Reader's Guide to the Muon Phase-Retrieval Manuscripts

Anonymous

July 3, 2026

Abstract

This guide explains how to read the Muon phase-retrieval manuscript bundle. The bundle has one main paper and three technical companions. The main paper is meant to be read first. The companions separate the constant-exponent spectral theory, the variable-exponent Boltzmann/HJB control layer, and the long Markovian-control derivation.

Pedagogical Contract

The bundle is written in the same order as the phenomenon itself. First comes the spectral picture: a bulk, a moving edge, and outliers. Then come the observables that can be measured on a run: captured masses, residual masses, Schur roots, residues, and BBP exit times. Only after that do the formulas enter: Paquette's Volterra equation moves the state, Schur/Dyson equations read the spectra, and HJB or Boltzmann chooses the exponent.

This order is deliberate. A reader should never have to decode a symbol before knowing what it is meant to explain. Every important equation is attached to a visible object: a learning curve, a bulk edge, an outlier branch, an overlap, or a trajectory of $a(t)$.

1 The Core Message

The problem is quadratic multi-index phase retrieval with a power-law teacher spectrum. Weak teacher modes are present from the beginning, but they are initially hidden inside a random spectral bulk. Learning becomes visible when the corresponding directions detach as outliers in one of three spectra: gradient, weights, and empirical Hessian.

Muon acts on the singular values of the gradient. Its exponent $a \in [0, 1]$ interpolates between ordinary gradient descent, $a = 1$, and the sign-SVD limit, $a = 0$. The thesis of the bundle is:

$$\boxed{\text{Muon is a dynamic spectral denoiser.}}$$

For fixed a, Paquette's Volterra equation predicts the learning curve and the spectral BBP exits. For variable $a(t)$, AMP gives the instantaneous spectral denoising objective, while HJB adds the future cost of changing today's exponent. Boltzmann is the entropy-regularized reduced control rule.

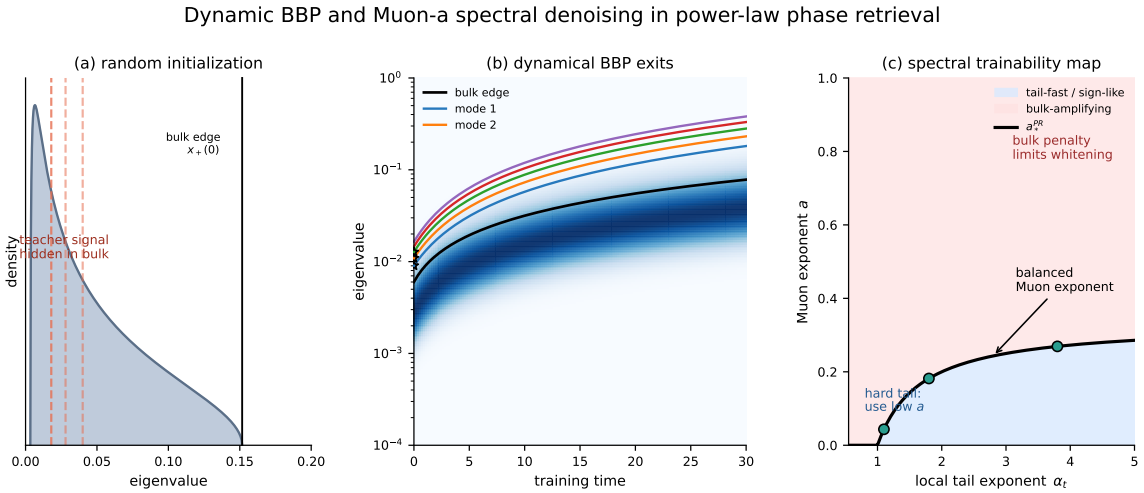

Figure 1: The mental picture for the whole bundle. A random bulk hides the teacher directions at initialization; training creates dynamical BBP exits; Muon changes the singular values of the gradient so that the power-law tail becomes visible at a better scale. The rest of the manuscript turns this picture into Volterra equations, Schur/BBP tests, and a control rule for the exponent a.

2 The Three Questions

Every document in the bundle answers the same three questions at a different level of detail.

Question	Mathematical object	Reader check
What is the state?	finite-frame variables Q, M, q_i, ρ_i and the Volterra scale $Q(t)$	whether the learning curve and mode masses are predicted by the same state.
What becomes visible?	Schur roots, Dyson edges, BBP residues and overlap formulas	whether spectral branches exit the bulk at the predicted times and carry the predicted teacher mass.
Which exponent should be used?	fixed a, local AMP projection, or Boltzmann/HJB policy over a	whether the selected exponent improves the future BBP margins, not only the instantaneous loss.

