# Sequential Feature Recovery in a Three-Layer Network: Exact Hermite Geometry, a Conditional DMFT–RMT–BBP Closure, and Muon Experiments

Janis Heran Aiad
janis.aiad@polytechnique.edu

July 26, 2026

**Abstract**

We develop a proof-level bridge between hierarchical feature learning, spectral optimization, dynamical mean-field theory, and random-matrix transitions in a three-layer quadratic network. The model composes two quadratic Hermite feature maps and therefore produces a degree-four polynomial of a latent Gaussian subspace. At fixed latent dimension we prove an exact population closure: the risk is a weighted Euclidean distance between multivariate Hermite coefficients, the parameter gradient is the pullback of this coefficient error, and the Hessian decomposes into a positive semidefinite Gauss–Newton term plus residual curvature. Residualized degree-2, 3, and 4 coefficient Jacobians yield positive semidefinite Schur matrices that quantify access to successively new feature sectors.

We then analyze a singular-value power family containing polar Muon (SignSVD) and prove that every nonzero update in the tested family is a strict first-order descent direction, while emphasizing that no universal power is implied for the three trainable matrix blocks. We also derive the exact Fréchet derivative of its nonseparable spectral channel. This is the input required by a matrix Onsager correction: ordinary Muon-$a$ is AMP-like as a denoiser, but the tested optimizer is not AMP because it has no Onsager reaction. To connect the finite population theory to empirical spectra, we embed a reduced state in ambient dimension and evaluate a fresh Gaussian Hessian. Conditional on the latent coordinates, its orthogonal direct-weight block is *exactly* matrix weighted block Wishart. This identity permits a rigorous import of the bounded-block global law and variational support-edge formulas of Montanari and Saeed; an exact finite Schur determinant isolates the remaining signal sector. We state separately the additional local-law, leave-one-out, and tail conditions needed for a genuine dynamic BBP theorem and for a reused-sample DMFT.

Beyond the polynomial testbed, we prove a Hermite-tail transfer theorem: degree-$R$ truncation errors in the coefficient map and its first two derivatives give explicit errors for population risk, gradient, and Hessian. This supplies a controlled extension to general links while making clear that data reuse, growing ranks, and unbounded empirical blocks still require model-specific analysis. We assemble these ingredients into a conditional DMFT–RMT–BBP closure whose four hypotheses are stated individually, and separate the MSRJD, dynamic-cavity, replica, and spectral-AMP routes. We also distinguish population, fresh-sample, same-data trajectory, and critical-point Hessian laws, placing solvable transient-BBP and BBP–Kac–Rice results as controls rather than silently importing them.

A pre-specified confirmatory campaign compares gradient descent, polar Muon, and two transferred singular-value powers on twenty paired datasets. It finds no endpoint dominance: the spectral rules win 9/20, 8/20, and 10/20 pairs, and every paired bootstrap interval for the median log-risk difference contains zero. Their sector clocks are nevertheless earlier: all three halve degree two at Kaplan–Meier median step 80, versus 160 for gradient descent, and move the degree-four median from 480 to 240–320, without uniformly improving terminal

event coverage. A nested sample-size panel and fresh-Hessian diagnostics remain exploratory. The paper thus supplies an exact deterministic core, a theorem-compatible random-matrix lift, an empirical speed–coverage tradeoff, and a falsifiable roadmap for the still-open joint three-layer DMFT.

**Keywords.** Hierarchical learning; three-layer neural networks; Hermite expansion; Muon; SignSVD; dynamical mean-field theory; block-Wishart matrices; BBP transition; approximate message passing; cavity method; replica method; sequential feature recovery; scaling laws.

**Proof-status convention.** Unqualified theorems and propositions are proved in this paper. Results imported from another source are labeled "imported corollary" and retain the source assumptions. Statements requiring a new high-dimensional argument are labeled assumption, conditional proposition, or conjecture. This convention is essential because the deterministic population model, the fresh-sample random-matrix lift, and reused-data training are different probability spaces.

## Contents

# 1 Introduction

Depth can be statistically useful only if training converts composition into a sequence of simpler estimation problems. This principle is explicit in recent theory for hierarchical Gaussian targets: controlled gradient descent reduces effective dimension one layer at a time [14]; layerwise three-layer procedures first recover a nonlinear feature and then fit its outer link [29, 37]; and a solvable spectral model turns sharp, feature-wise transitions into an aggregate power-law learning curve [38]. These works explain why depth may help, but they do not determine the joint dynamics of all matrices in a three-layer network trained by a spectral optimizer.

The optimizer creates a second gap. Muon and related matrix updates act on singular values rather than coordinatewise gradient entries. In a tractable matrix regression model, Paquette et al. [30] show how SignSVD changes the effective preconditioning and can outperform SignSGD in anisotropic regimes. For a compositional network, however, the first feature matrix, the second direct matrix, and the feature-mixing matrix have different dimensions and different gradient spectra. Transferring a scalar Muon exponent from a two-layer model is therefore a hypothesis, not a theorem.

A third gap concerns the Hessian. In multi-index problems, extensive finite-sample Hessian blocks are matrix weighted sample-covariance matrices. The global law and variational support-edge formulas for block-Wishart matrices are now available [24]. A support edge is not yet a BBP theorem: outlier locations and eigenvector overlaps require anisotropic resolvent control, and a training Hessian evaluated on reused observations requires a dynamic leave-one-out argument. Polynomial activations introduce an additional tail problem because the matrix weights are unbounded.

The surrounding literature also studies genuinely different ensembles. Classical BBP and finite-rank deformation theory concern a prescribed spike and random bath [5, 7]. Kac–Rice theory counts critical points conditioned on energy and overlaps [22, 25], while dynamic studies follow a Hessian or another spectral observable along gradient flow [11, 13]. None of these laws is automatically the same as a fresh Hessian conditional on a trained state. We make the probability-space dictionary explicit in subsection 7.8, rather than using "BBP" as a visual synonym for any moving extreme eigenvalue.

This paper places those three gaps in one model while keeping their proof levels separate. We study a two-hidden-block quadratic residual architecture, called three-layer because the trainable maps are the first feature map, the second feature map, and the scalar readout. In latent teacher coordinates $z \in \mathbb{R}^k$, the network is

$$\begin{aligned} h &= Vz + b, & q &= \mathrm{He}_2(h) = h^{\odot 2} - \mathbf{1}, \\ r &= Uz + Aq + c, & s &= \mathrm{He}_2(r) = r^{\odot 2} - \mathbf{1}, \\ f_\theta(z) &= a_0 + \alpha^\mathsf{T} q + \beta^\mathsf{T} s. \end{aligned} \tag{1}$$

It is nonlinear in the parameters but only degree four in $z$. This fact gives an exact finite-dimensional population theory without invoking a width or ambient-dimension limit.

## 1.1 Contributions

Our contributions are the following.

**Exact deterministic population theory.** We prove that (1) has a finite multivariate Hermite expansion through degree four. For a degree-four teacher, population risk, gradient flow, and the complete Hessian are exact algebraic functions of the coefficient map. Tensor Gauss–Hermite quadrature with at least five nodes per coordinate evaluates every coefficient and the risk exactly. This is stronger than a Monte Carlo approximation but should not be called a high-dimensional DMFT.

**A rigorous sectoral hierarchy.** Weighted degree-wise coefficient Jacobians define Gram blocks for degrees $2, 3, 4$. We identify their Schur complements with squared residualized Jacobians. These matrices are always positive semidefinite and measure the instantaneous parameter directions that a new Hermite sector adds after the earlier sectors have been projected out. Combining them with the exact residual-curvature Hessian gives a checkable lower bound on active curvature.

**Muon geometry without an unjustified exponent.** For a matrix gradient $G$, we analyze the same normalized singular-value power map used in the code. Polar Muon is the $a = 0$ member, while $a = 1$ recovers the gradient singular-value profile up to normalization. We prove a strict first-order descent identity and a sufficient smooth-step bound for all $a \geq 0$. The result licenses the family as an optimizer; it does not select one common $a$ for $V, U, A$. We derive its exact matrix Fréchet derivative, which exposes the response contraction required by an Onsager term.

**Exact fresh-sample block-Wishart lift.** After embedding a reduced state in $\mathbb{R}^d$, we derive the analytic preactivation Hessian. On a fresh Gaussian sample, the orthogonal direct-weight block is conditionally

$$\frac{1}{n}\sum_{\mu=1}^{n} T_\theta(z_\mu) \otimes \xi_\mu \xi_\mu^{\mathsf{T}},$$

with $\xi_\mu$ independent of the fixed-size self-adjoint block $T_\theta(z_\mu)$. We may therefore apply the Montanari–Saeed global block-Wishart theory after bounded spectral clipping. An exact finite Schur determinant identifies the equation that would become a dynamic BBP condition once an anisotropic local law is proved.

**A DMFT and scaling-law matching program.** We formulate the minimum two-time covariance and response kernels required by joint, reused-data training. The frozen-feature specialization matches the random-feature DMFT of Kramp et al. [20]; the population anisotropic limit matches the three-phase Volterra perspective of Braun et al. [12]; and the feature-wise sample threshold matches the sequential-recovery mechanism of Wortsman-Zurich et al. [38]. The joint closure itself is explicitly stated as a conjecture. We distinguish its MSRJD, dynamic-cavity, replica, and spectral-AMP formulations. In particular, Muon-$a$ supplies a nonseparable spectral channel but becomes an AMP algorithm only after a self-consistent matrix Onsager reaction is derived.

**Controlled extension beyond quadratic activations.** For an arbitrary twice differentiable coefficient map into Gaussian $L^2$, we prove nonasymptotic Hermite-tail bounds for population risk, gradient, and Hessian. The quadratic model is the zero-tail case. This separates the algebra that extends to smooth links from the local laws, moment estimates, and reused-data closure that must be reproved for each high-dimensional model.

**Reproducible empirical stress test.** We compare gradient descent, exact polar Muon, and powers $1/7$ and $1/3$ under paired initializations and datasets. We evaluate population risk exactly and keep raw polynomial Hessians separate from clipped, theorem-compatible ones. A three-pair pilot suggests an endpoint gain, but a pre-specified twenty-pair confirmation does not reproduce it. Instead it reveals earlier sector half-error clocks with no uniform improvement in event coverage or terminal risk. The fixed latent rank remains too small for a scaling exponent, and the spectral panel is exploratory.

### 1.2 Relationship to the two-layer theory

The two-layer phase-retrieval program closes on low-rank overlap matrices and, under anisotropy, on a Volterra/Stieltjes hierarchy. In the present model there are three distinct obstructions to copying that calculation.

1. The feature basis itself changes: $q$ depends on $V$, while $s$ depends jointly on $V, U, A$.
2. The coefficient map includes degrees $2, 3, 4$, so the active tangent space grows by residualized sectors rather than a single overlap.
3. The empirical direct-weight Hessian has block size $p + q$ and is coupled through a finite feature-mixing/readout sector.

Our exact coefficient and Schur formulas are the three-layer replacements for the two-layer overlap closure. The fresh block-Wishart identity is the three-layer replacement for the scalar weighted-Wishart bulk. What remains open is precisely the joint dynamic local law, rather than an unspecified "DMFT correction."

### 1.3 Organization

section 2 defines the model and probability spaces. section 3 gives the exact population closure. section 4 develops the sectoral Schur hierarchy. section 5 analyzes the spectral update. section 6 proves the fresh-sample block-Wishart identity and finite Schur reduction. section 7 states the high-dimensional matching program. section 8 proves the Hermite-tail transfer and states the conditional DMFT–RMT–BBP closure. section 9 reports the checked-in experiments, and section 10 gives the proof-status audit and open problems.

## 2 Model, notation, and three probability spaces

### 2.1 Multivariate Hermite basis

Let $Z = (Z_1, \dots, Z_k) \sim \mathcal{N}(0, \mathrm{I}_k)$. We use the probabilists' Hermite polynomials, determined by

$$\exp(tx - t^2/2) = \sum_{m \geq 0} \mathrm{He}_m(x) \frac{t^m}{m!}.$$

For a multi-index $\gamma = (\gamma_1, \dots, \gamma_k) \in \mathbb{N}^k$, write

$$|\gamma| = \sum_{j=1}^{k} \gamma_j, \qquad \gamma! = \prod_{j=1}^{k} \gamma_j!, \qquad \mathrm{He}_\gamma(z) = \prod_{j=1}^{k} \mathrm{He}_{\gamma_j}(z_j).$$

Gaussian orthogonality gives

$$\mathbb{E}[\mathrm{He}_\gamma(Z)\mathrm{He}_\eta(Z)] = \gamma! \, \mathbf{1}_{\{\gamma=\eta\}}. \tag{2}$$

If $g \in L^2(\mathcal{N}(0, \mathrm{I}_k))$, its coefficient convention is

$$c_\gamma(g) = \frac{1}{\gamma!} \mathbb{E}[g(Z)\mathrm{He}_\gamma(Z)]. \tag{3}$$

### 2.2 Three-layer quadratic residual network

Fix latent dimension $k$, first width $p$, and second width $q$. The parameters are

$$\theta = (V, b, U, A, c, \alpha, \beta, a_0)$$

with

$$V \in \mathbb{R}^{p\times k}, \quad b, \alpha \in \mathbb{R}^p, \quad U \in \mathbb{R}^{q\times k}, \quad A \in \mathbb{R}^{q\times p}, \quad c, \beta \in \mathbb{R}^q, \quad a_0 \in \mathbb{R}.$$

For $z \in \mathbb{R}^k$, define

$$h_\theta(z) = Vz + b, \qquad q_\theta(z) = h_\theta(z)^{\odot 2} - \mathbf{1}_p, \tag{4}$$

$$r_\theta(z) = Uz + Aq_\theta(z) + c, \qquad s_\theta(z) = r_\theta(z)^{\odot 2} - \mathbf{1}_q, \tag{5}$$

$$f_\theta(z) = a_0 + \alpha^\mathsf{T} q_\theta(z) + \beta^\mathsf{T} s_\theta(z). \tag{6}$$

The untied model treats $U$ and $V$ independently. A tied control may set $U = V$ when $p = q$, but none of the exact coefficient identities requires tying.

The teacher is

$$f_\star(z) = \sum_{|\gamma|\le 4} c_{\star,\gamma} \mathrm{He}_\gamma(z). \tag{7}$$

The experiments set degrees $0, 1$ to zero and include diagonal and mixed coefficients in degrees $2, 3, 4$. We keep the general degree-four notation because biases can generate lower student sectors.

**Definition 2.1** (Population risk)**.** The deterministic population objective is

$$\mathcal{R}(\theta) = \frac{1}{2}\mathbb{E}\big[(f_\theta(Z) - f_\star(Z))^2\big]. \tag{8}$$

### 2.3 Parameter and coefficient coordinates

Let

$$\Gamma_4 = \{\gamma \in \mathbb{N}^k : |\gamma| \le 4\}, \qquad N_4 = |\Gamma_4| = \binom{k+4}{4}.$$

Order $\Gamma_4$ once and identify the coefficient map

$$c(\theta) = (c_\gamma(f_\theta))_{\gamma\in\Gamma_4} \in \mathbb{R}^{N_4}.$$

Set

$$D = \mathrm{Diag}((\gamma!)_{\gamma\in\Gamma_4}), \qquad e(\theta) = c(\theta) - c_\star, \qquad J(\theta) = \nabla_\theta c(\theta) \in \mathbb{R}^{N_4\times P},$$

where $P = pk + p + qk + qp + q + p + q + 1$ in the untied model. The orientation is chosen so that $J_{\gamma,j} = \partial c_\gamma/\partial\theta_j$.

For degree $r$, let $I_r = \{\gamma \in \Gamma_4 : |\gamma| = r\}$, let $J_r$ be the corresponding rows of $J$, and let $D_r$ be the restriction of $D$. The weighted parameter-space Jacobian is

$$M_r = J_r^\mathsf{T} D_r^{1/2} \in \mathbb{R}^{P\times|I_r|}. \tag{9}$$

### 2.4 Population, fixed-sample training, and fresh-sample spectroscopy

We distinguish three experiments.

**Population dynamics.** The expectation in (8) is evaluated exactly and the parameter trajectory is deterministic after initialization.

**Fixed-sample training.** Given a training set $\mathcal{D}_n = \{(z_\mu, y_\mu)\}_{\mu=1}^n$, the empirical risk is

$$\widehat{\mathcal{R}}_n(\theta) = \frac{1}{2n}\sum_{\mu=1}^{n}(f_\theta(z_\mu) - y_\mu)^2. \tag{10}$$

The same sample is reused over optimization steps. This creates temporal dependence and is the setting that ultimately needs a joint DMFT or dynamic leave-one-out proof.

**Fresh-sample spectroscopy.** At a fixed checkpoint $\theta_t$, draw observations independent of the training history and form a Hessian. Conditional on $\theta_t$, this is a static random-matrix problem. The fresh construction is what makes the exact block-Wishart reduction in section 6 possible.

