Mathematical Details for Dynamic Spectral Denoising in Power-Law Phase Retrieval

Volterra equations, finite-rank readouts, and control

Abstract

This manuscript is the technical companion to the main paper *Muon as Dynamic Spectral Denoising for Power-Law Phase Retrieval.* It collects the mathematical calculations used there into one self-contained document. The exact part of the argument is the reduction of the population dynamics to the finite student–teacher frame. The projected-risk recursion of [1] then yields a Volterra state under an explicit fresh-sample channel hypothesis. Given this state and the required bulk resolvent limits, the gradient singular spectrum, the weight covariance spectrum, and the empirical Hessian spectrum are three finite-rank readouts of the same trajectory.

The paper is organized as a linear proof. We first derive the finite-frame population dynamics for the quadratic phase retrieval loss. We next state the channel hypothesis, insert the Muon power filter into the general spectral recursion, and derive the nonlinear Volterra equation. A power-law Laplace analysis gives an asymptotic fixed-exponent balance. The same variables enter finite Schur equations for the gradient, weight, and Hessian outliers. The last part formulates the variable-exponent problem as a control problem. At finite teacher rank its HJB equation is finite-dimensional; in the continuum limit it is functional, and a spectral-block model is a discretization. The same-sample local law and global AMP equivalence are not asserted here.

Contents

1 Purpose and logical structure

The main paper gives the conceptual narrative. This companion gives the mathematical derivations. It contains the Volterra reduction, the power-law asymptotics, the three spectral BBP readouts, and the variable-exponent control equations. A small number of numerical figures illustrate the observables appearing in the assumptions; they are not used as proof.

The proof has five layers.

$$\text{quadratic phase retrieval} \longrightarrow \text{finite-frame Muon dynamics} \longrightarrow \text{reduced Muon Volterra state}$$
$$\longrightarrow \text{gradient/weight/Hessian BBP readouts} \longrightarrow \text{online exponent selection.}$$

The first arrow is exact algebra. The second is conditional on the fresh-sample channel hypothesis in Assumption 5.1; it specializes the general spectral-filter recursion of [1] to a Muon power. The third requires convergence of the bulk resolvents and finite Schur matrices: outside the bulk, outliers are roots of a finite equation and their residues are derivatives of the same equation. The final arrow defines a reduced control problem on this state.

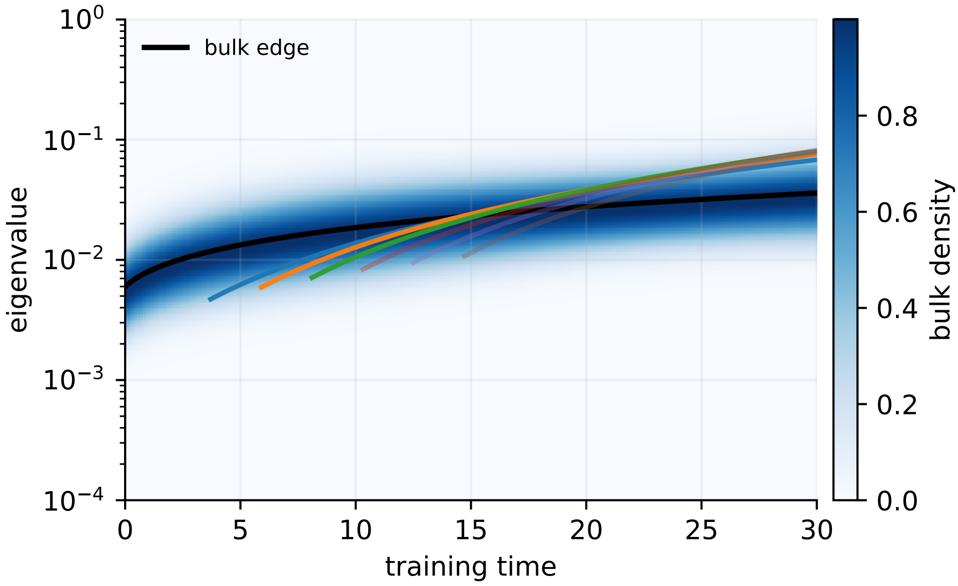


Figure 1: The moving spectral picture. The Volterra state predicts signal strengths; the bulk edge decides when those signals become visible. The remaining sections explain which equations produce the state and which finite-rank equations produce the outlier branches.

2 Notation and proof dependencies

The paper uses three kinds of objects. The first are physical phase-retrieval quantities, such as the teacher strengths and captured masses. The second are row-resolvent quantities, which describe the fresh spectral optimizer channel. The third are finite-rank BBP readouts, which decide whether a signal already present in the Volterra state is visible outside a bulk spectrum.

To avoid ambiguity, throughout this document

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

is a scalar residual physical mass, while

$$\varrho_{i,t}(\,\mathrm{d}x)$$

is the row spectral measure. The two are related only after the phase-retrieval front substitution of Section 5.2.

Here "RFA" means resolvent-frame averaging: the comparison is made only after projecting resolvents onto the finite frame generated by the student and teacher directions. It is therefore weaker than operator-norm convergence of the full random matrix, but stronger than a scalar risk comparison. This is the topology in which roots, residues, and branch masses are stable.

There are three levels of rigor. The finite-frame population identities are proved directly. The Volterra and fixed-path spectral statements are deterministic consequences of specified fresh-sample resolvent limits. The transfer from a fresh or frozen path to the same-sample online trajectory is a second probabilistic problem, isolated in Assumption 15.1 and in the projected-resolvent topology of Section 15.2. In particular, proving the same-sample comparison would not by itself prove the fresh nonlinear channel identification in Assumption 5.1; both inputs are stated separately.

The deterministic contract

The deterministic implications use four inputs. First, conditional row spectral measures $\varrho_{i,t}$ converge on the chosen fresh path. Second, the one- and two-resolvent observables defining $\delta_i^{(a)}$ and $\nu_i^{(a)}$ converge uniformly for the regularized filter. Third, the bulk resolvents and finite Schur matrices of the gradient, weight, and Hessian channels converge on the outlier contours. Fourth, the contacts are transversal, or are replaced by the smoothed spectral projectors defined in Section 15.1. These assumptions are denoted P1–P4 below.

Given P1–P4, the paper proves the following deterministic chain:

$$\begin{aligned}\varrho_{i,t} &\longrightarrow (\delta_i^{(a)}, \nu_i^{(a)}) \longrightarrow \mathfrak{q}(t)\\ &\longrightarrow (\mathcal{K}_t^G, \mathcal{K}_t^W, \mathcal{K}_t^H) \longrightarrow \text{BBP locations, residues, and crossing times.}\end{aligned}$$

The additional same-sample question is whether this chain remains valid when the resolvent is built from the data that generated the trajectory. That question is isolated in Section 15.2; it is not hidden inside the Volterra, power-law, or BBP calculations.

3 Quadratic phase retrieval and finite summaries

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

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

In the power-law regime,

$$\mu_i = \mu_0 i^{-\gamma}, \qquad \gamma > 1/2.$$

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

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

and

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

Set

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

The population loss is

$$R(W) = \frac{1}{2}\mathbb{E}(f_W(x) - f_\star(x))^2 = \operatorname{Tr}(E_W^2) + \frac{1}{2}\tau_W^2. \tag{3.1}$$

The empirical loss on samples $x_1, \dots, x_n$ is

$$R_n(W) = \frac{1}{2n}\sum_{\ell=1}^{n}(x_\ell^\top E_W x_\ell)^2. \tag{3.2}$$

The finite summaries used throughout are

$$Q = W^\top W, \qquad M = W^\top\Theta, \qquad C = M^\top Q^{-1}M.$$

The matrix C records the part of the teacher subspace 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 residual signal strength of the same mode.

Lemma 3.1 (Gaussian loss and gradient)**.** *The population gradient is*

$$G(W) = \nabla_W R(W) = \frac{2}{P}(2E_W + \tau_W I_d)W. \tag{3.3}$$

The empirical gradient is

$$G_n(W) = \frac{2}{Pn}\sum_{\ell=1}^{n}(x_\ell^\top E_W x_\ell)x_\ell x_\ell^\top W. \tag{3.4}$$

Proof. For a deterministic symmetric B, Wick's formula gives

$$\mathbb{E}(x^\top Bx)^2 = 2\operatorname{Tr}(B^2) + (\operatorname{Tr} B)^2.$$

Taking $B = E_W$ gives (3.1). If

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

then

$$\frac{\mathrm{d}}{\mathrm{d}t}R(W + t\dot{W})\big|_{t=0} = 2\operatorname{Tr}(E_W\dot{E}_W) + \tau_W\operatorname{Tr}(\dot{E}_W) = \operatorname{Tr}\left[\left(\frac{2}{P}(2E_W + \tau_W I_d)W\right)^\top \dot{W}\right].$$

This proves (3.3). The empirical formula follows by differentiating each sample term in (3.2). $\square$

4 Finite-frame Muon dynamics

The population gradient belongs to the span of the current student directions and the teacher directions. Define

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

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

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

Proposition 4.1 (Gradient-flow Riccati equation)**.** *Population gradient flow $\dot{W} = -G(W)$ closes on Q, M:*

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

Consequently

$$\dot{C} = \frac{4}{P}(\Lambda C + C\Lambda - 2C\Lambda C). \tag{4.1}$$

In a separated scalar window,

$$\dot{r}_i = \frac{8\mu_i}{P} r_i(1 - r_i). \tag{4.2}$$

Proof. The equations for Q and M follow from $\dot{Q} = \dot{W}^\top W + W^\top \dot{W}$ and $\dot{M} = \dot{W}^\top \Theta$. Differentiating $C = M^\top Q^{-1} M$ and substituting the two equations cancels the common left multiplication terms and gives (4.1). Diagonalizing the separated scalar mode gives (4.2). □

Muon acts by a singular-value power map. If

$$Y = U \operatorname{diag}(s_j) V^\top,$$

define

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

and $\psi_a(0) = 0$. The case $a = 1$ is ordinary gradient descent, while $a = 0$ is the polar or sign-SVD update.

Lemma 4.2 (Finite-frame functional calculus)**.** *Suppose first that $\Gamma = Z^\top Z$ is positive definite. If $Y = ZA$, then*

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

In particular, if $G(W) = ZA_{\mathrm{gd}}$, then

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

Proof. The nonzero singular values of $Y = ZA$ are the singular values of $\Gamma^{1/2}A$, because

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

Writing the singular-value map in the orthonormalized frame $Z\Gamma^{-1/2}$ gives the formula. □

If Γ is singular, the same statement is read on $\operatorname{Ran}\Gamma$: one may replace Z by any full-column-rank matrix spanning $\operatorname{Ran} Z$, or equivalently use the Moore–Penrose square root and discard null coordinates. This support formulation is essential once a teacher direction enters the student span. It preserves the finite-frame closure but does not assert differentiability through a changing rank; the regularized map $s(s^2+\varepsilon^2)^{(a-1)/2}$ supplies that continuation when needed.

