Muon as Dynamic Spectral Denoising for Power-Law Phase Retrieval

Anonymous

July 6, 2026

Abstract

We study a quadratic multi-index phase retrieval model with a power-law teacher spectrum. The weak teacher directions are already present in the population gradient, but at finite dimension they are useful only after the corresponding eigenvalues detach from the random bulk. This produces a hierarchy of BBP times during training.

The optimizer considered here is Muon with a singular-value power a. In the quadratic model the population dynamics are first reduced exactly to the finite frame spanned by the student weights and the teacher directions. After this reduction, the spectral-update formulas of Paquette et al. give a Volterra equation for the active risk scale. For a locally power-law front, a Laplace estimate of this Volterra equation gives the fixed-exponent balance

$$a_{\mathrm{PR},*}(\zeta) = \mathrm{clip}_{[0,1]}\left(\frac{\zeta - 1}{3\zeta - 1}\right),$$

where ζ is the local tail exponent of the modes currently close to the visibility threshold. The same state also predicts the gradient, weight, and Hessian BBP transitions by eliminating the random bulk and solving a finite outlier equation. Finally, allowing a to vary leads to a reduced control problem: the local term projects the linearized AMP/VAMP spectral denoiser onto the Muon power family, while the dynamic term prices future BBP margins.

1 Introduction

Many high-dimensional learning problems contain signal before that signal is spectrally visible. At initialization, the matrices seen by the optimizer are dominated by a random bulk. A direction becomes identifiable only when it creates an isolated eigenvalue outside that bulk. This is the Baik–Ben Arous–Péché transition. In static inference it is the detection threshold of a spike. In training dynamics it is the time at which an informative representation becomes visible.

The model in this paper is a quadratic teacher-student phase retrieval problem. The teacher has strengths μ_i, and we focus on the hierarchical regime $\mu_i \simeq i^{-\gamma}$. The large modes detach early. The weak modes form a long tail: they influence the population gradient, but finite matrices still hide them in the bulk. Thus the relevant question is not only how fast the loss decreases, but when each teacher direction becomes spectrally observable.

Muon acts directly on this visibility problem. Given a gradient matrix with singular values s, Muon applies a power map $s \mapsto s^a$. The case $a = 1$ is ordinary gradient descent, while $a = 0$ is the sign-SVD update that gives the same weight to all nonzero singular directions. Varying a therefore changes the conditioning of weak spectral directions. The purpose of the paper is to compute this effect in a setting where both the population dynamics and the finite-dimensional spectra can be followed.

The first simplification is algebraic. For the quadratic loss, the population gradient always lies in the span of the current student weights and the teacher directions. The same is true after applying the Muon singular-value map. Consequently the high-dimensional population

trajectory is described by a finite frame: the student Gram matrix and the student–teacher overlaps. No random-matrix input is needed for this reduction.

The second step is dynamical. Once the finite frame is fixed, Paquette's analysis of spectral optimizers gives the drift and volatility of each active mode for a general spectral filter. Specializing that result to the Muon power filter produces a Volterra equation for the active risk scale. On a locally power-law front this equation has two competing costs: one coming from the hard edge of the singular-value map, the other from resolving the tail. Balancing them gives

$$a_{\mathrm{PR},*}(\zeta) = \mathrm{clip}_{[0,1]}\left(\frac{\zeta - 1}{3\zeta - 1}\right),$$

where ζ is the local slope of the modes currently close to the visibility threshold. The formula is local in the moving front; if the front has several slopes, the corresponding object is blockwise or a full spectral filter.

The third step is spectral. A positive population overlap does not by itself produce a visible eigenvalue. From the same finite state we compute the bulk edges and the outlier branches in the gradient, weight, and Hessian spectra. The random bulk is eliminated by its resolvent, leaving a finite outlier equation. Its roots give the outlier locations, and its derivative gives the eigenvector residue, namely the teacher mass carried by the detached branch. This is the quantity that remains stable when empirical eigenvalue ranks exchange near a dense front.

Finally, if the exponent a is allowed to vary with time, it should be chosen on the same reduced state. Locally, the best spectral direction is the linearized AMP/VAMP denoiser, which amplifies singular values according to the signal-to-bulk likelihood ratio. When that ratio is locally log-linear, its projection onto Muon powers is again a Muon exponent. Dynamically, the current choice of exponent changes future BBP margins, so the reduced state leads to a Hamilton–Jacobi–Bellman equation. The Boltzmann rule used below is the entropy-regularized version of this selector.

All formulas are first stated for a fixed population path, or for a fresh spectral sample drawn after the path is fixed. This separates the deterministic mechanism from the same-sample dependence created when the training data are also used to form the spectra. The latter is a leave-one-out comparison, recorded in Appendix E. It transfers the same deterministic equations to the reused-sample setting without changing the Volterra state or the outlier equations derived in the main text.

2 Model, population loss, and empirical loss

Let $x \sim N(0, I_d)$. The teacher has k orthonormal directions $\Theta = (\theta_1, \dots, \theta_k) \in \mathbb{R}^{d\times k}$ and strengths

$$\Lambda = \mathrm{diag}(\mu_1, \dots, \mu_k), \qquad \mu_i = \mu_0 i^{-\gamma}, \qquad \gamma > 1/2.$$

The student has width P and weights

$$W = (w_1, \dots, w_P) \in \mathbb{R}^{d\times P}.$$

The teacher and student functions are

$$f_\star(x) = \sum_{i=1}^{k} \mu_i (\theta_i^\top x)^2, \qquad f_W(x) = \frac{1}{P}\sum_{a=1}^{P} (w_a^\top x)^2.$$

Equivalently,

$$f_\star(x) = x^\top A_\star x, \qquad A_\star = \Theta\Lambda\Theta^\top,$$

and

$$f_W(x) = x^\top \Sigma_W x, \qquad \Sigma_W = \frac{1}{P} W W^\top.$$

The error matrix and its trace are

$$E_W = \Sigma_W - A_\star, \qquad \tau_W = \operatorname{Tr} E_W.$$

The population loss is

$$R(W) = \frac{1}{2}\mathbb{E}\big(f_W(x) - f_\star(x)\big)^2 = \operatorname{Tr}(E_W^2) + \frac{1}{2}\tau_W^2. \tag{2.1}$$

Given samples $x_1, \dots, x_n$, the empirical loss is

$$R_n(W) = \frac{1}{2n}\sum_{\ell=1}^{n}\big(f_W(x_\ell) - f_\star(x_\ell)\big)^2 = \frac{1}{2n}\sum_{\ell=1}^{n}(x_\ell^\top E_W x_\ell)^2. \tag{2.2}$$

The gradients are

$$G(W) = \nabla_W R(W) = \frac{2}{P}(2E_W + \tau_W I_d)W, \tag{2.3}$$

and

$$G_n(W) = \nabla_W R_n(W) = \frac{2}{Pn}\sum_{\ell=1}^{n}(x_\ell^\top E_W x_\ell)x_\ell x_\ell^\top W. \tag{2.4}$$

Equations (2.1)–(2.4) fix the population and empirical objects, but they play different roles. The finite-frame ODE in the next section is a population statement, based on (2.3). Empirical matrices enter later as spectral probes of a path whose low-dimensional state has already been defined.

Three finite summaries recur throughout the paper. The student Gram matrix is $Q = W^\top W$, the student–teacher overlap is $M = W^\top \Theta$, and

$$C = M^\top Q^{-1} M$$

is the part of the teacher subspace already captured by the student span. In a separated scalar window, $r_i(t)$ denotes the captured overlap of mode i, and

$$\rho_i(t) = \mu_i(1 - r_i(t))$$

is the corresponding residual signal strength. In the Volterra reduction, $\chi_i(t)$ denotes the contribution of mode i to the active risk scale.

3 Gradient descent and Muon in the same finite frame

The key algebraic fact is that the population gradient lives in the finite span of the student and teacher directions. Define

$$Z = [W, \Theta], \qquad \Gamma = \begin{pmatrix} Q & M \\ M^\top & I_k \end{pmatrix}.$$

Then $G(W) = ZA_{\mathrm{gd}}$, where

$$A_{\mathrm{gd}} = \begin{pmatrix} \frac{2}{P}\left(\frac{2}{P}Q + \tau_W I_P\right) \\ -\frac{4}{P}\Lambda M^\top \end{pmatrix}.$$

Population gradient flow $\dot W = -G(W)$ gives

$$\dot Q = -\frac{4}{P}\left[2\left(\frac{1}{P}Q^2 - M\Lambda M^\top\right) + \tau_W Q\right],$$

and

$$\dot{M} = -\frac{2}{P}\left[2\left(\frac{1}{P}QM - M\Lambda\right) + \tau_W M\right].$$

Consequently

$$\dot{C} = \frac{4}{P}(\Lambda C + C\Lambda - 2C\Lambda C). \tag{3.1}$$

In a scalar separated regime this becomes

$$\dot{r}_i = \frac{8\mu_i}{P} r_i(1 - r_i).$$

Muon changes the update map: it replaces a matrix Y by a singular-value power. If $Y = U\,\mathrm{diag}(s_j)V^\top$, define

$$\mathsf{M}_a(Y) = U\,\mathrm{diag}(\psi_a(s_j))V^\top, \qquad \psi_a(s) = s^a \ (s > 0), \quad \psi_a(0) = 0.$$

Thus $a = 1$ is ordinary gradient descent and $a = 0$ replaces every nonzero singular value by one.

Since $G(W) = ZA_{\mathrm{gd}}$, write

$$\mathsf{M}_a(G(W)) = ZA_a, \qquad A_a = \Gamma^{-1/2}\mathsf{M}_a(\Gamma^{1/2}A_{\mathrm{gd}}).$$

Split

$$A_a = \begin{pmatrix} U_a \\ V_a \end{pmatrix}, \qquad U_a \in \mathbb{R}^{P\times P}, \quad V_a \in \mathbb{R}^{k\times P}.$$

The block U_a is the component of the update inside the student span. The block V_a is the component pointing toward teacher directions. The population Muon flow is

$$\dot{W} = -\eta_a(t)(WU_a + \Theta V_a).$$

It gives the exact finite ODE

$$\dot{Q} = -\eta_a\left(U_a^\top Q + QU_a + V_a^\top M^\top + MV_a\right),$$

and

$$\dot{M} = -\eta_a(U_a^\top M + V_a^\top).$$

Finally, with $D_a = V_a Q^{-1} M$,

$$\dot{C} = -\eta_a\left[(I_k - C)D_a + D_a^\top(I_k - C)\right]. \tag{3.2}$$

Equation (3.2) is the exact Muon counterpart of the gradient Riccati equation (3.1). The comparison between gradient flow and Muon therefore begins at the deterministic population level, before any random-matrix approximation is introduced.

The following proposition is the algebraic compression step. It replaces the moving matrix W_t by a finite state whose modewise components are the inputs to Paquette's recursion.