**Remark 2.2** (Why the distinction matters)**.** A population ODE cannot prove finite-sample generalization. A fresh-sample global spectral law cannot prove the spectrum of a reused-data training Hessian. A support-edge theorem cannot by itself prove a BBP outlier or its eigenvector overlap. We use a separate notation and claim level for each object.

### 2.5 Ambient embedding

For the random-matrix lift, fix $d > k$ and an isometry $\Theta \in \mathbb{R}^{d\times k}$, $\Theta^\mathsf{T}\Theta = \mathrm{I}_k$. Complete it to an orthogonal matrix $[\Theta, \Theta_\perp]$. Draw

$$X = \Theta Z + \Theta_\perp \Xi, \qquad Z \sim \mathcal{N}(0, \mathrm{I}_k), \quad \Xi \sim \mathcal{N}(0, \mathrm{I}_{d-k}), \quad Z \perp \Xi.$$

Embed the direct matrices by

$$W_1 = V\Theta^\mathsf{T}, \qquad W_2 = U\Theta^\mathsf{T}.$$

Then $W_1 X = VZ$ and $W_2 X = UZ$; the network output depends on $X$ only through $Z$, while the direct-weight Hessian still has $d$ columns per hidden unit. This split produces an extensive orthogonal bulk and a fixed-dimensional signal sector.

## 3 Exact Hermite population theory

**Proposition 3.1** (Finite Hermite support)**.** *For every parameter $\theta$, the function $f_\theta$ in (6) is a polynomial in $z$ of total degree at most four. Consequently,*

$$f_\theta(z) = \sum_{\gamma\in\Gamma_4} c_\gamma(\theta)\mathrm{He}_\gamma(z)$$

*as an exact polynomial identity.*

The point is elementary but decisive. The first feature $q_\theta$ has degree two, $r_\theta$ has degree two, and squaring $r_\theta$ creates degrees at most four. No uncontrolled Hermite tail exists in the reduced model.

**Theorem 3.2** (Exact coefficient risk, flow, and Hessian)**.** *Let $f_\star$ have degree at most four. Then*

$$\mathcal{R}(\theta) = \frac{1}{2}e(\theta)^\mathsf{T} D e(\theta), \tag{11}$$

$$\nabla_\theta \mathcal{R}(\theta) = J(\theta)^\mathsf{T} D e(\theta), \tag{12}$$

$$\nabla_\theta^2 \mathcal{R}(\theta) = J(\theta)^\mathsf{T} D J(\theta) + \sum_{\gamma\in\Gamma_4} \gamma!\, e_\gamma(\theta)\nabla_\theta^2 c_\gamma(\theta). \tag{13}$$

*Under population gradient flow $\dot{\theta}_t = -\nabla_\theta \mathcal{R}(\theta_t)$, the coefficient error obeys the exact, generally non-autonomous equation*

$$\dot{e}_t = -J_t J_t^\mathsf{T} D e_t. \tag{14}$$

*Equivalently, for the weighted error $u_t = D^{1/2} e_t$,*

$$\dot{u}_t = -B_t u_t, \qquad B_t = D^{1/2} J_t J_t^\mathsf{T} D^{1/2} \succeq 0. \tag{15}$$

*In particular,*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}(\theta_t) = -\|\nabla_\theta \mathcal{R}(\theta_t)\|_2^2 = -u_t^\mathsf{T} B_t u_t \le 0. \tag{16}$$

We write

$$H_{\mathrm{GN}} = J^\mathsf{T} D J = \sum_{r=0}^{4} M_r M_r^\mathsf{T}, \qquad H_{\mathrm{res}} = \sum_{\gamma \in \Gamma_4} \gamma!\, e_\gamma \nabla_\theta^2 c_\gamma, \tag{17}$$

so $H = H_{\mathrm{GN}} + H_{\mathrm{res}}$. The first term is positive semidefinite. The second term is the source of negative curvature away from interpolation and vanishes whenever the entire coefficient error vanishes.

**Corollary 3.3** (Geometry at interpolation)**.** *If $c(\theta) = c_\star$, then*

$$\nabla_\theta^2 \mathcal{R}(\theta) = J(\theta)^\mathsf{T} D J(\theta) \succeq 0.$$

*Its nonzero eigenvalues agree, including multiplicity, with those of $B(\theta) = D^{1/2} J J^\mathsf{T} D^{1/2}$.*

**Remark 3.4** (Closure in parameters versus coefficients)**.** Equation (14) is exact, but it is not autonomous in $e_t$ because $J_t$ depends on the current factorization $(V_t, U_t, A_t, \alpha_t, \beta_t, \ldots)$. Calling (14) a "DMFT" would hide precisely the feature-learning problem. It is a finite deterministic pullback flow.

### 3.1 Exact Gauss–Hermite evaluation

Let $\{x_i, w_i\}_{i=1}^{n_q}$ be the one-dimensional Gauss–Hermite rule scaled to $\mathcal{N}(0,1)$, and use its $k$-fold tensor product.

**Proposition 3.5** (Quadrature exactness)**.** *If $n_q \ge 5$, the tensor Gauss–Hermite rule evaluates exactly:*

*(i) every coefficient $c_\gamma(\theta)$, $|\gamma| \le 4$;*

*(ii) the population risk $\mathcal{R}(\theta)$;*

*(iii) every derivative obtained by differentiating these finite polynomial sums with respect to $\theta$.*

Indeed, $f_\theta \mathrm{He}_\gamma$ and $(f_\theta - f_\star)^2$ have coordinatewise degree at most eight, while an $n_q$-point Gaussian rule integrates degree $2n_q - 1$. The implementation uses $n_q = 7$, giving a safety margin through degree thirteen.

### 3.2 Degree-wise error observables

Define

$$E_r(\theta) = \sum_{|\gamma|=r} \gamma!\, e_\gamma(\theta)^2, \qquad \mathcal{R}(\theta) = \frac{1}{2}\sum_{r=0}^{4} E_r(\theta). \tag{18}$$

These observables are exact and orthogonal in function space. They are not independent under training because the coefficient kernel $B_t$ has off-diagonal degree blocks. The next section uses Schur residualization to separate genuinely new sector directions from parameter directions already used by earlier degrees.

### 3.3 Relation to general-link and alternating theories

For a general multi-index link $g(Wz)$, Hermite projection gives the same risk identity but an infinite coefficient hierarchy unless the link is polynomial. The present architecture is therefore a finite exact testbed for the general-link closure. Joint direction/link convergence in the single-index setting can be proved for an alternating RKHS scheme [31]; our simultaneous factorized flow is different, but (14) makes its additional coupling explicit.

## 4 Residualized Hermite sectors and sequential recovery

The target-bearing hierarchy is degree $2 \to 3 \to 4$. For $r \in \{2, 3, 4\}$, concatenate the earlier weighted Jacobians

$$M_{<r} = [M_2, \dots, M_{r-1}],$$

with $M_{<2}$ empty, and let

$$P_{<r} = M_{<r} M_{<r}^{\dagger}$$

be the orthogonal projector onto their parameter-space range. Define the new-sector Jacobian

$$\overline{M}_r = (\mathrm{I}_P - P_{<r}) M_r. \tag{19}$$

**Proposition 4.1** (Schur hierarchy)**.** *Let $G_{rs} = M_r^{\mathsf{T}} M_s$ and let $G_{<r,<r}$ denote the block Gram matrix of degrees $2, \dots, r-1$. Then*

$$S_r := G_{rr} - G_{r,<r} G_{<r,<r}^{\dagger} G_{<r,r} = \overline{M}_r^{\mathsf{T}} \overline{M}_r \succeq 0. \tag{20}$$

*In particular,*

$$\begin{aligned} S_2 &= G_{22}, \\ S_3 &= G_{33} - G_{32} G_{22}^{\dagger} G_{23}, \\ S_4 &= G_{44} - \begin{bmatrix} G_{42} & G_{43} \end{bmatrix} \begin{bmatrix} G_{22} & G_{23} \\ G_{32} & G_{33} \end{bmatrix}^{\dagger} \begin{bmatrix} G_{24} \\ G_{34} \end{bmatrix}. \end{aligned} \tag{21}$$

**Remark 4.2** (Interpretation)**.** $S_r$ does not measure the raw sensitivity of degree $r$. It measures the part of that sensitivity that cannot be synthesized from parameter directions already used by earlier target-bearing sectors. A zero eigenvalue can mean either a symmetry/redundancy or a genuine failure to expose a new feature direction.

Let $Q_r$ have orthonormal columns spanning $\mathrm{Ran}(\overline{M}_r)$, and define the projected residual-curvature radius

$$\rho_r(\theta) = \|Q_r^{\mathsf{T}} H_{\mathrm{res}}(\theta) Q_r\|_{\mathrm{op}}. \tag{22}$$

Write $\lambda_{\min}^{+}(S_r)$ for the smallest strictly positive eigenvalue, when it exists.

**Proposition 4.3** (Active curvature certificate)**.** *At every fixed state with $\mathrm{Ran}(\overline{M}_r) \neq \{0\}$,*

$$\lambda_{\min}\Big(Q_r^{\mathsf{T}} \nabla_\theta^2 \mathcal{R}(\theta) Q_r\Big) \geq \lambda_{\min}^{+}(S_r) - \rho_r. \tag{23}$$

*Hence $\lambda_{\min}^{+}(S_r) > \rho_r$ is a sufficient certificate of positive curvature throughout the active degree-$r$ tangent space.*

The certificate separates two mechanisms. $S_r$ is a Gauss–Newton access term: it records whether a parameter perturbation can change the new coefficient sector independently of earlier ones. $\rho_r$ is a non-interpolating correction: it can erase that access through negative residual curvature.

### 4.1 Conditional sector clocks

An instantaneous Schur matrix is not, by itself, a theorem that sectors learn in order. The coefficient equation (15) contains time-dependent cross blocks. The following elementary result states exactly which dynamical decoupling is needed to turn curvature into a clock.

**Proposition 4.4** (Conditional error clock)**.** *Let $u_r = D_r^{1/2} e_r$. Suppose that on an interval $[t_0, t_1]$ the weighted coefficient dynamics restricted to sector $r$ satisfy*

$$\dot{u}_r(t) = -B_{rr}(t) u_r(t) + \zeta_r(t), \qquad B_{rr}(t) \succeq \kappa_r(t) \mathrm{I},$$

*where $\zeta_r$ collects cross-sector forcing. Then*

$$\|u_r(t)\| \le e^{-\int_{t_0}^{t} \kappa_r(s)\mathrm{d}s} \|u_r(t_0)\| + \int_{t_0}^{t} e^{-\int_{s}^{t} \kappa_r(v)\mathrm{d}v} \|\zeta_r(s)\| \mathrm{d}s. \tag{24}$$

*In particular, if $\zeta_r = 0$, the half-error time is reached once $\int_{t_0}^{t} \kappa_r(s)\mathrm{d}s \ge (\log 2)/2$.*

The factor $1/2$ in the last statement appears because $E_r = \|u_r\|^2$. In a staged algorithm, freezing or fitting earlier blocks can make $\zeta_r$ small. In joint backpropagation this must be verified; it cannot be assumed from a plot.

### 4.2 Match to hierarchical depth theory

The controlled mechanism of Dandi et al. [14] reduces the effective latent dimension at successive layers. The layerwise guarantees of Nichani et al. [29], Wang et al. [37] first recover a lower-degree feature and then learn its outer polynomial. In our finite model, the ranks

$$d_r^{\mathrm{new}} = \mathrm{rank}(\overline{M}_r) = \mathrm{rank}(S_r)$$

are local, parameter-dependent analogues of "new effective dimension." This is a mathematical match at the level of tangent geometry, not an import of their sample-complexity theorem.

The experiment tracks $E_2, E_3, E_4$, the Schur spectra, and the projected residual radii. A publishable sequential-recovery claim should demonstrate all three:

1. earlier errors become small before later ones;
2. $\lambda_{\min}^{+}(S_r) - \rho_r$ becomes positive in the same order;
3. the forcing term implicit in (24) is small enough to predict the measured crossing times.

The checked-in pilot establishes the first item descriptively and verifies the Schur algebra, but it does not yet establish a uniform dynamical bound for the third.

## 5 Spectral optimization and Muon geometry

Let $G \in \mathbb{R}^{m \times n}$ be a nonzero matrix gradient and write a thin SVD

$$G = L \,\mathrm{Diag}(\sigma_1, \dots, \sigma_\ell) R^{\mathsf{T}}, \qquad \ell = \min(m, n), \qquad \sigma_i \ge 0.$$

Set $\bar{\sigma} = \ell^{-1} \sum_i \sigma_i > 0$. For $a \ge 0$ and numerical floor $\varepsilon \ge 0$, define

$$w_i^{(a,\varepsilon)} = \max\{\sigma_i/\bar{\sigma}, \varepsilon\}^a, \tag{25}$$

$$c_{a,\varepsilon}(G) = \left( \frac{1}{mn} \sum_{i=1}^{\ell} (w_i^{(a,\varepsilon)})^2 \right)^{1/2},$$

$$\Phi_{a,\varepsilon}(G) = c_{a,\varepsilon}(G)^{-1} L \,\mathrm{Diag}(w_1^{(a,\varepsilon)}, \dots, w_\ell^{(a,\varepsilon)}) R^{\mathsf{T}}. \tag{26}$$

Thus $\|\Phi_{a,\varepsilon}(G)\|_{\mathrm{F}}^2 = mn$, matching the unit-entry-RMS normalization in the implementation. At $a = 0$,

$$\Phi_{0,\varepsilon}(G) = \sqrt{\frac{mn}{\ell}}\, LR^{\mathsf{T}},$$

the normalized polar or SignSVD direction. Ignoring the floor, $a = 1$ preserves the gradient singular-value profile up to a scalar.

**Proposition 5.1** (Muon powers are descent directions)**.** *For every nonzero $G$, $a \geq 0$, and $\varepsilon > 0$,*

$$\langle G, \Phi_{a,\varepsilon}(G)\rangle_{\mathrm{F}} = \frac{\sum_{i=1}^{\ell} \sigma_i w_i^{(a,\varepsilon)}}{c_{a,\varepsilon}(G)} > 0. \tag{27}$$

*Let $F$ have an $L_F$-Lipschitz gradient in a neighborhood of $W$, with $G = \nabla F(W)$. Then*

$$F(W - \eta\Phi_{a,\varepsilon}(G)) \leq F(W) - \eta\langle G, \Phi_{a,\varepsilon}(G)\rangle_{\mathrm{F}} + \frac{L_F\eta^2}{2}mn. \tag{28}$$

*Consequently the update strictly decreases $F$ whenever*

$$0 < \eta < \frac{2\langle G, \Phi_{a,\varepsilon}(G)\rangle_{\mathrm{F}}}{L_F mn}. \tag{29}$$

The experiment uses backtracking, so it enforces actual loss decrease rather than assuming (29) with an unknown local smoothness constant.

**Proposition 5.2** (Fréchet derivative of the spectral channel)**.** *Let $G \in \mathbb{R}^{m\times n}$, $m \leq n$, have full row rank, set $S = GG^{\mathsf{T}} = L\,\mathrm{Diag}(\lambda_1, \ldots, \lambda_m)L^{\mathsf{T}}$, and put $b = (a-1)/2$. Ignoring the numerical floor, define*

$$\Psi_a(G) = S^b G, \qquad \Phi_a(G) = \sqrt{mn}\,\frac{\Psi_a(G)}{\|\Psi_a(G)\|_{\mathrm{F}}}.$$

*For a perturbation $\Delta$, let*

$$E = G\Delta^{\mathsf{T}} + \Delta G^{\mathsf{T}}$$

*and let $F_b^{[1]}$ be the Loewner matrix of $x^b$:*

$$(F_b^{[1]})_{ij} = \begin{cases} \dfrac{\lambda_i^b - \lambda_j^b}{\lambda_i - \lambda_j}, & \lambda_i \neq \lambda_j, \\ b\lambda_i^{b-1}, & \lambda_i = \lambda_j. \end{cases}$$

*Then*

$$D\Psi_a(G)[\Delta] = L\Big[F_b^{[1]} \odot (L^{\mathsf{T}}EL)\Big]\, L^{\mathsf{T}}G + S^b\Delta, \tag{30}$$

$$D\Phi_a(G)[\Delta] = q\, D\Psi_a(G)[\Delta] - q\,\Psi_a(G)\frac{\langle \Psi_a(G), D\Psi_a(G)[\Delta]\rangle_{\mathrm{F}}}{\|\Psi_a(G)\|_{\mathrm{F}}^2}, \tag{31}$$

*where $q = \sqrt{mn}/\|\Psi_a(G)\|_{\mathrm{F}}$.*

Theorem 5.2 is the object needed for a spectral Onsager reaction: an AMP correction uses an averaged Jacobian or response contraction of the denoising channel. The derivative is matrix valued and couples singular directions; replacing it by the scalar derivative of $\sigma \mapsto \sigma^a$ would discard the rotations of the singular vectors and the unit-RMS normalization.