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 population Muon flow is

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

It gives

$$\dot{Q} = -\eta_a(U_a^\top Q + QU_a + V_a^\top M^\top + MV_a), \tag{4.3}$$

$$\dot{M} = -\eta_a(U_a^\top M + V_a^\top), \tag{4.4}$$

and

$$\dot{C} = -\eta_a\left(D_a + D_a^\top - D_a^\top C - CD_a\right), \tag{4.5}$$

where $D_a = V_a Q^{-1} M$. Indeed, differentiating $C = M^\top Q^{-1} M$ cancels all terms containing U_a; the four remaining terms combine into the displayed expression. Since U_a, V_a are functions of Q, M, Λ, the Muon population flow is finite-dimensional.

5 Projected-risk recursion and the Muon filter

The projected-risk theory of [1] analyzes stochastic spectral optimizers of the form

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

For Muon-a, the regularized filter is

$$\varphi_{a,\varepsilon}(s) = (s+\varepsilon^2)^{(a-1)/2}, \qquad \phi_{a,\varepsilon}(\sigma) = \sigma(\sigma^2+\varepsilon^2)^{(a-1)/2}. \tag{5.1}$$

For $\varepsilon = 0$, $\phi_{a,0}(\sigma) = \sigma^a$. The parameter ε regularizes the hard edge. Uniform estimates in a are proved first with $\varepsilon > 0$, then passed to the pure Muon map if the small singular values are controlled.

The spectral recursion supplies, for each teacher mode i, a contraction coefficient and a volatility coefficient. In source notation these are row-resolvent moments. We denote them by

$$\delta_i^{(a)}, \qquad \nu_i^{(a)}.$$

For the Muon power family, the leading oversampled power-law scaling is

$$\delta_i^{(a)} \asymp \gamma_b^{(a-1)/2} \mu_i^{(a+1)/2}, \qquad \nu_i^{(a)} \asymp \gamma_b^a \mu_i^a, \tag{5.2}$$

where γ_b is the batch aspect parameter. The exact finite coefficients are the corresponding one- and two-resolvent integrals against the row spectral measure. In the notation of the line-by-line port,

$$\delta_i^{(a)} = \frac{1}{\gamma_b}\int x^{(a+1)/2}\varrho_i(\mathrm{d}x), \qquad \nu_i^{(a)} = \int x^a \varrho_i(\mathrm{d}x). \tag{5.3}$$

Formula (5.3) is the direct replacement of the SignSVD endpoint moments by the Muon-a moments.

5.1 Resolvent formulas for a half-anisotropic design

This subsection specializes the Appendix F calculation of [1] to the Muon-a phase-retrieval front. The random matrix fixed point itself is unchanged by the optimizer. The optimizer enters only after the row spectral measures have been computed.

The half-anisotropic fixed point can be written as

$$\sigma(z) = \frac{1}{B}\sum_i \frac{\mu_i}{\widetilde{s}_1(z)\mu_i - z}, \qquad \widetilde{\sigma}(z) = \frac{\widetilde{s}_1(z)\sigma(z) - \gamma_b}{z}, \tag{5.4}$$

together with the scalar closure

$$\widetilde{s}_1(z)\sigma(z) = f(\xi(z)), \qquad \xi(z) = (-2z\sigma(z)\widetilde{\sigma}(z))^{-1/2}. \tag{5.5}$$

Here B is the number of spectral rows in source notation and $\gamma_b = N/B$. In the present phase-retrieval model, the symbol μ_i in this resolvent equation is a generic row strength. The teacher strength μ_i of Section 3 enters the gradient channel through the cavity-normalized atom described below.

For each row i, define the row Stieltjes transform

$$R_i(z) = \frac{1}{\widetilde{s}_1(z)\mu_i - z} = \int_0^\infty \frac{\varrho_i(\mathrm{d}x)}{x - z}. \tag{5.6}$$

Thus, at regular points of the limiting density,

$$\varrho_i(x) = \frac{\mu_i\,\Im[-\widetilde{s}_1(x + i0)]}{\pi\,|\widetilde{s}_1(x + i0)\mu_i - x|^2}\,\mathrm{d}x. \tag{5.7}$$

Equations (5.4)–(5.7) are the part of Appendix F that is independent of the optimizer.

For a general spectral update $G\varphi(G^\top G)$, the row drift and volatility are

$$d_i[\varphi] = \frac{1}{\gamma_b}\int x\varphi(x)\varrho_i(\mathrm{d}x), \qquad v_i[\varphi] = \int x\varphi(x)^2\varrho_i(\mathrm{d}x). \tag{5.8}$$

Muon-a has $\varphi_a(x) = x^{(a-1)/2}$. Therefore

$$d_i[\varphi_a] = \delta_i^{(a)}, \qquad v_i[\varphi_a] = \nu_i^{(a)},$$

with $\delta_i^{(a)}, \nu_i^{(a)}$ given by (5.3). The SignSVD endpoint is $a = 0$:

$$\delta_i^{(0)} = \gamma_b^{-1}\int \sqrt{x}\,\varrho_i(\mathrm{d}x), \qquad \nu_i^{(0)} = \varrho_i((0,\infty)).$$

The gradient endpoint is $a = 1$:

$$\delta_i^{(1)} = \gamma_b^{-1}\int x\,\varrho_i(\mathrm{d}x), \qquad \nu_i^{(1)} = \int x\,\varrho_i(\mathrm{d}x).$$

In the oversampled regime $\gamma_b \le 1$, the positive part of ϱ_i concentrates at $x = \gamma_b\mu_i$. Hence

$$\delta_i^{(a)} \sim \gamma_b^{(a-1)/2}\mu_i^{(a+1)/2}, \qquad \nu_i^{(a)} \sim \gamma_b^a\mu_i^a, \tag{5.9}$$

which is the scaling used in the power-law Laplace estimate. In the undersampled regime $\gamma_b > 1$, the row measure has a zero atom and a positive mesoscopic component. The correct Muon extension is not obtained by replacing $\nu_i^{(0)}$ by the positive mass alone; one must keep the full moment integrals

$$\delta_i^{(a)} = \frac{1}{\gamma_b}\int x^{(a+1)/2}\varrho_i(\mathrm{d}x), \qquad \nu_i^{(a)} = \int x^a\varrho_i(\mathrm{d}x).$$

In the unresolved hard-edge scaling, the positive component has the Gamma-profile asymptotics of the same row measure:

$$\begin{aligned} \delta_i^{(a)} &\asymp \gamma_b^{-1}\lambda\mu_i\left(\frac{2\gamma_b}{\lambda}\right)^{(a+1)/2}\frac{\Gamma((a+2)/2)}{\sqrt{\pi}}, \\ \nu_i^{(a)} &\asymp \lambda\mu_i\left(\frac{2\gamma_b}{\lambda}\right)^{a}\frac{\Gamma(a+1/2)}{\sqrt{\pi}}. \end{aligned} \tag{5.10}$$

This is why the hard edge and the volatility budget appear together in the fixed-a optimization.

5.2 Phase-retrieval front substitution

The row variable is not literally the teacher strength in the phase-retrieval gradient channel. At a fresh or leave-one-out state, the finite signal atom seen by the gradient singular spectrum has the form

$$\sigma_i^{\mathrm{fr}}(t) \simeq c_{d,P}\sqrt{q_i(t)}\,\rho_i(t), \qquad c_{d,P} = \frac{4}{d(Pd)^{1/4}}, \tag{5.11}$$

up to normalization conventions. Thus the same resolvent formulas are used twice: first to compute the spectral optimizer coefficients from row measures, and then to read the phase-retrieval gradient atom as a finite-rank signal depending on the Volterra masses. This is the bridge between Appendix F and the gradient BBP equation (10.1).

Assumption 5.1 (Fresh projected-gradient channel)**.** *Fix a compact time interval and $\varepsilon > 0$. Conditional on the finite frame (Q_t, M_t), a fresh phase-retrieval gradient has the orthogonal Gaussian decomposition required by the projected-risk theorem of [1]. Uniformly over the finite mode family, the admissible exponents, and the resolvent contours used below, its one-resolvent drift and two-resolvent quadratic-variation observables converge to $\delta_i^{(a)}$ and $\nu_i^{(a)}$. These coefficients are obtained by inserting the conditional phase-retrieval covariance profile and $\varphi_{a,\varepsilon}$ into the general-filter formulas of that theorem.*

The source theorem proves this statement for its matrix least-squares channel. For nonlinear phase retrieval it is an additional channel-identification assumption. It concerns a fresh gradient and is therefore logically prior to the same-sample comparison in Section 15. All deterministic calculations below are conditional on these coefficients.

6 Mode equations and nonlinear Volterra law

Let $\chi_i(t)$ denote the contribution of mode i to the projected risk scale, and let

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

be the aggregate risk scale. The weights w_i depend on the chosen normalization; in phase retrieval one may take $w_i \propto \mu_i$.

Theorem 6.1 (Reduced Muon mode equation)**.** *Under Assumption 5.1, the scalar projected mode equation for fixed a has the form*

$$\dot\chi_i(t) = -2\eta\,\delta_i^{(a)}\mathfrak{q}(t)^{(a-1)/2}\chi_i(t) + \eta^2\nu_i^{(a)}\mathfrak{q}(t)^a. \tag{6.1}$$

Proof. The projected-risk recursion gives a deterministic contraction term linear in the projected mode scale and a quadratic fluctuation term. Inserting $\varphi_a(s) = s^{(a-1)/2}$ into the rescaled gradient $G^\circ = \sqrt{2\mathfrak{q}}\,G$ produces

$$\varphi_a(G^\circ(G^\circ)^\top)G^\circ = (2\mathfrak{q})^{a/2}\varphi_a(GG^\top)G.$$

After projecting on mode i, the linear drift is proportional to $\mathfrak{q}^{(a-1)/2}\chi_i$, while the variance term is proportional to $\mathfrak{q}^a$. The proportionality constants are exactly the row-resolvent moments $\delta_i^{(a)}$ and $\nu_i^{(a)}$. ☐

Introduce the Volterra clock

$$d\tau = \eta \mathfrak{q}(t)^{(a-1)/2}\, dt. \tag{6.2}$$

Then (6.1) becomes

$$\partial_\tau \chi_i(\tau) = -2\delta_i^{(a)}\chi_i(\tau) + \eta\nu_i^{(a)}\mathfrak{q}(\tau)^{(a+1)/2}. \tag{6.3}$$

Theorem 6.2 (Nonlinear Volterra equation)**.** *Let*

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

Then the aggregate risk scale satisfies

$$\mathfrak{q}(\tau) = F_a(\tau) + \eta \int_0^\tau K_a(\tau - u)\mathfrak{q}(u)^{(a+1)/2}\, du. \tag{6.4}$$

Proof. Solving (6.3) gives

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

Multiplying by w_i, summing over i, and exchanging the finite or dominated infinite sum with the integral gives (6.4). ☐

7 Two-resolvent volatility

The scalar volatility coefficient in (6.1) is not a fitted variance. It is the row-summed projection of the corrected two-resolvent kernel. This section records the matrix-valued object before any scalar or power-law reduction is made.

