 derives the full Hessian and transfers the block-Wishart variational formulas to its orthogonal weight sector. Section 5 derives the joint Schur BBP equations, including exact finite-dimensional residue formulas. Section 6 explains how the same construction enters an energy-conditioned Kac–Rice problem. Section 7 reports direct spectral and dynamical diagnostics.

The logical status of each statement is explicit. The population equations, rank criterion, Hessian blocks, conditional block-Wishart reduction, and finite Schur identities are exact. The spectral limit, edge, and log-potential formulas are imported theorems for bounded matrix weights. The limiting coupled Schur kernel is conditional on an anisotropic local law uniform in the spectral parameter. The effective-slack criterion is a conjecture, not a consequence of $p > k$. The unbounded quadratic link, same-sample training, and continuum Kac–Rice conditioning require additional tail, transport, and large-deviation arguments.

2 Exact joint population dynamics

2.1 Model and finite summaries

Let p, k be fixed. Let $\Theta = (\theta_1, \dots, \theta_k) \in \mathbb{R}^{d\times k}$ have orthonormal columns, and let

$$\Lambda = \operatorname{diag}(\mu_1, \dots, \mu_k), \qquad \mu_i > 0.$$

For $W = (w_1, \dots, w_p) \in \mathbb{R}^{d\times p}$ and $a = (a_1, \dots, a_p) \in \mathbb{R}^p$, write $A = \operatorname{diag}(a)$ and

$$B = \frac{1}{p} W A W^\top, \qquad A_\star = \Theta \Lambda \Theta^\top, \qquad E = B - A_\star. \tag{2.1}$$

For $x \sim N(0, I_d)$, the teacher is $f_\star(x) = x^\top A_\star x$ and the student is $f_{a,W}(x) = x^\top B x$. Put

$$\tau = \operatorname{Tr} E, \qquad L = 2E + \tau I_d. \tag{2.2}$$

Gaussian integration gives

$$R(a, W) = \frac{1}{2}\mathbb{E}\big(f_{a,W}(x) - f_\star(x)\big)^2 = \operatorname{Tr}(E^2) + \frac{1}{2}\tau^2. \tag{2.3}$$

The finite summaries are

$$Q = W^\top W, \qquad M = W^\top \Theta, \qquad \mathcal{G} = \begin{pmatrix} Q & M \\ M^\top & I_k \end{pmatrix}. \tag{2.4}$$

When Q is invertible,

$$C = M^\top Q^{-1} M = \Theta^\top P_{\operatorname{span}(W)} \Theta \tag{2.5}$$

measures capture of the teacher subspace.

Theorem 2.1 (Closed joint gradient flow)**.** *For the Euclidean metric in* (W, a)*,*

$$\nabla_W R = \frac{2}{p} L W A, \qquad \nabla_a R = \frac{1}{p} \operatorname{diag}(W^\top L W). \tag{2.6}$$

Define

$$S = W^\top L W = \frac{2}{p} Q A Q - 2 M \Lambda M^\top + \tau Q, \tag{2.7}$$

$$T = W^\top L \Theta = \frac{2}{p} Q A M - 2 M \Lambda + \tau M, \qquad \tau = \frac{1}{p} \operatorname{Tr}(AQ) - \operatorname{Tr} \Lambda. \tag{2.8}$$

Then the population gradient flow closes exactly on (a, Q, M):

$$\dot{Q} = -\frac{2}{p}(AS + SA), \qquad \dot{M} = -\frac{2}{p} A T, \qquad \dot{a} = -\frac{1}{p} \operatorname{diag} S. \tag{2.9}$$

The risk is non-increasing and obeys

$$\frac{\mathrm{d}}{\mathrm{d}t} R = -\frac{4}{p^2} \left\| L W A \right\|_F^2 - \frac{1}{p^2} \left\| \operatorname{diag} S \right\|_2^2 . \tag{2.10}$$

Proof. The differential of (2.3) is $\mathrm{d}R = \operatorname{Tr}(L \, \mathrm{d}B)$. Varying one column w_r gives $\mathrm{d}B = p^{-1} a_r (\mathrm{d}w_r w_r^\top + w_r \mathrm{d}w_r^\top)$, while varying a_r gives $\mathrm{d}B = p^{-1} w_r w_r^\top \mathrm{d}a_r$. This proves (2.6). Substitution in $\dot{Q} = \dot{W}^\top W + W^\top \dot{W}$ and $\dot{M} = \dot{W}^\top \Theta$ gives (2.9). The last identity is the Euclidean gradient-flow dissipation formula. □

Remark 2.2 (Loss learning rates)**.** *If the weight and output blocks use learning rates η_W and η_a, the right sides of the corresponding equations are multiplied by those rates and*

$$\dot{R} = -\eta_W \left\| \nabla_W R \right\|_F^2 - \eta_a \left\| \nabla_a R \right\|_2^2 .$$

Consequently comparisons across p must state whether raw time or a p-rescaled time is held fixed.

2.2 Why the output distribution matters

Proposition 2.3 (The learned matrix is not autonomous)**.** *Along the joint flow,*

$$\begin{aligned} \dot{B} = & -\frac{2}{p^2} \left(L W A^2 W^\top + W A^2 W^\top L \right) \\ & - \frac{1}{p^2} W \operatorname{diag}\big(\operatorname{diag}(W^\top L W) \big) W^\top . \end{aligned} \tag{2.11}$$

Thus B is not a closed state variable. For any invertible diagonal matrix D, the transformation

$$W \mapsto W D, \qquad A \mapsto D^{-2} A \tag{2.12}$$

leaves B and R invariant but does not, in general, preserve the Euclidean vector field.

