Muon-a Phase Retrieval at Constant Exponent: Volterra Learning Curves and Three Spectral BBP Tests

Anonymous

July 3, 2026

Abstract

This companion records the part of the Muon phase-retrieval theory in which the exponent a is fixed. Paquette's projected-risk recursion gives a deterministic Volterra equation. The same Volterra state then predicts the gradient singular spectrum, the spectrum of the weights, and the empirical Hessian spectrum. The figures compare these deterministic predictions with the observed spectral branches. The variable-a and Boltzmann/HJB control layer is kept separate.

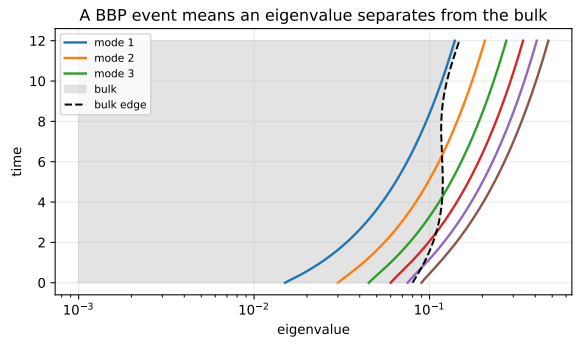
Relation to the main paper. This document is the fixed-exponent companion to *Muon as Dynamic Spectral Denoising for Power-Law Phase Retrieval*. It keeps the constant-a derivations and spectral checks in one place. The Volterra recursion is the Paquette Muon recursion [1] specialized to phase retrieval; the power-law phase-retrieval motivation comes from Braun–Loureiro–Minh–Imaizumi [3]; the Hessian readout uses the finite-summary effective spectral theory of Ben Arous–Gheissari–Huang–Jagannath [2]; and the outlier residue calculus is the standard finite-rank spectral calculus [4].

How to read it. Section 1 defines the state and gives the dictionary with Paquette's notation. Section 2 derives the fixed-a learning curve. Sections 3–5 then read the same state through three spectra: gradient, weights and Hessian. The last section states exactly what is closed and what still requires a same-sample RMT theorem.

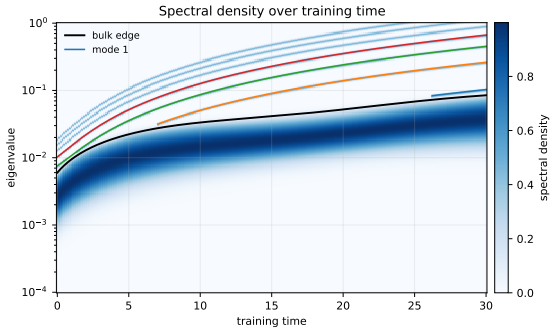
Summary and outline. The companion has one organizing principle:

$$\text{one fixed exponent } a \longrightarrow \text{one Volterra state} \longrightarrow \text{three spectral tests.}$$

The first arrow is Paquette's reduced recursion after inserting the Muon power filter. The second arrow is deterministic spectral calculus: the same state determines the gradient bulk, the weight Schur roots, and the Hessian Dyson–Schur contacts. This is why the figures are not independent experiments. They are three views of the same reduced trajectory. Agreement means that the theory predicts not only the loss curve, but also when and where the informative spectral branches become visible.



(a) A teacher direction is useful only after its outlier has separated from the random bulk. The fixed exponent a changes how fast the moving front reaches that condition.



(b) The same event is read dynamically: the bulk edge moves during training, and isolated branches mark visible modes.

Figure 1: The constant-a story before the formulas. The Volterra equation predicts the moving state, while Schur/BBP equations decide which modes are visible in the gradient, weight and Hessian spectra.

Order parameters and phase boundaries. For a fixed exponent the numerical phase diagram is read through the following observables. The learning-curve order parameters are $q_i^2(t)$, the part of teacher mode i captured by the student, and $\rho_i(t)$, the remaining residual mass. The spectral order parameters are the bulk edges and the Schur roots in the gradient, weight and Hessian spectra. A BBP time is the first time at which a root is outside the deterministic bulk edge with a positive residue. This is the fixed-a analogue of the standard rank-one picture: below the boundary the top eigenvalue follows the bulk edge, while above the boundary it follows the isolated-root prediction.

Object	Meaning in this companion	Why it matters
$q_i^2(t)$	captured mass of teacher mode i	gives the learned part of the mode in the Volterra state
$\rho_i(t)$	residual mass of teacher mode i	controls how much signal remains to be exposed by BBP branches
$Q(t)$	scalar risk or aggregate energy scale	sets the raw Muon clock and the Volterra nonlinearity
$\delta_i^{(a)}$	deterministic drift coefficient after the Muon filter	gives the signal part of the mode equation
$\nu_i^{(a)}$	volatility coefficient after the Muon filter	gives the fluctuation floor in the Volterra kernel
Schur root	finite outlier root after the bulk has been eliminated	predicts branch location and teacher overlap
Dyson edge	deterministic edge of a spectral bulk	decides when a Schur root is genuinely visible

Table 1: The fixed-a objects in plain language. The companion uses one state $X_t = (q_i^2, \rho_i, Q)$ and reads it through three spectra.