### 5.1 Blockwise update used in the experiments

For $B \in \{V, U, A\}$, let $G_B = \nabla_B F$. A spectral run updates

$$B^+ = B - \eta_B \, \Phi_{a_B,\varepsilon}(G_B). \tag{32}$$

Biases and readout vectors use ordinary gradient descent. The reported constant-power controls impose $a_V = a_U = a_A = a$, with

$$a \in \left\{0, \frac{1}{7}, \frac{1}{3}\right\}.$$

The values $1/7$ and $1/3$ come from distinct normalizations in the power-law phase-retrieval program and are *transfer controls*. They are not derived minimizers of a three-layer objective.

**Remark 5.3** (Different blocks have different spectra)**.** $V$ controls the first quadratic feature, $U$ injects a direct linear coordinate into the second block, and $A$ mixes already nonlinear features. Even when $p = q = k$, their gradients have different factorizations. A theoretically natural control is therefore

$$(a_V(t), a_U(t), a_A(t), \eta_V(t), \eta_U(t), \eta_A(t)),$$

not a single universal exponent.

### 5.2 Why spectral flattening may change sector clocks

Write the degree-$r$ coefficient derivative with respect to block $B$ as $J_{r,B}$. Under ordinary gradient flow, the block contributes

$$B_{rs}^{(B)} = D_r^{1/2} J_{r,B} J_{s,B}^{\mathsf{T}} D_s^{1/2}$$

to the coefficient kernel. Muon is nonlinear in the full gradient, so it does not correspond to replacing $J_{r,B} J_{s,B}^{\mathsf{T}}$ by a fixed preconditioner. Nevertheless, at one state its progress is exactly

$$- \langle \nabla_B F, \Phi_{a_B,\varepsilon}(\nabla_B F) \rangle_{\mathrm{F}}, \tag{33}$$

which weights singular channels by $\sigma_i^{1+a_B}$ up to normalization. Lower $a_B$ flattens the channel speeds and may expose weak late-sector directions sooner; it can also over-amplify noisy or unhelpful channels. This tradeoff explains why the pilot contains both large gains and paired failures.

### 5.3 Relation to the two-layer spectral theory

In matrix least squares, the spectral-optimizer dynamics can be reduced to deterministic spectral observables [30]. In a low-rank two-layer factorization, invariance can reduce Muon to a finite right action. For (6), the gradient of $V$ appears both directly through $q$ and indirectly through $Aq$; the action on $A$ changes the basis seen by the $U$-block. The exact Hermite map remains finite, but the Muon update is not a function of coefficient error alone. This is the precise extra state that a three-layer controlled DMFT must retain.

**Conjecture 5.4** (Block-specific spectral control)**.** *In a proportional growing-width and growing-rank limit, the asymptotically optimal power schedule is block specific and state dependent. Each $a_B(t)$ is determined by a local projection of the corresponding signal-to-noise spectral filter onto the power family $\sigma \mapsto \sigma^{a_B}$.*

This conjecture is falsifiable by measuring useful and bulk singular-vector overlaps separately for $G_V, G_U, G_A$. Endpoint risk alone cannot identify the proposed mechanism.

# 6 Fresh-sample Hessian, block Wishart, and the BBP interface

## 6.1 Analytic preactivation Hessian

At a fixed latent input $z$, regard the two direct preactivations

$$h = Vz + b \in \mathbb{R}^p, \qquad u = Uz \in \mathbb{R}^q$$

as independent variables and write

$$q_h = h^{\odot 2} - \mathbf{1}_p, \qquad r = u + Aq_h + c, \qquad f = a_0 + \alpha^\mathsf{T} q_h + \beta^\mathsf{T}(r^{\odot 2} - \mathbf{1}_q).$$

Define

$$d_h = \alpha + 2A^\mathsf{T}(\beta \odot r), \qquad D_h = \mathrm{Diag}(h), \qquad D_\beta = \mathrm{Diag}(\beta). \tag{34}$$

**Proposition 6.1** (Link gradient and Hessian)**.** *With $v = (h, u) \in \mathbb{R}^{p+q}$, the gradient of $f$ is*

$$g_\theta(z) = \nabla_v f = \begin{bmatrix} 2h \odot d_h \\ 2\beta \odot r \end{bmatrix}. \tag{35}$$

*The Hessian $K_\theta(z) = \nabla_v^2 f$ has blocks*

$$K_{hh} = 2\,\mathrm{Diag}(d_h) + 8D_h A^\mathsf{T} D_\beta A D_h, \tag{36}$$

$$K_{hu} = 4D_h A^\mathsf{T} D_\beta, \tag{37}$$

$$K_{uu} = 2D_\beta, \qquad K_{uh} = K_{hu}^\mathsf{T}. \tag{38}$$

*For square loss $\ell_\theta(z, y) = \frac{1}{2}(f_\theta(z) - y)^2$, its preactivation Hessian is*

$$T_\theta(z, y) := \nabla_v^2 \ell_\theta(z, y) = g_\theta(z) g_\theta(z)^\mathsf{T} + (f_\theta(z) - y) K_\theta(z). \tag{39}$$

## 6.2 Direct-weight Kronecker structure

Stack the ambient direct matrices as

$$W = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} \in \mathbb{R}^{(p+q)\times d}.$$

Since $v = Wx + (b, 0)$, the per-example direct-weight Hessian is, up to the fixed permutation associated with vectorization,

$$\nabla^2_{\mathrm{vec}(W)} \ell_\theta(x, y) = T_\theta(z, y) \otimes xx^\mathsf{T}. \tag{40}$$

Let $r_b = p + q$ denote the block size.

**Theorem 6.2** (Exact fresh orthogonal block-Wishart identity)**.** *Embed a fixed reduced state as in section 2. Draw a fresh independent sample*

$$x_\mu = \Theta z_\mu + \Theta_\perp \xi_\mu, \qquad (z_\mu, \xi_\mu) \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathrm{I}_k) \otimes \mathcal{N}(0, \mathrm{I}_{d-k}),$$

*and set $y_\mu = f_\star(z_\mu)$. Let $H^{(n)}_{WW}$ be the empirical direct-weight Hessian. Its principal block on $\mathbb{R}^{r_b} \otimes \mathrm{Ran}(\Theta_\perp)$ is exactly*

$$H^{(n)}_{\perp\perp} = \frac{1}{n} \sum_{\mu=1}^n T_\theta(z_\mu, y_\mu) \otimes \xi_\mu \xi_\mu^\mathsf{T}. \tag{41}$$

*Conditional on $\{z_\mu\}_{\mu \le n}$, the Gaussian vectors $\{\xi_\mu\}_{\mu \le n}$ are independent of the self-adjoint matrix weights $\{T_\theta(z_\mu, y_\mu)\}_{\mu \le n}$. Hence (41) is a matrix-weighted block-Wishart matrix with fixed block size $r_b$.*

This theorem is an identity, not an asymptotic approximation. Freshness is essential: if the same observations trained $\theta_t$, then $T_{\theta_t}(z_\mu, y_\mu)$ would depend on $\xi_\mu$ through the entire trajectory.

### 6.3 Bounded blocks and imported global spectrum

For $L > 0$, define spectral clipping

$$\Pi_L(T) = Q\,\mathrm{Diag}(\max\{-L, \min\{\lambda_i, L\}\})Q^{\mathsf{T}}$$

when $T = Q\,\mathrm{Diag}(\lambda_i)Q^{\mathsf{T}}$. Let

$$H^{(n)}_{\perp\perp,L} = \frac{1}{n}\sum_{\mu=1}^{n} \Pi_L(T_\mu) \otimes \xi_\mu\xi_\mu^{\mathsf{T}}.$$

**Lemma 6.3** (Deterministic clipping perturbation)**.** *For every realized sample,*

$$\|H^{(n)}_{\perp\perp} - H^{(n)}_{\perp\perp,L}\|_{\mathrm{op}} \le \frac{1}{n}\sum_{\mu=1}^{n} \|T_\mu - \Pi_L(T_\mu)\|_{\mathrm{op}}\,\|\xi_\mu\|_2^2. \tag{42}$$

The bound explains why a small clipping fraction need not imply a small edge perturbation: a rare polynomial block can have both a large excess norm and a large $\|\xi_\mu\|^2$.

Let $\nu_{\theta,L}$ be the law of $\Pi_L(T_\theta(Z, f_\star(Z)))$. For $S \succ 0$, define the matrix $K$-transform

$$\mathbb{K}_{\alpha,\nu}(S) = \int T(\mathrm{I}_{r_b} + ST)^{-1}\nu(\mathrm{d}T) - \alpha^{-1}S^{-1}. \tag{43}$$

Its derivative in a self-adjoint direction $\Delta$ is

$$\begin{aligned} D\mathbb{K}_{\alpha,\nu}(S)[\Delta] &= \alpha^{-1}S^{-1}\Delta S^{-1} \\ &\quad - \int T(\mathrm{I} + ST)^{-1}\Delta(\mathrm{I} + TS)^{-1}T\,\nu(\mathrm{d}T). \end{aligned} \tag{44}$$

Let $\mathcal{P}_-(\alpha,\nu)$ contain those $S \succ 0$ for which $S^{-1} + T \succ 0$ on $\operatorname{supp}\nu$ and $D\mathbb{K}_{\alpha,\nu}(S) \succ 0$ on the self-adjoint matrix space.

**Corollary 6.4** (Imported block-Wishart law and edges)**.** *Assume $r_b$ is fixed, $n/(d-k) \to \alpha \in (0,\infty)$, and use the clipped blocks with fixed $L$. Then the hypotheses of Montanari and Saeed [24] hold. The empirical spectral distribution of $H^{(n)}_{\perp\perp,L}$ converges almost surely to a deterministic compactly supported law $\mu_{\alpha,\nu_{\theta,L}}$. Its matrix Stieltjes coordinate $S(z)$ solves*

$$\mathbb{K}_{\alpha,\nu_{\theta,L}}(S(z)) = z\mathrm{I}_{r_b}. \tag{45}$$

*The left edge is*

$$E_- = \sup_{S\in\mathcal{P}_-(\alpha,\nu_{\theta,L})} \lambda_{\min}\big(\mathbb{K}_{\alpha,\nu_{\theta,L}}(S)\big), \tag{46}$$

*and the right edge follows by reflection $E_+(\alpha,\nu) = -E_-(\alpha,\check{\nu})$, where $\check{\nu} = \mathrm{Law}(-T)$.*

**Remark 6.5** (What is and is not imported)**.** Theorem 6.4 supplies a global law, support convergence, variational edges, and an exterior logarithmic potential. It does not supply the anisotropic local resolvent estimates needed to replace a finite signal-sector quadratic form by its deterministic limit.

### 6.4 Exact finite signal-sector Schur equation

Reorder all Hessian coordinates into

$$\underbrace{\mathbb{R}^{r_b} \otimes \mathrm{Ran}(\Theta_\perp)}_{\text{extensive orthogonal block}} \oplus \underbrace{[\mathbb{R}^{r_b} \otimes \mathrm{Ran}(\Theta)] \oplus \mathbb{R}^{P_{\mathrm{fin}}}}_{\text{finite signal sector}},$$

where $P_{\mathrm{fin}} = qp + 2p + 2q + 1$ contains $(A, b, c, \alpha, \beta, a_0)$. Write the complete empirical Hessian as

$$H_n = \begin{bmatrix} B_n & C_n \\ C_n^{\mathsf{T}} & D_n \end{bmatrix}, \qquad B_n = H^{(n)}_{\perp\perp}. \tag{47}$$

**Proposition 6.6** (Finite Schur determinant)**.** *For every $\lambda \notin \operatorname{spec}(B_n)$,*

$$\det(H_n - \lambda \mathrm{I}) = \det(B_n - \lambda \mathrm{I}) \det\Big[D_n - \lambda \mathrm{I} - C_n^{\mathsf{T}}(B_n - \lambda \mathrm{I})^{-1} C_n\Big]. \tag{48}$$

*Thus every eigenvalue of $H_n$ outside $\operatorname{spec}(B_n)$ is a zero of the finite-dimensional determinant*

$$F_n(\lambda) = \det\Big[D_n - \lambda \mathrm{I} - C_n^{\mathsf{T}}(B_n - \lambda \mathrm{I})^{-1} C_n\Big]. \tag{49}$$

Equation (49) is the exact BBP interface. To obtain a theorem along a trajectory, one must prove that the matrix-valued resolvent quadratic form inside $F_n$ converges uniformly for $\lambda$ outside the moving bulk and that its limiting zeros cross a regular edge transversely. This is the anisotropic, vector-resolvent level of control familiar from local-law theory [19]; a convergent empirical spectral distribution alone cannot supply it.

**Conjecture 6.7** (Dynamic BBP crossing)**.** *Under a fresh-sample protocol, bounded link derivatives, proportional asymptotics, and a regular time-uniform anisotropic local law, $F_n(t, \lambda)$ converges locally uniformly outside the bulk to a deterministic $F(t, \lambda)$. A simple zero of $F(t, \cdot)$ detaches from a regular support edge exactly when the corresponding finite-signal branch satisfies the edge contact and transversality equations.*

### 6.5 Raw polynomial blocks

For the quadratic composition, $T_\theta(Z, f_\star(Z))$ is an unbounded Gaussian polynomial matrix. The clipped corollary therefore cannot be silently applied to the raw Hessian. Removing clipping requires at least:

1. a tail estimate for the maximum block norm on the relevant $n, d$ scale;
2. a comparison showing that discarded blocks do not move a regular bulk edge at the target scale;
3. an anisotropic local law stable under this comparison;
4. a separate analysis if rare blocks create their own extreme-value eigenvalues.

The experiments consequently report raw empirical edges and eigenpairs in separate columns from clipped variational and clipped empirical edges.

## 7 DMFT matching and scaling-law predictions

This section is deliberately split between an exact finite-row identity and a conjectural reused-data mean-field closure.

### 7.1 Exact finite-row representation of the spectral map

**Proposition 7.1** (Muon as a finite Gram action)**.** *Let $G \in \mathbb{R}^{m \times d}$, $m \le d$, have full row rank, and ignore the numerical singular-value floor. Then the unnormalized singular-value power map has the exact representation*

$$L \operatorname{Diag}((\sigma_i/\bar{\sigma})^a) R^{\mathsf{T}} = \bar{\sigma}^{-a} (GG^{\mathsf{T}})^{(a-1)/2} G. \tag{50}$$

*For rank-deficient $G$, the same identity holds on $\operatorname{Ran}(G)$ with the Moore–Penrose functional calculus. The unit-RMS Muon map differs only by a scalar normalization.*

Thus, when $p, q$ remain fixed while $d \to \infty$, Muon on $W_1$ and $W_2$ depends on finite $p \times p$ and $q \times q$ gradient Gram matrices. This is the cleanest match to the finite-overlap reduction in the two-layer theory. The difficulty is not the spectral functional calculus; it is closing the stochastic evolution of those Gram matrices jointly with $A, \alpha, \beta$ and the reused sample.

### 7.2 Candidate low-rank DMFT state

Consider $n, d \to \infty$ with $n/d \to \alpha$, fixed $k, p, q$, and direct rows

$$w_a(t) \in \mathbb{R}^d, \qquad a = 1, \dots, p+q,$$

obtained by stacking the rows of $W_1(t)$ and $W_2(t)$. A minimal dynamic state contains:

$$Q_{ab}(t,s) = w_a(t)^\mathsf{T} w_b(s), \tag{51}$$

$$M_{aj}(t) = w_a(t)^\mathsf{T} \Theta_j, \qquad j = 1, \dots, k, \tag{52}$$

$$R_{ab}(t,s) = \frac{1}{d} \sum_{i=1}^{d} \frac{\delta w_{ai}(t)}{\delta \zeta_{bi}(s)}, \tag{53}$$

where $\zeta$ is an infinitesimal external field. The finite parameters

$$\varphi(t) = (A(t), b(t), c(t), \alpha(t), \beta(t), a_0(t))$$

must remain in the state because they determine the link derivatives (35)–(39).

For fresh online samples, one expects effective Gaussian preactivations whose two-time covariance is determined by $Q$, coupled to a finite deterministic evolution of $(M, \varphi)$. For repeated passes over a fixed dataset, the effective process additionally contains an Onsager memory term built from $R$ and colored noise whose covariance depends on the two-time residual correlations.

**Conjecture 7.2** (Joint three-layer reused-sample DMFT)**.** *Under Gaussian covariates, fixed bottleneck dimensions, proportional sample size, exchangeable initialization, and a regularized Lipschitz spectral map, the empirical trajectory of the finite observables*

$$(Q(t,s), M(t), R(t,s), \varphi(t))$$

*converges on every fixed time horizon to a deterministic causal solution. A representative training example is described by an effective non-Markovian process*

$$h(t) = h^{\mathrm{sig}}(M(t), Z) + \eta(t) + \int_0^t \Sigma_R(t,s)\, \mathcal{G}(h(s), \varphi(s), Y)\, \mathrm{d}s, \tag{54}$$

*where $\eta$ is centered colored Gaussian noise, its covariance $\Sigma_C$ is a functional of $Q$ and residual correlations, and $\Sigma_R$ is a response kernel. The matrix drifts for the direct blocks are obtained by applying the finite action (50) to their limiting gradient Gram matrices.*

This conjecture specifies the state and causal structure but not the final self-consistency equations. A proof requires dynamic cavity or generating functional methods, tightness, and uniqueness. The existing code does not claim to have supplied these steps.