Proof. Differentiate $B = p^{-1} W A W^\top$ and insert Theorem 2.1. Equation (2.12) follows by direct substitution. The right side of (2.11) contains $W A^2 W^\top$ and the individual atom scores $w_r^\top L w_r$, neither of which is determined by B. □

The captured-subspace matrix C is no longer autonomous. Indeed, differentiation of $M^\top Q^{-1} M$ leaves uncancelled factors of A. The Riccati equation

$$\dot{C} = \frac{4}{p}(\Lambda C + C \Lambda - 2 C \Lambda C)$$

is recovered only in the frozen-output model $A = I_p$, $\dot{a} = 0$.

The following observables distinguish decompositions that have the same risk:

$$K_\Theta = \frac{1}{p} M^\top A M, \qquad N_{\mathrm{eff}}(a) = \frac{(\sum_r |a_r|)^2}{\sum_r a_r^2}, \qquad N_{\mathrm{eff}}^+(a) = \frac{(\sum_r (a_r)_+)^2}{\sum_r (a_r)_+^2}. \tag{2.13}$$

Here K_Θ is the learned quadratic form restricted to the teacher subspace. The residual state therefore contains both $I_k - C$ and $\Lambda - K_\Theta$, including its off-diagonal entries.

3 Width, rank, and effective slack

Proposition 3.1 (Exact rank criterion)**.** *There exist $a \in \mathbb{R}^p$ and $W \in \mathbb{R}^{d\times p}$ such that $p^{-1}WAW^\top = A_\star$ if and only if $p \geq k$. The representation can be chosen with all nonzero coefficients positive.*

Proof. Every sum of p rank-one matrices has rank at most p, so $p \geq k$ is necessary. If $p \geq k$, choose $a_i = 1$ and $w_i = \sqrt{p\mu_i}\,\theta_i$ for $1 \leq i \leq k$, and set the remaining columns to zero. □

Proposition 3.1 does not imply that $p = k$ or $p > k$ determines spectral separation. A nominally wide model may concentrate almost all positive mass on k atoms, while a balanced decomposition may use its redundancy effectively. For a teacher subset $J \subset \{1, \dots, k\}$ define

$$b_r(J) = (a_r)_+ \left\| P_J \Theta^\top w_r \right\|_2^2, \qquad N^+_{\mathrm{eff},J} = \frac{(\sum_r b_r(J))^2}{\sum_r b_r(J)^2}. \tag{3.1}$$

This sector participation combines output mass and teacher alignment.

Conjecture 3.2 (Effective slack at a regular front)**.** *Consider a regular trajectory interval on which a teacher group J is spectrally active, the relevant Schur roots are simple, and the positive coefficients stay bounded away from zero and infinity on their active atoms. Generic separation of output-dominated and weight-dominated roots requires sector slack*

$$N^+_{\mathrm{eff},J} > |J|. \tag{3.2}$$

When $N^+_{\mathrm{eff},J} \downarrow |J|$, the coupled finite kernel is expected to exhibit avoided crossings or mixed residues. The inequality is not claimed to be sufficient without regularity and transversality of the Schur kernel.

The conjecture deliberately concerns an effective count, not p alone. It predicts that $p = k+1$ is the first width at which strict slack is possible, but also that increasing p can fail to help if the output distribution collapses or if runs are compared at incompatible clocks.

4 The full empirical Hessian

Let $x_1, \dots, x_n \sim N(0, I_d)$, independently of the parameter state, and let X have rows $x_\ell^\top$. For one observation write

$$h = W^\top x, \qquad y = \Theta^\top x, \qquad s(a,h,y) = \frac{1}{p}\sum_{r=1}^{p} a_r h_r^2 - y^\top \Lambda y. \tag{4.1}$$

The sample loss is $\ell = s^2/2$.

Proposition 4.1 (Exact Hessian blocks)**.** *Order the variables as $(w_1, \dots, w_p, a)$. The weight–weight blocks are*

$$H^{WW}_{rs} = \frac{1}{n} X^\top D^{WW}_{rs} X, \tag{4.2}$$

where

$$\Phi^{WW}_{rs}(a,h,y) = \frac{4a_r a_s}{p^2} h_r h_s + \frac{2a_r}{p} s(a,h,y)\delta_{rs}. \tag{4.3}$$

The output–output and mixed blocks are

$$H^{aa}_{rs} = \frac{1}{np^2}\sum_{\ell=1}^{n} h_{\ell r}^2 h_{\ell s}^2, \tag{4.4}$$

$$H^{Wa}_{r,s} = \frac{1}{n}\sum_{\ell=1}^{n} x_\ell \left[\frac{2a_r}{p^2} h_{\ell r} h_{\ell s}^2 + \frac{2}{p} s(a, h_\ell, y_\ell) h_{\ell r}\delta_{rs} \right]. \tag{4.5}$$

Proof. Use $\partial_{h_r} s = 2a_r h_r/p$, $\partial^2_{h_r h_s} s = 2a_r\delta_{rs}/p$, and $\partial_{a_r} s = h_r^2/p$, then apply the chain rule to $s^2/2$. □

4.1 A bounded link and the exact block-Wishart reduction

It is useful to state the differential identities for a general smooth link φ. Define

$$\widehat{R}_n^{\varphi}(W,a)=\frac{1}{2n}\sum_{i=1}^{n} r_i^2, \qquad r_i=\frac{1}{p}\sum_{r=1}^{p} a_r\varphi(h_{ir})-\sum_{j=1}^{k}\mu_j\varphi(y_{ij}), \tag{4.6}$$

where $h_i=W^\top x_i$ and $y_i=\Theta^\top x_i$. Put

$$g_i=\frac{1}{p}\,a\odot\varphi'(h_i), \qquad T_i=g_ig_i^\top+r_i\,\mathrm{diag}\!\left(\frac{1}{p}a\odot\varphi''(h_i)\right). \tag{4.7}$$