The fixed-a trainability diagram

For ordinary gradient descent in a rank-one linear model, the natural trainability diagram has axes such as learning rate and initial variance. In this companion the axes are different. The

exponent a is fixed, and the moving coordinate is the active power-law front: the index scale of teacher modes that are currently close to becoming visible. A fixed exponent is therefore a horizontal cut through a spectral trainability diagram.

There are three regimes along such a cut.

Regime	Spectral description	Observable signature
Hidden front	the mode has residual mass, but its Schur root is still inside the bulk	risk improves slowly; no stable outlier residue is visible
Visible BBP front	the Schur root has crossed the deterministic bulk edge	an outlier branch appears and carries nonzero teacher residue
Terminal floor	the signal term and volatility term in the Volterra law balance	risk decay slows and finite-dimensional calibration becomes most delicate

The first boundary is a BBP visibility boundary. In a spectral channel $c \in \{G, W, H\}$, write

$$M_i^c(t, a) = \lambda_i^c(t, a) - x_+^c(t, a), \tag{1}$$

with the left-edge sign convention for left Hessian branches. The equation $M_i^c(t, a) = 0$ is the moment at which mode i becomes visible in that channel. The same mode can have different visibility times in gradient, weight and Hessian spectra.

The second boundary is the Volterra tail-balance boundary. It does not say when one branch detaches. It says which constant exponent gives the best asymptotic compromise between hard-edge amplification and tail resolution. If

$$\mu_i \asymp i^{-\alpha}, \qquad q_i^2(0) \asymp i^{-\beta},$$

then the two powers are

$$\frac{1-a}{2}, \qquad \frac{\alpha(1+a)}{2(\alpha+\beta-1)}.$$

Their equality gives

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

Thus the constant-a optimum is a broad trainability region, not a single fragile decimal. The right validation is joint agreement of learning curves, bulk edges, Schur roots, residues and BBP exit times.

1 Model and Paquette dictionary

We work with the quadratic teacher

$$f_\star(x) = \sum_{i=1}^{k} \mu_i (\theta_i^\top x)^2$$

and a student matrix W. The important state variables are the captured mode masses $q_i^2(t)$, the residual masses $\rho_i(t)$, and the total risk

$$Q(t) = \frac{1}{2} \sum_i \mu_i q_i^2(t).$$

For a fixed Muon exponent a the spectral filter is

$$\phi_{a,\eta}(\sigma) = \sigma(\sigma^2 + \eta^2)^{(a-1)/2}.$$

In Paquette's notation this is the general spectral method

$$\widetilde{G} = G\,\varphi(G^\top G), \qquad \varphi_{a,\eta}(s) = (s+\eta^2)^{(a-1)/2}.$$

The exact fresh-minibatch projected recursion in Paquette's framework is

$$\mathbb{E}[q_{ij}(t+1) \mid \mathcal{F}_t] = q_{ij}(t) - 2\eta_t \mathcal{D}_{ij}(t) + \eta_t^2 \mathcal{V}_{ij}(t). \tag{2}$$

The two coefficients are the corresponding drift and volatility terms:

$$\mathcal{D}_{ij} = (u_i^\top \Delta_t v_j)\,\mathbb{E}[u_i^\top \widetilde{G}^\circ v_j \mid \mathcal{F}_t], \qquad \mathcal{V}_{ij} = \mathbb{E}[(u_i^\top \widetilde{G}^\circ v_j)^2 \mid \mathcal{F}_t].$$

Paquette then removes the changing risk scale by a deterministic rescaling. The drift and volatility are then written as contour integrals. Substituting $\varphi_{a,\eta}$ in those two integrals is the only change needed for fixed Muon-a.

Paquette object	location in the Paquette derivation	used here as
projected recursion	projected-risk recursion before the continuous limit	mode dynamics
drift/volatility definitions	Appendix definitions of D_{ij} and V_{ij}	$\delta_i^{(a)}, \nu_i^{(a)}$
risk rescaling	deterministic rescaling of the residual diagonal	remove $Q(t)$ from RMT core
one-resolvent fixed point	Step 4 fixed-point closure, including the Stieltjes transform $m(z)$	gradient bulk Stieltjes law
two-resolvent volatility	Step 8 variance kernel and P_3 two-resolvent term	noise floor and kernel
Volterra clock and kernel	Volterra appendix: clock ϕ, forcing F, kernel K	learning curve
power-law time bounds	Volterra appendix: sandwich estimate and power-law Laplace bounds	terminal scaling

Table 2: Line-by-line translation from Paquette to fixed Muon-a. The middle column names the exact part of the Paquette derivation being imported. The only specialization is the insertion of $\varphi_{a,\eta}$ into Paquette's general spectral filter. The imported steps are the projected-risk recursion, the one-resolvent fixed point for the Stieltjes transform, the two-resolvent variance kernel, and the Volterra clock/kernel construction.