### 7.3 MSRJD construction: the physics target

The appropriate statistical-physics derivation is a disorder-averaged generating functional, not a Gaussian ansatz imposed directly on the parameters. In a continuous-time regularization of the training rule, stack the direct rows in $W(t)$ and write schematically

$$\dot{W}(t) = -\mathcal{P}_a(G_W(t)) + \zeta(t), \tag{55}$$

where $G_W$ is the fixed-sample gradient and $\mathcal{P}_a$ is the regularized singular-value map. Introducing response fields $\widehat{W}$ gives an MSRJD functional of the form

$$\mathcal{Z}[\zeta] = \int \mathcal{D}W\, \mathcal{D}\widehat{W}\, \exp\Big\{ - \int \mathrm{d}t\, \Big\langle \widehat{W}(t), \dot{W}(t) + \mathcal{P}_a(G_W(t)) - \zeta(t) \Big\rangle \Big\}, \tag{56}$$

with the usual response-field contour convention suppressed. The Martin–Siggia–Rose–De Dominicis–Janssen construction and its effective action are the standard field-theoretic route to dynamic mean-field theory [17]. Averaging (56) over the reused Gaussian training sample creates nonlocal temporal interactions.

At the saddle, the required collective fields are precisely the two-time correlations and responses $Q, R$, the teacher overlaps $M$, and the finite state $\varphi$, together with pattern-level preactivation fields for both quadratic blocks. The latter are essential: the same training pattern enters $h = Wx$, the nonlinear features, the residual, and every later gradient. After disorder averaging, this repeated use becomes a retarded self-interaction and colored effective noise. The exact finite Gram action (50) shows that Muon does not introduce an extensive new order parameter at fixed row dimension; it changes the finite matrix functional applied to the saddle-point gradient Gram matrices.

This construction follows the field-theory lineage used to derive Gaussian process saddles and systematic finite-width corrections for deep and recurrent networks [35]. Here, however, the saddle is dynamic and task conditioned: its teacher overlaps and coefficient map move during training.

### 7.4 Cavity, replica, and the spectral-AMP interpretation

The three standard disordered-system routes answer different questions and should agree where their domains overlap.

**Dynamic cavity.** Remove one training pattern $\mu$, run the trajectory in the cavity system, and expand the full trajectory in the influence of that pattern. At leading order one expects a relation of the schematic form

$$h_\mu(t) = h_\mu^{\mathrm{cav}}(t) + \int_0^t \mathcal{O}(t,s)\, g_\mu\big(h_\mu^{\mathrm{cav}}(s), \varphi(s)\big)\, \mathrm{d}s + o_d(1), \tag{57}$$

where $\mathcal{O}$ is the Onsager reaction kernel. In a reused-data problem this reaction is generically retarded rather than a one-step scalar. Establishing (57) uniformly over patterns and time is the leave-one-out part of the joint DMFT proof.

**Replica.** For a Langevin or Bayesian equilibrium, the static object is

$$Z_\beta(\mathcal{D}) = \int \exp[-\beta \mathcal{L}_{\mathcal{D}}(\theta)]\, \mathrm{d}P_0(\theta), \qquad \mathbb{E}_{\mathcal{D}} \log Z_\beta = \lim_{r \downarrow 0} \frac{\mathbb{E}_{\mathcal{D}} Z_\beta^r - 1}{r}.$$

A replica-symmetric saddle would give a static potential in overlap and representation order parameters. It can locate information-theoretic and metastability transitions and should agree with stationary points of a valid Bayes-matched state evolution. The TAP/Onsager viewpoint also admits direct iterative constructions in canonical spin-glass models [10]. It does not determine the transient dynamics of deterministic Muon, and replica symmetry may fail. The Low-RAMP literature makes the AMP/TAP/replica correspondence explicit for low-rank matrix estimation [21]; the present hierarchical factor graph is not one of the models solved there.

**Muon-$a$ as an AMP-like spectral channel.** At fixed row dimension, (50) writes the update as a nonseparable spectral map of a finite Gram matrix. In that precise sense, Muon-$a$ is a natural matrix denoiser or channel for an AMP construction. Rigorous scalar state evolution begins with dense Gaussian AMP, and extends to nonseparable Lipschitz denoisers [6, 8]. VAMP extends scalar state evolution to right-orthogonally invariant linear operators [32].

The optimizer tested in section 9 is nevertheless *not* an AMP algorithm: it applies $\Phi_a$ to the ordinary reused-data gradient and contains no explicit Onsager correction. A candidate long-memory spectral-AMP iteration would instead have the structure

$$G^t_{B,\mathrm{cav}} = G^t_B - \sum_{s<t} \Omega_B(t,s)\,\Delta B^s, \tag{58}$$

$$B^{t+1} = B^t - \eta^t_B\,\Phi_{a^t_B,\varepsilon}\Big(G^t_{B,\mathrm{cav}}\Big), \qquad B \in \{V,U,A\}, \tag{59}$$

where the matrix memory kernels $\Omega_B$ are fixed by response self-consistency and contractions of $D\Phi_{a_B,\varepsilon}$, whose exact full-rank formula is (31). Because the same observations pass through two nonlinear blocks, a scalar one-step Onsager coefficient is not assumed to suffice.

**Conjecture 7.3** (Spectral-AMP state evolution)**.** *There exists a regularized, blockwise, long-memory AMP version of* (59) *whose finite-time state evolution closes on* $(Q, M, R, \varphi)$ *and the two layerwise effective fields. Its Onsager kernels cancel the leading reused-pattern self-interaction, and its state-evolution fixed points coincide with replica-symmetric stationary conditions whenever the latter are stable.*

Proving this conjecture requires more than observing Gaussian gradient histograms: one must derive the factor graph, the matrix-valued Onsager reaction, and a state evolution for the nonseparable spectral channel.

### 7.5 Lazy/kernel control and rich feature-learning target

Three limits constrain any valid closure.

**Frozen features: the lazy/kernel control.** If $V, U, A$ are frozen and only a linear readout is trained, the microscopic feature modes decouple conditionally on common response and noise kernels. The closure must reduce to the random-feature DMFT of Kramp et al. [20]. Their microscopic model is linear regression in a fixed Gaussian feature basis. It describes NNGP, NTK, and random-feature limits in which the representation does not acquire an $O(1)$ task-dependent deformation. Its disorder average yields independent eigenmode processes coupled through a common response kernel and colored noise. A proposed three-layer DMFT that does not reproduce this memory and noise after freezing $V, U, A$ is inconsistent. Conversely, copying that closure while allowing only the readout to move would remove the phenomenon studied here.

**Non-lazy equilibrium: the representation control.** The Bayesian mean-field theory of van Meegen and Sompolinsky [36] shows that wide networks in the non-lazy scaling can acquire $O(1)$, task-dependent representations and that their coding structure depends strongly on the activation: linear, sigmoidal, and ReLU links produce qualitatively different codes. This is a static posterior theory rather than a reused-data training DMFT, but it supplies a second consistency check. A rich-regime closure must retain activation-dependent single-site laws rather than close only on the kernel. Theorem 8.1 gives the corresponding finite population control in the present coefficient language.

**Population anisotropy.** With fresh population gradients and power-law input covariance, a quadratic single-index specialization must reproduce the infinite weighted-overlap hierarchy and Duhamel–Volterra phases of Braun et al. [12]: fast escape, macroscopic convergence, and spectral-tail learning. In the isotropic finite-$k$ reduced model, Theorem 3.2 replaces that infinite hierarchy by the exact finite Hermite coefficient flow.

### 7.6 A longitudinal signal coupled to a transverse bath

The block decomposition (47) has a direct physical interpretation. The extensive orthogonal block $B_n(t)$ is a transverse spectral bath. The finite teacher and feature-mixing block $D_n(t)$ is the longitudinal signal sector, and $C_n(t)$ couples the two. Integrating out the bath produces the exact, energy-dependent self-energy

$$\Sigma_n(t, \lambda) = C_n(t)^{\mathsf{T}}(B_n(t) - \lambda \mathrm{I})^{-1} C_n(t), \tag{60}$$

so the signal propagator is controlled by

$$D_n(t) - \lambda \mathrm{I} - \Sigma_n(t, \lambda).$$

This is precisely the finite Schur determinant (49). In a valid high-dimensional closure, the DMFT single-site process determines the equal-time law $\nu_t$ of the Hessian block $T_t$; the matrix Dyson equation turns $\nu_t$ into the transverse bath spectrum; and the anisotropic resolvent determines the self-energy seen by a teacher-aligned longitudinal mode.

A BBP transition is then a stability change of this longitudinal mode against the transverse continuum. Below threshold the Schur zero is absorbed by the bath edge and its eigenvector weight is delocalized. Above threshold a separated zero survives with nonvanishing teacher residue. This physical picture is exact at the level of the finite bath integration, but its deterministic dynamic limit still requires the local-law and data-reuse locks listed in section 10.

### 7.7 The solvable transient-BBP control

The closest closed dynamical control is the linear teacher–student model of Coeurdoux et al. [13]. Its gradient flow is an explicit matrix semigroup. A two-block input covariance produces a time-dependent Wigner-type bath described by a $2 \times 2$ Dyson equation, while the rank-one teacher becomes a time-dependent rank-two perturbation after symmetrization. The associated $2 \times 2$ determinant yields three regimes: no separated branch, a persistent branch, or a branch that separates for a finite time window and is then reabsorbed by the bulk. This is a genuine transient BBP mechanism and a useful null model for early stopping.

It is not a special case of the Hessian ensemble in this paper. The spectral observable there is a symmetrized trained weight matrix with a Wigner-type noise component; here it is a Hessian whose transverse block is matrix-weighted Wishart. The methodological correspondence is

| Ingredient | Solvable linear control | Three-layer target |
|---|---|---|
| training state | explicit semigroup | MSRJD/cavity state $(Q, M, R, \varphi)$ |
| spectral bath | two-block Wigner type | matrix-weighted block Wishart |
| bulk equation | $2 \times 2$ Dyson equation | matrix Dyson equation driven by $\nu_t$ |
| signal sector | time-dependent rank two | teacher coordinates plus finite parameters |
| outlier test | $2 \times 2$ determinant | finite Schur determinant $F_n$ |
| main obstruction | explicit coefficients | moving representation, reuse, tails, Muon response |

The absent/persistent/transient trichotomy is therefore a qualitative null hypothesis for the modes in Figure 3, not a label that can be assigned to them from a finite plot.

### 7.8 Kac–Rice, landscape complexity, and BBP

Kac–Rice and DMFT condition on different objects. If $\mathcal{N}_d(e, q)$ counts critical points at risk density $e$ and teacher overlap $q$, the landscape observables are the annealed and quenched complexities

$$\Sigma_{\mathrm{ann}}(e, q) = \lim_{d \to \infty} \frac{1}{d} \log \mathbb{E}\mathcal{N}_d(e, q), \qquad \Sigma_{\mathrm{que}}(e, q) = \lim_{d \to \infty} \frac{1}{d} \mathbb{E} \log \mathcal{N}_d(e, q).$$

The Kac–Rice identity itself is exact under regularity assumptions [1, 18, 33]. Evaluating the quenched complexity, or replacing it by an annealed calculation, is a separate step. Random matrix conditioning of the Hessian has made this route effective in spin glasses and non-Gaussian empirical-risk landscapes [4, 22].

Recent single-index results sharpen the interface relevant here. Asgari et al. [3] establish general high-dimensional Kac–Rice formulas for local minima under explicit regularity and concentration hypotheses. Montanari and Saeed [25] prove a finite-dimensional variational formula for the annealed exponential rate of local minima and sufficient conditions for rate trivialization. Maillard et al. [23] condition further on overlap and combine Kac–Rice with a low-rank Hessian analysis: a teacher-aligned BBP instability can appear before the annealed complexity of minima vanishes. At initialization and along phase-retrieval gradient flow, related studies find continuous or discontinuous informative Hessian branches and strong finite-size effects [2, 11].

These results do not turn the fresh Hessian of section 6 into a Kac–Rice Hessian. Four probability laws must be kept distinct:

1. the population Hessian at a prescribed parameter;
2. the fresh empirical Hessian conditional on a trained state;
3. the same-data Hessian along a random training trajectory;
4. the Hessian at a critical point conditioned on risk, overlaps, and representation invariants.

The fourth object requires a new three-layer Kac–Rice calculation. It must remove gauge zero modes, condition the joint gradient/Hessian law on all layerwise overlaps, evaluate the exponential determinant, and distinguish annealed from quenched complexity. Such a calculation would map the stationary topology. It would still not replace the causal response kernels needed for transient Muon dynamics.

**Continuous versus discontinuous detachment.** The classical BBP transition and general finite-rank deformation theory identify the continuous separation of a spike from a regular edge [5, 7]. A different mechanism occurs when the limiting density vanishes sufficiently rapidly at the relevant edge: the asymptotic eigenvector overlap can jump, with an extended finite-size precursor region [9]. Whether a three-layer Hessian contact is continuous, discontinuous, or multicritical cannot be decided from an outlier count. It requires the edge exponent, the derivative of the limiting Schur determinant, and the finite-size scaling of both the eigenvalue gap and teacher residue.

**Sparse cavity as a computational control.** For locally tree-like, non-Hermitian matrices, cavity recursions and population dynamics can recover spectral support and outliers [27, 28, 34]. Giammanco et al. [16] use this machinery for sparse antagonistic matrices with diagonal disorder and a Jacobian-like structure, finding five spectral phases and also a strong-disorder regime where the numerical population dynamics underestimates the support. This is relevant as a computational warning and as a template for distribution-valued resolvent order parameters. It is not an imported theorem for the dense Hermitian block-Wishart bath considered here.

### 7.9 Sequential BBP thresholds and aggregate scaling

Suppose a lifted hierarchical stage contains orthogonal latent features with strengths

$$\lambda_i \asymp i^{-\gamma}, \qquad \gamma > \frac{1}{2},$$

and effective ambient feature dimension $D \asymp d^{q_0}$. A spectral spike of strength $\lambda_i$ becomes visible when its signal exceeds the sample-covariance fluctuation scale, giving the feature-wise prediction

$$n_i \asymp \frac{D}{\lambda_i^2} \asymp d^{q_0} i^{2\gamma}. \tag{61}$$

Consequently the number of recovered features at sample size $n$ is

$$m(n) \asymp \left(\frac{n}{d^{q_0}}\right)^{1/(2\gamma)}. \tag{62}$$

If prediction error is the unrecovered squared-strength tail, then

$$\sum_{i>m(n)} \lambda_i^2 \asymp m(n)^{1-2\gamma} \asymp \left(\frac{n}{d^{q_0}}\right)^{-1+1/(2\gamma)}. \tag{63}$$

These are the sharp-threshold and aggregate-scaling mechanisms proved in the solvable hierarchical model of Wortsman-Zurich et al. [38], up to its precise normalizations.

**Conjecture 7.4** (Hierarchical feature-front scaling)**.** *For a growing-rank version of* (6) *with power-law teacher strengths, the dynamic Schur sectors and the fresh Hessian Schur determinant describe the same feature front: feature $i$ enters the residualized tangent space when a finite-signal branch crosses the appropriate block-Wishart edge. Aggregating these crossings yields* (63)*, with a block- and depth-dependent effective dimension $D$.*

The current experiment keeps $k = 3$. It can test ordering and finite-size smoothing, but no fit from six sample sizes at fixed rank can prove (63).

### 7.10 Three matched phases

The references suggest the following falsifiable dictionary:

| Phase | Exact three-layer observable | RMT observable | Expected phenomenon |
|---|---|---|---|
| Escape | $E_2$, $S_2$, first-layer overlap | first teacher-aligned branch | recovery of strongest latent subspace |
| Composition | $E_3$, $S_3$, evolution of $A$ | additional finite Schur zeros | nonlinear feature mixing |
| Tail/late sector | $E_4$, $S_4$, weak singular channels | late BBP contacts near moving edges | slow weak-feature recovery |

This table is a research hypothesis. The experiments should compare the event times quantitatively rather than infer causality from visually aligned curves.

## 8 Beyond the quadratic link: a controlled universality layer

The degree-four model is exactly closed, but "three layers" is not a universality class by itself. Changing the activation, teacher spectrum, width scaling, data covariance, or training-data protocol can change both the effective state and the random-matrix ensemble. This section gives a precise extension that is valid for any link with controlled Hermite tails, and then states the additional hypotheses needed to connect that extension to DMFT and BBP.