Proposition 4.2 (Empirical gradient and Hessian for a smooth link)**.** *The exact empirical gradients are*

$$\nabla_W\widehat{R}_n^{\varphi}=\frac{1}{n}\sum_{i=1}^{n} x_i(r_ig_i)^\top, \qquad \nabla_a\widehat{R}_n^{\varphi}=\frac{1}{np}\sum_{i=1}^{n} r_i\varphi(h_i). \tag{4.8}$$

The weight–weight blocks and the remaining two blocks are

$$H_{rs}^{WW}=\frac{1}{n}X^\top\,\mathrm{diag}\,\big((T_i)_{rs}:i\le n\big)X, \tag{4.9}$$

$$H^{aa}=\frac{1}{np^2}\sum_{i=1}^{n}\varphi(h_i)\varphi(h_i)^\top, \tag{4.10}$$

$$H_{r,s}^{Wa}=\frac{1}{n}\sum_{i=1}^{n} x_i\left[\frac{a_r}{p^2}\varphi'(h_{ir})\varphi(h_{is})+\frac{r_i}{p}\varphi'(h_{ir})\delta_{rs}\right]. \tag{4.11}$$

For $\varphi(u)=u^2$, these formulas reduce exactly to Proposition 4.1 and

$$T_i=A^{WW}(a,h_i,y_i)=\frac{4}{p^2}(Ah_i)(Ah_i)^\top+\frac{2s(a,h_i,y_i)}{p}A. \tag{4.12}$$

Proof. The derivatives of r_i with respect to h_i are g_i and $\mathrm{diag}(p^{-1}a\odot\varphi''(h_i))$. Applying the chain rule to $r_i^2/2$ gives T_i and all three Hessian blocks. Differentiation with respect to W contributes the factor x_i. □

The compact-support hypothesis of the block-Wishart theorem can be met by an explicit model rather than by a formal cutoff. For $L<\infty$, let

$$\varphi_L(u)=L^2\tanh(u^2/L^2), \qquad \varphi_L'(u)=2u\,\mathrm{sech}^2(u^2/L^2), \tag{4.13}$$

$$\varphi_L''(u)=2\,\mathrm{sech}^2(u^2/L^2)\left[1-4\frac{u^2}{L^2}\tanh(u^2/L^2)\right]. \tag{4.14}$$

All three functions are bounded. Hence, for fixed bounded a and μ, the matrix weight T_i in (4.7) ranges over a compact subset of the real symmetric matrices.

Let $\mathcal{V}=\mathrm{span}(W,\Theta)$, let $d_0=d-\dim\mathcal{V}$, and let $U\in\mathbb{R}^{d\times d_0}$ be an orthonormal basis of $\mathcal{V}^\perp$. For Gaussian covariates,

$$\xi_i=U^\top x_i\sim N(0,I_{d_0}) \quad \text{is independent of} \quad (h_i,y_i). \tag{4.15}$$

Therefore, conditionally on the finite projections, the Hessian restricted to orthogonal weight perturbations is exactly

$$H_\perp^{WW}=\frac{1}{n}\sum_{i=1}^{n}(I_p\otimes\xi_i)T_i(I_p\otimes\xi_i^\top). \tag{4.16}$$

This is the block-Wishart ensemble, not merely an analogy with it.

Theorem 4.3 (Variational weight-spectrum formulas)**.** *Assume p, k are fixed, the Hessian sample is fresh, $d_0, n \to \infty$ with $n/d_0 \to \alpha \in (0, \infty)$, and*

$$\nu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{T_i} \Longrightarrow \nu \in \mathcal{P}_0(\mathbb{S}^p). \tag{4.17}$$

Here $\mathcal{P}_0(\mathbb{S}^p)$ denotes the compactly supported probability measures on the real symmetric $p \times p$ matrices. Assume also Hausdorff convergence of the finite supports to $\operatorname{supp} \nu$*. Define the matrix K-transform*

$$\mathscr{K}_{\alpha,\nu}(S) = \int T(I_p + ST)^{-1} \, \nu(\mathrm{d}T) - \frac{1}{\alpha} S^{-1}. \tag{4.18}$$

For every $z \in \mathbb{C}_+$, the equation

$$\mathscr{K}_{\alpha,\nu}(S(z)) = zI_p, \qquad \Im S(z) \succ 0, \tag{4.19}$$

has a unique solution. The empirical spectral distribution of $H_\perp^{WW}$ converges almost surely to a deterministic measure $\mu_{\alpha,\nu}$ with Stieltjes transform and density

$$m(z) = \frac{\alpha}{p} \operatorname{Tr} S(z), \qquad \rho(x) = \frac{1}{\pi} \lim_{\eta \downarrow 0} \Im m(x + \mathrm{i}\eta). \tag{4.20}$$

The same limiting empirical measure holds for the full weight–weight block; the omitted finite projection sector can only create finitely many outliers.