The order matters. First identify the state. Then read the spectra from the state. Only after that ask whether a fixed or variable exponent is optimal.

Figure-first reading. The fastest route through the bundle is to start from the spectral figures, not from the longest derivations. A figure should answer four questions: where is the bulk edge, where is the predicted Schur root, when does the branch detach, and how much teacher overlap does the branch carry? Once these four quantities are clear, the equations are mostly bookkeeping: Paquette gives the moving state, Schur gives the outlier equation, and the residue gives the overlap.

Parallel with the trainability paper. The intended reading order mirrors the phase-transition presentation in the reference trainability manuscript. There, the opening figure shows a Wishart bulk, a dynamical BBP detachment, and a phase diagram in learning-rate coordinates. Here the same three roles are played by the random spectral bulk at initialization, the three BBP

readouts during training, and the Muon trainability map in (a, α_t) coordinates. The dictionary is:

Reference paper	Muon bundle	Check to perform
initial Wishart or MP bulk	initial gradient, weight and Hessian bulks	compare empirical bulk edges with the deterministic edges
dynamical BBP detachment	Schur roots crossing the moving edge	compare predicted and observed exit times
ferromagnetic trainable region	visible BBP regime with positive residue	check that the outlier eigenvector carries teacher mass
large-step or disordered regimes	hidden-tail or bulk-amplifying Muon regimes	check whether risk progress happens without stable spectral visibility
learning-rate phase boundary	$a_{\mathrm{raw},*}$, Boltzmann and HJB curves	compare the selected exponent with BBP margins, not only final loss

3 Document Map

The presentation follows the same pedagogical order throughout the bundle: first the spectral picture, then a solvable deterministic state equation, then the BBP readouts, and only after that the probabilistic transport conditions. In the main paper this is carried by the early summary figure and the logical map table. In the companions the same order is repeated in a focused form, so that the reader can check the constant-a theory, the variable-a control layer, or the longer Markovian derivation without rebuilding the full manuscript.

Document	Role	Best use
Main paper	Main manuscript. It contains the model, finite-frame Muon dynamics, Volterra learning curves, three BBP spectra, variable $a(t)$, AMP, HJB, and the RFA condition.	Read first. This is the main publication manuscript and contains the deterministic spine theorem.
Constant-a companion	Fixed-exponent companion. It isolates the Paquette/Volterra calculation and the gradient, weight and Hessian spectral tests.	Use when checking the fixed-exponent closure theorem, BBP formulas, and figure checks.
Variable-a companion	Variable-exponent companion. It defines b, c, F_t, A_t, Boltzmann, HJB, and the AMP-to-Muon power projection.	Use when discussing the reduced online-control theorem and what is, or is not, globally optimal.
Markovian companion	Long-form technical derivation. It preserves the side-by-side SGD/Muon calculations and the staged numerical verification program.	Use for intuition, detailed derivations, and supporting evidence; cite the main paper or focused companions for compact theorem statements.

The corresponding file names are:

Main paper: `muon_powerlaw_phase_retrieval_master_paper.pdf`,
Constant-a: `muon_constant_a_spectral_paper.pdf`,
Variable-a: `muon_variable_a_boltzmann_hjb_paper.pdf`,
Markovian companion: `markovian_p_optimal_phase_retrieval.pdf`.

4 Citation-Level Claims

The bundle contains several strengths of statement. When writing a paper or talking to a referee, it is useful to keep them separate.

Claim type	Safe formulation	Where to point
Finite algebra	The quadratic model, finite-frame Muon dynamics, and finite Schur complements are exact once the state matrices are formed.	Main paper and constant-a companion.
Fresh/frozen deterministic limit	Paquette's spectral recursion, after inserting the Muon filter, gives the Volterra learning curve and the deterministic spectral coefficients.	Main paper and constant-a companion.
Finite spectral validation	The observed branches are best compared to Schur roots and residues, not to raw eigenvalue ranks.	Constant-a companion and Markovian companion.
Reduced online control	Boltzmann is the entropy-regularized HJB policy for the reduced cost $F_t(a) + A_t(a)$.	Variable-a companion.
AMP bridge	On a locally power-law active front, the linearized AMP/VAMP spectral denoiser projects to a Muon power.	Main paper and variable-a companion.
Global optimality	Not claimed. The current statement is a reduced-model and spectral projection statement.	Variable-a companion, final status section.
Same-sample theorem	The remaining RMT input is $\varepsilon_{\mathrm{RFA,d}} = o_{\mathbb{P}}(1)$ for reused-data resolvents.	Main paper and Markovian companion.