### 8.1 Hilbert-space coefficient calculus

Let $\mathcal{H} = L^2(\gamma_k)$, with its orthonormal Hermite basis, and identify a function with its weighted coefficient vector. Let $c : \Theta \subset \mathbb{R}^P \to \mathcal{H}$ be the coefficient map of a general three-layer network and $c_\star \in \mathcal{H}$ the teacher. Write $P_R$ for orthogonal projection onto total Hermite degree at most $R$, set

$$\delta = c - c_\star, \qquad \mathcal{R} = \frac{1}{2}\|\delta\|_{\mathcal{H}}^2, \qquad \mathcal{R}_R = \frac{1}{2}\|P_R\delta\|_{\mathcal{H}}^2,$$

and define, at a fixed parameter value,

$$\varepsilon_0(R) = \|(\mathrm{I} - P_R)\delta\|_{\mathcal{H}}, \tag{64}$$

$$\varepsilon_1(R) = \|(\mathrm{I} - P_R)Dc\|_{\mathrm{op}(\mathbb{R}^P,\mathcal{H})}, \tag{65}$$

$$\varepsilon_2(R) = \sup_{\|u\|=\|v\|=1} \|(\mathrm{I} - P_R)D^2c[u,v]\|_{\mathcal{H}}. \tag{66}$$

**Theorem 8.1** (Hermite-tail transfer)**.** *Suppose $c$ is twice Fréchet differentiable as an $\mathcal{H}$-valued map. Then*

$$\mathcal{R} - \mathcal{R}_R = \frac{1}{2}\varepsilon_0(R)^2, \tag{67}$$

$$\|\nabla\mathcal{R} - \nabla\mathcal{R}_R\|_2 \le \varepsilon_0(R)\varepsilon_1(R), \tag{68}$$

$$\|\nabla^2\mathcal{R} - \nabla^2\mathcal{R}_R\|_{\mathrm{op}} \le \varepsilon_1(R)^2 + \varepsilon_0(R)\varepsilon_2(R). \tag{69}$$

*The same bounds hold uniformly on a compact parameter set $K$ after taking the supremum of each $\varepsilon_j$ over $K$.*

The two terms in (69) have different meanings. $\varepsilon_1^2$ is the omitted Gauss–Newton tail; $\varepsilon_0\varepsilon_2$ is the omitted residual curvature. For the quadratic architecture, $\varepsilon_j(R) = 0$ for $R \ge 4$, so Theorem 3.2 is recovered exactly. For a smooth non-polynomial activation, Theorem 8.1 turns a Hermite approximation into an explicit $C^2$ approximation rather than an informal appeal to universality.

**Corollary 8.2** (Stability of finite Hessian branches)**.** *At a fixed state, let $\eta_R = \varepsilon_1(R)^2 + \varepsilon_0(R)\varepsilon_2(R)$. Every ordered eigenvalue of $\nabla^2\mathcal{R}_R$ differs from the corresponding eigenvalue of $\nabla^2\mathcal{R}$ by at most $\eta_R$. Consequently, an isolated finite eigenvalue cluster separated from the rest of the truncated spectrum by a gap larger than $2\eta_R$ has the same multiplicity after restoring the Hermite tail.*

This corollary concerns the finite population Hessian. It does not remove unbounded per-example blocks from a high-dimensional empirical Hessian: moment, truncation, and local-law estimates remain separate requirements.

**Distributional universality is a separate interface.** Montanari and Saeed [26] prove universality of the minimum value, and under additional hypotheses train and test error, for fixed-index empirical-risk problems when the relevant minimizers satisfy a delocalization condition. This justifies Gaussian replacement in the models covered by their theorem. It does not imply universality of a training trajectory, Hessian edge, BBP residue, or Muon response kernel. Extending the present Gaussian theory to non-Gaussian covariates therefore requires both their optimization-level hypotheses and new derivative/local-law control for the spectral observables.

### 8.2 The DMFT–RMT–BBP closure diagram

The physical link can be stated without identifying objects that live in different probability spaces. At time $t$, a reused-data DMFT would determine the law of the effective preactivations and residuals. Through (39), this law induces a time-dependent block law

$$\nu_t = \mathrm{Law}\left(T_{\theta_t}(Z, Y)\right).$$

For a fresh probe, $\nu_t$ is the input to the block-Wishart matrix Dyson equation and therefore determines the orthogonal bulk. The finite teacher-sector Schur kernel then tests whether a signal branch lies outside that bulk. In short,

$$\boxed{\text{DMFT state } (Q, M, R, \varphi)} \implies \boxed{\text{block law } \nu_t} \implies \boxed{\text{RMT bulk } [E_-(t), E_+(t)]} \xrightarrow{\text{finite Schur equation}} \boxed{\text{BBP branch}}. \tag{70}$$

The first arrow is conjectural for jointly trained, reused observations; the fresh conditional block-Wishart identity behind the second arrow is exact; the bounded global edge theorem is imported; and the final outlier arrow requires an anisotropic local law.

**Proposition 8.3** (Conditional dynamic spectral closure)**.** *Fix a compact time interval. Assume:*

*(i) the joint training state converges uniformly to a deterministic causal DMFT solution and induces a deterministic block law $\nu_t$;*

*(ii) the fresh orthogonal Hessian obeys a time-uniform matrix Dyson equation with regular support edges;*

*(iii) an anisotropic local law holds uniformly away from those edges and the finite signal-sector matrices converge;*

*(iv) either the blocks are uniformly bounded or truncation removal is controlled in operator norm at the local-law scale.*

*Then the limiting bulk edges and every simple separated teacher-sector branch are deterministic functions of the DMFT state. A branch contact is a BBP transition only when the limiting Schur zero reaches a regular edge and the contact is transverse.*

The proposition is deliberately conditional: it packages the four analytic locks into one reusable statement but does not claim that the locks have been proved for reused-data Muon.

### 8.3 What "many three-layer models" requires

The framework extends by substitution only after the following objects have been identified for the new model:

| Model choice | Object that changes | Proof obligation |
|---|---|---|
| Activation/link | coefficient map $c(\theta)$, block $T_\theta$ | Hermite $C^2$ tails and per-example moments |
| Teacher spectrum | target coefficients and active ranks | signal-strength scaling and sector forcing |
| Data covariance | Hermite basis or orthogonal polynomials | population closure and anisotropic bulk law |
| Widths/ranks | dimension of the finite signal sector | uniform rank control or a growing-block RMT theorem |
| Optimizer | drift and response kernels | descent, stability, and DMFT well-posedness |
| Data reuse | dependence of $T_{\theta_t}$ on its own sample | dynamic leave-one-out and time-uniform local law |

Thus the present paper gives a master reduction, not a theorem for every possible three-layer network. The exact finite algebra survives broadly; the asymptotic closure is model- and protocol-dependent.

This dependence is physical rather than cosmetic. In the non-lazy Bayesian theory of van Meegen and Sompolinsky [36], changing the activation changes the symmetry-breaking pattern and the internal coding scheme even when kernel-level observables can look similar. The Hermite-tail quantities $\varepsilon_0, \varepsilon_1, \varepsilon_2$ therefore control approximation to a declared link; they do not imply an activation-independent representation law.

Table 1: Main pilot endpoint exact population risks. "Wins" counts paired runs with smaller risk than GD. With only three pairs, all summaries are descriptive.

| Optimizer | risks by pair | mean | median | wins/GD |
|---|---|---|---|---|
| GD | $0.7569, 0.9416, 0.04620$ | 0.5816 | 0.7569 | – |
| Muon $a = 0$ | $0.2712, 0.08062, 0.08712$ | 0.1463 | 0.08712 | 2/3 |
| Muon $a = 1/7$ | $0.4114, 0.06553, 0.6195$ | 0.3655 | 0.4114 | 2/3 |
| Muon $a = 1/3$ | $0.08887, 0.05081, 0.2975$ | 0.1457 | 0.08887 | 2/3 |

# 9 Experiments

## 9.1 Questions and protocol

The experiments address four questions:

1. Do the exact degree errors move in the order $2 \to 3 \to 4$?
2. Does a singular-value update accelerate a late sector at matched data and initialization?
3. Do teacher-aligned eigenmodes move relative to the fresh orthogonal block-Wishart bulk?
4. Does increasing sample size improve exact population risk in a manner consistent with sequential feature recovery?

The main pilot uses

$$k = p = q = 3, \quad n_{\text{train}} = 1024, \quad 1200 \text{ steps}, \quad n_q = 7, \quad R = 3$$

paired repetitions. The first and second direct matrices are untied. The teacher includes diagonal and mixed Hermite coefficients in degrees $2, 3, 4$. Gradient descent uses learning rate 0.02. Spectral matrix blocks use base rate 0.004; their vector and bias parameters use 0.02. All methods use loss-decreasing backtracking. For each repetition, every optimizer receives the same initialization and training sample.

Population test risk is evaluated by exact tensor Gauss–Hermite quadrature, not a finite test set. Fresh Hessians use $d = 32$, $d_\perp = 29$, $n_{\text{fresh}} = 145$, and checkpoints $0, 400, 800, 1200$. The block cap is $L = 100$.

## 9.2 Paired endpoint risk

Polar Muon lowers the median by a factor 0.115, but the paired ratios relative to GD are $0.358, 0.0856, 1.886$. The third pair is easy for GD and reverses the optimizer order. The nonzero transferred powers are even less stable on that pair. The defensible conclusion is evidence for optimizer–state interaction, not universal dominance.

## 9.3 Sequential sector times

For the first pair, define the half-error time of sector $r$ as the first recorded step with $E_r(t) \leq E_r(0)/2$.

The degree-four sector is the late bottleneck in this trajectory. Polar Muon changes its access time substantially while leaving the degree-two clock unchanged. This is consistent with spectral flattening helping a weak late-sector channel, but a single trajectory cannot identify the mechanism. The confirmatory campaign therefore computes sector times for every pair and reports paired uncertainty rather than selecting repeat zero.

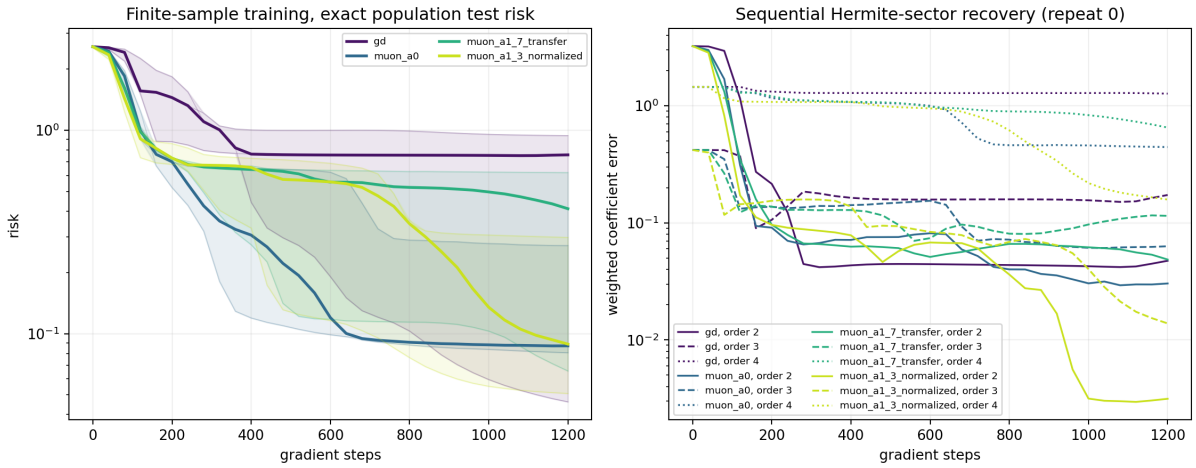


Figure 1: Left: exact population test risk during fixed-sample training. Thin curves are individual paired runs; thick curves are medians and bands show the min–max range. Right: exact degree-wise Hermite errors for the first pair. The figure is descriptive because $R = 3$.

Table 2: Sector half-error times on the first paired trajectory.

| Optimizer | degree 2 | degree 3 | degree 4 |
|---|---|---|---|
| GD | 120 | 160 | $> 1200$ |
| Muon $a = 0$ | 120 | 120 | 680 |
| Muon $a = 1/7$ | 80 | 120 | 1160 |
| Muon $a = 1/3$ | 80 | 80 | 800 |

### 9.4 Fresh Hessian spectroscopy

At initialization all methods share the same state. The clipped variational bulk is $[-1.153, 1.112]$ and the clipped empirical orthogonal bulk is $[-0.954, 0.886]$. The raw informative-outlier counts over checkpoints are

| | 0 | 400 | 800 | 1200 |
|---|---|---|---|---|
| GD | 4 | 1 | 3 | 2 |
| $a = 0$ | 4 | 0 | 1 | 2 |
| $a = 1/7$ | 4 | 0 | 1 | 1 |
| $a = 1/3$ | 4 | 1 | 2 | 3 |

under the stated empirical bulk and teacher-overlap criterion.

The counts are not monotone, and clipping is consequential. Across the sixteen optimizer/checkpoint states, the maximum clipped-block fraction is 6.21%. The median fraction is about 2.1%, but the median operator-norm difference between raw and clipped full Hessians is about 32.9. This observation directly motivates Theorem 6.3: a small contamination fraction is not an edge-stability theorem.

### 9.5 One clock for population geometry and fresh spectroscopy

The denser coupled run stores 25 parameter states for each optimizer. At each state it evaluates two different objects:

1. exact population coefficients, Hessian decomposition, Schur matrices, and residual radii at the trained parameter;

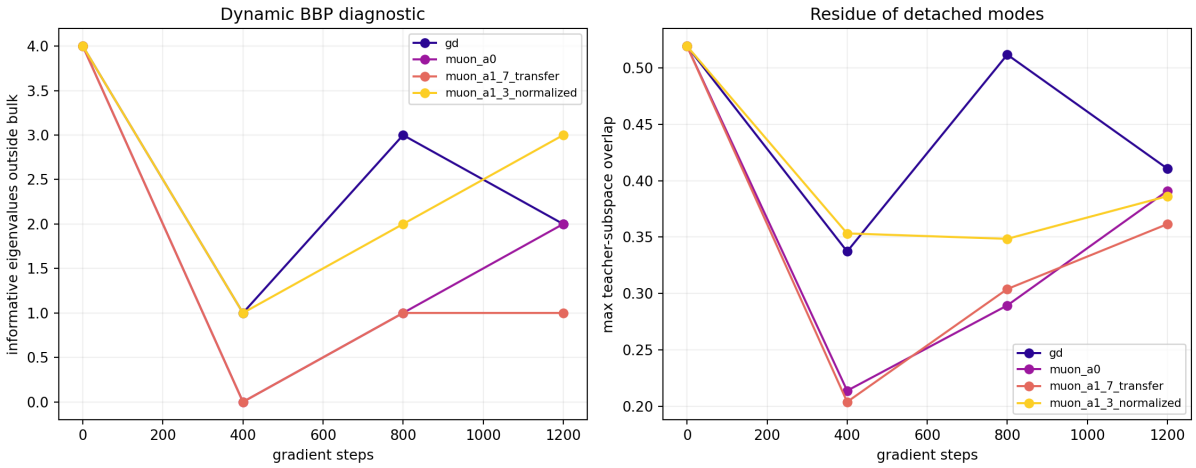


Figure 2: Fresh-sample spectral diagnostic. Left: number of raw full-Hessian eigenvectors outside the raw orthogonal empirical bulk and above the teacher-overlap threshold. Right: maximum teacher-subspace overlap among those modes. This is not a proof of asymptotic detachment.

2. a raw direct-weight Hessian from one independent Gaussian probe frozen across all states and optimizers.

The first object is deterministic finite algebra conditional on the trained state. The second is a finite-size fresh-sample diagnostic. Placing them on one horizontal axis does not merge their probability spaces.

On this instance, GD reaches the degree-2 and degree-3 half-error events at steps 120 and 260, but not degree 4 by step 1200. Polar Muon reaches the three events at steps $80, 120, 140$. Nevertheless, no Schur ratio in Row B exceeds one: the largest degree-2 ratio is about 0.28 for both methods, while the degree-3 and degree-4 ratios are much smaller. The active-curvature inequality is therefore not the explanation of the observed crossings on this trajectory. Either the bound is conservative, cross-sector forcing is essential, or both.

The fresh spectral behavior also differs without producing a proved event match. The number of teacher-aligned detached modes under GD falls from four at initialization to zero after step 850; under polar Muon it settles at two from step 100 onward. This persistent finite-size separation is a useful target for the conditional chain (70), but the plot cannot identify a limiting branch or establish that a spectral contact caused a coefficient event. In particular, it does not distinguish the absent, persistent, and transient branches of the solvable linear control [13], nor continuous from discontinuous detachment [9].