Let

$$R_t(z) = (H_t - zI)^{-1}$$

be the resolvent of the fresh or leave-one-out gradient covariance block used in the projected-risk recursion. The two-resolvent calculation in Appendix H of [1] requires the two-resolvent bilinear form

$$P_{3,i,t}(w,z) = \frac{1}{B} u_i^\top R_t(w)\widetilde{G}_t \vec{X}\vec{D}^3\vec{Y}^\top R_t(z)u_i. \tag{7.1}$$

This term is the reason why one cannot replace the volatility by a product of two one-resolvent deterministic equivalents. The companion object is

$$E_{0,i,t}(w,z) = u_i^\top R_t(w)\Sigma_{\text{out},t}R_t(z)u_i = \frac{\mu_{i,t}}{d_{w,i,t}d_{z,i,t}(1 - L_t(w,z)\sigma_{wz,t})}, \tag{7.2}$$

where

$$d_{\zeta,r,t} = \widetilde{s}_{1,t}(\zeta)\mu_{r,t} - \zeta, \qquad \sigma_{wz,t} = \frac{1}{B}\sum_r \frac{\mu_{r,t}^2}{d_{w,r,t}d_{z,r,t}}.$$

After inserting the self-consistent two-resolvent corrections derived in the source appendix, the variance kernel is

$$\boxed{V_{ij,t}^{\text{corr}}(z,w) = \frac{-w\lambda_{j,t}}{B(s_{1,t}(z)\lambda_{j,t} - z)(s_{1,t}(w)\lambda_{j,t} - w)}\left[\frac{\widetilde{s}_{1,t}(z)}{\widetilde{\sigma}_t(z)}E_{0,i,t}(w,z) + z\sigma_t(w)P_{3,i,t}(w,z)\right].} \tag{7.3}$$

The remaining symbol $P_{3,i,t}$ is explicit once the one-resolvent fixed point has been solved. In the notation of the two-resolvent calculation,

$$\begin{aligned} P_{3,i,t}(w,z) = &\left(\gamma_{\mathrm{in},t} - \frac{\sigma_t(z)\Gamma_{E,t}(w,z)}{1-\Gamma_{C,t}(w,z)}\right) E_{0,i,t}(w,z) I_4(\nu_{z,t}) \\ &+ \frac{\nu_{z,t}^2 \overline{\Lambda}_t^{(3)}(\nu_{z,t}) - \nu_{w,t}^2 \overline{\Lambda}_t^{(3)}(\nu_{w,t})}{\nu_{z,t}^2 - \nu_{w,t}^2}. \end{aligned} \tag{7.4}$$

Here

$$\sigma_t, \quad \widetilde{\sigma}_t, \quad s_{1,t}, \quad \widetilde{s}_{1,t}$$

are the fixed-point scalars of the fresh/frozen deterministic equivalent at time t. In the notation of the source appendix, the ingredients correspond to the final E_0, P_3, and corrected variance-kernel identities. The important point for the present paper is structural: Muon changes the contour test function, not the fixed-point equation itself.

For Muon-a,

$$\varphi_a(x) = x^{(a-1)/2}.$$

The corrected projected volatility before row summation is

$$\mathcal{V}_{ij,t}^{(a)} = \frac{1}{(2\pi i)^2} \oint_{\Gamma_t} \oint_{\Gamma_t} \varphi_a(z)\varphi_a(w) V_{ij,t}^{\mathrm{corr}}(z,w)\, \mathrm{d}z\, \mathrm{d}w. \tag{7.5}$$

The contour Γ_t surrounds the active spectral support and is kept at a regularized distance from the hard edge until the uniform bound is proved.

Proposition 7.1 (Row-summed two-resolvent reduction)**.** *In the half-anisotropic row-summed regime, after the corrected kernel* (7.3) *has been inserted and the corresponding Cauchy identities have been applied,*

$$\sum_j \mathcal{V}_{ij,t}^{(a)} = \int x^a \varrho_{i,t}(\mathrm{d}x) + o(1). \tag{7.6}$$

For phase retrieval with a finite Schur/RFA front, the corresponding statement is

$$\mathcal{V}_{i,\mathrm{PR},t}^{(a)} = \int x^a \varrho_{i,t}(\mathrm{d}x) + b_i(t)^\top c_t^{\mathrm{fr}}(a) + r_{i,t}^{(a)}, \tag{7.7}$$

with a front-norm remainder

$$\sup_{a \in I_t} \sum_{i \in \mathrm{front}(t)} w_i(t) |r_{i,t}^{(a)}| = o_{\mathbb{P}}(1)$$

on resolved windows.

Proof. Equation (7.5) is the volatility contour with (7.3) inserted. In the row-summed half-anisotropic setting, the terms containing E_0 and P_3 combine under the Cauchy identities into the row spectral moment of the deterministic measure $\varrho_{i,t}$. This gives (7.6). In phase retrieval, the finite teacher/front coordinates are not eliminated by the row sum; they span a low-dimensional Schur/RFA basis $b_i(t)$, producing (7.7). The remainder is exactly the unresolved part of the same contour projection. □

The coefficient in the finite correction is computable from the corrected kernel. Define

$$U_{i,t}^{(a)} = \sum_j \mathcal{V}_{ij,t}^{(a)}, \qquad v_{i,t}^{(a)} = \int x^a \varrho_{i,t}(\mathrm{d}x), \qquad S_{i,t}^{(a)} = U_{i,t}^{(a)} - v_{i,t}^{(a)}.$$

If B_t is the Schur/RFA front basis and W_t^{fr} is the associated front weight matrix, then

$$c_t^{\mathrm{fr}}(a) = (B_t^\top W_t^{\mathrm{fr}} B_t)^{-1} B_t^\top W_t^{\mathrm{fr}} S_t^{(a)}. \tag{7.8}$$

Thus the scalar coefficient $\nu_i^{(a)}$ used in (6.1) is the resolved-window limit of a two-resolvent calculation, while (7.8) is the finite correction that must be kept near terminal/front-floor windows.

8 Power-law Laplace estimates and fixed exponent

The Volterra equation becomes explicit on a local power-law front. Suppose that on the active window

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

Using (5.2), the forcing has the form

$$F_a(\tau) \asymp \sum_i i^{-(\zeta+\beta)} \exp\{-c_a \tau i^{-\zeta(a+1)/2}\}.$$

Lemma 8.1 (Power-law Laplace shell)**.** *Let $p > 1$, $q > 0$, and $c > 0$. Then, as $\tau \to \infty$,*

$$\sum_{i \geq 1} i^{-p} e^{-c\tau i^{-q}} \asymp \tau^{-(p-1)/q}.$$

Proof. The active shell is defined by $c\tau i^{-q} = O(1)$, hence $i \asymp \tau^{1/q}$. More explicitly, compare the sum with the integral

$$\int_1^\infty x^{-p} e^{-c\tau x^{-q}}\, \mathrm{d}x.$$

With $u = c\tau x^{-q}$, this integral equals

$$\frac{1}{q}(c\tau)^{-(p-1)/q} \int_0^{c\tau} u^{(p-1)/q-1} e^{-u}\, \mathrm{d}u,$$

which is $\asymp \tau^{-(p-1)/q}$. Monotone integral comparison gives the same order for the sum. $\square$

Applying Lemma 8.1 gives the Volterra tail exponent

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

The hard-edge and tail-resolution costs used in the main text are

$$e_{\text{hard}}(a) = \frac{1-a}{2}, \qquad e_{\text{tail}}(a) = \frac{\zeta(1+a)}{2(\zeta+\beta-1)}. \tag{8.1}$$

The bottleneck exponent is

$$e_{\text{bal}}(a) = \max\{e_{\text{hard}}(a), e_{\text{tail}}(a)\}.$$

Proposition 8.2 (Fixed-exponent balance)**.** *On a local power-law front with exponents (ζ, β), the exponent that minimizes $e_{\text{bal}}(a)$ over $a \in [0, 1]$ is*

$$a_{\text{bal},*}(\zeta, \beta) = \operatorname{clip}_{[0,1]}\left(\frac{\beta - 1}{2\zeta + \beta - 1}\right). \tag{8.2}$$

In the near-zero phase retrieval regime $\beta = \zeta$,

$$a_{\text{PR},*}(\zeta) = \operatorname{clip}_{[0,1]}\left(\frac{\zeta - 1}{3\zeta - 1}\right). \tag{8.3}$$

For $\zeta > 1/2$, the minimized bottleneck exponent is

$$e_*^{PR}(\zeta) = \begin{cases} \dfrac{\zeta}{2(2\zeta - 1)}, & 1/2 < \zeta \leq 1, \\[2ex] \dfrac{\zeta}{3\zeta - 1}, & \zeta > 1. \end{cases} \tag{8.4}$$

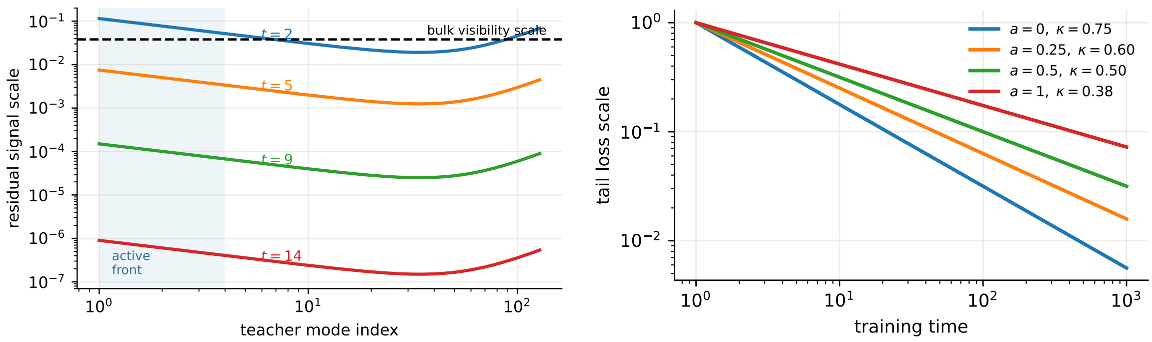
Proof. The function e_{hard} decreases in a, and e_{tail} increases in a. The minimum of their maximum is obtained at their intersection, unless the intersection lies outside $[0, 1]$. Solving

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

gives

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

Clipping to $[0, 1]$ gives (8.2). The formula (8.3) follows by substituting $\beta = \zeta$. If $1/2 < \zeta \leq 1$, the unconstrained intersection is nonpositive and the minimum is attained at $a = 0$, where the tail branch equals $\zeta/[2(2\zeta - 1)]$ and dominates the hard-edge branch. If $\zeta > 1$, the intersection is interior and either branch there equals $\zeta/(3\zeta - 1)$. This proves (8.4). $\square$



(a) Local front slope.