The words *fresh* and *frozen* have a simple meaning in this companion. The trajectory W_t, or equivalently the reduced state X_t, is regarded as already fixed when the random matrix or resolvent is evaluated. Equivalently, the data used to measure the spectrum is independent of the data that produced the state. This is the natural setting of the Paquette calculation. The real online experiment can reuse the same data; transferring the formulas to that setting is exactly the Schur/RFA transport statement at the end of the theorem.

Theorem 1.1 (Fixed-exponent spectral closure)**.** *Fix $a \in [0,1]$ and assume the Paquette fresh/frozen spectral coefficients for the filter $\varphi_{a,\eta}$ exist uniformly on the regular part of the trajectory. Then the reduced state*

$$X_t = (q_i^2(t), \rho_i(t), Q(t))$$

determines four objects:

Volterra	*learning curve,*
Gradient	*bulk and finite atoms,*
Weights	*finite Schur roots,*
Hessian	*Dyson–Schur branches.*

If the corresponding Schur roots are simple and separated from the bulk edge, their locations, residues and exit times are stable deterministic functions of X_t. Passing from the fresh/frozen construction to the exact reused training sample requires only the Schur/RFA estimate

$$\varepsilon_{\mathrm{RFA,d}} = o_{\mathbb{P}}(1).$$

Proof. The Paquette projected-risk recursion gives a linear equation for each mode after the Muon filter is inserted. Summing these equations gives the Volterra law. The gradient, weight and Hessian spectra are then obtained by applying the appropriate resolvent or Schur complement to the same reduced state. Simple roots are stable under uniform perturbations of the Schur kernel, and residues are obtained by differentiating the same finite resolvent. The RFA estimate is exactly the statement that the reused-data Schur matrices have the same limit as the fresh/frozen ones. □

2 Learning curve

From (2), after passing to the continuous clock, the scalar constant-a reduction is

$$\dot{q}_i^2(t) = -2\eta\, \delta_i^{(a)} Q(t)^{(a-1)/2} q_i^2(t) + \eta^2 \nu_i^{(a)} Q(t)^a. \tag{3}$$

The raw Muon clock is

$$\mathrm{d}\tau = \eta Q(t)^{(a-1)/2} \mathrm{d}t. \tag{4}$$

Solving (3) mode by mode and summing gives the Volterra equation

$$Q(\tau) = F_a(\tau) + \eta \int_0^\tau K_a(\tau - u) Q(u)^{(a+1)/2} \mathrm{d}u, \tag{5}$$

which is the Muon-a version of Paquette's Volterra equation. The forcing and kernel are the same Paquette objects listed in Table 2, with δ_i, ν_i replaced by $\delta_i^{(a)}, \nu_i^{(a)}$.

For a power-law front

$$\mu_i \asymp i^{-\alpha}, \qquad q_i^2(0) \asymp i^{-\beta}, \qquad \alpha + \beta > 1,$$

Laplace estimates in Paquette Appendix H give the two competing powers

$$e_{\mathrm{raw}}(a) = \max\left\{ \frac{1-a}{2}, \frac{\alpha(1+a)}{2(\alpha+\beta-1)} \right\}. \tag{6}$$

Balancing them gives

$$a_{\mathrm{raw},*}(\alpha,\beta) = \left[\frac{\beta - 1}{2\alpha + \beta - 1} \right]_{[0,1]}. \tag{7}$$

For phase retrieval near the origin we use $\beta = \alpha$, hence

$$a_{\mathrm{raw},*}^{PR}(\alpha) = \frac{\alpha - 1}{3\alpha - 1}.$$

This is an optimization of the deterministic Volterra law, not an empirical calibration constant.

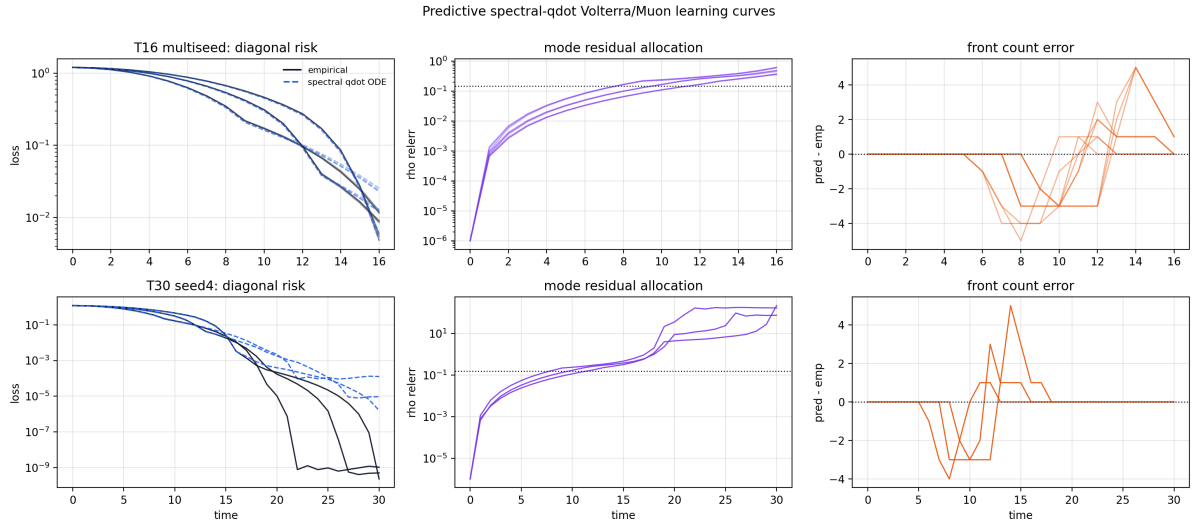


Figure 2: Learning curve check. The Paquette/Volterra q-dot prediction tracks the empirical risk well on the early and middle windows; the terminal window is more sensitive to finite dimension and discretisation.