The Fréchet derivative of (4.18) *is the self-adjoint map*

$$D\mathscr{K}(S)[\Delta] = \frac{1}{\alpha} S^{-1} \Delta S^{-1} - \int T(I_p + ST)^{-1} \Delta (I_p + TS)^{-1} T \, \nu(\mathrm{d}T). \tag{4.21}$$

Let

$$\mathcal{P}_- = \{S \succ 0 : S^{-1} + T \succ 0 \ \forall T \in \operatorname{supp} \nu, \ D\mathscr{K}(S) \succ 0\}. \tag{4.22}$$

Let $\check{\nu} = (-\operatorname{Id})_\# \nu$ be the law obtained by replacing every block T by $-T$. Then the two bulk edges are

$$\zeta_- = \sup_{S \in \mathcal{P}_-} \lambda_{\min}(\mathscr{K}_{\alpha,\nu}(S)), \qquad \zeta_+(\alpha, \nu) = -\zeta_-(\alpha, \check{\nu}). \tag{4.23}$$

For $u < \zeta_-$, define

$$\mathfrak{g}_{\alpha,\nu}(S; u) = -\alpha u \operatorname{Tr} S + \alpha \int \log\det(I_p + TS) \, \nu(\mathrm{d}T) - \log|\det S| - p(\log \alpha + 1). \tag{4.24}$$

The normalized logarithmic potential is

$$\int \log|\lambda - u| \, \mu_{\alpha,\nu}(\mathrm{d}\lambda) = \frac{1}{p} \inf_{S \in \mathcal{P}_-} \mathfrak{g}_{\alpha,\nu}(S; u). \tag{4.25}$$

For $u > \zeta_+(\alpha, \nu)$, the potential equals the left potential of $\check{\nu}$ evaluated at $-u$. Finally, $\operatorname{supp} \mu_{\alpha,\nu} \subset (0, \infty)$ *if and only if there exist $S, B \succ 0$ such that*

$$D\mathscr{K}(S)[B] \succ 0, \qquad \mathscr{K}_{\alpha,\nu}(S) \succ 0, \qquad S^{-1} + T \succ 0 \quad \forall T \in \operatorname{supp} \nu. \tag{4.26}$$

Proof. Conditional independence gives (4.16). The standard asymptotic characterization, the two edge formulas, the logarithmic-potential formula, and the positivity criterion are respectively Proposition 1, Theorems 1–2, and Corollary 1 of [5], after identifying their block size with p and their ambient dimension with d_0. Removing or restoring $p \dim \mathcal{V} = O(1)$ coordinates does not change the empirical spectral limit, by interlacing. □

For comparison with the effective-spectral normalization, set $R(z) = \alpha S(z)$. Equation (4.19) becomes

$$-R(z)^{-1} = zI_p - \int T(I_p + \alpha^{-1} R(z)T)^{-1} \nu(\mathrm{d}T), \tag{4.27}$$

which is the matrix Dyson equation used below. For the pure quadratic link, the exact weight is (4.12), but its Gaussian law is unbounded and does not satisfy the compact-support hypothesis as written. Passing from φ_L to u^2 requires a uniform tail-and-stability argument or a growing spectral truncation; none of the bounded-link conclusions is used to hide that remaining step. The p output coordinates have vanishing relative dimension and do not change the limiting empirical measure, but they do change the finite outlier problem.

4.2 Joint innovations and compact response

The joint output–feature Hessian has exact sample diagonals, but it is not a Wigner deformation. Define

$$D_{rs}^{WW} = \mathrm{diag}(\Phi_{rs}^{WW}(a, h_1, y_1), \dots, \Phi_{rs}^{WW}(a, h_n, y_n)). \tag{4.28}$$

For the mixed block put

$$\begin{aligned} \psi_{rs,\ell}^{Wa} &= \frac{2a_r}{p^2} h_{\ell r} h_{\ell s}^2 + \frac{2}{p} s(a, h_\ell, y_\ell) h_{\ell r} \delta_{rs}, \\ D_{rs}^{Wa} &= \mathrm{diag}(\psi_{rs,1}^{Wa}, \dots, \psi_{rs,n}^{Wa}). \end{aligned} \tag{4.29}$$

Then the two high-dimensional observation maps are exactly

$$H_{rs}^{WW} = \frac{1}{n} X^\top D_{rs}^{WW} X, \qquad H_{r,s}^{Wa} = \frac{1}{n} X^\top D_{rs}^{Wa} \mathbf{1}_n. \tag{4.30}$$

The output–output block is the finite Gram matrix

$$H^{aa} = \frac{1}{n} \sum_{\ell=1}^{n} z_\ell z_\ell^\top, \qquad z_\ell = p^{-1}(h_{\ell 1}^2, \dots, h_{\ell p}^2)^\top.$$

Thus D^{WW} contains the weighted-covariance innovations, D^{Wa} contains the mixed innovations, and the output features enlarge the finite response sector. Their conditional means contribute to the deterministic signal blocks; their centered parts generate correlated fluctuations.

After the orthogonal weight coordinates are eliminated, the relevant compact response is the teacher–student–output projection of these weighted design maps. Its bare singular modes order output, teacher-weight, and mixed candidates. Their observable BBP locations are nevertheless determined by the resolvent-dressed joint kernel $\mathcal{K}_d(z)$:

$$\det(\mathcal{K}_d(x_\pm) - x_\pm I_{\mathcal{F}}) = 0. \tag{4.31}$$

The derivative of the same kernel gives the two attribution residues. A scalar threshold is justified only if the weight-bulk resolvent is isotropic on every compact response mode; neither the weighted covariance bulk nor the mixed block has this property automatically.