(b) Fixed-a tail law.

Figure 2: Power-law inputs for Proposition 8.2. The local front supplies ζ, β; the Volterra/Laplace estimate supplies the tail clock.

9 A common BBP readout for three spectra

The Volterra state describes how much signal exists. A spectrum determines whether that signal is visible at finite dimension. The same deterministic calculation applies to the gradient, the weights, and the Hessian.

There are therefore three layers, not one. The first layer is dynamical:

$$\mathfrak{V}_t = \{\chi_i(t), \rho_i(t), q_i(t), \mathfrak{q}(t)\}_{i \leq k}.$$

It is obtained from the reduced Muon Volterra equation. The second layer is kinematic: each measured matrix has a bulk law and a finite signal frame,

$$\mathfrak{S}_t^c = (\nu_{B,t}^c,\ A_{\text{sig},t}^c,\ L_t^c,\ B_{\text{bulk},t}^c), \qquad c \in \{G, W, H\},$$

where G is the gradient, W is the weight covariance, and H is the Hessian. The third layer is spectral: Schur or BBP equations turn $\mathfrak{S}_t^c$ into visible roots, residues, and crossing times. The logical dependence is

$$\mathfrak{V}_t \longmapsto \mathfrak{S}_t^G, \mathfrak{S}_t^W, \mathfrak{S}_t^H \longmapsto \{(\lambda_i^c(t), \Omega_i^c(t), t_i^c)\}_{c \in \{G,W,H\}}. \tag{9.1}$$

Thus the Hessian calculations do not replace the weight or gradient calculations. They are the third readout of the same deterministic state.

The three readouts have different meanings, so the same BBP word should not be read as the same event in all channels. For later reference, the channel dictionary is

$$\begin{aligned} G \text{ channel}: &\quad 1 = \theta_i^G(t)^2 D_t^G(\lambda_i^G(t)), \\ W \text{ channel}: &\quad \det\Big(\lambda I - A^W_{\mathrm{sig},t} - L_t^W(\lambda I - B^W_{\mathrm{bulk},t})^{-1}(L_t^W)^\top\Big) = 0, \\ H \text{ channel}: &\quad \det\Big(\lambda I - \mathcal{K}_t^H(\lambda)\Big) = 0. \end{aligned} \tag{9.2}$$

All three equations are evaluated along the same Volterra path $\mathfrak{V}_t$. What changes from one line to the next is the bulk transform and the finite signal frame.

Let $\mathcal{H}_t^c$ be one of the self-adjoint matrices whose outliers are being studied, where $c \in \{G, W, H\}$. For a rectangular gradient, $\mathcal{H}_t^G$ is the Hermitian dilation. The weight covariance and Hessian are self-adjoint already. Split the space into a finite signal subspace and its orthogonal bulk complement:

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

Theorem 9.1 (Finite outlier equation and residue)**.** *For z outside the bulk spectrum of $B^c_{\mathrm{bulk,t}}$, define*

$$\mathcal{K}_t^c(z) = A^c_{\mathrm{sig,t}} + L_t^c(zI - B^c_{\mathrm{bulk,t}})^{-1}(L_t^c)^*. \tag{9.3}$$

Then z is an outlier of $\mathcal{H}_t^c$ if and only if

$$\det(zI - \mathcal{K}_t^c(z)) = 0. \tag{9.4}$$

If u is the normalized finite vector solving $\mathcal{K}_t^c(\lambda)u = \lambda u$, the signal-space residue of the corresponding outlier eigenvector is

$$\Omega^c(\lambda) = \frac{1}{u^\top(I - \partial_z \mathcal{K}_t^c(\lambda))u}. \tag{9.5}$$

Proof. Write the eigenvector as $(v_{\mathrm{sig}}, v_{\mathrm{bulk}})$. The bulk component of $(\mathcal{H}_t^c - zI)v = 0$ gives

$$v_{\mathrm{bulk}} = -(B^c_{\mathrm{bulk,t}} - zI)^{-1}(L_t^c)^* v_{\mathrm{sig}}.$$

Substituting into the signal component gives

$$(\mathcal{K}_t^c(z) - zI)v_{\mathrm{sig}} = 0,$$

which proves (9.4). For the residue, differentiate the finite eigenvalue equation with respect to z. Equivalently, compare the pole of the projected resolvent

$$P_{\mathrm{sig}}(\mathcal{H}_t^c - zI)^{-1}P^*_{\mathrm{sig}} = (\mathcal{K}_t^c(z) - zI)^{-1}$$

with its spectral decomposition. The pole coefficient is (9.5). □

If $x_+^c(t)$ is the right bulk edge, the right margin is

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

For a left outlier with left edge $x_-^c(t)$, use

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

The BBP exit time is the first sign change of this margin.

Proposition 9.2 (Stability of BBP times)**.** *Assume that* $\mathcal{K}_t^c(z)$ *and* $\partial_z \mathcal{K}_t^c(z)$ *converge uniformly on compact contours outside the bulk, and that the relevant BBP contact is simple:*

$$\frac{\mathrm{d}}{\mathrm{d}t} \Delta_i^c(t_i^c) \neq 0.$$

Then the outlier location, residue, and BBP exit time are deterministic continuous functions of the Volterra state.

Proof. Uniform convergence on a contour surrounding one simple zero of $\det(zI - \mathcal{K}_t^c(z))$ gives convergence of the zero by Rouche's theorem. The residue is a rational function of the same finite matrix and derivative, therefore it converges as well. A simple sign change of the margin is stable under uniform perturbation, hence the exit time is stable. □

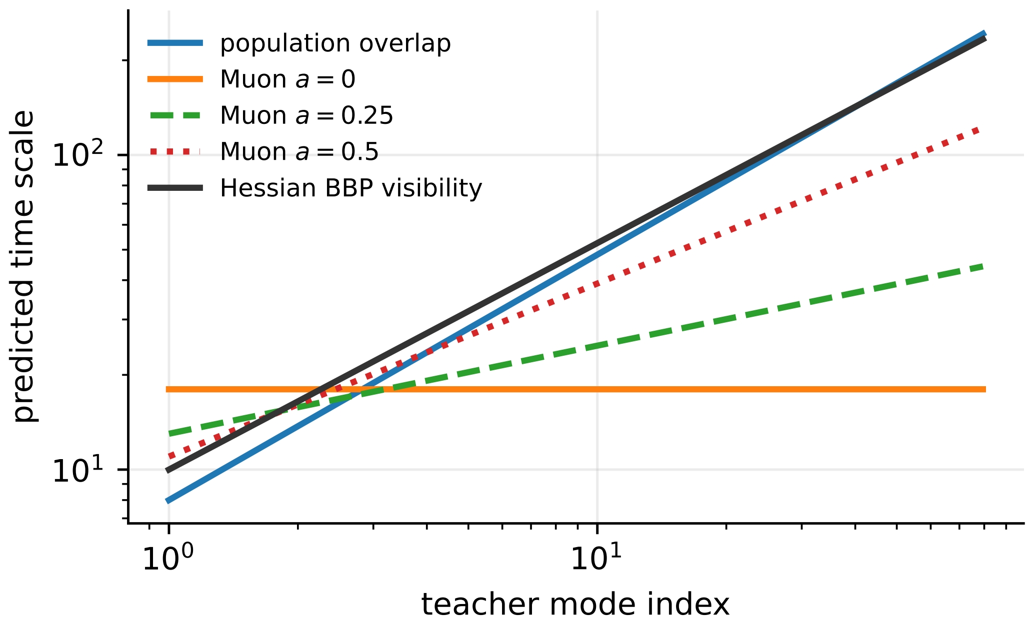


Figure 3: Three notions of escape. Population overlap, Muon-amplified drift, and BBP visibility are distinct. The formulas above identify the BBP clock through finite roots and residues.

10 Gradient channel

The gradient channel is the spectrum on which Muon directly acts. Hermitize the gradient:

$$\widetilde{G}_t = \begin{pmatrix} 0 & G_t \\ G_t^\top & 0 \end{pmatrix} = B^G_{\text{bulk,t}} + S_t^G,$$

where S_t^G is finite rank in the signal coordinates. The bulk singular law is denoted by $\nu_{B,t}$. If D_t^G is the rectangular D-transform of the bulk singular-value law, a rank-one signal strength $\theta_i(t)$ creates an outlier $\lambda_i^G(t)$ through

$$1 = \theta_i(t)^2 D_t^G(\lambda_i^G(t)). \tag{10.1}$$

The right-edge margin is

$$M_i^G(t) = \theta_i(t)^2 D_t^G(x_{+,G}(t)) - 1. \tag{10.2}$$

The gradient mode is visible when $M_i^G(t) > 0$.

In the phase-retrieval front, the effective gradient atom has the form

$$\theta_i(t) \asymp c_{d,P}\sqrt{q_i(t)}\,\rho_i(t)$$

up to the projected-risk clock and normalization constants. Thus the gradient BBP condition is a deterministic function of the Volterra masses.

The preceding line is the compressed scalar form. In the finite-frame notation used throughout the paper, the empirical gradient is

$$G_{n,t} = \frac{2}{P\,n}\sum_{\ell=1}^{n} r_t(x_\ell)\,x_\ell h_t(x_\ell)^\top, \qquad r_t(x) = P^{-1}\left\|W_t^\top x\right\|^2 - y^\top \Lambda y.$$

After conditioning on the finite Gaussian projections $(W_t^\top x, \Theta^\top x)$, the orthogonal Gaussian component gives the rectangular bulk law $\nu^G_{B,t}$. The finite conditional mean gives a signal frame

$$S^G_t = \sum_{i\le k}\theta^G_i(t)\,u^G_i(t)(v^G_i(t))^\top + \text{higher finite-frame couplings.}$$

For a separated mode, this reduces to (10.1). For a cluster J, the scalar D-transform is replaced by the finite matrix

$$\mathcal{K}^G_{J,t}(z) = (\Theta^G_{J,t})^\top \mathcal{D}^G_t(z)\Theta^G_{J,t}, \qquad \det(I - \mathcal{K}^G_{J,t}(\lambda)) = 0. \tag{10.3}$$

The branch residue is read from the same Schur derivative as in (9.5). Hence gradient visibility, gradient singular-vector overlap, and the Muon drift window are all deterministic readouts of $\mathfrak{V}_t$, provided the fresh or LOO bulk law is used.