Proposition 3.1 (Finite-frame closure)**.** *For the quadratic phase retrieval loss, both gradient flow and population Muon-a flow remain in the finite span of the current student directions and the teacher directions. Consequently their deterministic population dynamics close on the finite matrices Q and M. For gradient flow the captured teacher matrix $C = M^\top Q^{-1} M$ obeys* (3.1)*; for Muon-a it obeys* (3.2)*.*

Proof. The population gradient has the form $G(W) = ZA_{\mathrm{gd}}$ with $Z = [W, \Theta]$. The Muon update is a singular-value functional calculus of the same finite-frame matrix, hence can be written ZA_a. Differentiating $Q = W^\top W$, $M = W^\top\Theta$, and $C = M^\top Q^{-1} M$ gives the displayed equations. No high-dimensional limit is used in this step. □

This closure is the entry point of the rest of the paper. It replaces a moving $d \times P$ matrix by a finite deterministic state. The Volterra equation in the next section is obtained by projecting this state on the active teacher modes.

4 Learning curves at a fixed exponent

The finite-frame Gram matrix remains denoted by $Q = W^\top W$. The scalar quantity that drives the Volterra clock is denoted by $\mathfrak{q}(t)$. It is the aggregate risk scale obtained after projecting the dynamics onto the active mode variables. With this convention, Q is always a matrix and $\mathfrak{q}$ is always a scalar.

For a fixed a, we use the deterministic recursion of Paquette et al. [3] for spectral updates of the form $G\varphi(G^\top G)$. In the present model, this recursion has a concrete modewise form. Once the current finite frame is fixed, their resolvent calculation gives two scalar coefficients for each teacher mode: a mean contraction coefficient and a quadratic fluctuation coefficient. Projecting the recursion on mode i therefore gives an equation for the mode scale seen by the risk. In the single active-front reduction, this equation has the form

$$\dot{\chi}_i(t) = -2\eta\, \delta_i^{(a)} \mathfrak{q}(t)^{(a-1)/2} \chi_i(t) + \eta^2 \nu_i^{(a)} \mathfrak{q}(t)^a. \tag{4.1}$$

Here $\chi_i(t)$ is the contribution of mode i to the risk scale and $\mathfrak{q}(t)$ is the scalar aggregate risk scale. The coefficient $\delta_i^{(a)}$ is the deterministic contraction of that mode, while $\nu_i^{(a)}$ is the variance injected by the stochastic spectral update. Paquette's theorem expresses these two coefficients by one- and two-resolvent formulas for a general filter φ. The present specialization evaluates those formulas at the Muon filter and then translates the resulting mode-by-mode drift and variance into the phase-retrieval front variables.

Paquette studies updates of the form

$$\widetilde{G} = G\,\varphi(G^\top G).$$

Muon-a corresponds to

$$\varphi_{a,\varepsilon}(s) = (s+\varepsilon^2)^{(a-1)/2}, \qquad \phi_{a,\varepsilon}(\sigma) = \sigma(\sigma^2+\varepsilon^2)^{(a-1)/2}. \tag{4.2}$$

The small ε regularizes the hard edge, namely the singular behavior of $s^{(a-1)/2}$ near $s = 0$ when $a < 1$. The formulas in this section are first read with this regularization; in regimes where the hard-edge bounds are uniform, one then lets $\varepsilon \downarrow 0$. With this substitution, Paquette's one-resolvent coefficient is $\delta_i^{(a)}$, and the two-resolvent coefficient is $\nu_i^{(a)}$. The forcing F_a and the kernel K_a in the Volterra equation (4.3) are then fixed by the mode equation: explicitly, Appendix C gives

$$F_a(\tau) = \sum_i w_i e^{-2\delta_i^{(a)}\tau} \chi_i(0), \qquad K_a(s) = \sum_i w_i \nu_i^{(a)} e^{-2\delta_i^{(a)} s}.$$

Thus the Volterra law contains no fitted coefficient once the Paquette resolvent coefficients have been evaluated.

Introduce the rescaled time

$$\mathrm{d}\tau = \eta \mathfrak{q}(t)^{(a-1)/2}\,\mathrm{d}t.$$

Solving (4.1) mode by mode and summing produces the Volterra equation

$$\mathfrak{q}(\tau) = F_a(\tau) + \eta \int_0^\tau K_a(\tau - u)\mathfrak{q}(u)^{(a+1)/2}\,\mathrm{d}u. \tag{4.3}$$

The forcing F_a is the propagated initial condition. The kernel K_a is the accumulated fluctuation response. This equation is the main deterministic object for fixed Muon-a: all learning-curve predictions in this paper are obtained from it, and the spectral predictions later read their signal strengths from the same variables χ_i and $\mathfrak{q}$.

For a power-law teacher, the single-front Laplace tail calculation gives

$$\mathcal{E}_T(a) \asymp T^{-\kappa(a)}, \qquad \kappa(a) = \frac{2\gamma - 1}{\gamma(a+1)}$$

in the ideal tail regime. Thus, if all exponents are equally stable, the smallest safe a has the fastest tail decay. Finite-dimensional BBP constraints can nevertheless favor a positive exponent before the asymptotic tail regime is reached.

The form used below is local in the active front, not global in the whole spectrum. At a given time, the active front is the group of modes whose signal scale is comparable with the current bulk visibility scale. On that window we write

$$\mu_i \asymp i^{-\zeta}, \qquad \chi_i(0) \asymp i^{-\beta}, \qquad \zeta + \beta > 1.$$

For the pure power-law teacher in Section 2, the local exponent ζ is simply γ. We keep a separate letter because a moving front may have a local slope that differs from the global tail exponent. This is also how the local power law is assessed in finite experiments: one plots the residual signal scale against the mode index and estimates the slope only on the window that is close to the bulk scale. Modes that are already well separated and modes that are far below the bulk are excluded from the local estimate of ζ. The same log–log window gives β, the slope of the initial mode scale entering the Volterra forcing. Thus the balance formula has a direct empirical input: the active front determines (ζ, β), and the deterministic Volterra equation then predicts both the preferred exponent and the learning curve. If the selected window is not close to linear, the front is split into approximately linear blocks, or the full spectral filter is kept. This local fit is part of the reduction. When the selected window is close to a straight segment on the log–log scale, a scalar exponent is meaningful. When the front has visible curvature, the same calculation is applied block by block, or replaced by the full spectral filter.

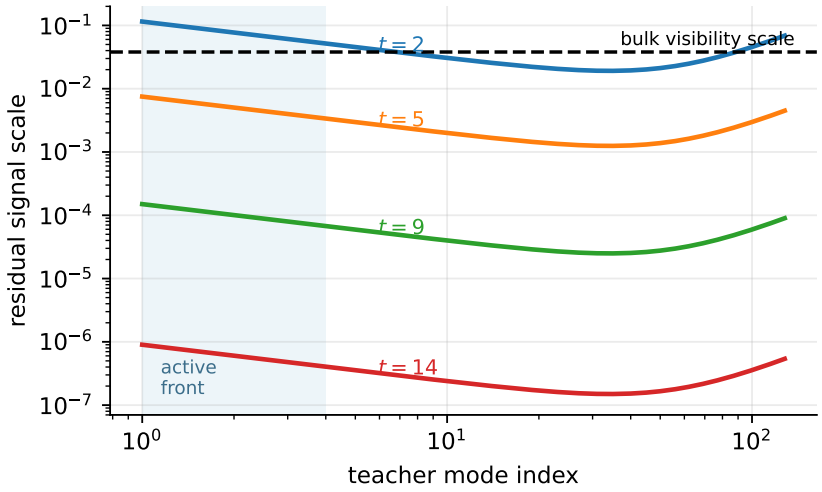


Figure 1: Local active front. The power-law approximation is local: the slope is estimated on the shaded group of modes near the current visibility scale, not on the whole tail. This is the quantity denoted by ζ in (4.4). If the selected window is not close to a straight segment on this log–log scale, the scalar exponent is replaced by a blockwise or full spectral filter; this figure is the empirical check of that reduction.

The single-front Laplace estimate of the Volterra equation gives two competing resolution costs. The first comes from the hard edge of the filter: when a is small, weak singular directions are strongly amplified, but the hard-edge regularization has to work harder. The second comes from the power-law tail: when a is large, the update is more stable, but the weak tail is resolved more slowly. On the local front these two costs have powers

$$e_{\mathrm{hard}}(a) = \frac{1-a}{2}, \qquad e_{\mathrm{tail}}(a) = \frac{\zeta(1+a)}{2(\zeta+\beta-1)}. \tag{4.4}$$

The bottleneck exponent is therefore

$$e_{\mathrm{bal}}(a) = \max\{e_{\mathrm{hard}}(a), e_{\mathrm{tail}}(a)\}. \tag{4.5}$$

Smaller values of e_{bal} are better. Balancing the two terms gives

$$a_{\mathrm{bal},*}(\zeta, \beta) = \mathrm{clip}_{[0,1]}\left(\frac{\beta - 1}{2\zeta + \beta - 1}\right), \tag{4.6}$$

where $\mathrm{clip}_{[0,1]}$ denotes clipping to the interval $[0, 1]$. In the phase-retrieval near-zero regime we use $\beta = \zeta$, hence

$$a_{\mathrm{PR},*}(\zeta) = \mathrm{clip}_{[0,1]}\left(\frac{\zeta - 1}{3\zeta - 1}\right). \tag{4.7}$$

The equality $\beta = \zeta$ means that, before a mode is captured, the initial scale entering the active front has the same local power as the teacher residual. If an initialization or a previous training phase changes that slope, the formula to use is the two-parameter version (4.6). The exponent is obtained by equalizing the two powers in (4.4). This is the fixed-a prediction associated with a locally power-law front: once the local tail slope is known, the Volterra equation gives both a candidate exponent and a whole learning curve. At finite dimension and finite horizon the optimum can be broad; the stable prediction is the shape of the learning curve and the BBP times, rather than the second decimal of the minimizing exponent. The hypotheses behind the balance are displayed in the same order in which they enter the proof. Figure 1 shows the local front slope. Figure 2 shows the corresponding Volterra clock and tail exponent. The later spectral figures then compare the same clock with the bulk edge and the BBP exit times. In this way the fixed-a prediction is tested through intermediate quantities, not only through a terminal loss. The finite-dimensional comparisons therefore have a fixed order. The active front supplies the local slopes; the Volterra equation supplies the risk clock; the reduced outlier equations supply the spectral exit times. The main figures are arranged in this order, so that each observable is displayed before it is used.

We record the fixed-a balance in the form used by the spectral analysis. The random-matrix input is the computation of the mode coefficients $\delta_i^{(a)}$ and $\nu_i^{(a)}$ in (4.1). Once the local Volterra/Laplace costs are (4.4), the optimization over the scalar exponent is elementary.