3 Gradient spectrum

The gradient channel is a bulk plus a finite signal part,

$$G_t^{\mathrm{cav}} = B_t + S_t.$$

The bulk is read from the hermitized resolvent. Paquette Step 4 gives the closed one-resolvent fixed point and its Stieltjes transform. If $x = \sigma^2$, then

$$\rho_B^{\sigma}(\sigma) = \frac{2\sigma}{\pi} \Im m_B(\sigma^2 + i0). \tag{8}$$

The log-slope used for the AMP/Muon comparison is

$$\alpha_B^{\mathrm{Paq}} = 2 + 2x \, \frac{\Im m_B'(x + i0)}{\Im m_B(x + i0)}. \tag{9}$$

The finite gradient atoms solve a Schur, or equivalently a D-transform, equation

$$1 = \theta_i(t)^2 D_t^G(\lambda_i^G(t)), \qquad M_i^G(t) = \theta_i(t)^2 D_t^G(x_{+,G}(t)) - 1. \tag{10}$$

The transition is $M_i^G(t) = 0$.

The word AMP is used here in this precise sense. We do not use vanilla AMP for an arbitrary non-i.i.d. spectrum. The primitive statement is spectral matched filtering. If ν_B is the bulk spectral law and μ_S the infinitesimal signal/front measure, any spectral update f has local Rayleigh quotient

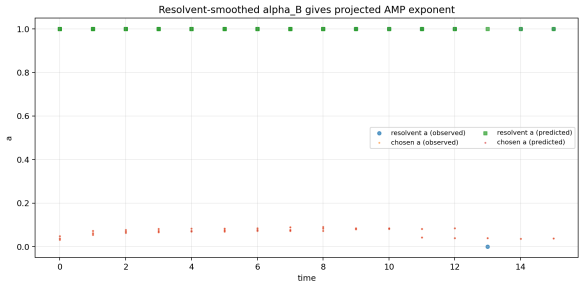
$$\mathcal{H}_t(f) = \frac{\langle f, R_t \rangle_{\nu_B}^2}{\langle f^2 \rangle_{\nu_B}}, \qquad R_t = \frac{\mathrm{d}\mu_{S,t}}{\mathrm{d}\nu_{B,t}}. \tag{11}$$

By Cauchy–Schwarz, $f_{\mathrm{opt}} \propto R_t$. This is the linearized AMP/VAMP spectral denoiser. The general-spectrum justification is the orthogonally-invariant AMP line of work, especially Rangan–Schniter–Fletcher, VAMP, Ma–Ping, OAMP, and Takeuchi, EP/OAMP; the associated spiked spectral calculus is the Benaych-Georges–Nadakuditi and OptShrink picture. For multi-index models, the closest reference is Defilippis–Dandi–Mergny–Krzakala–Loureiro, where the linearized message-passing transition is spectral.

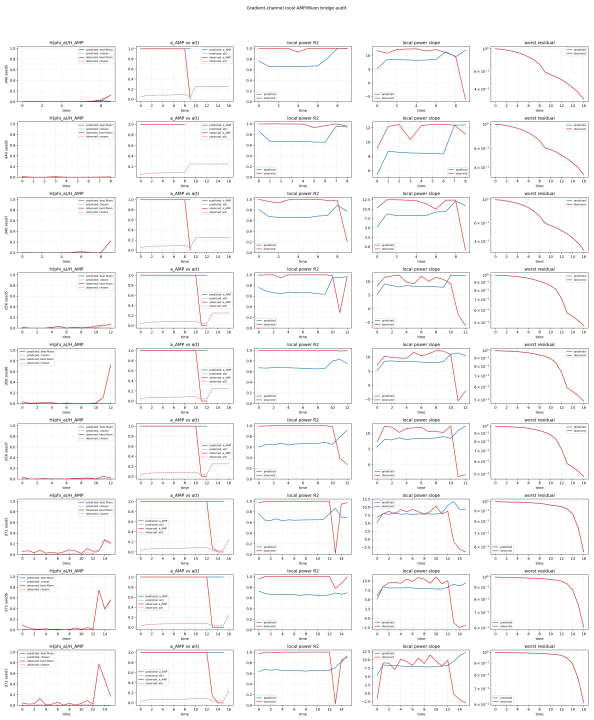
Muon-a is therefore a projection of f_{opt}, not the whole AMP algorithm. If, on the active front,

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

then $R_t(\sigma) \simeq C_t \sigma^{a_{\mathrm{loc}}}$, and the optimal local spectral filter is a power. This is the power-law situation. If the active front has several slopes or oscillations, a single scalar a is only a one-dimensional projection; the correct extension is blockwise Muon or the full spectral filter R_t.