5 A joint Schur equation for weight, output, and mixed branches

5.1 Exact finite-dimensional reduction

Identify the weight space with $\mathbb{R}^d \otimes \mathbb{R}^p$. Let

$$\mathcal{F}_W = \mathrm{span}(W, \Theta) \otimes \mathbb{R}^p, \qquad \mathcal{F} = \mathcal{F}_W \oplus \mathbb{R}_a^p, \tag{5.1}$$

let Π_Θ project $\mathcal{F}_W$ onto $\mathrm{span}(\Theta) \otimes \mathbb{R}^p$, and let $\mathcal{F}^\perp$ be the orthogonal weight sector. Including $\mathrm{span}(W)$ removes the complete finite summary sector, including non-informative student rotations, from the random-matrix bulk. In this decomposition write the full Hessian as

$$H_d = \begin{pmatrix} H_{\perp\perp} & H_{\perp F} \\ H_{F\perp} & H_{FF} \end{pmatrix}. \tag{5.2}$$

Theorem 5.1 (Joint finite Schur equation and residue)**.** *For $z \notin \mathrm{spec}(H_{\perp\perp})$, define*

$$\mathcal{K}_d(z) = H_{FF} - H_{F\perp}(H_{\perp\perp} - zI)^{-1} H_{\perp F}. \tag{5.3}$$

Then z is an eigenvalue of H_d if and only if

$$\det\left(\mathcal{K}_d(z) - zI_{\mathcal{F}}\right) = 0. \tag{5.4}$$

If $c = (c_W, c_a) \in \mathcal{F}$ is a unit vector in the kernel of $\mathcal{K}_d(z) - zI$, the corresponding normalized eigenvector has finite-sector mass

$$\Omega_F(z, c) = \frac{1}{c^\top (I - \partial_z \mathcal{K}_d(z)) c}. \tag{5.5}$$

Its output and teacher-weight masses are

$$\Omega_a = \Omega_F \left\| c_a \right\|_2^2, \qquad \Omega_\Theta = \Omega_F \left\| \Pi_\Theta c_W \right\|_2^2. \tag{5.6}$$

Proof. The determinant identity is the Schur complement formula. If $c \in \ker(\mathcal{K}_d(z) - zI)$, the orthogonal component of the eigenvector is

$$u = -(H_{\perp\perp} - zI)^{-1} H_{\perp F} c.$$

Since

$$\partial_z \mathcal{K}_d(z) = -H_{F\perp}(H_{\perp\perp} - zI)^{-2} H_{\perp F},$$

we have $\|u\|_2^2 = c^\top(-\partial_z \mathcal{K}_d(z))c$. Normalizing (u, c) gives (5.5) and (5.6). □

Definition 5.2 (Branch attribution)**.** *A sequence of isolated eigenpairs is output-dominated if $\liminf \Omega_a > 0$ and $\Omega_\Theta \to 0$, teacher-weight-dominated if $\liminf \Omega_\Theta > 0$ and $\Omega_a \to 0$, and mixed if both masses have positive lower limits. A finite-sector root with both masses vanishing is a non-informative student branch. This definition is invariant under rotations within the teacher and output finite sectors.*

Theorem 5.1 is preferable to tracking sorted eigenvalue ranks. At a crossing or avoided crossing, rank labels can exchange while the Schur root and its two residues remain meaningful.

5.2 Conditional limiting BBP theorem

Assumption 5.3 (Uniform joint Schur local law)**.** *Fix a compact set $\mathcal{D} \subset \mathbb{C}$ at positive distance from the limiting bulk support. Along a deterministic state path $(a(t), Q(t), M(t))$ independent of the Hessian sample:*

1. *the truncated weight bulk obeys* (4.27)*;*
2. *there is a deterministic analytic matrix $\mathcal{K}(z, t)$ such that, uniformly on $\mathcal{D}$ and in probability,*

$$\mathcal{K}_d(z, t) \longrightarrow \mathcal{K}(z, t), \qquad \partial_z \mathcal{K}_d(z, t) \longrightarrow \partial_z \mathcal{K}(z, t);$$

3. *the resolvents are uniformly bounded on the contours under consideration.*

Theorem 5.4 (Conditional joint BBP limit)**.** *Under Assumption 5.3, every simple real root* $\lambda_\star(t)$ *of*

$$\det\left(\mathcal{K}(\lambda,t)-\lambda I_{\mathcal{F}}\right)=0 \tag{5.7}$$

that is separated from the limiting bulk is accompanied, with probability tending to one, by exactly one empirical eigenvalue converging to $\lambda_\star(t)$*. If* $c_\star$ *is its normalized finite-kernel vector, the output and teacher-weight masses converge to*

$$\Omega_a=\frac{\|(c_\star)_a\|_2^2}{c_\star^\top(I-\partial_z\mathcal{K}(\lambda_\star,t))c_\star},\qquad \Omega_\Theta=\frac{\|\Pi_\Theta(c_\star)_W\|_2^2}{c_\star^\top(I-\partial_z\mathcal{K}(\lambda_\star,t))c_\star}. \tag{5.8}$$

A BBP contact occurs when the root reaches a regular edge of the measure defined by (4.27)*.*

Proof. Uniform convergence of the determinant on a contour containing one simple root gives root convergence by Rouché's theorem. Uniform convergence of the derivative and Theorem 5.1 give the residue limits. □

Remark 5.5 (What remains probabilistic)**.** *Theorem 4.3 supplies the weight-bulk deterministic equivalent and its edges for the bounded fresh-sample model. The new work needed for an unconditional joint theorem is the anisotropic control of the mixed vectors* (4.5) *inside the same resolvent, uniformly near the contours used for roots and derivatives. Assumption 5.3 records exactly this missing input. Reusing the training sample requires a further leave-one-out comparison.*

6 Energy-conditioned extension

At a critical point of prescribed energy, the finite state is now (a,Q,M) rather than (Q,M). The Kac–Rice order parameter is a law ν on $(h,y)\in\mathbb{R}^{p+k}$ with Gaussian reference $N(0,\mathcal{G})$. For the square loss, output stationarity adds the p constraints

$$\int s(a,h,y)h_r^2\,\mathrm{d}\nu(h,y)=0,\qquad 1\le r\le p. \tag{6.1}$$

Projected weight stationarity is

$$\int\begin{pmatrix}h\\y\end{pmatrix}\left(\frac{2}{p}\,s(a,h,y)Ah\right)^\top\mathrm{d}\nu(h,y)=0. \tag{6.2}$$

The orthogonal weight gradient is conditionally Gaussian with covariance

$$\Sigma_g(\nu)=\int g(a,h,y)g(a,h,y)^\top\,\mathrm{d}\nu,\qquad g(a,h,y)=\frac{2}{p}\,s(a,h,y)Ah. \tag{6.3}$$

Its density at zero contributes $-\frac{1}{2}\log\det\Sigma_g(\nu)$ to the annealed complexity.