10.1 AMP/VAMP denoising interpretation

Let $R_t(\sigma)$ be the smoothed signal-to-bulk likelihood ratio on the active gradient singular window. For a spectral filter f, define

$$\mathcal{H}_t(f) = \frac{\langle f, R_t\rangle^2_{\nu_{B,t}}}{\langle f, f\rangle_{\nu_{B,t}}}. \tag{10.4}$$

Lemma 10.1 (Linearized AMP/VAMP filter)**.** *The unrestricted maximizer of $\mathcal{H}_t(f)$ is $f_{\mathrm{opt}} \propto R_t$.*

Proof. By Cauchy–Schwarz,

$$\langle f, R_t\rangle^2_{\nu_{B,t}} \le \langle f, f\rangle_{\nu_{B,t}}\,\langle R_t, R_t\rangle_{\nu_{B,t}}\,,$$

with equality if and only if f is proportional to R_t on the active window. □

Restricting f to Muon powers $f_a(\sigma) = \sigma^a$ gives the local free energy

$$F_t(a) = -\log \mathcal{H}_t(f_a).$$

If

$$\log R_t(\sigma_t e^u) = \kappa_t + a_{\mathrm{loc}}(t)u + o(1)$$

on the active window, then the AMP/VAMP filter is locally proportional to $\sigma^{a_{\mathrm{loc}}(t)}$. This is the precise sense in which the local AMP/VAMP direction projects to a Muon exponent.

11 Weight channel

The weight covariance $WW^\top/P$ records which teacher directions have entered the representation. The captured masses generated by the Volterra equation produce finite signal blocks in this covariance. The outlier equation is

$$\det\Big(\lambda I - A^W_{\text{sig,t}} - L^W_t(\lambda I - B^W_{\text{bulk,t}})^{-1}(L^W_t)^\top\Big) = 0. \tag{11.1}$$

Let $\lambda^W_i(t)$ be the branch labeled by its residue in the i-th teacher direction. The weight BBP exit time is

$$t^W_i = \inf\{t : \lambda^W_i(t) > x_{+,W}(t)\}. \tag{11.2}$$

For a simple root with normalized teacher component u_i, the finite Schur residue is

$$\Omega^W_i(t) = \frac{1}{1 + \left\|(\lambda^W_i I - B^W_{\text{bulk,t}})^{-1}(L^W_t)^\top u_i\right\|^2}. \tag{11.3}$$

This formula is useful in practice because it labels branches by teacher mass rather than by raw eigenvalue rank.

In a scalar separated window, the branch is a function of the captured mass $r_i(t)$. Thus the weight BBP is not a second training equation: it is a visibility margin evaluated along the same Volterra trajectory. More explicitly, if the update atom satisfies

$$\sigma^{\text{upd}}_i(t) \asymp m_t\, s_i^{3/2}\sqrt{r_i(t)}(1 - r_i(t)),$$

then a rectangular D-transform description of the bulk gives the contact condition

$$(\sigma^{\text{upd}}_i(t^W_i))^2 D^W_t(x_{+,W}(t^W_i)) = 1. \tag{11.4}$$

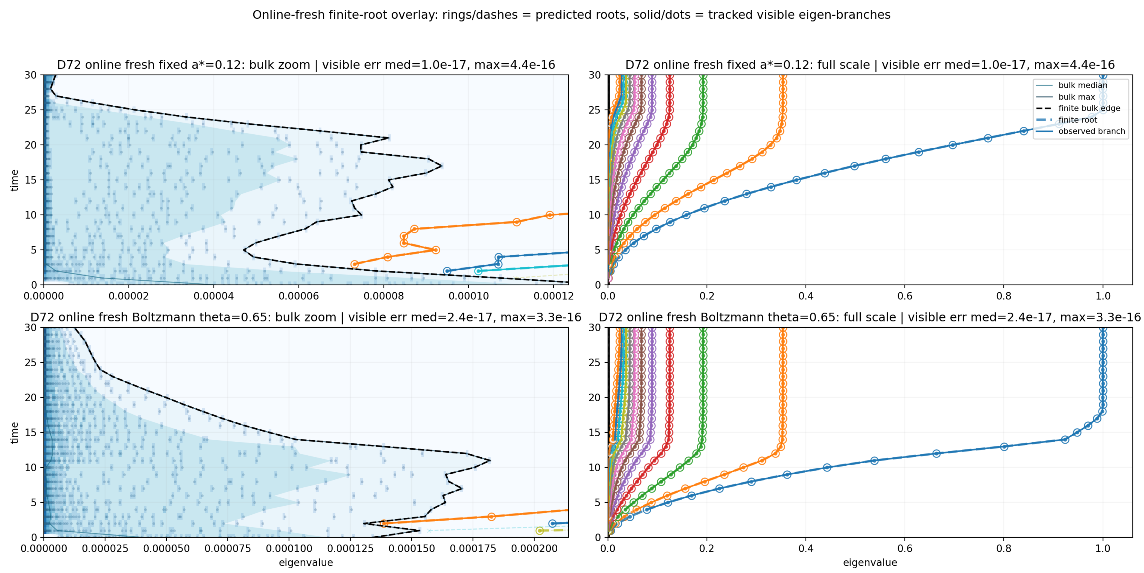


Figure 4: Weight spectrum over time. The black curve is the finite bulk edge, the colored curves are reduced roots, and the stars are matched empirical branches. This is the strongest branch-level BBP verification: the labels are root/residue labels, not raw eigenvalue ranks.

12 Hessian channel

The empirical Hessian is the most detailed spectral readout. For the square loss, the block Hessian has weighted sample-covariance form:

$$H_{n,bc}(W) = \frac{1}{n}\sum_{\ell=1}^{n} \Phi_{bc}(W^\top x_\ell, \Theta^\top x_\ell) x_\ell x_\ell^\top, \tag{12.1}$$

where

$$\Phi_{bc}(h,y) = \frac{4}{P^2} h_b h_c + \frac{2}{P}\left(\frac{1}{P}\|h\|^2 - y^\top \Lambda y\right)\delta_{bc}. \tag{12.2}$$

Let $A_t^H(h,y)$ be the $P \times P$ matrix with entries $\Phi_{bc}(h,y)$, and let $\alpha_H = n_H/d$. The fresh-sample bulk is given by the matrix Dyson equation

$$-S_t^H(z)^{-1} = zI - \mathbb{E}\Big[A_t^H(I + \alpha_H^{-1} S_t^H(z) A_t^H)^{-1}\Big]. \tag{12.3}$$

Residual teacher coordinates produce left or unstable branches. Parallel teacher coordinates produce right or stable branches. Projecting the same Dyson resolvent onto these two coordinate systems gives finite matrices $\mathcal{K}_t^{H,-}$ and $\mathcal{K}_t^{H,+}$, and the branches solve

$$\det(\lambda I - \mathcal{K}_t^{H,-}(\lambda)) = 0, \qquad \det(\lambda I - \mathcal{K}_t^{H,+}(\lambda)) = 0. \tag{12.4}$$

The matrices themselves are

$$\mathcal{K}_t^{H,\pm}(z) = \mathbb{E}_{\Gamma_t}\left[\xi_t^\pm (\xi_t^\pm)^\top \otimes A_t^H(h,y)(I + \alpha_H^{-1} S_t^H(z) A_t^H(h,y))^{-1}\right], \tag{12.5}$$

with the amplitude projection applied on the right branch when rotational tangent directions are present. If c is the normalized finite eigenvector at a simple root, the Hessian residue is

$$\Omega_i^H(t) = \frac{1}{c^\top (I - \partial_z \mathcal{K}_t^H(\lambda_i^H(t)))c}. \tag{12.6}$$

In the large-α_H scalar BBP window, the residual contact scale is

$$\mu_i(1 - r_i(t_{i,\pm})) = \frac{c_i^\pm}{\sqrt{\alpha_H}} + O(\alpha_H^{-1}). \tag{12.7}$$

The constants $c_i^\pm$ are obtained by evaluating the edge equation (12.3) and the finite kernels (12.4) at the current Volterra state. Equivalently,

$$t_i^{H,-}: \quad \sqrt{\alpha_H}\mu_i(1 - r_i(t_i^{H,-})) = c_i^- + o(1), \qquad t_i^{H,+}: \quad \sqrt{\alpha_H}\mu_i(1 - r_i(t_i^{H,+})) = c_i^+ + o(1). \tag{12.8}$$

The left time marks residual unstable curvature becoming bulk-hidden; the right time marks aligned amplitude curvature becoming visibly detached.

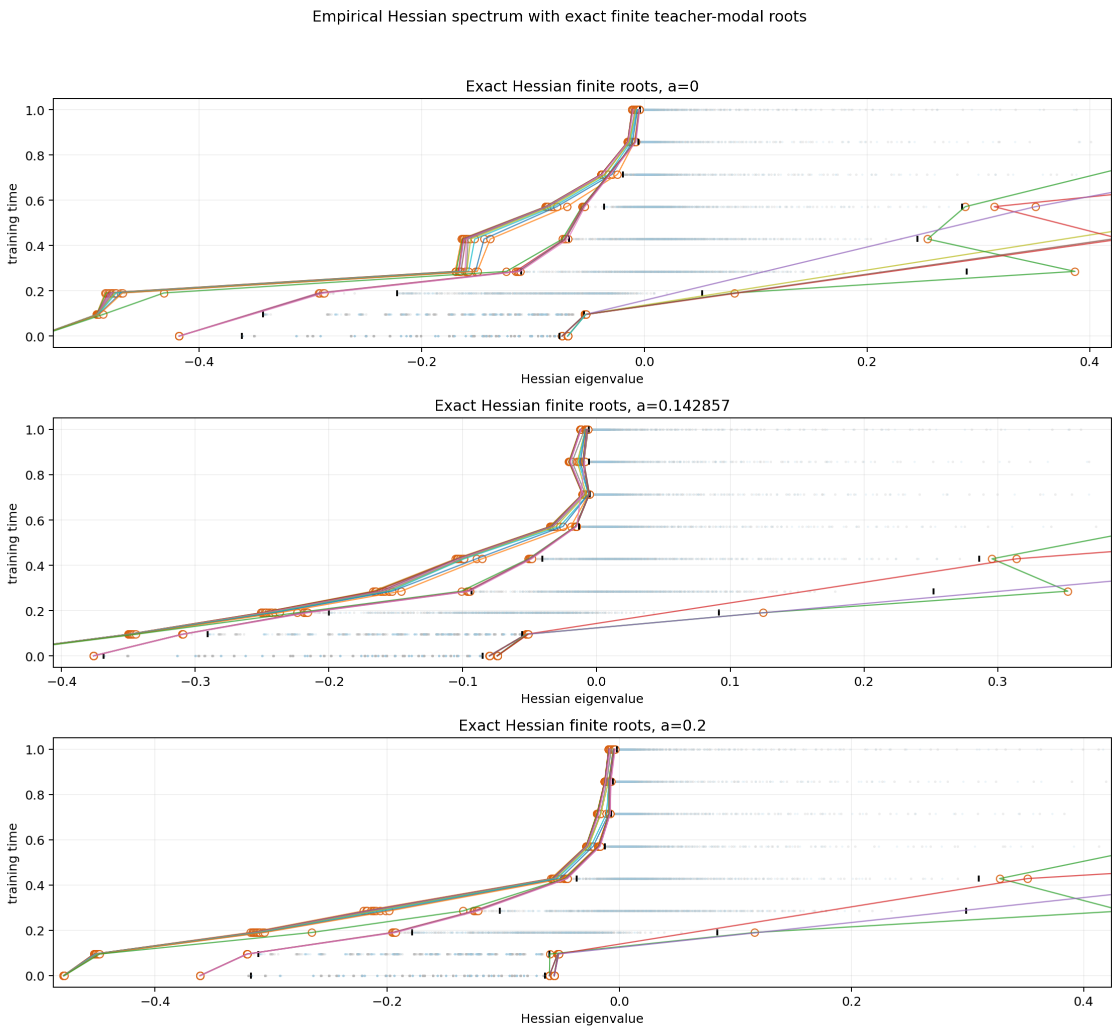


Figure 5: Hessian outlier roots over time. The plot shows residual and parallel Hessian branches as readouts of the same Volterra state.