(a) Bulk side. The effective exponent is computed from the Stieltjes derivative, so the comparison uses a resolvent observable rather than a histogram slope.



(b) Signal side. The moving teacher front is projected onto the one-parameter Muon family, giving the local AMP/Muon exponent comparison.

Figure 3: Gradient spectrum checks. The bulk side is not estimated by a histogram slope; it is the Stieltjes derivative in (9).

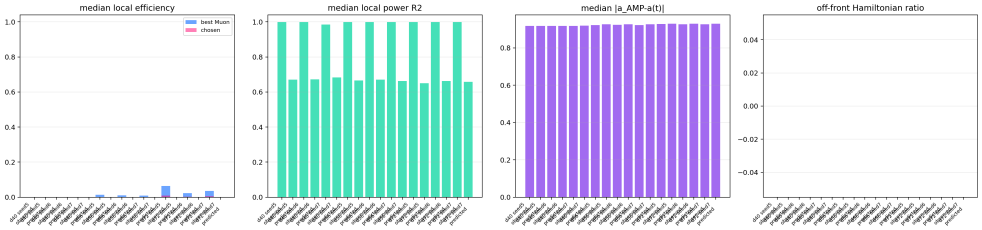


Figure 4: Summary of the gradient AMP/Muon bridge. The relevant comparison is local and spectral: Muon-a is the one-parameter projection of the AMP likelihood-ratio filter.

4 Weight spectrum

The weight spectrum is the most direct finite-dimensional test. The theorem object is not a raw visible label. One eliminates the bulk coordinates and solves the finite Schur equation

$$\det(I - K_t^W(\lambda)) = 0. \tag{12}$$

The right edge of the finite bulk is $x_{+,W}(t)$, and the exit time is

$$t_i^W = \inf\{t : \lambda_i^W(t) > x_{+,W}(t)\}. \tag{13}$$

The residue of the same Schur resolvent predicts teacher mass and overlap. This is why branch switching in a visible eigenvalue tracker does not contradict the theory: the stable object is the matched Schur branch.

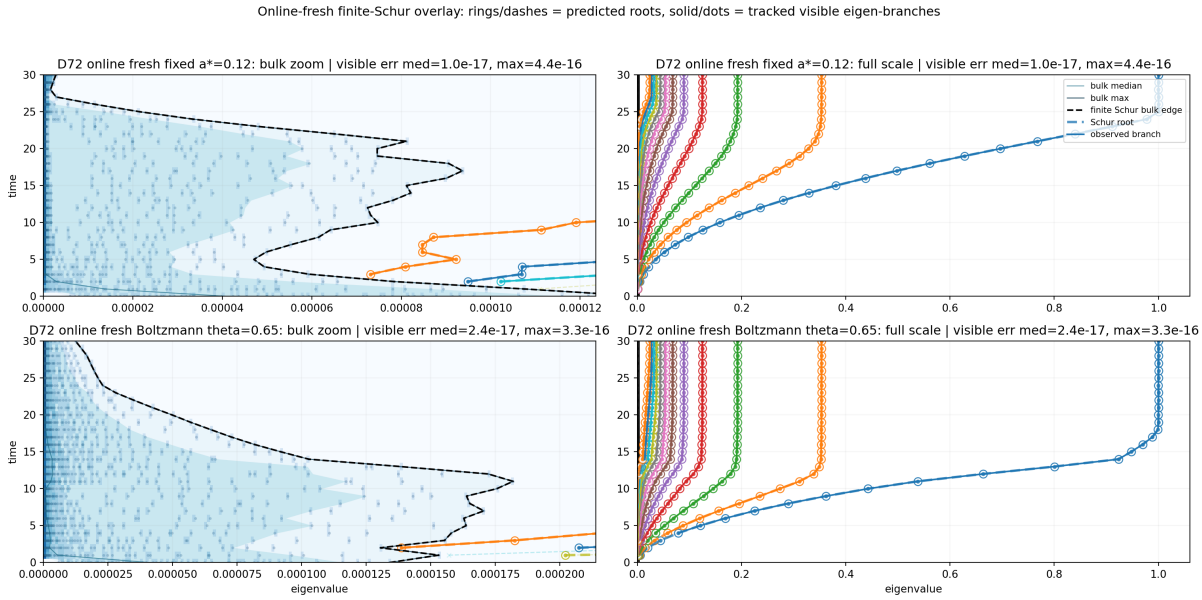
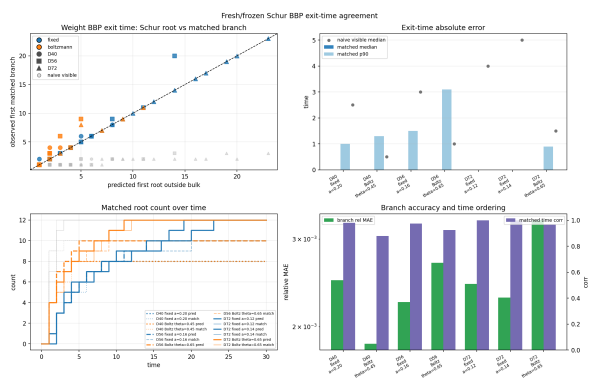
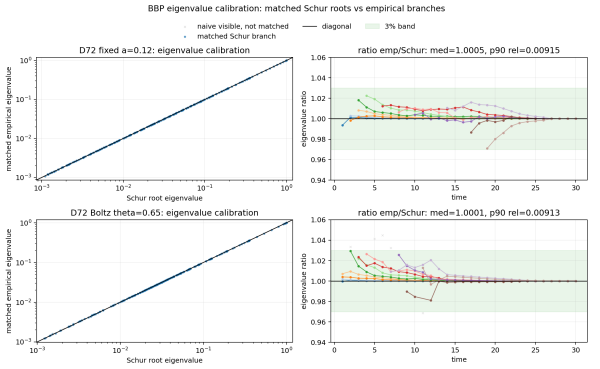