The output variables have fixed dimension and therefore do not add an order-d geometric entropy. They alter the admissible saddle through (6.1) and alter every Hessian outlier through the finite kernel (5.3). Conditional on existence of a regular saddle law $\nu_{e,a,Q,M}$ and on a gradient-conditioned version of Assumption 5.3, the energy-resolved spectrum is therefore

$$(e,a,Q,M)\longmapsto(\rho_{e,a,Q,M},\mathcal{K}_{e,a,Q,M}), \tag{6.4}$$

where the density comes from the conditioned analogue of (4.27) and all output, weight, and mixed outliers come from the single finite determinant (5.4). Establishing the continuum saddle, determinant/index asymptotics, and the correlated local law is outside the exact reduction and remains a proof target, as in the frozen-output multi-index problem [4].

If the gradient conditioning could be shown to preserve an asymptotic block-Wishart law with matrix-weight distribution $\nu_{e,a,Q,M}$, then Theorem 4.3 would identify both the Hessian bulk and its normalized log-determinant through (4.25). This implication is exact once that conditional law is proved; the fresh-sample theorem alone does not prove it.

7 Finite-dimensional diagnostics

7.1 Direct test of the block-Wishart formulas

We first test the part of the theory that is unconditional under compact support. We use the saturated link (4.13) with $L = 2$, $p = k = 3$, $\alpha = 4$, output coefficients $(0.72, 1, 1.28)$, and teacher strengths $\mu_j = j^{-0.75}$. The joint Gaussian law of (h, y) has $Q = I_3$ and

$$M = \mathrm{diag}(\sqrt{0.28}, \sqrt{0.52}, \sqrt{0.76}). \tag{7.1}$$

For each orthogonal dimension $d_0 \in \{36, 72, 120\}$ and each of three seeds, we draw $n = 4d_0$ matrix weights T_i and independent Gaussian orthogonal vectors ξ_i, then form (4.16). The theoretical density is computed from (4.19) with imaginary regulator $\eta = 0.035$. Each regular left edge is computed by solving

$$\mathscr{K}(S) = \zeta I_3, \qquad \lambda_{\min}(D\mathscr{K}(S)) = 0, \tag{7.2}$$

on the positive branch. The right edge is the reflected left edge for $T_i \mapsto -T_i$. The empirical extreme is used only to initialize the nonlinear solver and does not enter either equation. The logarithmic potential is evaluated from (4.25) to the left of the support.

The analytic gradient and full Hessian of the saturated model were checked against centered directional differences. Their relative discrepancies are $5.67 \cdot 10^{-11}$ and $2.66 \cdot 10^{-8}$, respectively. Across all nine spectral runs, the largest K-equation residual is $8.02 \cdot 10^{-16}$ and the largest edge-stability residual is $9.18 \cdot 10^{-15}$. At $d_0 = 120$, the Cauchy-smoothed density has relative L^1 error 0.0297. The mean absolute errors of the left and right edges are 0.1634 and 0.0834, while the mean log-potential error is 0.00174. For comparison, the corresponding mean errors at $d_0 = 36$ are 0.2920, 0.1299, and 0.00729. The edge errors are larger because they compare one fluctuating extreme eigenvalue with a deterministic edge; the density and logarithmic potential average over the complete spectrum.

7.2 Joint output–feature dynamics

We integrate the exact population flow with a fourth-order Runge–Kutta scheme and form the full Hessian from an independent Gaussian sample. The teacher has $k = 3$ power-law modes $\mu_i = i^{-0.75}$, the aspect ratio is $n/d = 6$, and the initial output coefficients are lognormal with log-standard deviation 0.8, normalized to mean one. We compare $p \in \{3, 4, 6\}$ at $d = 28$ over three independent seeds. The raw integration time is 2.1, with 60 steps and seven spectral snapshots. This is a finite diagnostic, not an asymptotic estimator.

At each snapshot, the empirical reference interval is the interval between the 2% and 98% quantiles of the weight–weight eigenvalues, enlarged on both sides by 4% of its width. A full-Hessian eigenvalue outside this interval is classified from its normalized eigenvector. It is called output-dominated when the output mass is at least 0.45 and the teacher-weight mass is below 0.20, or when its output mass is at least 0.30 and no stronger label applies. The symmetric rule labels a teacher-weight branch. Mass at least 0.20 in both sectors gives a mixed label. These fixed thresholds define a reproducible finite-size diagnostic; they are not used in Theorem 5.4, where attribution is a limiting residue statement.

The analytic Hessian was checked against centered second differences along a random joint direction. The mean relative discrepancy was $2.5 \cdot 10^{-6}$, $3.8 \cdot 10^{-6}$, and $1.7 \cdot 10^{-4}$ for $p = 3, 4, 6$, respectively. At initialization, the mean numbers of output-dominated outliers were 3.0, 3.67, and 6.0. At the final snapshot all three means were zero; weight-dominated outliers remained, with means 1.67, 2.33, and 2.67. For $p = 6$, output-dominated branches persisted for more snapshots before entering the empirical weight-bulk interval.