Proposition 4.1 (Fixed-exponent balance)**.** *Work on a time window where the active front is described by one local power-law slope, $\mu_i \asymp i^{-\zeta}$, $\chi_i(0) \asymp i^{-\beta}$, and $\zeta + \beta > 1$. Suppose the local Volterra/Laplace estimate has a hard-edge cost and a tail-resolution cost with exponents* (4.4)*. Then the exponent that minimizes the bottleneck exponent $e_{\mathrm{bal}}(a)$ over $a \in [0, 1]$ is* (4.6)*. In the near-zero phase retrieval regime $\beta = \zeta$, this reduces to* (4.7)*.*

Proof. The function e_{hard} is decreasing in a, while e_{tail} is increasing in a. Therefore the minimum of $\max(e_{\mathrm{hard}}, e_{\mathrm{tail}})$ over an interval is either the unique intersection or an endpoint. The intersection is obtained from

$$\frac{1 - a}{2} = \frac{\zeta(1 + a)}{2(\zeta + \beta - 1)}.$$

Multiplying by $2(\zeta + \beta - 1)$ gives

$$(1 - a)(\zeta + \beta - 1) = \zeta(1 + a).$$

Expanding and collecting the a-terms gives

$$\beta - 1 = a(2\zeta + \beta - 1).$$

Thus the unconstrained minimizer is

$$a = \frac{\beta - 1}{2\zeta + \beta - 1}.$$

Restricting to Muon exponents $a \in [0, 1]$ clips this value to the interval, which gives (4.6). In the near-zero phase retrieval window the initial mode scale has the same local exponent as the signal, $\beta = \zeta$, and substitution gives

$$a_{\mathrm{PR},*}(\zeta) = \mathrm{clip}_{[0,1]}\left(\frac{\zeta - 1}{3\zeta - 1}\right). \qquad \square$$

For constant a, Proposition 4.1 gives the tail law shown in Figure 2. For variable $a(t)$, the same balance becomes local in time and is tracked as the active front moves. If Figure 1 shows several slopes at once, the scalar formula is a local block approximation. The global replacement is then a blockwise exponent or the full spectral filter.

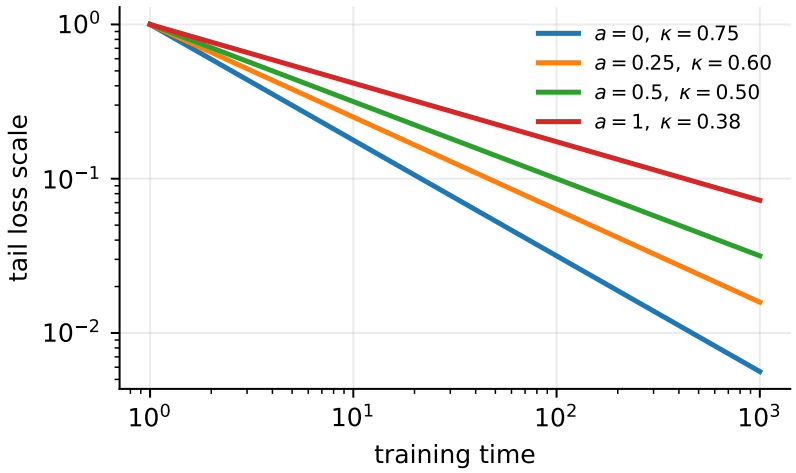


Figure 2: Fixed-a tail law. The exponent a changes the predicted decay $T^{-\kappa(a)}$ in the power-law tail. Smaller admissible values of a learn the tail faster, until finite-size or BBP-visibility constraints become active. This plot is the Volterra/Laplace prediction after the local front slope has been fixed; the finite runs are then checked against this clock and against the spectral exit times, not against a terminal-point fit.

5 Visibility in the three spectra

The Volterra state tells us how much signal exists at time t. A spectrum tells us whether that signal is visible at finite dimension. This distinction is essential: a teacher mode can already have positive population overlap while its associated eigenvalue is still buried inside the random bulk.

We work on intervals where the relevant branch is either separated from the bulk or crosses the edge with nonzero relative velocity. Near a dense group of modes, empirical eigenvalue ranks can exchange. The label of a branch is therefore not its instantaneous rank but the pair consisting of its reduced outlier root and its residue. This regular-crossing condition has a concrete numerical signature: the predicted margin must change sign over the same time window where the matched empirical branch separates from the moving bulk edge. Figures 3–6 test exactly this signature for the bulk scale, the weight spectrum, and the Hessian spectrum.

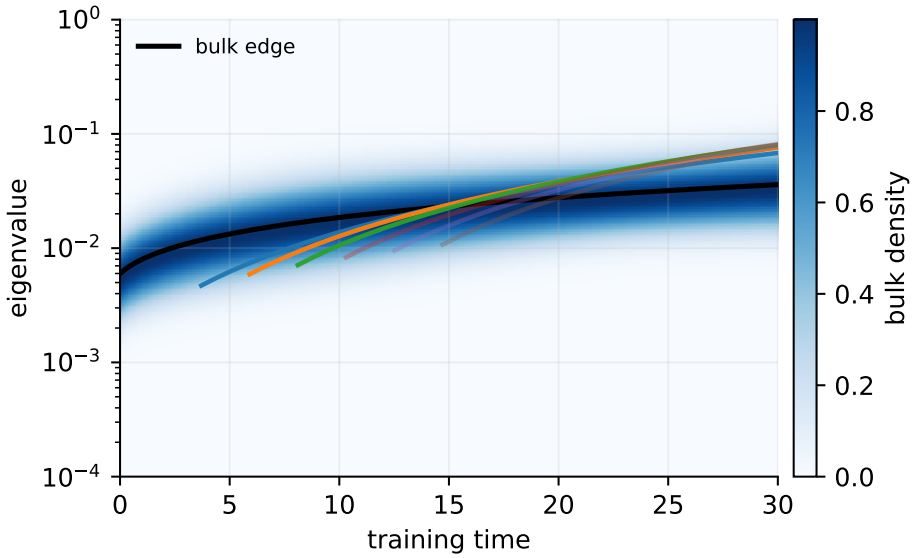


Figure 3: Bulk and outlier branches. The heat map is the moving empirical bulk, the black curve is the deterministic edge, and colored curves are predicted outlier branches. This is the check that the bulk scale used in the BBP calculation is the one seen by the finite spectrum.

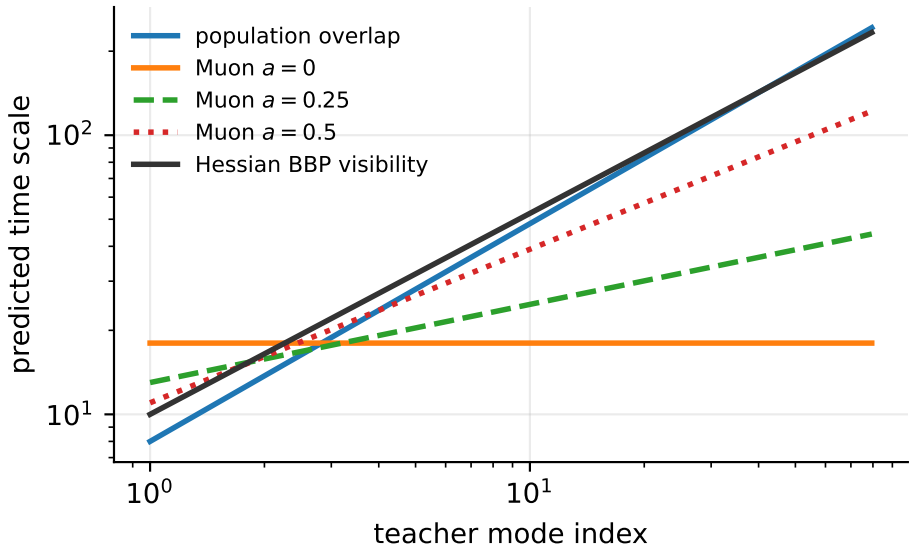


Figure 4: Three exit-time notions. Population gradient flow asks when an overlap becomes $O(1)$. Muon asks how the same overlap changes after a singular-value power map. BBP asks when an eigenvalue leaves the current bulk. The plot checks that the observed spectral exits follow the BBP clock, not the raw population-overlap clock.

5.1 Outliers after eliminating the bulk

The same outlier calculation is used for all three matrices. The label $c \in \{G, W, H\}$ refers respectively to three matrices seen during training: the current gradient, the weight covariance, and the empirical Hessian. The gradient is rectangular, so we replace it by its Hermitian dilation. The weight covariance and the Hessian are already self-adjoint. In all three cases, denote the self-adjoint matrix whose outliers are being studied by $\mathcal{H}_t^c$.

For $c = G$, an outlier is a singular direction of the gradient that the optimizer can amplify.

For $c = W$, it is a teacher direction that has entered the learned representation. For $c = H$, it is a visible curvature direction of the local landscape. The bulk laws differ in the three cases, but once the bulk resolvent is known, the outlier equation has the same finite form.

At a fixed time, the only deterministic directions distinguished by the model are the teacher directions and the student directions. We therefore split the space into a finite signal subspace and its orthogonal complement. In this decomposition,

$$\mathcal{H}_t^c = \begin{pmatrix} A_{\mathrm{sig},t}^c & L_t^c \\ (L_t^c)^* & B_{t,\mathrm{bulk}}^c \end{pmatrix}.$$

The upper-left block is finite-dimensional and acts on the signal coordinates. The lower-right block is the bulk block. The off-diagonal block records how the signal coordinates couple to the bulk. Depending on c, this bulk is a Hermitian dilation bulk, a covariance bulk, or a Hessian Dyson bulk. The probabilistic theorem that identifies the bulk is different in each case, but the deterministic outlier calculation is the same.

An eigenvalue z outside the bulk can be tested without diagonalizing the full matrix. Solving the bulk component of the eigenvector equation and substituting it into the finite component gives the matrix

$$\mathcal{K}_t^c(z) = A_{\mathrm{sig},t}^c + L_t^c(zI - B_{t,\mathrm{bulk}}^c)^{-1}(L_t^c)^*. \tag{5.1}$$

We call $\mathcal{K}_t^c(z)$ the finite outlier matrix. It contains the direct signal block and the correction induced by coupling that block to the bulk resolvent. Thus, outside the bulk spectrum, the high-dimensional eigenvalue problem is equivalent to a finite-dimensional one. The outliers are the real solutions of

$$\det\big(zI - \mathcal{K}_t^c(z)\big) = 0. \tag{5.2}$$

Let $x_-^c(t) \le x_+^c(t)$ be the deterministic left and right bulk edges. We denote by $\lambda_i^c(t)$ the outlier branch associated with teacher mode i, when such a branch exists. If the branch lies to the right of the bulk, its signed margin is

$$\Delta_i^c(t) = \lambda_i^c(t) - x_+^c(t).$$

If it lies to the left, we set

$$\Delta_i^c(t) = x_-^c(t) - \lambda_i^c(t).$$

Thus $\Delta_i^c(t) > 0$ means that the mode is visible in spectrum c. If $u_i^c(t)$ is the normalized finite vector solving $\mathcal{K}_t^c(\lambda_i^c)u_i^c = \lambda_i^c u_i^c$, the corresponding signal residue is

$$\Omega_i^c(t) = \frac{1}{(u_i^c)^\top(I - \partial_z\mathcal{K}_t^c(\lambda_i^c))u_i^c}. \tag{5.3}$$

The residue is the squared mass of the full outlier eigenvector in the finite signal coordinates. It is the quantity that identifies which teacher direction a branch carries. This distinction matters near a dense group of contacts: empirical eigenvalue ranks can switch even when the analytic branch is continuous. A root of (5.2), together with its residue, gives a stable label. It says both where the outlier is and which signal direction it represents. We now specialize this construction to the gradient, weight, and Hessian spectra.