### 9.6 Nested sample-size panel

A second experiment uses five paired repetitions, 800 steps, and nested sample sizes

$$n \in \{64, 128, 256, 512, 1024, 2048\}.$$

### 9.7 Algebra and reproducibility checks

Six focused tests accompany the campaign:

1. (39) matches automatic differentiation of the per-example preactivation loss Hessian to numerical precision;

2. the selected orthogonal principal block of $n^{-1} \sum_\mu T_\mu \otimes x_\mu x_\mu^\mathsf{T}$ equals $n^{-1} \sum_\mu T_\mu \otimes \xi_\mu \xi_\mu^\mathsf{T}$;

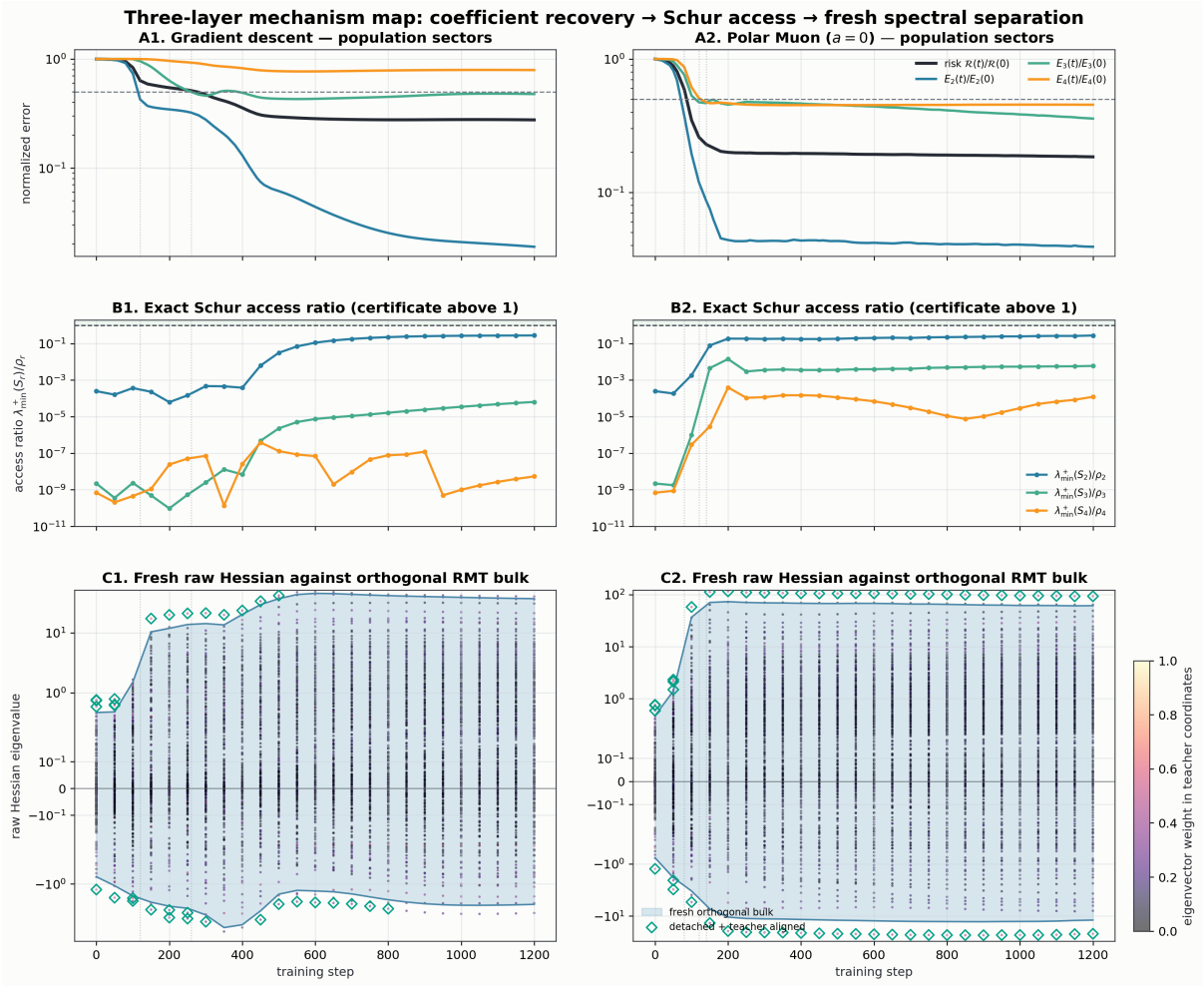


Figure 3: **How to read the mechanism panel.** Columns compare gradient descent (left) and polar Muon (right) on the same initialization and training sample. Row A plots exact population risk and exact degree-$2, 3, 4$ Hermite errors, each divided by its initial value; the dashed horizontal line is the half-error threshold. Row B plots the exact finite ratio $\lambda^+_{\min}(S_r)/\rho_r$. Values above one would certify positive curvature throughout the residualized degree-$r$ tangent space; a value below one means only that this sufficient certificate is inconclusive. Row C plots every raw full-Hessian eigenvalue against training step. The blue band is the empirical orthogonal spectrum from the frozen fresh probe, point color is eigenvector mass in teacher coordinates, and green diamonds mark points satisfying the declared outside-bulk and overlap thresholds. Dotted vertical lines repeat the sector half-error events.

Table 3: Nested-sample endpoint summary. The slope is a descriptive log–log fit over the large-$n$ tail at fixed $k = 3$, not the growing-rank exponent in (63).

| Optimizer | mean risk, $n = 2048$ | median risk, $n = 2048$ | tail slope |
|---|---|---|---|
| GD | 0.4285 | 0.2592 | $-0.536$ |
| Muon $a = 0$ | 0.1281 | 0.1038 | $-0.381$ |
| Muon $a = 1/7$ | 0.1116 | 0.06948 | $-0.779$ |
| Muon $a = 1/3$ | 0.1539 | 0.08954 | $-0.521$ |

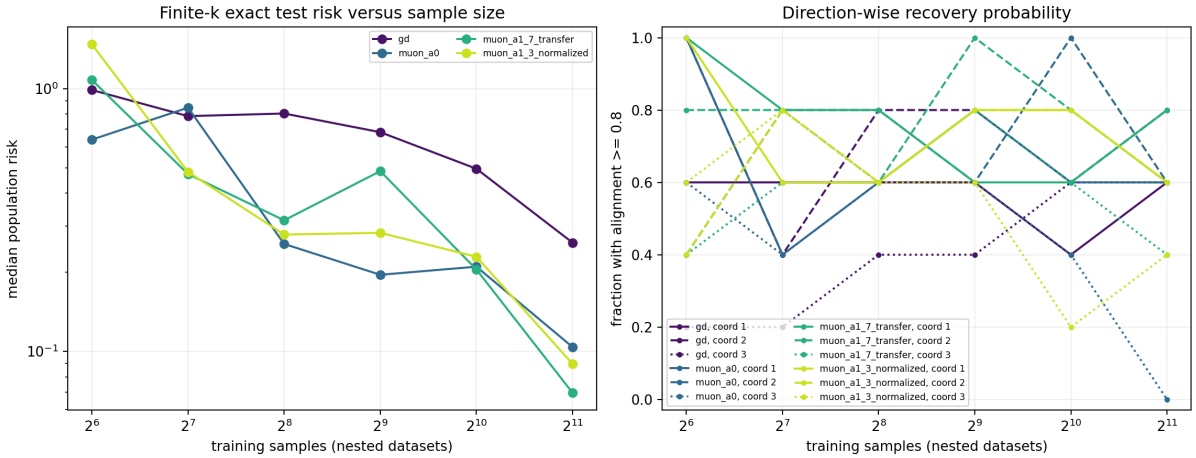


Figure 4: Nested-sample panel. Exact population risk generally improves with sample size, but fixed latent rank, five repetitions, and nonmonotone coordinate recovery probabilities preclude a sharp-threshold or universal scaling claim.

3. the $a = 0$ implementation has a flat nonzero singular spectrum and unit entry RMS;
4. the SVD implementation agrees with the finite Gram action in (50) for $a \in \{0, 1/7, 1/3\}$;
5. every tested nonzero matrix update has positive Frobenius inner product with its gradient;
6. the exact sign-test and right-censored Kaplan–Meier utilities reproduce hand-checkable cases.

The exact coefficient risk, Jacobian Gram blocks, Hessian decomposition, and Schur identities are also computed by the older reduced-model verifier.

### 9.8 Confirmatory design

The pre-specified confirmatory run uses twenty paired repetitions at the main configuration and a new seed. Its primary estimand is the paired log-risk difference

$$\Delta_a^{(j)} = \log \mathcal{R}_a^{(j)}(T) - \log \mathcal{R}_{\mathrm{GD}}^{(j)}(T).$$

We report its median, mean, nonparametric paired bootstrap interval, and sign count. Secondary estimands are the degree-wise half-error time differences, with unreached events treated as right-censored rather than numerically equal to the terminal step. Fresh spectral probes remain exploratory until repeated across checkpoints and independent Hessian samples.

### 9.9 Confirmatory results: faster clocks without endpoint dominance

The twenty-pair result does *not* reproduce the pilot's apparent endpoint advantage. Table 4 reports the exact population risk and the paired log-risk estimand. The geometric risk ratios of

all three spectral rules are slightly above one. Every paired bootstrap interval contains zero, and none of the exact sign tests rejects equal win probability. In particular, polar Muon wins 9/20 pairs and has median paired log difference 0.200, while $a = 1/3$ splits 10/20 and has median difference $-0.0460$.

Table 4: Pre-specified twenty-pair endpoint analysis. $\widetilde{\Delta}$ is the median paired log-risk difference relative to GD; intervals are percentile intervals from 50,000 paired bootstrap resamples. Sign-test $p$-values are exact, two-sided, and unadjusted.

| Optimizer | median risk | geometric ratio/GD | $\widetilde{\Delta}$ | 95% CI | wins/GD ($p$) |
|---|---|---|---|---|---|
| GD | 0.2651 | 1 | 0 | $[0, 0]$ | – |
| Muon $a = 0$ | 0.3436 | 1.128 | 0.1998 | $[-0.731, 1.209]$ | 9/20 (0.824) |
| Muon $a = 1/7$ | 0.2491 | 1.109 | 0.1718 | $[-0.358, 0.789]$ | 8/20 (0.503) |
| Muon $a = 1/3$ | 0.2645 | 1.048 | $-0.0460$ | $[-0.511, 0.451]$ | 10/20 (1.000) |

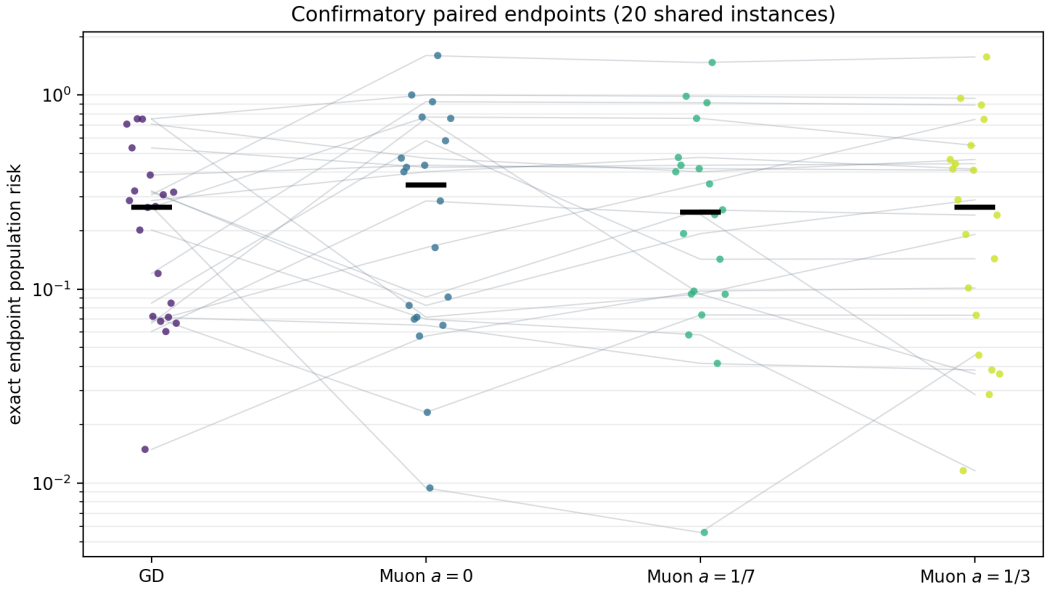


Figure 5: All twenty shared instances on a log-risk scale. Gray lines preserve pairing; black bars are marginal medians. The extensive crossings explain why the three-pair pilot did not yield a stable optimizer ranking.

The sector clocks nevertheless differ. All methods halve degree two on every pair, but the Kaplan–Meier median is 160 steps for GD and 80 for each spectral rule. For degree four, the corresponding medians are 480 for GD, 240 for $a = 0$, 240 for $a = 1/7$, and 320 for $a = 1/3$. This faster clock is accompanied by incomplete coverage: degree four is reached in 16/20 GD runs, compared with 14/20, 16/20, and 15/20 for the three spectral rules. Thus the stable empirical message is a *speed–coverage tradeoff*, not lower endpoint risk.

Table 5: Right-censored sector half-error summaries. Each cell gives events reached out of twenty, followed by the Kaplan–Meier median in parentheses.

| Optimizer | degree 2 | degree 3 | degree 4 |
|---|---|---|---|
| GD | 20/20 (160) | 19/20 (360) | 16/20 (480) |
| Muon $a = 0$ | 20/20 (80) | 17/20 (240) | 14/20 (240) |
| Muon $a = 1/7$ | 20/20 (80) | 17/20 (280) | 16/20 (240) |
| Muon $a = 1/3$ | 20/20 (80) | 17/20 (320) | 15/20 (320) |

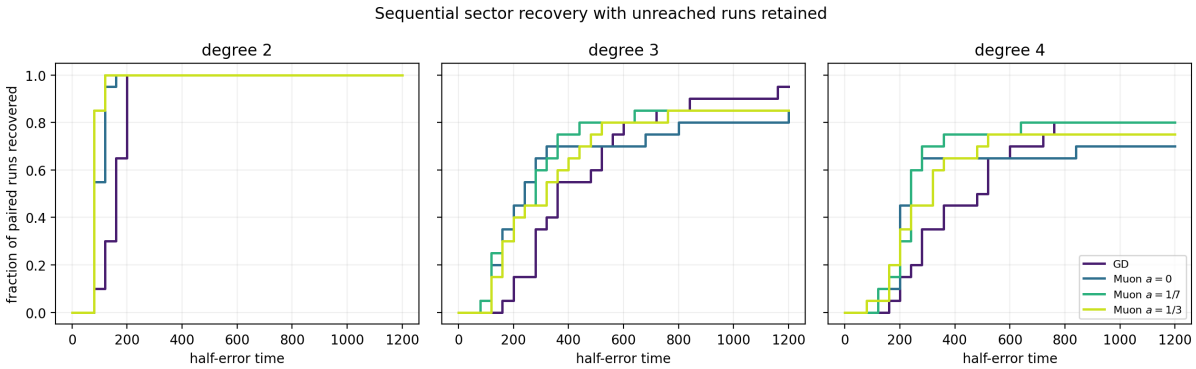


Figure 6: Empirical cumulative half-error events for degrees $2, 3, 4$. Unreached runs remain right-censored at step 1200. Spectral updates move the median event earlier, while their terminal event fractions are not uniformly larger.

The confirmatory campaign therefore falsifies the strongest pilot interpretation: under this shared learning-rate protocol, no tested spectral power has a demonstrated endpoint advantage over GD. It retains a narrower mechanistic observation—singular-value reshaping changes when sectors become accessible—and shows that this timing change need not improve terminal risk.

# 10 Discussion, limitations, and theorem frontier

## 10.1 What is established

The paper establishes a deterministic core and a static random-matrix interface.

1. The reduced three-layer model is exactly finite in Hermite degree. Population risk, gradient, coefficient dynamics, and Hessian follow from one coefficient map.

2. Residualized degree-wise Jacobians give exact positive semidefinite Schur matrices. Together with residual curvature they provide a local, checkable access certificate for each target-bearing sector.

3. The normalized Muon-$a$ family used by the code consists of strict first-order descent directions. For fixed row dimension it is an exact finite action of the gradient Gram matrix.

4. On a fresh sample at a reduced embedded state, the orthogonal direct-weight Hessian is exactly block Wishart. Bounded clipping puts it inside the hypotheses of the Montanari–Saeed edge theorem.

5. The complement is finite dimensional and satisfies an exact Schur determinant. This determinant is the correct starting point for BBP outliers.

6. For a general twice differentiable link in Gaussian $L^2$, explicit Hermite-tail norms control population risk, gradient, and Hessian truncation. The polynomial architecture is the exact zero-tail case.

### 10.2 What remains conditional

Table 6: Proof-status table.

| Statement | Status | Missing ingredient |
|---|---|---|
| Exact degree-four Hermite population closure | proved | none |
| Schur residualization and active curvature bound | proved | none |
| Muon first-order descent and finite Gram action | proved | none |
| Fresh clipped orthogonal block-Wishart global law | imported theorem | bounded fixed-size blocks and proportional asymptotics |
| Fresh Hessian outlier locations and residues | conditional | anisotropic local law plus finite-Schur convergence |
| Raw polynomial Hessian edge/outlier theorem | open | tail removal and extreme-block control |
| Same-data dynamic BBP | open | time-uniform leave-one-out/local law |
| Three-layer Kac–Rice complexity | open | gauge fixing, conditioned Hessian law, determinant asymptotics, annealed/quenched control |
| Continuous versus discontinuous contact | open | edge exponent, Schur contact order, residue and finite-size scaling |
| Joint three-layer reused-sample DMFT | conjectural | dynamic cavity, response/noise closure, uniqueness |
| Muon-$a$ spectral AMP | open | matrix Onsager kernels, nonseparable state evolution, and factor-graph derivation |
| Universal Muon exponent | contradicted as a current claim | block-specific controlled theory and larger experiments |
| Growing-rank sample scaling exponent | open in this model | joint $d, n, k$ asymptotics with power-law strengths |
| General-link population truncation | proved under $C^2$ tail control | estimate $\varepsilon_0, \varepsilon_1, \varepsilon_2$ for the chosen activation and parameter set |
| DMFT–RMT–BBP closure | conditional implication | verify the four hypotheses of Theorem 8.3 |

### 10.3 Connection to depth and renormalization

The useful renormalization picture is operational rather than metaphorical. At degree $r$, projection by $P_{<r}$ integrates out parameter directions already used by earlier sectors, leaving the effective Gram matrix $S_r$. At the next layer, learned quadratic features become inputs to $A$, reducing the representation problem to a new, lower-dimensional coefficient sector. This is the finite-model analogue of effective-dimension reduction in Dandi et al. [14].