13 One theorem for the three spectral readouts

Theorem 13.1 (Conditional three-spectrum prediction)**.** *Assume:*

1. *the reduced Muon Volterra state* $\chi_i, \rho_i, \mathfrak{q}$ *is regular on a compact time interval;*
2. *the bulk laws for the gradient, weight covariance, and Hessian converge uniformly on the contours used to read outliers;*
3. *the finite outlier matrices* $\mathcal{K}_t^G, \mathcal{K}_t^W, \mathcal{K}_t^H$ *and their* z*-derivatives converge uniformly away from exact non-transversal contacts;*
4. *branch contacts are either separated or treated by a smoothed visible mass.*

Then the gradient singular outliers, weight covariance outliers, and Hessian left/right outliers have deterministic locations, residues, and BBP exit times given by the same Volterra state.

Proof. The Volterra state determines the finite signal strengths, residual masses, and captured masses. The three bulk laws determine the corresponding resolvents. The finite outlier matrices are therefore deterministic functions of the same state. The conclusion follows by applying Theorem 9.1 and Proposition 9.2 to each of $c = G, W, H$. $\square$

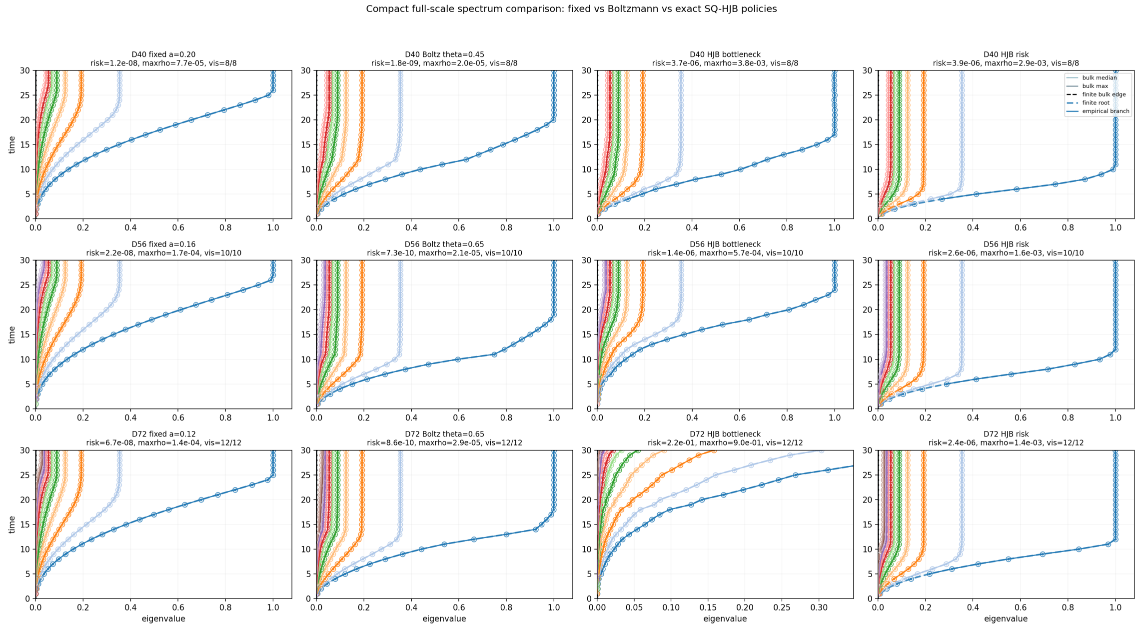


Figure 6: Three-spectrum outlier overlay for online exponent rules. The same Volterra/HJB state predicts the visible roots, and root/residue labels keep the comparison stable through rank exchanges.

Theorem 13.2 (Weight–Hessian–gradient bridge)**.** *Fix a finite family of teacher modes and a compact time interval avoiding non-transversal contacts, or replace hard labels by the smoothed partition observable of Section 15.1. Assume the fresh/frozen deterministic equivalents hold for the Muon-a contour family, the Volterra state converges uniformly, and the finite Schur/RFA comparison* (15.6) *is* $o_{\mathbb{P}}(1)$*. Then the empirical gradient singular outliers, weight covariance outliers, and Hessian outliers converge jointly to the deterministic roots obtained from* (10.1)*,* (11.1)*, and* (12.4)*. Their residues converge to* (9.5)*,* (11.3)*, and* (12.6)*. If the three margins cross transversally, the three families of crossing times converge to the deterministic times.*

Proof. The Volterra state fixes $r_i(t)$, the residual masses, the captured masses, and the finite signal atoms. The fresh/frozen deterministic equivalent fixes the gradient bulk and the Muon drift/volatility coefficients. The weight and Hessian bulks are then deterministic functions of the same summaries. Uniform convergence of the corresponding finite matrices on outlier contours allows Theorem 9.1 to be applied simultaneously to the three channels. The residues are derivatives of the same finite equations. A transversal zero of each margin is stable under uniform perturbation, giving the crossing-time convergence. □

In a separated front, the theorem gives the following physical dictionary:

$$
\begin{aligned}
t_i^G &: \quad \theta_i(t)^2 D_t^G(x_{+,G}(t)) = 1, \\
t_i^W &: \quad (\sigma_i^{\mathrm{upd}}(t))^2 D_t^W(x_{+,W}(t)) = 1, \\
t_i^{H,-} &: \quad \sqrt{\alpha_H}\mu_i(1 - r_i(t)) = c_i^-, \\
t_i^{H,+} &: \quad \sqrt{\alpha_H}\mu_i(1 - r_i(t)) = c_i^+.
\end{aligned}
$$

Thus gradient BBP measures whether the update channel sees the mode, weight BBP measures whether the representation has detached from its bulk, and Hessian BBP measures whether residual or aligned curvature is visible. They are different clocks, but all are read along the same Volterra trajectory.

Theorem 13.3 (End-to-end conditional prediction). *Consider centered multi-index phase retrieval on a power-law active front*

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

Assume the fresh/frozen deterministic equivalents, the two-resolvent volatility reduction, the Volterra stability assumptions of Section 6, and the finite Schur/Dyson regularity assumptions of Theorem 13.1. Then, for fixed a, the Muon trajectory is described by the nonlinear Volterra equation (6.4)*. Its raw power-law bottleneck exponent is*

$$e_{\mathrm{raw}}(a) = \max\left\{\frac{1-a}{2}, \frac{\zeta(1+a)}{2(\zeta+\beta-1)}\right\}. \tag{13.1}$$

The optimal constant exponent is

$$a_{\mathrm{raw},*}(\zeta,\beta) = \mathrm{clip}_{[0,1]}\left(\frac{\beta-1}{2\zeta+\beta-1}\right). \tag{13.2}$$

For the near-zero phase-retrieval front $\beta = \zeta$,

$$a^{PR}_{\mathrm{raw},*}(\zeta) = \mathrm{clip}_{[0,1]}\left(\frac{\zeta-1}{3\zeta-1}\right), \qquad e^{PR}_{*}(\zeta) = \begin{cases} \dfrac{\zeta}{2(2\zeta-1)}, & 1/2 < \zeta \le 1, \\ \dfrac{\zeta}{3\zeta-1}, & \zeta > 1. \end{cases} \tag{13.3}$$

Along this same Volterra state, the gradient, weight, and Hessian BBP exits are the finite roots described in (10.1)*,* (11.4)*, and* (12.8)*. If the fresh-to-reused Schur/RFA error* (15.6) *vanishes, the same predictions hold for the same-sample online trajectory in the finite Schur/RFA topology.*

Proof. The fresh/frozen deterministic equivalent supplies the row measures ϱ_i and hence the Muon moments $\delta_i^{(a)}, \nu_i^{(a)}$. The mode equation (6.1) and the clock change (6.2) give the nonlinear Volterra equation by Theorem 6.2. The power-law Laplace estimate of Lemma 8.1 gives the tail branch in (13.1), while the raw clock/hard-edge normalization gives the branch $(1-a)/2$. Balancing the decreasing and increasing branches gives (13.2); substituting $\beta = \zeta$ gives (13.3). Finally, the finite Schur/BBP theorem applies to the gradient, weight, and Hessian readouts, because their finite signal strengths are deterministic functions of the same Volterra state. The last statement is exactly Proposition 15.2. ☐

14 Variable exponent and HJB/Boltzmann control

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

$$\dot{\chi}_i(t) = -2\eta\, \delta_i^{(a_t)} \mathfrak{q}(t)^{(a_t-1)/2} \chi_i(t) + \eta^2 \nu_i^{(a_t)} \mathfrak{q}(t)^{a_t}. \tag{14.1}$$

At finite teacher rank, the exponent is a control and the complete state is

$$X_t = (\chi_i(t), \rho_i(t), \mathfrak{q}(t), \text{bulk laws, roots, residues, margins}).$$

Away from non-simple contacts,

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

This is a finite-dimensional Markov state only because every mode variable is retained. In a continuum power-law limit, $(\chi_i)_i$ becomes a measure-valued state and the dynamic-programming equation below is infinite-dimensional. A finite collection of spectral blocks is a numerical discretization of that state, not an exact memoryless reduction unless its convergence is proved.

The local spectral price is the negative logarithm of a linearized signal-to-bulk Rayleigh gain. Let $\nu_{B,t}$ be the active bulk singular-value law in the gradient channel, and let $R_t = \mathrm{d}\mu_{S,t}/\,\mathrm{d}\nu_{B,t}$ be the smoothed signal-to-bulk likelihood ratio. For $f_a(\sigma) = \sigma^a$, set

$$H_t(f_a) = \frac{\langle f_a, R_t\rangle_{\nu_{B,t}}^2}{\langle f_a, f_a\rangle_{\nu_{B,t}}}, \qquad F_t(a) = -\log H_t(f_a). \tag{14.2}$$

Equivalently,

$$F_t(a) = \log Z_{B,t}(a) - 2\log Z_{S,t}(a), \quad Z_{S,t}(a) = \langle \sigma^a, R_t\rangle_{\nu_{B,t}}, \quad Z_{B,t}(a) = \left\langle \sigma^{2a}, 1\right\rangle_{\nu_{B,t}}. \tag{14.3}$$

If a scalar step amplitude c is optimized locally, the quadratic AMP/SQ expansion has improvement

$$cZ_{S,t}(a) - \frac{c^2}{2} Z_{B,t}(a),$$

whose optimum is $Z_{S,t}(a)^2/(2Z_{B,t}(a))$. Minimizing $F_t(a)$ is therefore the myopic spectral-denoising decision.