5.2 Gradient spectrum: the denoising view

The gradient spectrum is the spectrum to which Muon is directly applied. It therefore answers a local question: which singular directions of the current gradient should be amplified by the update? After the finite-signal/bulk decomposition, the hermitized gradient has the form

$$\widetilde{G}_t = B_{t,\mathrm{bulk}}^G + S_t^G,$$

where $B^G_{t,\mathrm{bulk}}$ is the high-dimensional bulk and S^G_t is the finite-rank signal carried by the teacher coordinates at the scale being resolved. The bulk law is read from the Hermitian-dilation resolvent. If $m_B(z,t)$ is the Stieltjes transform of the squared singular-value law, then the singular-value density is

$$\rho^\sigma_B(\sigma,t) = \frac{2\sigma}{\pi}\Im m_B(\sigma^2 + i0, t). \tag{5.4}$$

When a local bulk slope is needed, it is read from the resolvent derivative rather than from one noisy empirical histogram:

$$\alpha_B(x,t) = 2 + 2x\,\frac{\Im \partial_x m_B(x+i0,t)}{\Im m_B(x+i0,t)}. \tag{5.5}$$

This quantity tests whether a single power law is a reasonable local compression of the spectral window. The AMP comparison in Section 7 uses the analogous signal-to-bulk likelihood ratio; when that ratio is locally log-linear, its slope gives the Muon exponent.

Let $\theta_i(t)$ be the effective singular strength of mode i in the finite signal block. For a rectangular finite-rank deformation, the same reduced outlier equation can be written through the usual D-transform of the bulk singular-value law, as in the finite-rank outlier theory of Benaych-Georges and Nadakuditi [11]:

$$1 = \theta_i(t)^2 D^G_t(\lambda^G_i(t)), \qquad \Delta^G_i(t) = \theta_i(t)^2 D^G_t(x_{+,G}(t)) - 1. \tag{5.6}$$

Here $x_{+,G}(t)$ is the gradient bulk edge. The BBP time is the first time at which $\Delta^G_i(t) > 0$.

5.3 Weight spectrum: representation outliers

The weight covariance $WW^\top/P$ records the representation carried by the student. It is the most direct place to locate learned teacher directions: if a teacher direction has become part of the student span, it appears here as an eigenvalue leaving the weight bulk. After splitting the finite teacher block from the bulk block, the outlier equation is

$$\det\Big(\lambda I - A^W_{\mathrm{sig},t} - L^W_t(\lambda I - B^W_{t,\mathrm{bulk}})^{-1}(L^W_t)^\top\Big) = 0. \tag{5.7}$$

The residue of the same reduced resolvent gives the teacher overlap of the outlier. This residue is the stable label of a visible branch. Unmatched empirical eigenvalue ranks may switch near a dense front, while finite outlier roots with their residues remain well defined.

If $x_{+,W}(t)$ denotes the weight bulk edge, the spectral exit time of mode i is

$$t^W_i = \inf\{t : \lambda^W_i(t) > x_{+,W}(t)\}. \tag{5.8}$$

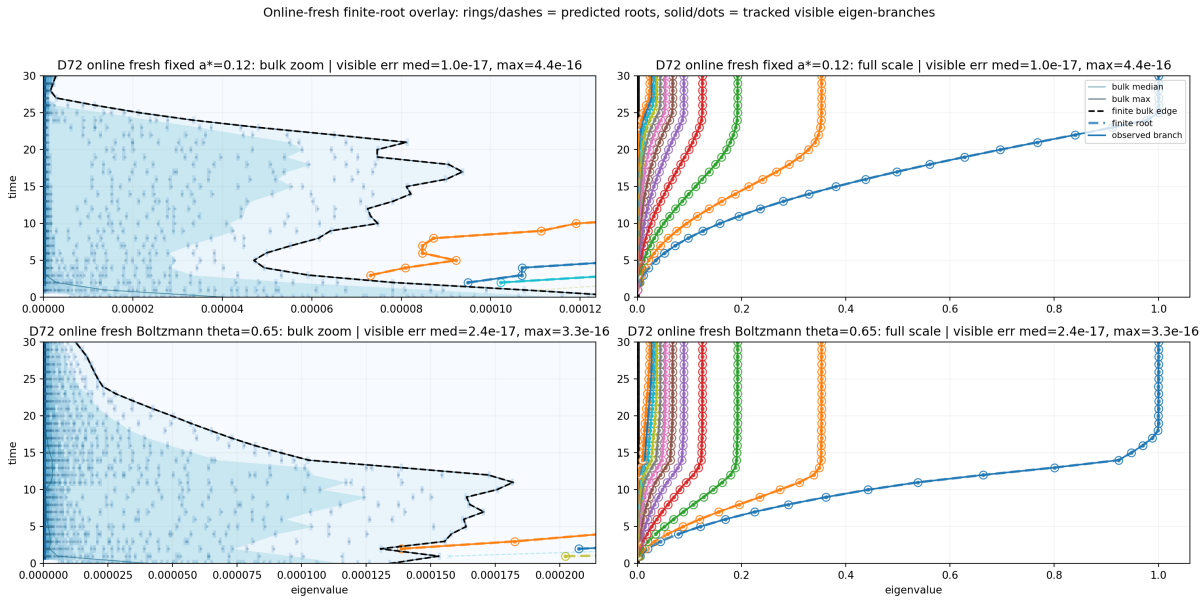


Figure 5: Weight spectrum over time for $d = 72$. The black curve is the finite bulk edge, colored curves are reduced outlier roots, and stars mark matched empirical branches. The finite-dimensional BBP comparison is made by matching empirical branches to roots and residues, not by following raw eigenvalue ranks through branch exchanges.

5.4 Hessian spectrum: local geometry and residual branches

Here the Hessian serves as a probe of the local landscape around the current iterate. Residual teacher directions create left or unstable branches; directions already aligned with the student span create right or stable branches. This is why the Hessian gives a more delicate visibility test than the weight spectrum.

The empirical Hessian block has weighted sample-covariance form:

$$H_{n,bc}(W) = \frac{1}{n}\sum_{\ell=1}^{n} \Phi_{bc}(W^\top x_\ell, \Theta^\top x_\ell) x_\ell x_\ell^\top, \tag{5.9}$$

where

$$\Phi_{bc}(h,y) = \frac{4}{P^2} h_b h_c + \frac{2}{P}\left(\frac{1}{P}\|h\|^2 - y^\top \Lambda y\right)\delta_{bc}. \tag{5.10}$$

Let $A_t^H(h,y)$ be the $P \times P$ matrix with entries $\Phi_{bc}(h,y)$, evaluated under the Gaussian law of $(h,y) = (W_t^\top x, \Theta^\top x)$. If $\alpha_H = n_H/d$ is the Hessian sample ratio, then the effective spectral theory of Ben Arous, Gheissari, Huang and Jagannath [1] gives the following matrix Dyson equation for the fresh-sample bulk:

$$-S_t^H(z)^{-1} = zI - \mathbb{E}\Big[A_t^H (I + \alpha_H^{-1} S_t^H(z) A_t^H)^{-1}\Big], \tag{5.11}$$

where the expectation is over the finite Gaussian coordinates (h,y). The bulk edges of the Hessian are the regular real edges of this Dyson equation. The Hessian has two signal blocks with different meanings. Residual teacher coordinates, not yet represented by the student, give the left or unstable branches. Parallel teacher coordinates, already partially represented, give the right or stable branches. Projecting the same Dyson resolvent onto these two coordinate systems gives finite matrices $\mathcal{K}_t^{H,-}$ and $\mathcal{K}_t^{H,+}$. Left and right branches solve the same reduced outlier equation as in (5.2):

$$\det(\lambda I - \mathcal{K}_t^{H,-}(\lambda)) = 0, \qquad \det(\lambda I - \mathcal{K}_t^{H,+}(\lambda)) = 0. \tag{5.12}$$

In the large-α_H scalar window, the outlier criterion reduces to a simple contact scale. A residual or parallel Hessian branch touches the bulk when

$$\mu_i(1 - r_i(t_{i,\pm})) = \frac{c_i^{\pm}}{\sqrt{\alpha_H}} + O(\alpha_H^{-1}). \tag{5.13}$$

The constants $c_i^{\pm}$ are state-dependent. They are obtained by evaluating the edge equation (5.11) and the finite outlier matrices (5.12) at the current Volterra state.

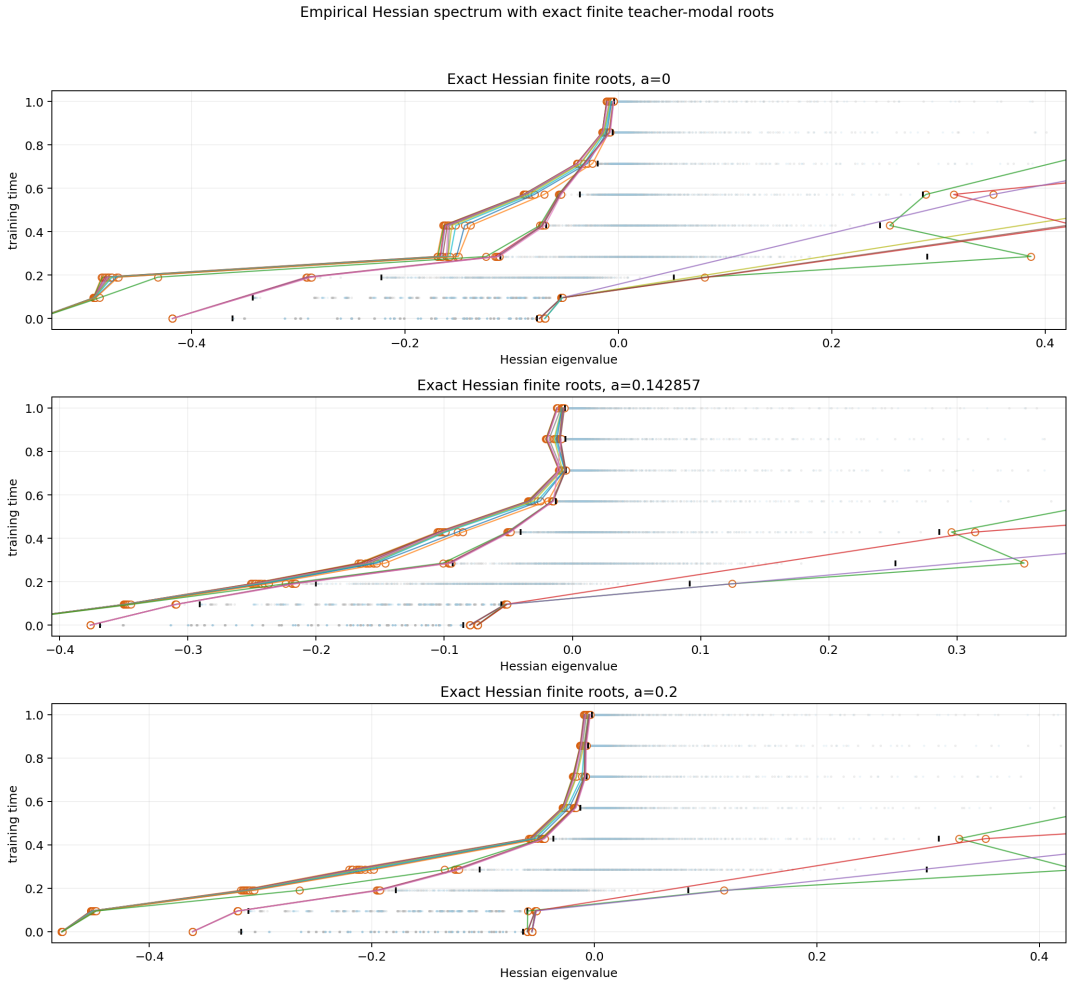


Figure 6: Hessian outlier roots over time. This is the Hessian analogue of the weight overlay: the bulk gives the visibility scale, while the reduced roots give the residual and parallel branches. The plot checks the finite Hessian BBP readout before replacing it by its asymptotic Dyson limit.