The random-feature DMFT of Kramp et al. [20] supplies a second coarse-graining: microscopic kernel eigenmodes couple only through collective response and noise kernels. In the present feature-learning model, those kernels must be coupled to the moving coefficient Jacobian and

finite parameters. Calling the frozen theory the final answer would remove the renormalization of the representation itself.

There are therefore three distinct coarse-grainings. Schur residualization integrates out coefficient directions already used by earlier sectors. The MSRJD disorder average integrates the extensive training coordinates into collective correlation, response, and noise kernels. Finally, the Hessian Schur complement integrates the transverse spectral bath into the energy-dependent self-energy (60). This hierarchy is compatible with effective-action and renormalized field-theory approaches to neural fluctuations [15, 17]; the present paper does not claim a renormalization-group fixed point or universal critical exponent.

### 10.4 Why the empirical result is informative but not final

The three-pair pilot produces a large median gain for polar Muon. The twenty-pair confirmation overturns that optimizer ranking: no spectral rule has a significant paired endpoint advantage, and the geometric risk ratio is slightly above one for all three. What does replicate at a more defensible scale is earlier sector timing, together with equal or lower terminal event coverage. This speed–coverage tradeoff, the nonmonotone outlier counts, and the large raw-versus-clipped spectral discrepancy are not nuisances to hide; they identify the mechanisms a final theorem must explain.

The next decisive experimental matrix is:

1. compute- and tuning-matched learning-rate sweeps on a held-out calibration set;
2. layerwise/frozen controls matching the algorithms covered by existing hierarchical theorems;
3. a factorial grid over $(a_V, a_U, a_A)$, not only a shared exponent;
4. growing $d$, growing latent rank, and power-law teacher strengths;
5. multiple independent fresh Hessian samples at every checkpoint;
6. clipped bounded-link and raw polynomial-link controls;
7. held-out, cross-fit, and same-sample Hessian comparisons.
8. energy- and overlap-conditioned stationary-point sampling to test a future three-layer Kac–Rice prediction.

### 10.5 Path to a full dynamic BBP theorem

A complete proof can be organized into four locks.

**Dynamic state lock.** Prove the joint causal closure of $(Q, M, R, \varphi)$ by MSRJD or dynamic cavity, and derive the matrix Onsager reaction for the spectral channel.

**Bulk lock.** Establish a time-uniform global law and regular support edges for the orthogonal Hessian blocks.

**Local-law lock.** Control finite-signal resolvent quadratic forms down to the distance from a candidate outlier to the edge.

**Tail/LOO lock.** Remove polynomial clipping and, for reused data, quantify the dependence of each observation on its own trajectory.

Only after all four locks should a moving eigenvalue be called a proved dynamic BBP transition.

### 10.6 Conclusion

The mathematical match between two and three layers is not a copied exponent. It is a replacement dictionary:

| | | |
|---|---|---|
| two-layer overlap state | $\longrightarrow$ | three-layer Hermite coefficient map and factor state, |
| single tangent channel | $\longrightarrow$ | residualized Schur sectors $S_2, S_3, S_4$, |
| scalar weighted-Wishart bulk | $\longrightarrow$ | matrix block-Wishart orthogonal Hessian, |
| finite spike equation | $\longrightarrow$ | finite signal-sector Schur determinant, |
| single spectral policy | $\longrightarrow$ | block-specific Muon control. |

This dictionary yields exact theorems where finite algebra suffices, imports global random-matrix results only under their hypotheses, and exposes the precise work required for the full joint DMFT. The accompanying experiments show that this frontier is empirically nontrivial: spectral flattening can accelerate sector half-error clocks, but the confirmation detects no endpoint dominance and instead exposes a speed–coverage tradeoff. The optimizer order depends on the realized state, while the polynomial Hessian tail remains spectrally consequential.

## A Proofs of the exact statements

### A.1 Hermite closure

*Proof of Theorem 3.1.* Each coordinate of $h = Vz + b$ is affine in $z$, hence each coordinate of $q = h^{\odot 2} - \mathbf{1}$ has degree at most two. Each coordinate of $r = Uz + Aq + c$ therefore has degree at most two, and each coordinate of $s = r^{\odot 2} - \mathbf{1}$ has degree at most four. The readout is a linear combination of $q, s$, and a constant, so $f_\theta$ has degree at most four. The multivariate Hermite polynomials with $|\gamma| \le 4$ form a basis for polynomials of total degree at most four. □

*Proof of Theorem 3.2.* By Theorem 3.1 and (7),

$$f_\theta - f_\star = \sum_{\gamma \in \Gamma_4} e_\gamma \mathrm{He}_\gamma.$$

Squaring, taking expectation, and using (2) gives

$$\mathcal{R} = \frac{1}{2} \sum_{\gamma \in \Gamma_4} \gamma! e_\gamma^2 = \frac{1}{2} e^\mathsf{T} D e.$$

The chain rule yields

$$\nabla_\theta \mathcal{R} = J^\mathsf{T} D e.$$

Differentiating once more gives

$$\nabla_\theta^2 \mathcal{R} = J^\mathsf{T} D J + \sum_\gamma \gamma! e_\gamma \nabla_\theta^2 c_\gamma.$$

Along gradient flow,

$$\dot{e} = J\dot{\theta} = -JJ^\mathsf{T} D e.$$

Multiplication by $D^{1/2}$ yields (15). Finally,

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{R} = \langle \nabla_\theta \mathcal{R}, \dot{\theta} \rangle = -\|\nabla_\theta \mathcal{R}\|^2 = -u^\mathsf{T} B u.$$

□

*Proof of Theorem 3.3.* If $e = 0$, the residual-curvature sum in (13) vanishes. The agreement of nonzero spectra follows from the standard fact that $X^\mathsf{T}X$ and $XX^\mathsf{T}$ have the same nonzero eigenvalues, applied to $X = D^{1/2}J$. □

*Proof of Theorem 3.5.* An $n_q$-point one-dimensional Gaussian quadrature rule is exact for polynomials through degree $2n_q - 1$. Tensorization preserves this coordinatewise property. For $|\gamma| \le 4$, $f_\theta \mathrm{He}_\gamma$ has coordinatewise degree at most eight, as does $(f_\theta - f_\star)^2$. Hence $n_q \ge 5$ is sufficient. Parameter differentiation changes coefficients but does not increase polynomial degree in $z$, so differentiating the exact finite quadrature sum gives the exact derivatives. □

## A.2 Schur hierarchy and clocks

*Proof of Theorem 4.1.* The orthogonal projector onto $\mathrm{Ran}(M_{<r})$ can be written

$$P_{<r} = M_{<r}(M_{<r}^\mathsf{T} M_{<r})^\dagger M_{<r}^\mathsf{T}.$$

Therefore

$$\begin{aligned}\overline{M}_r^\mathsf{T}\overline{M}_r &= M_r^\mathsf{T}(\mathrm{I} - P_{<r})M_r \\ &= M_r^\mathsf{T}M_r - M_r^\mathsf{T}M_{<r}(M_{<r}^\mathsf{T}M_{<r})^\dagger M_{<r}^\mathsf{T}M_r,\end{aligned}$$

which is (20). It is positive semidefinite because it is a Gram matrix. Expanding the earlier-sector blocks gives (21). □

*Proof of Theorem 4.3.* Because $Q_r$ is orthogonal to $\mathrm{Ran}(M_{<r})$,

$$Q_r^\mathsf{T}M_r = Q_r^\mathsf{T}\overline{M}_r.$$

The Gauss–Newton Hessian is a sum of positive semidefinite terms, so

$$Q_r^\mathsf{T}H_{\mathrm{GN}}Q_r \succeq Q_r^\mathsf{T}\overline{M}_r\overline{M}_r^\mathsf{T}Q_r.$$

The matrix on the right is positive definite on the columns of $Q_r$, and its eigenvalues are the strictly positive eigenvalues of $\overline{M}_r^\mathsf{T}\overline{M}_r = S_r$. Weyl's inequality and $\|Q_r^\mathsf{T}H_{\mathrm{res}}Q_r\|_{\mathrm{op}} = \rho_r$ give (23). □

*Proof of Theorem 4.4.* Let $\mathcal{U}(t,s)$ be the evolution operator generated by $-B_{rr}(t)$. The lower bound on $B_{rr}$ gives

$$\|\mathcal{U}(t,s)\|_{\mathrm{op}} \le \exp\!\left(-\int_s^t \kappa_r(v)\mathrm{d}v\right).$$

Variation of constants yields

$$u_r(t) = \mathcal{U}(t,t_0)u_r(t_0) + \int_{t_0}^t \mathcal{U}(t,s)\zeta_r(s)\mathrm{d}s.$$

Taking norms proves (24). If $\zeta_r = 0$, then

$$E_r(t) = \|u_r(t)\|^2 \le e^{-2\int_{t_0}^t \kappa_r(s)\mathrm{d}s}E_r(t_0),$$

which is at most $E_r(t_0)/2$ under the stated integrated-curvature condition. □

### A.3 Spectral update

*Proof of Theorem 5.1.* SVD orthogonality and (26) give

$$\langle G, \Phi_{a,\varepsilon}(G)\rangle_{\mathrm{F}} = c_{a,\varepsilon}(G)^{-1} \sum_i \sigma_i w_i^{(a,\varepsilon)}.$$

At least one $\sigma_i$ is positive, and every $w_i^{(a,\varepsilon)}$ is positive, proving strict positivity. The standard descent lemma for an $L_F$-smooth function gives

$$F(W - \eta\Phi) \le F(W) - \eta\langle G, \Phi\rangle_{\mathrm{F}} + \frac{L_F\eta^2}{2}\|\Phi\|_{\mathrm{F}}^2.$$

The unit-RMS normalization gives $\|\Phi\|_{\mathrm{F}}^2 = mn$. The quadratic upper bound is strictly below $F(W)$ under (29). □

*Proof of Theorem 5.2.* The derivative of $S(G) = GG^{\mathsf{T}}$ in direction $\Delta$ is

$$DS(G)[\Delta] = G\Delta^{\mathsf{T}} + \Delta G^{\mathsf{T}} = E.$$

The Daleckii–Krein formula for the spectral matrix function $S^b$ gives

$$D(S^b)[E] = L\Big[F_b^{[1]} \odot (L^{\mathsf{T}}EL)\Big]\, L^{\mathsf{T}}.$$

Applying the product rule to $\Psi_a(G) = S(G)^b G$ proves (30). Finally,

$$\Phi_a(G) = q(G)\Psi_a(G), \qquad q(G) = \sqrt{mn}\,\|\Psi_a(G)\|_{\mathrm{F}}^{-1},$$

and

$$Dq(G)[\Delta] = -q(G)\,\frac{\langle \Psi_a(G), D\Psi_a(G)[\Delta]\rangle_{\mathrm{F}}}{\|\Psi_a(G)\|_{\mathrm{F}}^2}.$$

Another product rule gives (31). □

*Proof of Theorem 7.1.* Since $GG^{\mathsf{T}} = L\,\mathrm{Diag}(\sigma_i^2)L^{\mathsf{T}}$,

$$(GG^{\mathsf{T}})^{(a-1)/2}G = L\,\mathrm{Diag}(\sigma_i^{a-1})L^{\mathsf{T}}L\,\mathrm{Diag}(\sigma_i)R^{\mathsf{T}} = L\,\mathrm{Diag}(\sigma_i^a)R^{\mathsf{T}}.$$

Multiplication by $\bar\sigma^{-a}$ gives (50). The pseudoinverse functional calculus gives the same identity on the positive singular subspace. □

### A.4 Analytic and empirical Hessians

*Proof of Theorem 6.1.* Because

$$\frac{\partial f}{\partial q} = \alpha + 2A^{\mathsf{T}}(\beta \odot r) = d_h, \qquad \frac{\partial q}{\partial h} = 2\,\mathrm{Diag}(h),$$

we obtain $\nabla_h f = 2h \odot d_h$. Direct differentiation in $u$ gives $\nabla_u f = 2\beta \odot r$, proving (35).

Next,

$$\frac{\partial d_h}{\partial h} = 2A^{\mathsf{T}}D_\beta A(2D_h) = 4A^{\mathsf{T}}D_\beta A D_h.$$

Differentiating $2h \odot d_h$ gives

$$K_{hh} = 2\,\mathrm{Diag}(d_h) + 2D_h(4A^{\mathsf{T}}D_\beta A D_h),$$

which is (36). Since $\partial d_h/\partial u = 2A^{\mathsf{T}}D_\beta$,

$$K_{hu} = 2D_h(2A^{\mathsf{T}}D_\beta),$$

and $K_{uu} = 2D_\beta$. Finally, the Hessian chain rule for $\ell = \frac{1}{2}(f-y)^2$ gives

$$\nabla_v^2\ell = (\nabla_v f)(\nabla_v f)^{\mathsf{T}} + (f-y)\nabla_v^2 f.$$

□

*Proof of Theorem 6.2.* For one observation, $\mathrm{d}v = (\mathrm{d}W)x$. Therefore the Hessian bilinear form in direct-weight perturbations $\Delta W, \widetilde{\Delta W}$ is

$$\langle \Delta W x, T_\theta \widetilde{\Delta W} x \rangle,$$

which is represented by $T_\theta \otimes xx^\mathsf{T}$ under vectorization. Projecting both ambient factors by $\Theta_\perp$ replaces $x$ by $\Theta_\perp^\mathsf{T} x = \xi$, proving (41). At the embedded reduced state, $T_\theta$ depends only on $z$ and $y = f_\star(z)$. Gaussian orthogonal coordinates make $z$ and $\xi$ independent, both within and across observations. □

*Proof of Theorem 6.3.* By the triangle inequality and the multiplicativity of the operator norm under Kronecker products,

$$\begin{aligned} \|H_{\perp\perp}^{(n)} - H_{\perp\perp,L}^{(n)}\|_{\mathrm{op}} &\le \frac{1}{n}\sum_\mu \|(T_\mu - \Pi_L(T_\mu)) \otimes \xi_\mu \xi_\mu^\mathsf{T}\|_{\mathrm{op}} \\ &= \frac{1}{n}\sum_\mu \|T_\mu - \Pi_L(T_\mu)\|_{\mathrm{op}} \, \|\xi_\mu \xi_\mu^\mathsf{T}\|_{\mathrm{op}}, \end{aligned}$$

and $\|\xi\xi^\mathsf{T}\|_{\mathrm{op}} = \|\xi\|^2$. □

*Justification of Theorem 6.4.* The clipped blocks are iid, self-adjoint, uniformly bounded, and independent of iid Gaussian $\xi_\mu$. The block size $r_b$ is fixed and the aspect ratio converges to $\alpha$. These are the block-Wishart hypotheses of Montanari and Saeed [24]. Their matrix Stieltjes fixed point and variational edge theorems give (45)–(46) after matching the normalization $n/(d-k) = \alpha$. □

*Proof of Theorem 6.6.* For $\lambda \notin \mathrm{spec}(B_n)$, multiply $H_n - \lambda \mathrm{I}$ on the left by

$$\begin{bmatrix} \mathrm{I} & 0 \\ -C_n^\mathsf{T}(B_n - \lambda \mathrm{I})^{-1} & \mathrm{I} \end{bmatrix},$$

whose determinant is one. The product is block upper triangular with diagonal blocks $B_n - \lambda \mathrm{I}$ and

$$D_n - \lambda \mathrm{I} - C_n^\mathsf{T}(B_n - \lambda \mathrm{I})^{-1} C_n.$$

Taking determinants proves (48). Outside the bulk spectrum, the first determinant is nonzero, so zeros are exactly those of $F_n$. □

### A.5 Hermite-tail transfer

*Proof of Theorem 8.1.* Orthogonality of $P_R$ gives

$$\|\delta\|_{\mathcal{H}}^2 = \|P_R \delta\|_{\mathcal{H}}^2 + \|(\mathrm{I} - P_R)\delta\|_{\mathcal{H}}^2,$$

which proves (67). Hilbert-space differentiation gives

$$D\mathcal{R}[u] = \langle \delta, Dc[u] \rangle_{\mathcal{H}}, \qquad D^2\mathcal{R}[u,v] = \langle Dc[u], Dc[v] \rangle_{\mathcal{H}} + \langle \delta, D^2 c[u,v] \rangle_{\mathcal{H}}.$$