Let $V(t, X)$ be the value function for a terminal objective penalizing final risk and negative BBP margins. The hard reduced HJB equation is

$$-\partial_t V(t, X) = \inf_{a\in[0,1]} \left\{F_t(a) + \nabla_X V(t, X) \cdot b(t, X, a)\right\}. \tag{14.4}$$

The future price of exponent a along the trajectory is

$$A_t(a) = \nabla_X V(t, X_t) \cdot b(t, X_t, a) - c_t,$$

where c_t is an arbitrary additive gauge. The entropy-regularized selector is

$$\pi_t(\,\mathrm{d}a) = \frac{1}{Z_t} \exp[-\beta_t(F_t(a) + A_t(a))]\,\mathrm{d}a, \qquad a_t = \int_0^1 a\,\pi_t(\,\mathrm{d}a). \tag{14.5}$$

This Gibbs law is the exact optimizer of the entropy-regularized reduced one-step control objective. It is not, without the preceding state reduction and the Bellman value function, a theorem of global optimality for empirical phase-retrieval training. The policy only sees the total energy

$$E_t(a) = F_t(a) + A_t(a).$$

The HJB definition of A_t contains only an additive, a-independent gauge through c_t. Algebraically one may rewrite the same total energy as

$$F_t(a) \mapsto F_t(a) + c_t + \kappa_t a, \qquad A_t(a) \mapsto A_t(a) - c_t - \kappa_t a.$$

This reallocation leaves the Gibbs policy unchanged, but it is only a bookkeeping identity: it does not make the physically defined Rayleigh cost (14.2) ambiguous, and it does not imply that the future action is affine. When the future action is smooth on the Gibbs support, its best affine approximation has slope

$$\lambda_t^{\mathrm{proj}} = \frac{\mathrm{Cov}_{\pi_t}(a, A_t(a))}{\mathrm{Var}_{\pi_t}(a)}. \tag{14.6}$$

This is the Bellman shadow-price interpretation of the empirical linear Boltzmann action. Let S_t carry asymptotically all the mass of the full and affine Gibbs laws. Its validity requires

$$\beta_t \operatorname{osc}_{S_t}\left(A_t(a) - c_t - \lambda_t^{\mathrm{proj}} a\right) = o(1), \tag{14.7}$$

so that the nonlinear residual is uniformly invisible on the selected support. If $r_t = A_t - c_t - \lambda_t^{\mathrm{proj}} a$, the total-variation distance between the two conditional Gibbs laws is at most

$$\frac{1}{2}\{\exp(\beta_t \operatorname{osc}_{S_t} r_t) - 1\}.$$

The RMS residual reported in finite-dimensional comparisons is not by itself sufficient; it needs a concentration or uniform-integrability bound that upgrades it to (14.7). Without (14.7), the full function $A_t(a)$ must be kept. Differentiating the Gibbs law gives the exact identity

$$\dot{a}_t = -\beta_t \operatorname{Cov}_{\pi_t}(a, \dot{E}_t(a)) - \dot{\beta}_t \operatorname{Cov}_{\pi_t}(a, E_t(a)). \tag{14.8}$$

For constant β_t, only the first covariance remains. In the zero-temperature interior limit,

$$\partial_a E_t(a_t) = 0, \qquad \dot{a}_t = -\frac{\partial_{ta} E_t(a_t) + \partial_a \nabla_X E_t(a_t) \cdot \dot{X}_t}{\partial_{aa} E_t(a_t)}.$$

If the active front has local exponent ζ_t and the near-zero phase-retrieval relation $\beta_t = \zeta_t$ holds, the adiabatic selector is

$$a_{\mathrm{PR},*}(t) = \operatorname{clip}_{[0,1]}\left(\frac{\zeta_t - 1}{3\zeta_t - 1}\right), \qquad \dot{a}_{\mathrm{PR},*}(t) = \frac{2\dot{\zeta}_t}{(3\zeta_t - 1)^2} \tag{14.9}$$

on the interior branch. More generally, for a local power-law front with residual exponent β_t, the raw Volterra timing boundary is

$$a_{\mathrm{raw},*}(t) = \operatorname{clip}_{[0,1]}\left(\frac{\beta_t - 1}{2\zeta_t + \beta_t - 1}\right). \tag{14.10}$$

On an interior branch,

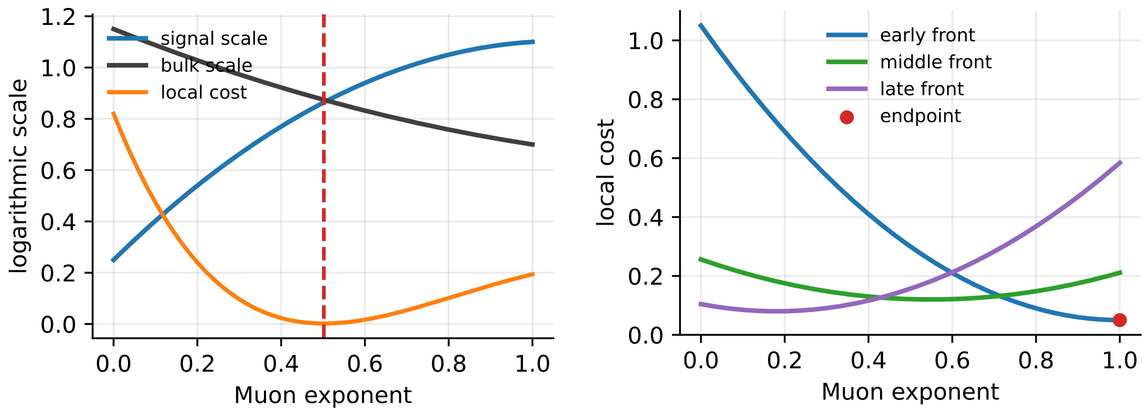
$$\dot{a}_{\mathrm{raw},*}(t) = -\frac{2(\beta_t - 1)}{(2\zeta_t + \beta_t - 1)^2}\dot{\zeta}_t + \frac{2\zeta_t}{(2\zeta_t + \beta_t - 1)^2}\dot{\beta}_t. \tag{14.11}$$

The phase-retrieval formula (14.9) is obtained by setting $\beta_t = \zeta_t$. The linear action that makes the Gibbs mean track this boundary is the unique $\lambda_{\mathrm{sat}}(t)$ solving

$$\int_0^1 a \, \frac{\exp[-\beta_t(F_t(a) + \lambda_{\mathrm{sat}}(t)a)]}{\int_0^1 \exp[-\beta_t(F_t(b) + \lambda_{\mathrm{sat}}(t)b)]\,\mathrm{d}b}\,\mathrm{d}a = a_{\mathrm{raw},*}(t), \tag{14.12}$$

because

$$\partial_\lambda \mathbb{E}_{\pi_t^\lambda}[a] = -\beta_t \operatorname{Var}_{\pi_t^\lambda}(a) < 0.$$



(a) Boltzmann matching.

(b) Endpoint effect.

Figure 7: Boltzmann selector. The local term compares signal and bulk logarithmic scales; the dynamic term prices future BBP margins.

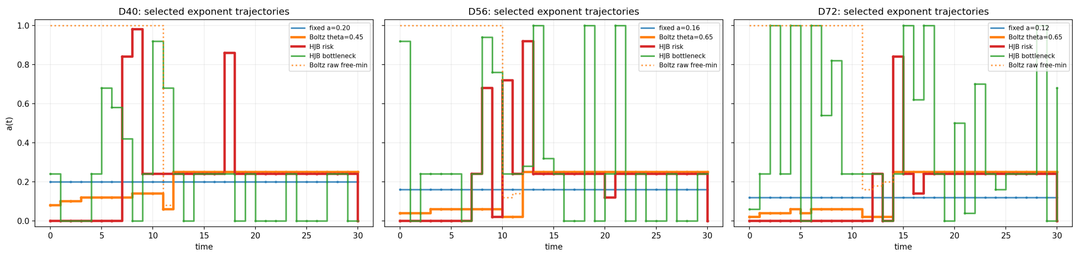


Figure 8: Online exponent trajectories. The smooth Boltzmann trajectory is the observable consequence of the Gibbs rule (14.5).

15 Fresh, reused, and leave-one-out comparison

The deterministic equations above are cleanest for a fresh spectral sample: the path is fixed first, and the matrices used to form resolvents are sampled afterward. Reusing the training sample introduces dependence between the path and the resolvent. The needed comparison is finite-dimensional.

Let $\mathcal{K}_d^{\mathrm{reuse}}(z,t)$ be the finite outlier matrix computed on the reused sample, and $\mathcal{K}_d^{\mathrm{fr}}(z,t)$ the corresponding fresh or leave-one-out matrix.

Assumption 15.1 (Fresh-to-reused outlier comparison)**.** *On compact time intervals and contours separated from exact non-transversal BBP tangencies,*

$$\sup_{t,z}\left(\left\|\mathcal{K}_d^{\mathrm{reuse}}(z,t)-\mathcal{K}_d^{\mathrm{fr}}(z,t)\right\|+\left\|\partial_z\mathcal{K}_d^{\mathrm{reuse}}(z,t)-\partial_z\mathcal{K}_d^{\mathrm{fr}}(z,t)\right\|\right)=o_{\mathbb{P}}(1). \tag{15.1}$$

Proposition 15.2 (Transfer to reused samples)**.** *Assume the fresh/frozen deterministic limit and Assumption 15.1. Then reused-data roots, residues, and BBP exit times have the same deterministic limits as the fresh/frozen readouts, away from non-transversal contacts.*

Proof. On a contour surrounding one simple root, the determinants of the fresh and reused finite equations differ by $o_{\mathbb{P}}(1)$ uniformly. Rouche's theorem gives the same number of roots in the contour. The derivative control gives the same residue limit by (9.5). A transversal margin crossing is stable under uniform perturbation, so the exit times converge. □

The standard proof strategy for Assumption 15.1 is leave-one-out. Freeze the finite frame $Z_t=[W_t,\Theta]$, decompose each sample into its projection on this frame plus an orthogonal Gaussian component, prove a local law for the orthogonal block uniformly on the contours, and compare W_t to $W_t^{(\ell)}$, the path trained without the ℓ-th sample. If the one-row influence on the averaged finite matrix is $o_{\mathbb{P}}(1)$, then (15.1) follows.