5.5 One Volterra state, three spectra

The preceding formulas fix the role of the Volterra state. The mode scales $\chi_i(t)$, residual signal strengths $\rho_i(t)$, and scalar scale $\mathfrak{q}(t)$ determine the finite signal strengths and the bulk transforms entering the three spectra. The gradient spectrum gives the denoising direction, the weight spectrum shows which teacher directions have entered the representation, and the Hessian shows the residual and parallel local-geometry branches. These are three spectral views of the same deterministic state.

To turn the preceding formulas into predictions, one uses the standard stability of isolated outlier roots. This is the same analytic principle behind the usual BBP calculation. An isolated

zero of the reduced outlier equation cannot jump under a small perturbation, and its eigenvector mass is read from the derivative of the same equation. The following proposition records the precise stability statement used to pass from Volterra quantities to root locations, residues, and exit times.

Proposition 5.1 (Stability of the spectral predictions)**.** *Fix a time interval on which the Volterra state is regular and the relevant BBP contacts are simple, meaning that the branch has nonzero velocity relative to the bulk edge at contact. For each matrix $c \in \{G, W, H\}$, suppose the finite outlier matrices obtained from* (5.1) *converge, together with their first z-derivatives, uniformly on compact contours away from the bulk edges, to limits $z \mapsto \mathcal{K}_t^c(z)$ determined by the Volterra state. Then the corresponding outlier locations, eigenvector residues and BBP exit times are deterministic functions of that Volterra state.*

Proof. For each spectrum c, define

$$D_t^c(z) = \det(zI - \mathcal{K}_t^c(z)).$$

An outlier is a zero of D_t^c outside the deterministic bulk. By assumption, $\mathcal{K}_t^c$ and $\partial_z \mathcal{K}_t^c$ converge uniformly on a compact contour surrounding one simple zero and no other zeros. Hence D_t^c and its derivative also converge uniformly on that contour. Rouché's theorem gives exactly one nearby zero for the perturbed problem, and the analytic implicit-function theorem gives continuous dependence of that zero on the Volterra state.

The residue formula (5.3) is a rational expression of the same finite eigenvector and the derivative $\partial_z \mathcal{K}_t^c$. Therefore the eigenvector residue converges with the outlier location. If the branch has a positive margin from the bulk edge except at a simple crossing time, then the sign change of $\Delta_i^c(t)$ is stable. The BBP exit time is consequently a deterministic consequence of the Volterra state. □

Thus the comparison with finite data is made through labeled roots and residues rather than through unlabeled empirical eigenvalue ranks. Near a dense front those ranks can switch. The stable object is the pair "root plus residue": the root gives the branch, and the residue says which teacher direction that branch carries. This is the reason the finite-dimensional comparisons use reduced roots, residues, and contact times, rather than raw eigenvalue labels.

6 Choosing the exponent over time

The fixed-a theory answers: if the exponent is held fixed, what happens? The online question is different: which exponent should be used now, knowing that this choice changes future BBP margins?

The reduced control problem has two terms. The first term is local. It asks what a spectral denoiser would do if it only saw the current gradient spectrum. At time t, let $\nu_{B,t}$ be the bulk singular-value law of the gradient. The active teacher front gives a finite signal measure $\mu_{S,t}$: an atom is placed at the predicted singular location of each currently relevant mode, with mass equal to its outlier residue. Since this signal measure is finite whereas the bulk law is continuous, both are smoothed at the same spectral resolution before they are compared. Denote the smoothed signal density relative to the smoothed bulk law by

$$R_t(\sigma) = \frac{\mathrm{d}\mu_{S,t}^{\mathrm{sm}}}{\mathrm{d}\nu_{B,t}^{\mathrm{sm}}}(\sigma)$$

on the active window. Thus R_t is the local signal-to-bulk likelihood ratio. If a filter f is applied to the gradient singular values, its linearized signal-to-noise ratio is measured by the Rayleigh quotient

$$\mathcal{H}_t(f) = \frac{\langle f, R_t \rangle_{\nu_{B,t}^{\mathrm{sm}}}^2}{\langle f^2 \rangle_{\nu_{B,t}^{\mathrm{sm}}}}. \tag{6.1}$$

By Cauchy–Schwarz, the unconstrained local optimum is

$$f_{\mathrm{opt,t}} \propto R_t.$$

This is the linearized AMP/VAMP spectral denoiser on the current spectral window: amplify directions in proportion to how signal-rich they are relative to the bulk. Muon restricts this filter to the one-parameter family $f_a = \phi_{a,\varepsilon}$, so the instantaneous Muon cost is

$$\mathcal{J}_t(a) = -\log \mathcal{H}_t(\phi_{a,\varepsilon}). \tag{6.2}$$

Thus $\mathcal{J}_t$ is a projection loss: it measures, on the active spectral window, how much signal-to-noise is lost when the best linearized denoiser is restricted to a single Muon exponent.

The second term is dynamic. The exponent chosen now changes the Volterra state, hence the future spectral margins. This is where the control problem enters. Let X_t first collect the Volterra coordinates $(\chi_i(t), \rho_i(t), \mathfrak{q}(t))$, together with the finite bulk transforms needed to evaluate the reduced outlier equations. Away from non-simple contacts, the roots, residues, and BBP margins are smooth functions of these coordinates by Proposition 5.1. We may therefore augment X_t by those spectral quantities without adding new dynamics. The controlled state equation is

$$\dot{X}_t = b(t, X_t, a_t). \tag{6.3}$$

Here b is the vector field induced by the Volterra mode equations, with the spectral quantities updated as differentiable functions of the same coordinates. For the scalar mode scales its i-th component is

$$b_i(t, X, a) = -2\eta\, \delta_i^{(a)} \mathfrak{q}^{(a-1)/2} \chi_i + \eta^2 \nu_i^{(a)} \mathfrak{q}^a. \tag{6.4}$$

The remaining components are obtained by differentiating the bulk edges, roots, residues, and margins with respect to these coordinates.

Now choose a terminal objective expressed in these same variables. For that reduced problem, let $V(t, X)$ be the value function. Thus $V(t, X)$ is the best remaining cost if, at time t, the mode scales and the spectral margins are equal to X. The control problem is closed on the quantities already computed by the Volterra equations and by the reduced outlier reductions. The associated deterministic control problem has the HJB equation

$$-\partial_t V(t, X) = \inf_{a \in [0,1]} \left\{ \mathcal{J}_t(a) + \nabla_X V(t, X) \cdot b(t, X, a) \right\}. \tag{6.5}$$

Thus, along the current trajectory, the reduced future price of choosing a is

$$A_t(a) = \nabla_X V(t, X_t) \cdot b(t, X_t, a) - c_t. \tag{6.6}$$

This is the marginal value, measured through the chosen terminal objective, of applying the vector field corresponding to exponent a. Within the reduced spectral-control problem, this is the shadow price against which the local denoising cost $\mathcal{J}_t(a)$ is compared. The scalar c_t is a gauge: adding the same number to all prices does not change the selected exponent. Only differences in $A_t(a)$, or the slope of A_t on the active support, have meaning.

The entropy-regularized version of this HJB selector is

$$\pi_t(\mathrm{d}a) = \frac{1}{Z_t} \exp[-\beta_t(\mathcal{J}_t(a) + A_t(a))]\, \mathrm{d}a, \qquad a_t = \int a\, \pi_t(\mathrm{d}a). \tag{6.7}$$

The constant Z_t is chosen as

$$Z_t = \int_0^1 \exp[-\beta_t(\mathcal{J}_t(u) + A_t(u))]\, \mathrm{d}u$$

so that π_t is a probability measure. This is a Gibbs distribution over exponents. The term Boltzmann denotes this entropy-regularized minimization over exponents: the energy is the

instantaneous spectral-denoising cost plus the future HJB price, and the inverse temperature β_t controls how sharply the rule concentrates around the minimizer.

Setting the future price to zero gives the local selector obtained by ignoring later BBP-margin costs. It is the one-step balance of the spectral denoising problem. Let $\mathcal{I}_t$ denote the currently active teacher front, let $g_i(t)$ be the effective gradient singular scale of mode i, and let $\omega_i(t)$ be its residue or importance weight. Define the signal and bulk partition functions

$$Z_S(a,t) = \sum_{i \in \mathcal{I}_t} \omega_i(t) g_i(t)^a, \qquad Z_B(a,t) = \int s^{2a} \nu_{B,t}(\,\mathrm{d}s).$$

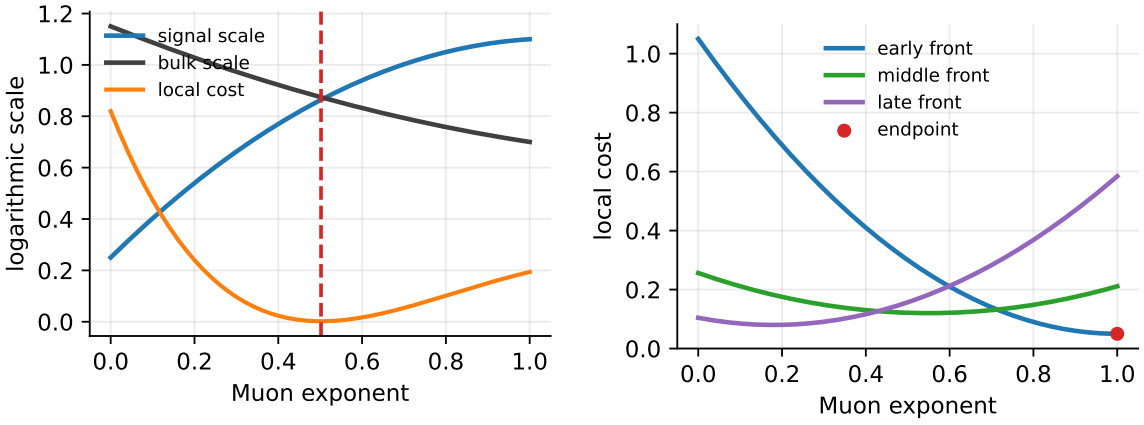
Then

$$L_S(a,t) = \partial_a \log Z_S(a,t), \qquad L_B(a,t) = \frac{1}{2}\partial_a \log Z_B(a,t),$$

and the Boltzmann matching equation is

$$L_S(a,t) = L_B(a,t). \tag{6.8}$$

It says that the typical logarithmic signal scale selected by the filter matches the typical logarithmic bulk scale selected by the same filter.