Subtracting the corresponding projected identities leaves only the orthogonal tails. Cauchy–Schwarz yields

$$|D\mathcal{R}[u] - D\mathcal{R}_R[u]| \le \varepsilon_0 \varepsilon_1 \|u\|,$$

and, for unit $u, v$,

$$|D^2\mathcal{R}[u,v] - D^2\mathcal{R}_R[u,v]| \le \varepsilon_1^2 + \varepsilon_0 \varepsilon_2.$$

Taking the appropriate suprema proves (68)–(69). The pointwise argument commutes with a supremum over a compact set. □

*Proof of Theorem 8.2.* Theorem 8.1 bounds the operator-norm perturbation of the two real symmetric $P \times P$ Hessians by $\eta_R$. Weyl's eigenvalue perturbation inequality gives the ordered-eigenvalue claim. If a truncated cluster has gap greater than $2\eta_R$, the closed $\eta_R$-neighborhoods of the cluster and of the remaining spectrum are disjoint, so the ordered eigenvalue count in the cluster is unchanged. □

*Proof of Theorem 8.3.* Assumption (i) makes $\nu_t$ a deterministic functional of the limiting state. Assumption (ii) therefore makes the regular bulk edges deterministic and time dependent. By (iii), each finite-signal resolvent quadratic form in (49) converges locally uniformly outside the bulk, so the finite determinants converge locally uniformly as well. Simple separated zeros are stable under locally uniform convergence. Assumption (iv) licenses the same conclusion for the raw rather than truncated blocks. At contact, regularity of the edge and transversality of the zero distinguish a genuine branch crossing from tangency or spectral sticking. □

# B Reproducibility protocol

## B.1 Reference implementation

The source tree contains:

- `experiments/brainstorm/resnet13/hierarchical3_muon_bbp.py`: paired population/fixed-sample training and fresh Hessian spectroscopy;
- `experiments/brainstorm/resnet13/hierarchical3_sample_scaling.py`: nested sample-size campaign;
- `experiments/brainstorm/resnet13/test_hierarchical3_muon_bbp.py`: focused algebra tests;
- `experiments/brainstorm/resnet13/analyze_hierarchical3_confirmatory.py`: paired bootstrap, exact sign tests, right-censored sector summaries, and confirmatory figures;
- `experiments/brainstorm/resnet13/hierarchical3_spectral_time_panel.py`: dense coupled fresh-probe spectroscopy, exact Schur geometry at every spectral checkpoint, and the master mechanism panel;
- `experiments/brainstorm/resnet13/resnet_order4_full_verify_v2.py`: exact Hermite setup, coefficient projection, and reduced model;
- `experiments/muon/lowrank_block_wishart_spectral_experiments.py`: block-Wishart construction, clipping, and variational-edge solver.

Public static copies of the principal scripts, CSV tables, JSON summaries, and the scientific audit are available at

https://janisaiad.github.io/assets/files/hierarchical3/.

## B.2 Training algorithm

## B.3 Fresh spectral protocol

At a checkpoint, the state is frozen before drawing the fresh sample. Common fresh random numbers are used across optimizers at the same checkpoint to reduce comparison variance. The implementation stores:

**Algorithm 1** Paired three-layer optimizer comparison

**for** $j = 1, \dots, R$ **do**
  draw one initialization $\theta_0^{(j)}$ and one training sample $\mathcal{D}_n^{(j)}$
  **for** $a \in \{\mathrm{GD}, 0, 1/7, 1/3\}$ **do**
    reset $\theta \leftarrow \theta_0^{(j)}$
    **for** $t = 0, \dots, T-1$ **do**
      compute the empirical gradient on the fixed sample
      **if** $a = \mathrm{GD}$ **then**
        apply ordinary gradient descent to every parameter
      **else**
        apply $\Phi_{a,\varepsilon}$ to $V, U, A$ and GD to vectors/biases
      **end if**
      backtrack until the empirical objective does not increase
      at diagnostic steps, evaluate exact population Hermite risk and degree errors
    **end for**
  **end for**
**end for**

- raw and clipped block matrices;
- raw full and orthogonal empirical spectral edges;
- clipped full and orthogonal empirical spectral edges;
- clipped Montanari–Saeed variational edges;
- raw and clipped outlier counts and teacher-subspace overlaps;
- clipping fraction and full-Hessian operator-norm perturbation.

The edge solver never substitutes a raw polynomial block law into the bounded variational theorem.

### B.4 Statistical reporting

Paired comparisons use log risk because optimizer effects are multiplicative and the observed risks span orders of magnitude. For optimizer $a$,

$$\Delta_a^{(j)} = \log \mathcal{R}_a^{(j)} - \log \mathcal{R}_{\mathrm{GD}}^{(j)}.$$

Every stored exact population risk is strictly positive, so no data-dependent pseudocount is introduced. The confirmatory report includes:

1. all paired values, not only means;
2. median and mean paired log ratio;
3. a deterministic-seed percentile paired bootstrap interval;
4. win count and exact two-sided sign-test $p$-value;
5. sector half-error times with an explicit "not reached" state;
6. no universal-exponent fit at fixed latent rank.

### B.5 Commands

From the repository root, the focused tests are

```
PYTHONPATH=$PWD python -m unittest \
  experiments.brainstorm.resnet13.test_hierarchical3_muon_bbp -v
```

and the main pilot is reproduced by

```
PYTHONPATH=$PWD python \
  experiments/brainstorm/resnet13/hierarchical3_muon_bbp.py
```

The confirmatory run changes only the output directory, seed, and number of paired repetitions:

```
PYTHONPATH=$PWD python \
  experiments/brainstorm/resnet13/hierarchical3_muon_bbp.py \
  --outdir experiments/brainstorm/resnet13/results/\
hierarchical3_muon_bbp_confirmatory_r20 \
  --seed 424242 --repeats 20 --steps 1200 --diag-every 40
```

The stored trajectories are analyzed by

```
PYTHONPATH=$PWD python \
  experiments/brainstorm/resnet13/\
analyze_hierarchical3_confirmatory.py \
  experiments/brainstorm/resnet13/results/\
hierarchical3_muon_bbp_confirmatory_r20/training_trajectories.csv
```

The time-resolved population/Schur/spectral panel is reproduced by

```
PYTHONPATH=$PWD python \
  experiments/brainstorm/resnet13/\
hierarchical3_spectral_time_panel.py
```

Its tables include every plotted eigenvalue, every empirical orthogonal edge, and the exact finite Schur eigenvalues and residual-curvature radii.

### B.6 Numerical claim boundaries

All reported endpoint statistics are computed from stored CSV rows and checked against the JSON summary. Figures are regenerated from those rows. Numerical agreement of an analytic formula with automatic differentiation verifies the implementation of that identity; it is not a proof of an asymptotic theorem. Conversely, the exact finite identities in section 3, section 4, and section 6 do not depend on the displayed finite simulations.

## References


[1] Robert J. Adler and Jonathan E. Taylor. *Random Fields and Geometry*. Springer Monographs in Mathematics. Springer, 2007. doi: 10.1007/978-0-387-48116-6.

[2] Brandon Livio Annesi, Dario Bocchi, and Chiara Cammarota. Overparametrization bends the landscape: BBP transitions at initialization in simple neural networks. *arXiv preprint arXiv:2510.18435*, 2025. doi: 10.48550/arXiv.2510.18435. URL https://arxiv.org/abs/2510.18435.

[3] Kiana Asgari, Andrea Montanari, and Basil Saeed. Local minima of the empirical risk in high dimension: General theorems and convex examples. *arXiv preprint arXiv:2502.01953*, 2025. doi: 10.48550/arXiv.2502.01953. URL https://arxiv.org/abs/2502.01953.

[4] Antonio Auffinger, Gérard Ben Arous, and Jiří Černý. Random matrices and complexity of spin glasses. *Communications on Pure and Applied Mathematics*, 66(2):165–201, 2013. doi: 10.1002/cpa.21422. URL https://arxiv.org/abs/1003.1129.

[5] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33 (5):1643–1697, 2005. doi: 10.1214/009117905000000233.

[6] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2): 764–785, 2011. doi: 10.1109/TIT.2010.2094817. URL https://arxiv.org/abs/1001.3448.

[7] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1): 494–521, 2011. doi: 10.1016/j.aim.2011.02.007. URL https://arxiv.org/abs/0910.2120.

[8] Raphaël Berthier, Andrea Montanari, and Phan-Minh Nguyen. State evolution for approximate message passing with non-separable functions. *Information and Inference*, 9(1):33–79, 2020. doi: 10.1093/imaiai/iay021. URL https://arxiv.org/abs/1708.03950.

[9] Dario Bocchi, Giulio Biroli, Chiara Cammarota, and Federico Ricci-Tersenghi. Discontinuous BBP transitions. *arXiv preprint arXiv:2604.27992*, 2026. doi: 10.48550/arXiv.2604.27992. URL https://arxiv.org/abs/2604.27992.

[10] Erwin Bolthausen. An iterative construction of solutions of the TAP equations for the sherrington–kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366, 2014. doi: 10.1007/s00220-013-1862-3. URL https://arxiv.org/abs/1201.2891.

[11] Tony Bonnaire, Giulio Biroli, and Chiara Cammarota. The role of the time-dependent hessian in high-dimensional optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(8):083401, 2025. doi: 10.1088/1742-5468/adf4bc. URL https://arxiv.org/abs/2403.02418.

[12] Guillaume Braun, Bruno Loureiro, Ha Quang Minh, and Masaaki Imaizumi. Fast escape, slow convergence: Learning dynamics of phase retrieval under power-law data. *arXiv preprint arXiv:2511.18661*, 2025. doi: 10.48550/arXiv.2511.18661. URL https://arxiv.org/abs/2511.18661.

[13] Florentin Coeurdoux, Grégoire Ferré, and Jean-Philippe Bouchaud. Random matrix theory of early-stopped gradient flow: A transient BBP scenario. *arXiv preprint arXiv:2604.18450*, 2026. doi: 10.48550/arXiv.2604.18450. URL https://arxiv.org/abs/2604.18450.

[14] Yatin Dandi, Luca Pesce, Lenka Zdeborová, and Florent Krzakala. The computational advantage of depth: Learning high-dimensional hierarchical functions with gradient descent. *Advances in Neural Information Processing Systems*, 2025. doi: 10.48550/arXiv.2502.13961. URL https://arxiv.org/abs/2502.13961. Spotlight; arXiv:2502.13961v4.

[15] Michael Dick, Alexander van Meegen, and Moritz Helias. Linking network- and neuron-level correlations by renormalized field theory. *Physical Review Research*, 6:013183, 2024. doi: 10.1103/PhysRevResearch.6.013183. URL https://arxiv.org/abs/2309.14973.

[16] Luca Giammanco, Pietro Valigi, and Chiara Cammarota. Spectral properties and phase diagrams of sparse antagonistic random matrices with diagonal disorder and jacobian-like structure. *arXiv preprint arXiv:2606.25925*, 2026. doi: 10.48550/arXiv.2606.25925. URL https://arxiv.org/abs/2606.25925.

[17] Moritz Helias and David Dahmen. *Statistical Field Theory for Neural Networks*, volume 970 of *Lecture Notes in Physics*. Springer, 2020. doi: 10.1007/978-3-030-46444-8. URL https://arxiv.org/abs/1901.10416.

[18] Mark Kac. A correction to "on the average number of real roots of a random algebraic equation". *Bulletin of the American Mathematical Society*, 49(12):938, 1943. doi: 10.1090/S0002-9904-1943-08069-X.

[19] Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169:257–352, 2017. doi: 10.1007/s00440-016-0730-4. URL https://arxiv.org/abs/1410.3516.

[20] Jakob Kramp, Javed Lindner, and Moritz Helias. Dynamics of neural scaling laws in random feature regression with powerlaw-distributed kernel eigenvalues. *arXiv preprint arXiv:2602.23039v2*, 2026. doi: 10.48550/arXiv.2602.23039. URL https://arxiv.org/abs/2602.23039.

[21] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7):073403, 2017. doi: 10.1088/1742-5468/aa7284. URL https://arxiv.org/abs/1701.00858.

[22] Antoine Maillard, Gérard Ben Arous, and Giulio Biroli. Landscape complexity for the empirical risk of generalized linear models. In *Mathematical and Scientific Machine Learning*, volume 107 of *Proceedings of Machine Learning Research*, pages 287–327, 2020. URL https://arxiv.org/abs/1912.02143.

[23] Antoine Maillard, Tony Bonnaire, and Giulio Biroli. Topological exploration of high-dimensional empirical risk landscapes: General approach, and applications to phase retrieval. *arXiv preprint arXiv:2602.17779*, 2026. doi: 10.48550/arXiv.2602.17779. URL https://arxiv.org/abs/2602.17779.

[24] Andrea Montanari and Basil Saeed. Variational formulas for the spectrum of block wishart matrices. *arXiv preprint arXiv:2606.27774*, 2026. doi: 10.48550/arXiv.2606.27774. URL https://arxiv.org/abs/2606.27774.

[25] Andrea Montanari and Basil Saeed. Topological trivialization in non-convex empirical risk minimization. *arXiv preprint arXiv:2602.14969*, 2026. doi: 10.48550/arXiv.2602.14969. URL https://arxiv.org/abs/2602.14969.

[26] Andrea Montanari and Basil Saeed. Universality of empirical risk minimization. *arXiv preprint arXiv:2202.08832v3*, 2026. doi: 10.48550/arXiv.2202.08832. URL https://arxiv.org/abs/2202.08832.

[27] Izaak Neri and Fernando L. Metz. Spectra of sparse non-hermitian random matrices: An analytical solution. *Physical Review Letters*, 109:030602, 2012. doi: 10.1103/PhysRevLett.109.030602. URL https://arxiv.org/abs/1205.0702.

[28] Izaak Neri and Fernando L. Metz. Eigenvalue outliers of non-hermitian random matrices with a local tree structure. *Physical Review Letters*, 117:224101, 2016. doi: 10.1103/PhysRevLett.117.224101.

[29] Eshaan Nichani, Alex Damian, and Jason D. Lee. Provable guarantees for nonlinear feature learning in three-layer neural networks. *arXiv preprint arXiv:2305.06986v3*, 2025. doi: 10.48550/arXiv.2305.06986. URL https://arxiv.org/abs/2305.06986.

[30] Elliot Paquette, Noah Marshall, Lucas Benigni, Guangyuan Wang, Atish Agarwala, and Courtney Paquette. Phases of muon: When muon eclipses SignSGD. *arXiv preprint arXiv:2605.09552*, 2026. doi: 10.48550/arXiv.2605.09552. URL https://arxiv.org/abs/2605.09552.

[31] Loucas Pillaud-Vivien and Adrien Schertzer. Joint learning in the gaussian single index model. *arXiv preprint arXiv:2505.21336*, 2025. doi: 10.48550/arXiv.2505.21336. URL https://arxiv.org/abs/2505.21336.

[32] Sundeep Rangan, Philip Schniter, and Alyson K. Fletcher. Vector approximate message passing. *IEEE Transactions on Information Theory*, 65(10):6664–6684, 2019. doi: 10.1109/TIT.2019.2916359. URL https://arxiv.org/abs/1610.03082.

[33] Stephen O. Rice. Mathematical analysis of random noise. *Bell System Technical Journal*, 24(1):46–156, 1945. doi: 10.1002/j.1538-7305.1945.tb00453.x.

[34] Tim Rogers, Isaac Pérez Castillo, Reimer Kühn, and Koujin Takeda. Cavity approach to the spectral density of sparse symmetric random matrices. *Physical Review E*, 78:031116, 2008. doi: 10.1103/PhysRevE.78.031116. URL https://arxiv.org/abs/0803.1553.

[35] Kai Segadlo, Bastian Epping, Alexander van Meegen, David Dahmen, Michael Krämer, and Moritz Helias. Unified field theoretical approach to deep and recurrent neuronal networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(10):103401, 2022. doi: 10.1088/1742-5468/ac8e57. URL https://arxiv.org/abs/2112.05589.

[36] Alexander van Meegen and Haim Sompolinsky. Coding schemes in neural networks learning classification tasks. *Nature Communications*, 16:3354, 2025. doi: 10.1038/s41467-025-58276-6. URL https://arxiv.org/abs/2406.16689.

[37] Zihao Wang, Eshaan Nichani, and Jason D. Lee. Learning hierarchical polynomials with three-layer neural networks. *arXiv preprint arXiv:2311.13774*, 2023. doi: 10.48550/arXiv.2311.13774. URL https://arxiv.org/abs/2311.13774.

[38] Arie Wortsman-Zurich, Hugo Tabanelli, Yatin Dandi, Florent Krzakala, and Bruno Loureiro. Scaling laws from sequential feature recovery: A solvable hierarchical model. *arXiv preprint arXiv:2605.14567*, 2026. doi: 10.48550/arXiv.2605.14567. URL https://arxiv.org/abs/2605.14567.