At the final time, the mean positive participation numbers were 2.63 ± 0.17, 3.42 ± 0.20, and 5.35 ± 0.38. The corresponding minimum captured fractions were 0.15 ± 0.19, 0.32 ± 0.21, and

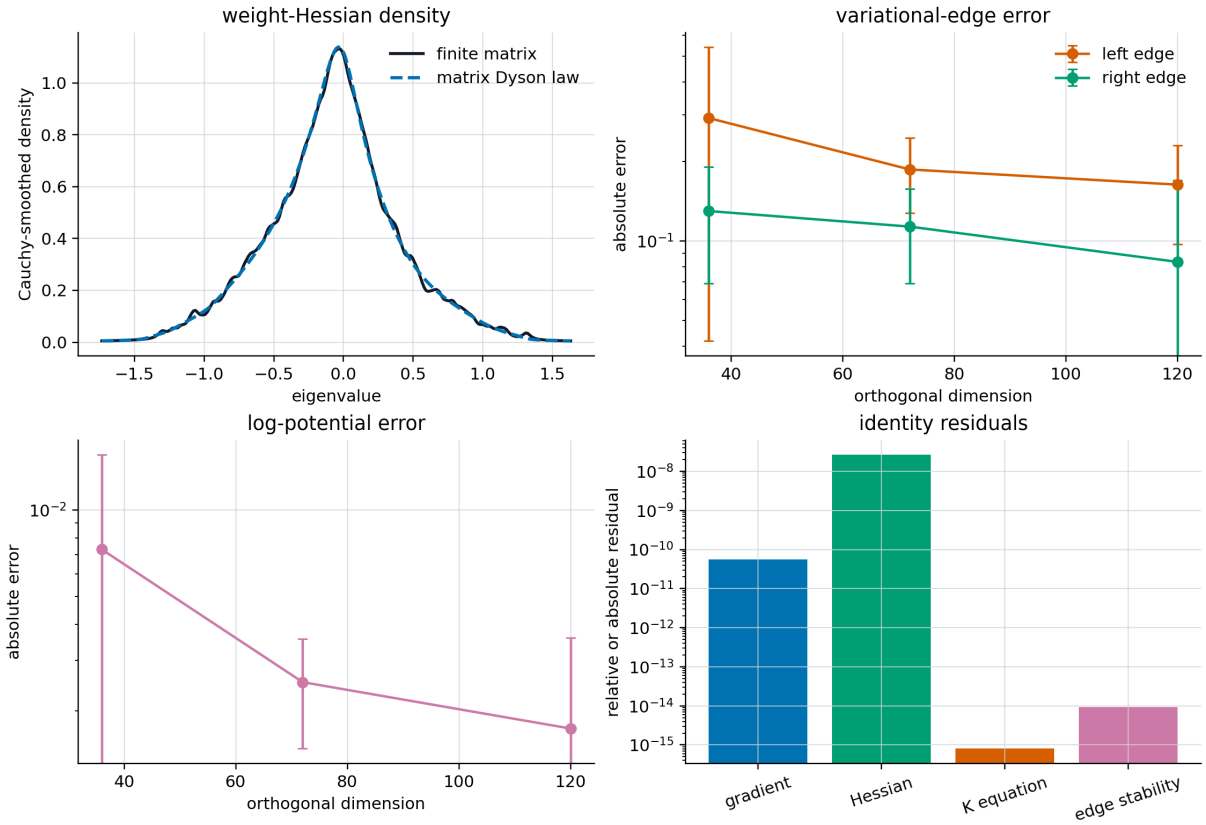

Figure 1: Direct validation of the block-Wishart weight-spectrum formulas. Top left: finite Cauchy-smoothed density and the matrix-Dyson prediction at $d_0 = 120$. Top right: absolute errors of the two variational edges; points and bars are seed means and one standard deviation. Bottom left: error of the variational logarithmic potential. Bottom right: directional derivative and edge-equation residuals. The experiment uses bounded matrix weights and therefore tests Theorem 4.3 under its stated compact-support hypothesis.

0.27 ± 0.19. Thus $p = 6$ has the largest output participation but does not dominate $p = 4$ at this raw time. This observation supports the weaker conclusion that p alone is not a sufficient control variable. It does not establish Conjecture 3.2: that test requires dimension scaling, a clock-matched design, more seeds, and direct comparison with the deterministic joint Schur kernel.