(a) Matched logarithmic scales. The rule balances the logarithmic signal scale selected by the filter with the logarithmic bulk scale selected by the same filter.

(b) Endpoint effect. Without future BBP-margin constraints, the unregularized local cost can prefer an endpoint such as $a = 1$.

Figure 7: The Boltzmann rule. The local matching equation is directly interpretable, while the dynamic price prevents a purely local choice from delaying later spectral escapes.

6.1 How future visibility enters the value function

If the terminal loss penalizes uncaptured spectral visibility, for example

$$\mathcal{L}_T(X_T) = \sum_{c,i} w_i^c \,\mathrm{softplus}_{\epsilon_{\mathrm{loss}}}(-\Delta_i^c(X_T)) + \omega \mathfrak{q}(T),$$

then $A_t(a)$ is a weighted price of future escapes. Here $\mathrm{softplus}_{\epsilon_{\mathrm{loss}}}$ is any smooth approximation of the positive part, so the loss is large when a margin is still negative. The index c denotes the spectrum: gradient, weight, residual Hessian or parallel Hessian. If

$$t_j^c = \inf\{t : \Delta_j^c(X_t) = 0\}$$

is a BBP contact time, then a first-order perturbation at time $s < t_j^c$ gives, at a simple contact,

$$\frac{\delta t_j^c}{\delta a_s} = -\frac{\nabla_X \Delta_j^c(X_{t_j^c}) \Phi(t_j^c, s) \partial_a b(s, X_s, a_s)}{\frac{d}{dt}\Delta_j^c(X_t)|_{t=t_j^c}}. \tag{6.9}$$

The matrix $\Phi(t,s)$ is the linearized flow of the state equation (6.3). This formula explains how BBP margins enter the HJB price: a change of exponent at time s is valuable if it moves later contact times in the favorable direction, and costly if it delays or destroys future separated branches.

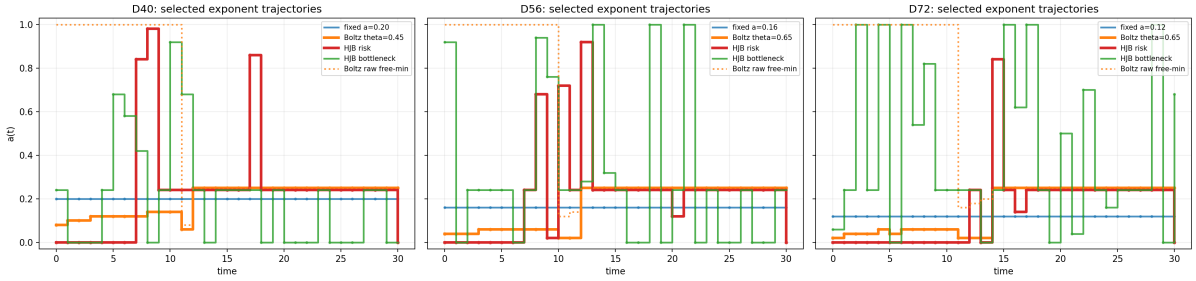


Figure 8: Exponent trajectories. Boltzmann is smoother than a hard minimizer because it averages the local spectral cost and the future price over a Gibbs distribution. The width and smoothness of the curve indicate whether the selected support is narrow enough for the affine reduction of the HJB price to be accurate on the active support.

6.2 Variable exponent as a non-autonomous Volterra equation

For nonconstant a_t, the fundamental equation is not an autonomous ODE for the exponent. It is the non-autonomous Volterra state:

$$\dot{\chi}_i(t) = -2\eta\,\delta_i^{(a_t)}\mathfrak{q}(t)^{(a_t-1)/2}\chi_i(t) + \eta^2\nu_i^{(a_t)}\mathfrak{q}(t)^{a_t}. \tag{6.10}$$

In a slowly moving power-law front, (6.10) reduces to the adiabatic formula

$$a_{\mathrm{PR},*}(t) = \mathrm{clip}_{[0,1]}\left(\frac{\zeta_t - 1}{3\zeta_t - 1}\right), \qquad \dot{a}_{\mathrm{PR},*}(t) = \frac{2\dot{\zeta}_t}{(3\zeta_t-1)^2}. \tag{6.11}$$

The derivative displayed in (6.11) is the interior derivative, valid while the unclipped value lies in $(0,1)$. Thus the ODE for $a_*(t)$ is a consequence of the moving Volterra front, not an independent modeling assumption.

6.3 When the future price is effectively affine

On intervals where the selected exponents occupy a narrow range, the full HJB price may be replaced by an affine shadow price,

$$A_t(a) \simeq c_t + \lambda_t a.$$

This affine replacement is valid on the Gibbs support when the nonlinear residual is invisible:

$$\beta_t\sqrt{\mathrm{Var}_{\pi_t}(A_t - c_t - \lambda_t a)} = o(1). \tag{6.12}$$

A sufficient local condition is

$$\beta_t\,\mathrm{Var}_{\pi_t}(a)\sup_a|\partial_{aa}A_t(a)| = o(1).$$

For the Muon filter, the second derivative brings a squared logarithm,

$$\partial_{aa}\phi_{a,\varepsilon}(\sigma) = \frac{1}{4}\log^2(\sigma^2+\varepsilon^2)\phi_{a,\varepsilon}(\sigma).$$

Thus the mathematical question is whether the curvature of the future price is small on the part of the front selected by the Gibbs rule. When (6.12) holds, the affine shadow price is an accurate compression of the reduced control problem on the selected support. When it does not hold, the control problem remains the same but the full future price $A_t(a)$ must be kept. In finite dimension, the affine explanation is tested by the width of the Gibbs support, the smoothness of the selected trajectory in Figure 8, and whether the BBP exit times remain predicted by the same Volterra state. These quantities distinguish the online rule itself from the stronger affine-price approximation. The Boltzmann rule is defined by (6.7); the interpretation as a low-dimensional shadow of the HJB price is the regime described by (6.12). Figures 7, 8, and 9 show these quantities in the runs below.

7 AMP projection and non-power-law spectra

The AMP comparison is made after linearization, at the early state where no teacher direction is yet macroscopically separated from the bulk. At that point AMP/VAMP selects a spectral denoising direction. The question here is whether the selected direction can be represented by a Muon power.

For i.i.d. Gaussian channels, AMP gives an optimal denoising iteration with an Onsager correction that cancels the leading self-interaction [5]. For orthogonally invariant spectra, the correct relatives are VAMP and OAMP [9, 10]. When these methods are linearized around the uninformative fixed point, the decision of which direction grows is a spectral denoising problem. That is the level at which Muon is compared to AMP.

The comparison becomes explicit when the likelihood ratio R_t is locally a power law on the active front:

$$\log R_t(\sigma_t e^u) = \kappa_t + a_{\mathrm{loc}}(t)u + o(1), \qquad |u| \le L. \tag{7.1}$$

Then $R_t(\sigma) \simeq C_t \sigma^{a_{\mathrm{loc}}(t)}$, so the unrestricted linearized AMP/VAMP filter is locally a Muon power. This comparison is made at the level of the spectral denoiser selected by the linearized iteration. The Onsager correction governs the full AMP trajectory; the local projection identifies the direction that the linearized denoiser wants to amplify. If the active spectrum has several slopes, oscillations, or separated front blocks, a scalar $a(t)$ is the one-dimensional compression. The natural extension is either a blockwise exponent $a_b(t)$ or the full spectral filter R_t. The local-log-linearity condition is assessed from the graph of $\log R_t(\sigma)$ against $\log \sigma$ on the active window. A straight segment gives a Muon exponent; curvature or multiple slopes indicate that a single exponent is a coarse compression of the AMP/VAMP denoiser. The active-front and Boltzmann matching figures display the corresponding spectral window. On a curved front the same statement is read blockwise, or with the full filter R_t, rather than as a single scalar exponent.

For the Rayleigh quotient (6.1), the linearized AMP/VAMP comparison supplies the unrestricted spectral denoiser; Muon keeps its best one-parameter power approximation. The next proposition records when this projection is again a Muon power.

Proposition 7.1 (Local AMP/VAMP projection onto Muon powers)**.** *Assume the active spectral front is locally log-linear in the sense of* (7.1)*, and that the Muon family is restricted to that front with the Rayleigh quotient* (6.1)*. Then the unrestricted linearized AMP/VAMP filter is locally proportional to a Muon power, and the best scalar Muon exponent satisfies*

$$a_{\mathrm{AMP}}(t) = a_{\mathrm{loc}}(t) + o(1),$$

up to clipping at the endpoints of the allowed interval.

Proof. The Rayleigh quotient (6.1) is maximized over all square integrable filters by $f_{\mathrm{opt}} \propto R_t$, by Cauchy–Schwarz. The log-linearity assumption says that on the active front

$$R_t(\sigma) = C_t \sigma^{a_{\mathrm{loc}}(t)}(1 + o(1)).$$

Restricting the quotient to $f_a(\sigma) = \sigma^a$, the numerator and denominator are evaluated on a one-parameter family that contains a function proportional to R_t at $a = a_{\mathrm{loc}}(t)$, up to the $o(1)$-error in the local log-linear approximation. Since the unrestricted maximizer is unique up to multiplication on the active window, the restricted maximizer satisfies $a = a_{\mathrm{loc}}(t) + o(1)$, unless this value lies outside the allowed interval. □

Thus a regularly varying front makes a power filter the local projection of the spectral denoiser. Full AMP state evolution and the Onsager correction remain part of the AMP/VAMP theory; the statement above extracts the one-parameter spectral projection that Muon can implement.

8 Summary of the reduction

The preceding sections reduce the training problem to a small number of deterministic objects. The reduction is

$$\text{population quadratic loss} \to \text{finite-frame Muon ODE} \to \text{Volterra risk equation}$$
$$\to \text{reduced outlier equations} \to \text{online exponent selection.}$$

The first arrow is algebraic and exact for the population loss: both the gradient update and the Muon update stay in the span of W and Θ. The second arrow inserts the Muon filter into the spectral-update recursion and gives the Volterra equation (4.3). The third arrow converts the Volterra state into spectral branches by the reduced outlier equation (5.2) and residue formula (5.3). The last arrow is the reduced control step: if the exponent is allowed to move, the same state evolves, but the vector field is chosen by the local denoising cost and by the value of future BBP margins.

For a fixed exponent, the deterministic reduction therefore produces a learning curve and a set of BBP exit times. For a moving exponent, it produces the state equation on which the exponent selector acts, through the current active front and the later spectral margins. These formulas are deterministic once the path is fixed. If the same samples are used both to train the path and to form the resolvents, the deterministic formulas are transferred through the leave-one-out comparison stated in Appendix E.