Figure 5: Weight spectrum over time for $d = 72$. The black curve is the finite bulk edge, colored curves are Schur roots, and stars mark matched empirical branches. This is the direct fixed-a BBP readout.



(a) Timing check. Schur-predicted branch exits are compared with matched empirical visible exits.



(b) Location check. Eigenvalue calibration is displayed separately because finite-size horizontal errors are more delicate than exit-time errors.

Figure 6: Weight BBP evidence. Across 72 branch rows all final Schur roots are matched; median matched exit-time error is 0, p90 is one checkpoint, and the p90 relative eigenvalue error is about $9.1 \cdot 10^{-3}$.

5 Hessian spectrum

For a fresh Hessian sample the bulk is the Dyson equation

$$-S_t^H(z)^{-1} = zI - \mathbb{E}\Big[A_t^H(I + \alpha_H^{-1} S_t^H(z) A_t^H)^{-1}\Big]. \tag{14}$$

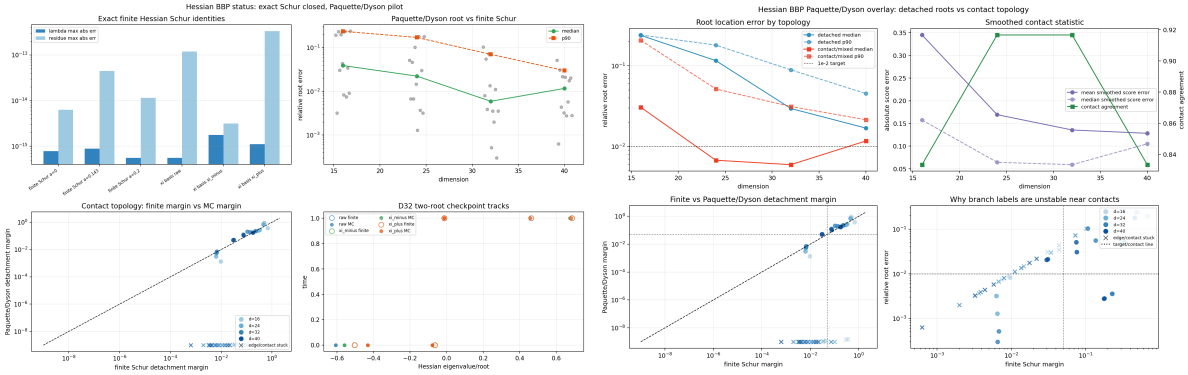
Finite left and right branches solve

$$\det(I - K_t^{H,-}(\lambda)) = 0, \qquad \det(I - K_t^{H,+}(\lambda)) = 0. \tag{15}$$

In the scalar large-α_H window,

$$\mu_i(1 - r_i(t_{i,\pm})) = \frac{c_i^{\pm}}{\sqrt{\alpha_H}} + O(\alpha_H^{-1}). \tag{16}$$

The constants $c_i^{\pm}$ are not universal; they are obtained by evaluating the edge equation (14) and the finite kernels (15) at the current Volterra state.



(a) Schur/Dyson layer. Finite roots are computed from the measured state and compared with the deterministic Hessian bulk edge.

(b) Contact layer. Detached branches are separated from near-edge contact windows, where dense BBP handoff is read through smoothed margins.

Figure 7: Hessian evidence. The finite Schur algebra is exact at a fixed state, and its numerical evaluation resolves detached branches. Detached Paquette/Dyson roots are predictive; dense contacts are read as smoothed contact statistics, not persistent labels.

6 One state, three spectra

The constant-a claim can be written as a deterministic map

$$X_t = (q_i^2(t), \rho_i(t), Q(t), m_{B,t}^G, K_t^W, S_t^H, K_t^{H,\pm}) \longmapsto (G_t, W_t, H_t).$$

The gradient channel gives the denoising direction, the weights show the most direct Schur BBP exits, and the Hessian shows residual/parallel left-right contacts. These are three views of the same Volterra state.