15.1 Label-free handoff topology

Near a dense front, individual eigenvalue labels can exchange. The stable object is therefore not a labeled curve but a finite partition of visible branches. Let $K_t(z)$ be one of the finite outlier matrices, with edge $x_+(t)$, and let

$$D_t(z)=\det(zI-K_t(z)).$$

For a simple root $\lambda_j(t)>x_+(t)$, define

$$m_j(t)=\lambda_j(t)-x_+(t),\qquad \Omega_j(t)=\frac{1}{1-\langle u_j,\partial_z K_t(\lambda_j)u_j\rangle}.$$

The resolved visible object is the multiset

$$\mathcal{P}_t = \{(\lambda_j(t), \Omega_j(t), v_j(t)) : \lambda_j(t) > x_+(t)\}, \tag{15.2}$$

where $v_j(t)$ is the teacher-mass vector carried by the branch. A permissible observable is any bounded Lipschitz symmetric function of $\mathcal{P}_t$. Symmetry is what removes the need to follow a branch label through a collision.

For a margin $\delta > 0$, define the handoff set

$$\begin{aligned} \mathcal{H}_\delta =& \{t : \min_j |m_j(t)| \le \delta\} \cup \{t : \min_{i \ne j} |\lambda_i(t) - \lambda_j(t)| \le \delta\} \\ & \cup \{t : \text{the active contour or KKT set changes within } \delta\}. \end{aligned} \tag{15.3}$$

Outside $\mathcal{H}_\delta$, roots and residues are C^1 functions of the Volterra state. Inside $\mathcal{H}_\delta$, one uses smoothed partition observables rather than hard labels.

Lemma 15.3 (Deterministic handoff window bound)**.** *Assume $K_t(z)$ and $x_+(t)$ are uniformly Lipschitz in t, all non-handoff roots are simple with*

$$1 - \langle u_j, \partial_z K_t(\lambda_j) u_j \rangle \ge c_0 > 0,$$

and each contact or collision is transversal with velocity at least v_0. If $N_{\text{hand}}(T)$ is the number of contacts and switches up to time T, then

$$|\mathcal{H}_\delta| \le C_T N_{\text{hand}}(T) \frac{\delta}{v_0}. \tag{15.4}$$

Moreover, for any bounded Lipschitz partition observable O,

$$\begin{aligned} \left| \int_0^T O(\mathcal{P}_t^{(d)})\, \mathrm{d}t - \int_0^T O(\mathcal{P}_t)\, \mathrm{d}t \right| \le & \, C_T \sup_{t \notin \mathcal{H}_\delta} d_{\text{part}}(\mathcal{P}_t^{(d)}, \mathcal{P}_t) \\ & + C_T |\mathcal{H}_\delta|. \end{aligned} \tag{15.5}$$

Proof. Each contact, root collision, or active-set switch contributes a time interval of length $O(\delta/v_0)$ by transversality. Summing over the $N_{\text{hand}}(T)$ events gives (15.4). On the complement, the observable is Lipschitz in the partition metric. Inside the handoff set, boundedness gives a contribution at most $C_T|\mathcal{H}_\delta|$. □

For a power-law front with $N_F(d)$ active crossings, a sufficient deterministic condition is

$$N_F(d)\delta_d / v_0 = o(1).$$

When this labelwise condition is too strong, the theorem target is the convergence of smoothed Schur point-process observables.

15.2 Same-sample transport budget

The full same-sample comparison is expressed in a finite test family. Let $\mathfrak{T}_{d,t}^{\text{obs}}$ contain the observables used in the proof:

$$\begin{aligned} \mathfrak{T}_{d,t}^{\text{obs}} = \{ & \text{partition, Stieltjes, and energy statistics;} \\ & t\text{- and } a\text{-derivatives;} \\ & \text{projected causal updates;} \\ & \dot{\mathfrak{q}}, \text{ Schur roots, and residues;} \\ & \text{corrected two-resolvent coefficients;} \\ & DR[\dot{H}] = -R\dot{H}R \}. \end{aligned}$$

Define

$$\varepsilon_{\mathrm{RFA,d}}(T) = \sup_{t\le T}\sup_{a\in I_t}\sup_{\psi\in\mathfrak{T}^{\mathrm{obs}}_{d,t}} |\psi^{\mathrm{same}}_{d,t,a} - \psi^{\mathrm{det}}_{t,a}|. \tag{15.6}$$

This is a finite Schur/RFA topology, not an operator-norm topology.

The dependency order is

$$\text{frozen block-MDE local law} \Rightarrow \text{one-row expansion} \Rightarrow \text{dynamic LOO recursion}$$
$$\Rightarrow \text{same-sample RFA} \Rightarrow \text{same-sample finite spectral readout.}$$

Let υ_d denote the imaginary resolution of the local law. Suppose the frozen law and one-row influence have rates

$$\varepsilon_{\mathrm{fr}}(d) = O_{\mathbb{P}}(d^{-\zeta_0}\upsilon_d^{-q}), \qquad B_{\mathrm{row}}(d,\upsilon_d) = O_{\mathbb{P}}(\upsilon_d^{-r}).$$

Collect the approximation terms in

$$\begin{aligned}\mathcal{E}_d &= d^{-1}\upsilon_d^{-r} + d^{-\zeta_0}\upsilon_d^{-q} + \varepsilon_{\mathrm{net}} + \varepsilon_{\mathrm{stat}} \\ &\quad + \sup_{t\le T}\frac{\upsilon_d}{\ell_d(t)} + \varepsilon_{\mathrm{tail}} + \varepsilon_{\mathrm{smooth}}.\end{aligned} \tag{15.7}$$

Here $\varepsilon_{\mathrm{net}}$ is the time/control discretization error, $\varepsilon_{\mathrm{stat}}$ is the projected empirical-statistics error, $\ell_d(t)$ is the contact-smoothing scale, and the last two terms control the omitted spectral tail and the use of smoothed branch observables.

Proposition 15.4 (Transport bound)**.** *Let $e_{n,d}$ be the maximum same-to-deterministic discrepancy over $\mathfrak{T}^{\mathrm{obs}}_{d,t_n}$. Suppose the dynamic leave-one-out expansion gives*

$$e_{n+1,d} \le (1+\Delta t_n L_{n,d})e_{n,d} + \rho_{n,d}, \qquad e_{0,d} = o_{\mathbb{P}}(1), \tag{15.8}$$

uniformly over the admissible powers, with

$$\sum_{n:t_n\le T}\rho_{n,d} \le C\mathcal{E}_d + o_{\mathbb{P}}(1).$$

If

$$\Gamma_d(T) = \exp\left(\sum_{n:t_n\le T}\Delta t_n L_{n,d}\right) \le d^{\chi},$$

then

$$\varepsilon_{\mathrm{RFA,d}}(T) \le d^{\chi}\{e_{0,d} + C\mathcal{E}_d + o_{\mathbb{P}}(1)\}. \tag{15.9}$$

Proof. Iterating (15.8) and bounding every product $\prod_{j=m}^{n}(1+\Delta t_j L_{j,d})$ by $\Gamma_d(T)$ gives

$$\max_{n:t_n\le T} e_{n,d} \le \Gamma_d(T)\left(e_{0,d} + \sum_{n:t_n\le T}\rho_{n,d}\right).$$

Insert the assumed remainder bound and the definition (15.6). □

For $\upsilon_d = d^{-\sigma}$, a sufficient polynomial rate condition is

$$1-\chi-r\sigma > 0, \qquad \zeta_0 - q\sigma - \chi > 0, \tag{15.10}$$

together with the vanishing of the smoothed contact and tail terms after multiplication by d^{χ}. Proposition 15.4 is deterministic. The model-specific work is to prove the frozen local law, derive (15.8), and bound the terms in (15.7).

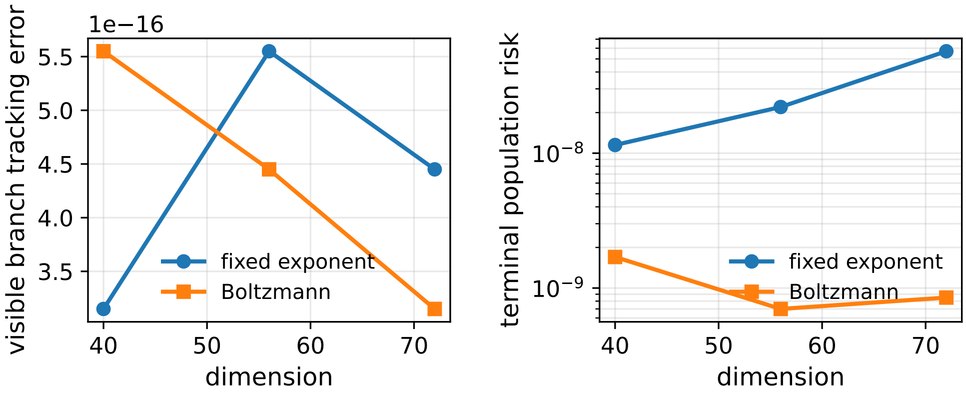


Figure 9: Fresh/frozen comparison. The plotted quantities are exactly the finite objects controlled by (15.1): roots, residues, and branch labels.

16 Logical status of the results

The population loss, the gradient formula, the finite-frame functional calculus, and the equations for Q, M, and C are exact finite-rank identities. No high-dimensional limit is used in their proofs. Once the coefficients $\delta_i^{(a)}$ and $\nu_i^{(a)}$ are given, the mode equations and the nonlinear Volterra representation are also exact deterministic consequences.

The use of those coefficients for phase retrieval is conditional on Assumption 5.1. The power-law value $a_{\mathrm{raw},*}$ is then the minimizer of the two asymptotic exponents in (13.1); it is not claimed to minimize every finite-time or finite-dimensional loss. Likewise, the three BBP readouts are consequences of the bulk-resolvent and finite-Schur convergence hypotheses in Theorem 13.1. The Schur identities themselves are exact; their deterministic limits are the random-matrix input.

For a moving exponent, the Gibbs policy (14.5) is exact for the stated entropy-regularized reduced control problem. The affine action is only valid under (14.7), and the AMP/VAMP calculation identifies a local linearized target filter rather than an algorithmic equivalence. In the continuum limit the HJB state is functional unless a convergent block discretization is supplied.

Finally, reused training data require Assumption 15.1. A complete proof must control projected resolvents and their derivatives through the regularized hard edge and through smoothed dense BBP contacts. The finite-dimensional comparisons shown in this paper test the quantities in that assumption, but they do not establish its uniform local law or strong almost-sure convergence of the full empirical trajectory.

References

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

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

[3] F. Benaych-Georges and R. Nadakuditi, *The singular values and vectors of low rank perturbations of large rectangular random matrices*, Journal of Multivariate Analysis, 2012.

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

[5] 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.

[6] S. Rangan, P. Schniter and A. Fletcher, *Vector approximate message passing*, IEEE Transactions on Information Theory, 2019.

[7] J. Ma and L. Ping, *Orthogonal AMP*, IEEE Access, 2017.