9 Numerical comparisons

The deterministic reduction produces several observable quantities, not only a final loss. The local front gives the slope ζ. The Volterra equation gives the risk clock and the modewise drift. The bulk equations give moving spectral edges. The reduced outlier equations give roots, residues, and BBP contact times. The numerical comparisons are made at exactly these levels.

This matters in a dense front. Empirical eigenvalue ranks can exchange, so a single ordered eigenvalue is not a reliable label. A branch is instead matched by the root of the reduced outlier equation and by its residue. The residue identifies the teacher direction carried by the branch. Thus the same object that predicts the outlier location also labels the empirical eigenvector.

The figures are organized according to this chain. Figure 1 checks that the active front is locally close to a power law. Figure 2 displays the corresponding Volterra/Laplace tail clock. Figures 3 and 4 compare the predicted bulk scale and BBP exit times with the finite spectra. Figure 5 performs the root-residue matching for the weight spectrum, and Figure 6 performs the same comparison for the Hessian branches. The overlay in Figure 9 shows the same matching for the online exponent rules. Finally, Figure 10 isolates the finite quantity controlled by the fresh-to-reused comparison that transfers the fresh-sample calculation to the same-sample setting.

Each figure corresponds to one input of the reduction. A curved active front leads to a blockwise filter rather than a scalar exponent. A displaced bulk edge changes the BBP clock.

Incorrect root-residue labels identify the wrong empirical branches. The numerical section therefore follows the same order as the proof: Figure 1 gives the local power-law front; Figure 2 gives the Volterra clock; Figures 3 and 4 give the BBP timing; Figures 5, 6, and 9 give the root-residue branch labels; and Figure 10 records the finite comparison needed before passing from a fresh spectral sample to the reused-sample setting.

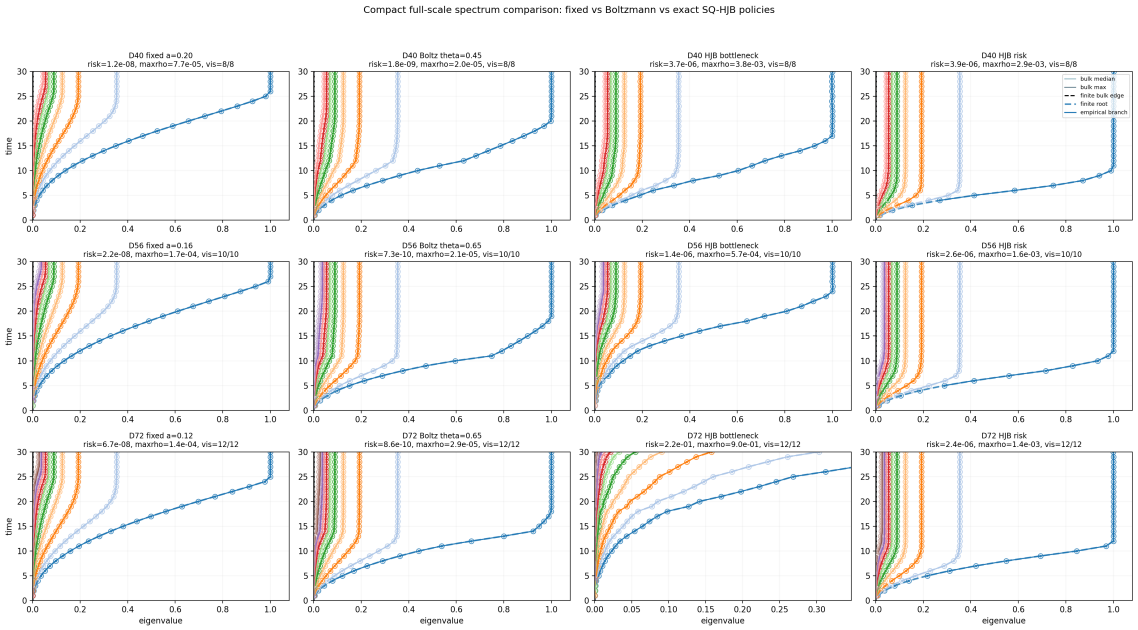


Figure 9: Spectral overlays for online exponent rules. The reduced outlier equation predicts the visible branches, and the same branches are seen in the finite spectra once their margins separate from the bulk. The labels are determined by the root and its residue, so the comparison remains meaningful even when nearby empirical ranks switch.

10 Related work

The closest phase-retrieval reference is the analysis of phase retrieval under power-law anisotropic data by Braun, Loureiro, Minh and Imaizumi [2]. Their work shows that anisotropy turns phase retrieval into a hierarchy of summary statistics, with an initial escape, a slow spectral front, and a tail-learning regime. We keep the same obstruction, but ask how a spectral optimizer changes the hierarchy. Muon enters by modifying the singular values of the gradient before the update is applied.

The direct optimization backbone is the analysis of Muon phases by Paquette and coauthors [3]. Their framework studies stochastic spectral optimizers through a general filter φ, projected-risk recursions, and resolvent formulas for drift and volatility. The present paper uses this mechanism with $\varphi_{a,\varepsilon}$ in (4.2); the resulting coefficients $\delta_i^{(a)}$ and $\nu_i^{(a)}$ are then summed into the Volterra equation (4.3).

The Hessian and information-matrix side is closest to the effective spectral theory of Ben Arous, Gheissari, Huang and Jagannath [1]. Once the state is summarized by finitely many Gaussian projections, the Hessian bulk, outliers, and overlaps are computed by deterministic Dyson equations and reduced outlier equations. In this paper those equations are used to convert the Volterra state into predicted gradient, weight, and Hessian branches. The reused-data version is handled by the comparison step described in Appendix E.

The AMP comparison is first-order and spectral. In the multi-index setting, Defilippis, Dandi, Mergny, Krzakala and Loureiro [4] identify linearized message-passing spectral algorithms reaching optimal reconstruction thresholds. For non-i.i.d. or orthogonally invariant spectra, the

appropriate language is VAMP/OAMP [9, 10], together with the finite-rank outlier calculus of Benaych-Georges and Nadakuditi [11]. The comparison made here is the linearized spectral one: near a regularly varying active front, the AMP/VAMP spectral denoiser is locally a power law, and its projection onto the Muon family gives the exponent in Proposition 7.1.

The organization follows the state-evolution philosophy of Bayati and Montanari [5]: identify the low-dimensional state and then compute the quantities revealed by the algorithm. In the present paper the low-dimensional state is the Volterra risk state; the quantities computed from it are the gradient, weight, and Hessian spectral branches.

This is also close in spirit to the landscape program of Asgari, Montanari and Saeed [6] and Montanari and Saeed [7]. Their Kac–Rice analysis identifies finite objects describing positions, errors, and Hessian spectra of minimizers. Here the finite object is dynamic rather than static: it is the Volterra state along training, and the spectral events are BBP transitions in the gradient, weight, and Hessian.

High-dimensional SGD dynamics for multi-index models are described by dynamical mean-field or effective-dynamics limits, for example in Fan and Wang [8]. We use the same philosophy after the finite-frame reduction: the optimizer is summarized by a small deterministic state, and the spectra are deterministic functions of that state.

11 Conclusion

The central lesson is that Muon is a dynamic spectral denoiser. In a power-law phase retrieval problem, weak teacher modes are not learned all at once. They become visible through a moving sequence of BBP events. A fixed exponent a captures part of this behavior, and Paquette's spectral-update recursion gives a Volterra equation for its learning curve. A time-dependent exponent is better understood as a local spectral balance: increase a when signal separation needs protection from the bulk, and decrease a when the visible group is stable and the tail can be learned aggressively. This gives a coherent bridge between population dynamics, empirical spectra, random-matrix outliers, and AMP-like spectral denoising.

A Gaussian identities for the quadratic loss

This appendix records the elementary Gaussian calculations used in Section 2. Let B be a deterministic symmetric matrix and let $x \sim N(0, I_d)$. Wick's formula gives

$$\mathbb{E}(x^\top B x) = \operatorname{Tr} B, \qquad \mathbb{E}(x^\top B x)^2 = 2\operatorname{Tr}(B^2) + (\operatorname{Tr} B)^2. \tag{A.1}$$

With $B = E_W = \Sigma_W - A_\star$, the square loss is

$$\frac{1}{2}\mathbb{E}(x^\top E_W x)^2 = \operatorname{Tr}(E_W^2) + \frac{1}{2}(\operatorname{Tr} E_W)^2,$$

which is (2.1). Differentiating

$$E_W = \frac{1}{P} W W^\top - A_\star$$

in the direction $\dot{W}$ gives

$$\dot{E}_W = \frac{1}{P}(\dot{W} W^\top + W \dot{W}^\top).$$

Therefore

$$\frac{\mathrm{d}}{\mathrm{d}t} R(W + t\dot{W})\Big|_{t=0} = 2\operatorname{Tr}(E_W \dot{E}_W) + \tau_W \operatorname{Tr}(\dot{E}_W).$$

Using cyclicity of the trace,

$$2\operatorname{Tr}(E_W \dot{E}_W) = \frac{4}{P}\operatorname{Tr}(W^\top E_W \dot{W}), \qquad \tau_W \operatorname{Tr}(\dot{E}_W) = \frac{2\tau_W}{P}\operatorname{Tr}(W^\top \dot{W}).$$

Thus

$$\nabla_W R(W) = \frac{2}{P}(2E_W + \tau_W I_d)W,$$

which is (2.3). The empirical formula (2.4) follows by differentiating each sample term

$$\frac{1}{2}(x_\ell^\top E_W x_\ell)^2.$$

B Finite-frame derivation of the Muon ODE

The finite-frame formulas in Section 3 are useful because they show that the high-dimensional population dynamics is really a finite matrix dynamics. Write

$$Z = [W, \Theta], \qquad \Gamma = Z^\top Z = \begin{pmatrix} Q & M \\ M^\top & I_k \end{pmatrix}.$$

Any matrix $Y = ZA$ with columns in the span of Z has Gram representation

$$Y^\top Y = A^\top \Gamma A.$$

If $Y = U\,\mathrm{diag}(s_j)V^\top$, then applying the Muon map $\mathsf{M}_a(Y) = U\,\mathrm{diag}(s_j^a)V^\top$ can be done inside the finite frame. Indeed, set

$$\widehat{A} = \Gamma^{1/2} A.$$

The singular values of Y are the singular values of $\widehat{A}$, and

$$\mathsf{M}_a(Y) = Z\Gamma^{-1/2}\mathsf{M}_a(\widehat{A}).$$

This is the origin of

$$A_a = \Gamma^{-1/2}\mathsf{M}_a(\Gamma^{1/2}A_{\mathrm{gd}}).$$

Splitting $A_a = (U_a^\top, V_a^\top)^\top$, the flow is

$$\dot{W} = -\eta_a(WU_a + \Theta V_a).$$

Then

$$\dot{Q} = \dot{W}^\top W + W^\top \dot{W} = -\eta_a(U_a^\top Q + QU_a + V_a^\top M^\top + MV_a),$$

and

$$\dot{M} = \dot{W}^\top \Theta = -\eta_a(U_a^\top M + V_a^\top).$$

Finally,

$$C = M^\top Q^{-1} M$$

gives

$$\dot{C} = \dot{M}^\top Q^{-1} M + M^\top Q^{-1}\dot{M} - M^\top Q^{-1}\dot{Q}Q^{-1}M.$$

Substituting the two preceding equations and collecting the terms containing

$$D_a = V_a Q^{-1} M$$

gives (3.2). This calculation is exact whenever Q is invertible. If Q is singular, the same formula is read on the support of the student span or after a harmless ridge regularization, followed by a limit.