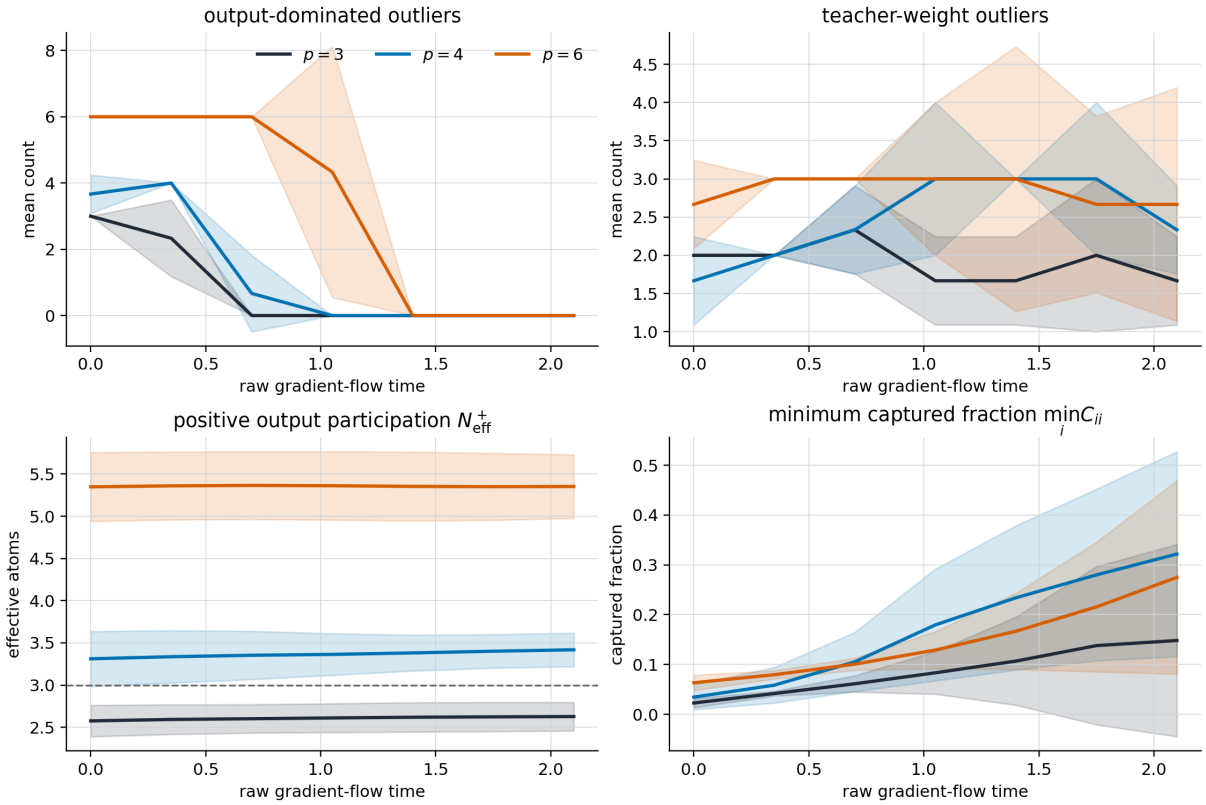

Figure 2: Joint output–weight diagnostics over three seeds. Curves show seed means and shaded bands show one standard deviation. Output-dominated branches are transient, whereas teacher-weight branches remain visible. The positive output participation grows with nominal width, but this does not monotonically improve finite-time capture at a fixed raw clock.

8 Discussion and proof frontier

Trainable output coefficients do not merely append p harmless coordinates. They change the finite state of the population dynamics, break the autonomous captured-subspace equation, and enlarge the signal Schur sector. The correct dynamic spectral object is consequently the pair consisting of the weight-bulk Dyson equation and the joint finite kernel on $(\mathrm{span}(W,\Theta)\otimes\mathbb{R}^p)\oplus\mathbb{R}^p$.

Five proof obligations remain before the complete quadratic asymptotic picture becomes unconditional. First, the passage from the bounded link to $\varphi(u) = u^2$ requires a uniform tail-and-stability theorem. Second, the mixed empirical vectors in (4.5) require an anisotropic deterministic equivalent jointly with the weighted covariance bulk. Third, convergence must hold for the spectral derivative in order to identify residues, not only eigenvalue locations. Fourth, a same-sample trajectory requires leave-one-out transport uniform over time. Fifth, the effective-slack conjecture requires a non-degeneracy theorem for the joint finite kernel; rank counting alone cannot provide it. Energy-conditioned Kac–Rice adds the separate task of proving that the conditioned orthogonal design has the block-Wishart limit used by the log-potential formula.

The next decisive experiment is therefore not simply a larger width sweep. It is a clock-matched dimension-and-seed scaling study that computes the deterministic joint kernel, follows its labelled roots, and compares their output and teacher-weight residues with the empirical eigenvectors. That study would directly test the missing probabilistic theorem and distinguish genuine mixed contacts from finite-size ambiguity.

A Frozen-output comparison

If $a_r = 1$ is imposed and the output coefficients are not trained, the population flow reduces to

$$\dot{W} = -\frac{2}{p}(2E + \tau I)W.$$

In that model

$$\dot{C} = \frac{4}{p}(\Lambda C + C\Lambda - 2C\Lambda C), \tag{A.1}$$

and, when $C(0)$ is invertible,

$$C(t) = \left[I_k + e^{-4\Lambda t/p}(C(0)^{-1} - I_k)e^{-4\Lambda t/p}\right]^{-1}. \tag{A.2}$$

This formula is not valid for the joint flow unless its extra amplitude terms happen to cancel on a special invariant manifold.

B A two-block output Schur formula

If one first treats the entire weight block as the reference matrix, then for $z \notin \mathrm{spec}(H^{WW})$ the output-only Schur function is

$$K_d^a(z) = H^{aa} - H^{aW}(H^{WW} - zI)^{-1}H^{Wa}. \tag{B.1}$$

The roots solve $\det(K_d^a(z) - zI_p) = 0$, and an output kernel vector c_a has output mass

$$\Omega_a = \frac{1}{c_a^\top(I - \partial_z K_d^a(z))c_a}. \tag{B.2}$$

This formula is useful away from weight-block outliers. Near a weight-teacher branch, the joint finite sector of Theorem 5.1 is the regular formulation because it does not hide a pole of the reference resolvent.

References

[1] J. Baik, G. Ben Arous and S. Péché, Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices, *Annals of Probability* **33** (2005), 1643–1697.

[2] 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.

[3] F. Benaych-Georges and R. R. Nadakuditi, The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices, *Advances in Mathematics* **227** (2011), 494–521.

[4] A. Maillard, T. Bonnaire and G. Biroli, Topological exploration of high-dimensional empirical risk landscapes: general approach, and applications to phase retrieval, arXiv:2602.17779, 2026.

[5] A. Montanari and B. Saeed, Variational formulas for the spectrum of block Wishart matrices, arXiv:2606.27774, 2026.