7 Status

The fixed-a deterministic layer is closed in the following sense:

Paquette fresh/frozen Volterra ⇒ gradient, weight and Hessian BBP predictions.

The strongest experimental layer is the weight Schur spectrum. The gradient bulk agrees at the Stieltjes-derivative level. The Hessian finite Schur algebra is exact, while the asymptotic Dyson branch test is most delicate near contacts.

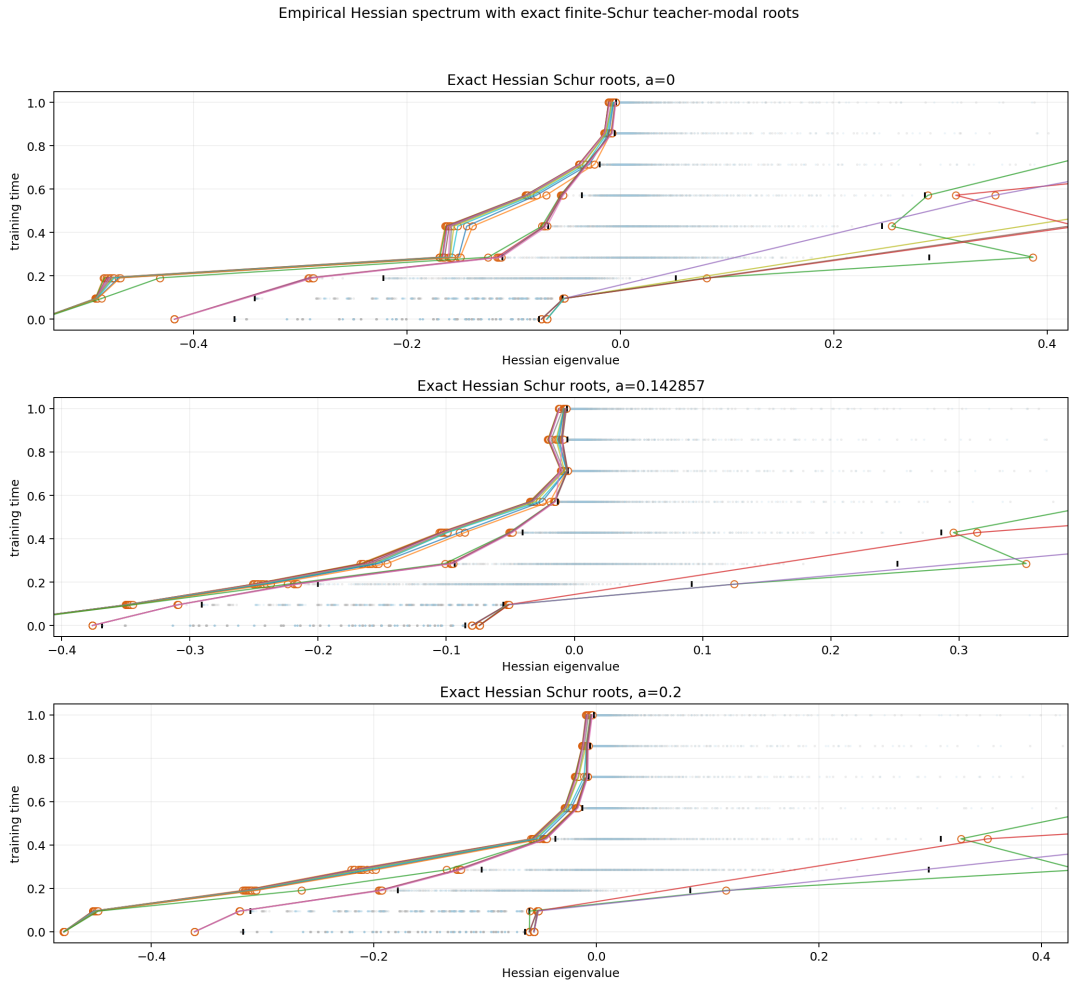


Figure 8: Exact finite Hessian Schur roots over time. The plot is the Hessian analogue of the weight Schur overlay: time is vertical, eigenvalue is horizontal, and visible branches are matched through Schur roots rather than raw rank labels.

Layer	What is established here	How to read the evidence
Volterra state	fixed-a Paquette recursion with the Muon power filter	predicts the risk clock and the mode masses
Gradient spectrum	resolvent slope and AMP/Muon projection from the same state	checks the predicted bulk and signal-front scale
Weight spectrum	finite Schur roots and residues are computed at the measured state	strongest finite check: branch identity and exit time are matched directly
Hessian spectrum	Dyson edge and finite left/right Schur roots are evaluated together	stable away from dense contacts; contacts use smoothed margins
Same-sample transport	separate RFA input	requires the Schur/RFA local law stated in the main paper

The reused-data local law is a separate theorem. To transfer every fresh/frozen statement to the exact same-sample online trajectory one needs

$$\varepsilon_{\mathrm{RFA,d}} = o_{\mathbb{P}}(1)$$

in the finite Schur/RFA topology. This is a separate RMT theorem; it is not part of the deterministic constant-a optimization calculation.