C From Paquette recursion to Volterra

The fixed-a Volterra equation follows by integrating the mode equations once the drift and variance coefficients have been computed by Paquette's resolvent formulas. Start from

$$\dot{\chi}_i(t) = -2\eta\,\delta_i^{(a)}\mathfrak{q}(t)^{(a-1)/2}\chi_i(t) + \eta^2\nu_i^{(a)}\mathfrak{q}(t)^a.$$

Introduce the clock

$$\mathrm{d}\tau = \eta\mathfrak{q}(t)^{(a-1)/2}\,\mathrm{d}t.$$

Then, as a function of τ,

$$\partial_\tau\chi_i(\tau) = -2\delta_i^{(a)}\chi_i(\tau) + \eta\nu_i^{(a)}\mathfrak{q}(\tau)^{(a+1)/2}. \tag{C.1}$$

Solving this scalar linear equation gives

$$\chi_i(\tau) = e^{-2\delta_i^{(a)}\tau}\chi_i(0) + \eta\nu_i^{(a)}\int_0^\tau e^{-2\delta_i^{(a)}(\tau-u)}\mathfrak{q}(u)^{(a+1)/2}\,\mathrm{d}u.$$

If the risk scale is a weighted sum

$$\mathfrak{q}(\tau) = \sum_i w_i\chi_i(\tau)$$

for deterministic weights w_i, then summing the preceding identity gives

$$\mathfrak{q}(\tau) = F_a(\tau) + \eta\int_0^\tau K_a(\tau-u)\mathfrak{q}(u)^{(a+1)/2}\,\mathrm{d}u,$$

with

$$F_a(\tau) = \sum_i w_i e^{-2\delta_i^{(a)}\tau}\chi_i(0), \qquad K_a(s) = \sum_i w_i\nu_i^{(a)}e^{-2\delta_i^{(a)}s}.$$

This is (4.3). The random-matrix input not rederived in this appendix is the computation of $\delta_i^{(a)}$ and $\nu_i^{(a)}$, which Paquette expresses through one- and two-resolvent formulas for a general spectral filter.

D Reduced outlier roots and eigenvector residues

The outlier equations used for the three spectra all have the same finite-rank form. The object that matters is the finite matrix obtained after the bulk resolvent has been inserted. After separating a finite signal block from a bulk block, write the hermitized or symmetric matrix schematically as

$$\mathcal{H} = \begin{pmatrix} A & L \\ L^\top & B_{\mathrm{bulk}} \end{pmatrix}.$$

For $\lambda \notin \mathrm{spec}(B_{\mathrm{bulk}})$, an eigenvector (u, v) satisfies

$$(A-\lambda I)u + Lv = 0, \qquad L^\top u + (B_{\mathrm{bulk}} - \lambda I)v = 0.$$

The second equation gives

$$v = (\lambda I - B_{\mathrm{bulk}})^{-1}L^\top u.$$

Substituting in the first equation gives the reduced outlier equation

$$\left[\lambda I - A - L(\lambda I - B_{\mathrm{bulk}})^{-1}L^\top\right]u = 0. \tag{D.1}$$

Thus an outlier is a zero of the determinant of the bracketed matrix.

The residue is obtained from the same equation. Let

$$S(\lambda) = A + L(\lambda I - B_{\mathrm{bulk}})^{-1} L^{\top}$$

and suppose $\lambda_\star$ is a simple solution with normalized finite vector $u_\star$. Near $\lambda_\star$, the projected resolvent has a simple pole, and the squared mass in the finite signal coordinates is

$$\Omega_\star = \frac{1}{u_\star^{\top}(I - \partial_\lambda S(\lambda_\star))u_\star}. \tag{D.2}$$

This is the finite-dimensional version of the usual BBP residue formula. It explains why the figures track roots and residues, rather than unmatched empirical ranks: the residue is what tells us whether a detached branch is actually informative.

E Fresh and reused spectral samples

The main text uses the clean setting where the trajectory is fixed before the spectral sample is drawn. This is the natural setting for the deterministic formulas: once the finite variables are fixed, the random bulk is independent of the finite frame, and the reduced outlier roots are obtained from finite matrices with deterministic limits.

If the same samples are used both to train the trajectory and to form the resolvent, one additional probabilistic comparison is needed. The issue is the dependence between W_t and the Gaussian noise appearing in the Hessian or gradient resolvent. The comparison below is the statement needed to transfer the fresh-sample outlier calculation to the reused-sample one. It is separated from the main text because it does not change the deterministic Volterra or outlier equations.

Let $\mathcal{K}_d^{\mathrm{reuse}}(z,t)$ be the finite outlier matrix computed on the reused training sample, and let $\mathcal{K}_d^{\mathrm{fr}}(z,t)$ be the corresponding fresh or leave-one-out matrix. The estimate needed is that these two finite matrices, and their z-derivatives, are asymptotically the same on the contours where outliers are read.

Assumption E.1 (Fresh-reused outlier comparison)**.** *On a compact time interval, and on spectral contours separated from exact BBP tangencies by a margin $m_\star > 0$,*

$$\sup_{t,z} \left(\left\| \mathcal{K}_d^{\mathrm{reuse}}(z,t) - \mathcal{K}_d^{\mathrm{fr}}(z,t) \right\| + \left\| \partial_z \mathcal{K}_d^{\mathrm{reuse}}(z,t) - \partial_z \mathcal{K}_d^{\mathrm{fr}}(z,t) \right\| \right) = o_{\mathbb{P}}(1).$$

This estimate is the probabilistic bridge from the fresh calculation to the reused training sample. Before proving it, one can already compare the finite objects that it controls: the roots, their residues, and the branch labels. These are the observable counterparts of (E.1). The root tests the outlier location, the residue tests the teacher mass carried by the eigenvector, and the branch label tests that the same analytic mode is followed through nearby eigenvalue exchanges. Figure 10 records this comparison in the fresh/frozen experiment. A full reused-sample proof upgrades the same finite comparison to the uniform estimate (E.1).

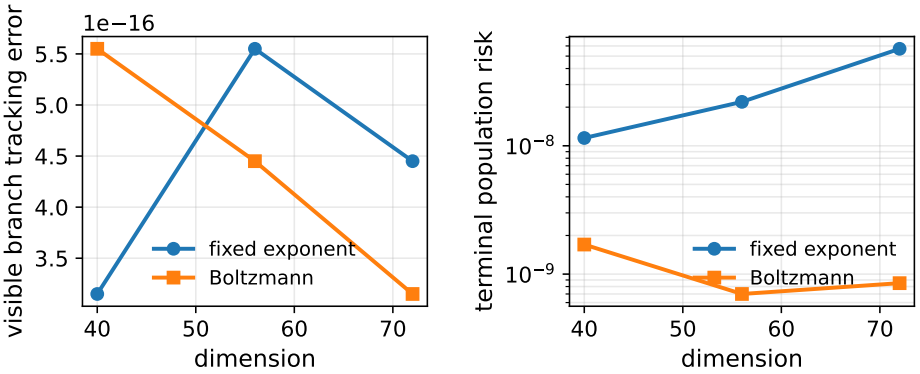


Figure 10: Fresh and frozen outlier-root comparison. The relevant empirical quantity is the stability of the outlier labels and residues, not the final loss alone. The reused-sample comparison estimate (E.1) is reduced to the leave-one-out local law described after the figure; the plot records the finite-dimensional quantity controlled by that law.

Proposition E.2 (Consequence for reused samples)**.** *Assume the fresh/frozen deterministic finite-outlier limit and Assumption E.1. Then the reused-data roots, residues, and BBP exit times have the same deterministic limits as their fresh/frozen counterparts, away from non-transversal contacts.*

Proof. On a contour enclosing one simple deterministic root, the reduced characteristic function is analytic in z. Uniform convergence of $\mathcal{K}_d^{\mathrm{reuse}}$ and $\partial_z \mathcal{K}_d^{\mathrm{reuse}}$ to their fresh counterparts implies uniform convergence of the determinant and of its derivative. Rouché's theorem gives one empirical root in the contour, and the implicit function theorem gives convergence of that root. The residue is a rational function of the same finite outlier matrix and derivative, hence it converges as well. A positive margin from the bulk edge makes the thresholded exit time stable. □

The usual proof strategy is a leave-one-out local law. One freezes the finite frame $Z_t = [W_t, \Theta]$, decomposes each sample into its projection on this frame plus an orthogonal Gaussian component, and proves a local law for the orthogonal block uniformly on the spectral contours used above. The reused trajectory is then compared with a leave-one-out trajectory $W_t^{(\ell)}$, trained without the ℓ-th sample. If replacing W_t by $W_t^{(\ell)}$ changes the averaged finite matrix by $o_{\mathbb{P}}(1)$, the fresh and reused outlier equations have the same roots and residues. Near a dense cluster of BBP contacts, the stable quantities are contour integrals, roots with positive margins, and smoothed visible masses. This is the role of the reused-sample comparison in the paper.

References

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

[2] G. Braun, B. Loureiro, H. Q. Minh and M. Imaizumi, *Fast Escape, Slow Convergence: Learning Dynamics of Phase Retrieval under Power-Law Data*, arXiv:2511.18661.

[3] E. Paquette, N. Marshall, L. Benigni, G. Wang, A. Agarwala and C. Paquette, *Phases of Muon: When Muon Eclipses SignSGD*, arXiv:2605.09552.

[4] L. Defilippis, Y. Dandi, P. Mergny, F. Krzakala and B. Loureiro, *Optimal Spectral Transitions in High-Dimensional Multi-Index Models*, arXiv:2502.02545.

[5] M. Bayati and A. Montanari, *The dynamics of message passing on dense graphs, with applications to compressed sensing*, IEEE Transactions on Information Theory, 57(2), 2011.

[6] K. Asgari, A. Montanari and B. Saeed, *Local minima of the empirical risk in high dimension: General theorems and convex examples*, arXiv:2502.01953.

[7] A. Montanari and B. Saeed, *Topological trivialization in non-convex empirical risk minimization*, arXiv:2602.14969.

[8] Z. Fan and L. Wang, *High-dimensional learning dynamics of multi-pass Stochastic Gradient Descent in multi-index models*, arXiv:2601.21093.

[9] S. Rangan, P. Schniter and A. Fletcher, *Vector approximate message passing*, IEEE Transactions on Information Theory, 65(10), 2019.

[10] J. Ma and L. Ping, *Orthogonal AMP*, arXiv:1602.06509.

[11] F. Benaych-Georges and R. Nadakuditi, *The singular values and vectors of low rank perturbations of large rectangular random matrices*, arXiv:1103.2221.