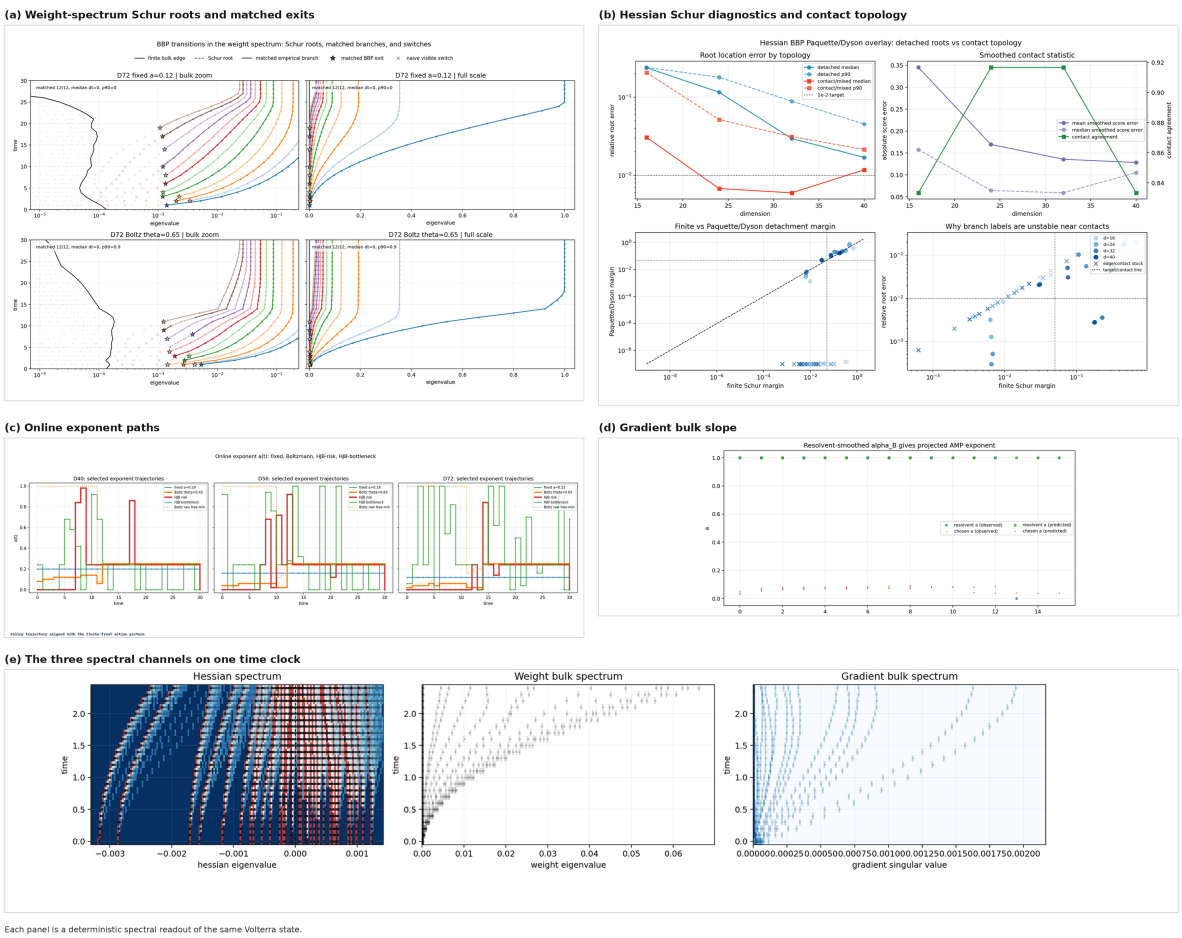


Figure 9: Compact transition board: gradient, weight and Hessian evidence are read together.

8 Conclusion

For a fixed Muon exponent, the story is complete at the deterministic level: one Paquette/Volterra state predicts one learning curve and three spectral readouts. The gradient spectrum measures the local denoising geometry, the weight spectrum gives the cleanest finite Schur branch test, and the Hessian spectrum checks whether the same residual directions are visible in local curvature. Thus constant-a Muon is not only a different learning-rate schedule. It is a fixed spectral preconditioner whose effect can be read simultaneously in risk, weights, gradients, and Hessians.

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] 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.

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

Figure or spectral check	Equation tested	Interpretation
Learning curve panel	Volterra equation (5)	The Paquette drift and volatility give the correct reduced clock for fixed a.
Gradient spectrum panels	Stieltjes slope (9) and atom condition (10)	The gradient channel is read as a spectral denoising problem.
Weight Schur overlay	finite Schur equation (12)	Branch identity and exit time are matched through Schur roots, not raw rank labels.
Hessian panels	Dyson–Schur system (14)–(15)	Curvature sees residual and parallel branches generated by the same Volterra state.
Three-spectrum board	deterministic map $X_t \mapsto (G_t, W_t, H_t)$	The spectra are three readouts of one fixed-a state, not three unrelated experiments.

Table 3: How to read the fixed-a figures. The target is agreement of learning curves, branch identities, residues and BBP times.