5 Where the Proofs Start

Each document has a theorem-style entry point.

Document	Entry theorem	What it tells the reader
Main paper	Deterministic spine of the theory	Finite-frame Muon, Volterra, Schur/BBP, and RFA transport are separated into their logical roles.
Constant-a companion	Fixed-exponent spectral closure	For fixed a, one reduced state predicts learning curves and the three spectral BBP readouts.
Variable-a companion	Reduced online control layer	AMP gives the local spectral cost, HJB gives the future action, and Boltzmann is the entropy-regularized policy in the reduced model.

The figure-check tables in the main paper and in the companions are part of the proof narrative. They say which equation each plot tests and what kind of agreement is expected.

6 Referee Checklist

A reader checking the manuscript should be able to verify the theory in the following order. The checklist is deliberately mechanical: each line names the mathematical object, the evidence used in the bundle, and the reason that the check matters.

Checkpoint	Evidence in the bundle	Why it matters
State closure	finite-frame equations for Q, M, C, q_i^2, ρ_i	proves that the optimizer can be described by a small deterministic state before any spectral approximation is used
Volterra clock	fixed-a Paquette recursion and learning-curve comparisons	checks that the reduced state moves at the predicted training speed
Gradient denoising	gradient bulk, Stieltjes/Rayleigh quantities, AMP projection	checks that Muon is acting on the signal-to-bulk geometry, not only changing the scalar loss
Weight BBP branches	finite Schur roots, matched empirical branches, residues	checks that branch identity is predicted by the Schur equation rather than by unstable eigenvalue ranks
Hessian BBP branches	Dyson edge, left/right Schur contacts, smoothed dense handoff	checks that the local curvature sees the same residual and parallel modes as the Volterra state
Online exponent	$F_t(a)$, $A_t(a)$, Boltzmann/HJB trajectories and spectral overlays	checks that $a(t)$ improves future BBP visibility, not just a one-step surrogate
Same-sample transport	fresh/frozen and leave-one-out comparisons, plus the stated RFA condition	identifies the only remaining probabilistic local-law input

If these checks pass, the deterministic optimization story is coherent. If one of them fails, the failure is localized: state closure, Volterra clock, one of the three spectral readouts, online control, or same-sample transport.

7 Reader Landmarks

The manuscript bundle now contains several tables whose role is to prevent the reader from having to reconstruct the whole story from memory.

Landmark	Where it appears	What it is for
Plain-language dictionary	Main paper, before the theorem	Defines Volterra, BBP, Schur, Dyson, RFA, AMP/VAMP, HJB and Boltzmann in the language of this paper.
Results-at-a-glance table	Main paper, after the bundle map	Lists the five citation-level outputs: fixed-a clock, power-law balance, weight BBP, Hessian BBP, and variable exponent control.
Proof roadmap	Main paper, after the theorem	Shows the finite algebra, the Paquette/Volterra step, the Schur/BBP step, and where the RFA input enters.
Fixed-a object dictionary	Constant-a companion	Defines q_i^2, ρ_i, Q, $\delta_i^{(a)}$, $\nu_i^{(a)}$, Schur roots and Dyson edges.
Control-layer dictionary	Variable-a companion	Defines X_t, b, F_t, A_t, c_t, β_t and Z_t before Boltzmann and HJB are used.
Figure-check tables	Main paper and both companions	Attach each numerical plot to the equation or theorem that it tests.

For a first reading, these landmarks are part of the mathematical text, not as decorative summaries. They are the shortest path from the physical story to the proofs.

8 Notation That Must Not Be Confused

The student width is denoted by P in the Markovian companion and by p in some derivations. The Muon exponent is now always denoted by a. The earlier notation $p(t)$ for the optimizer exponent is read as $a(t)$.

There is one harmless legacy convention. The focused papers write the captured Volterra mass as $q_i^2(t)$, following Paquette's projected-risk notation. The long Markovian companion sometimes writes the same non-negative mass as $q_i(t)$ in empirical BBP sections. In both cases the object is a mass, not a signed amplitude.

Symbol	Meaning
a	Muon exponent. $a = 1$ is raw gradient descent, $a = 0$ is sign-SVD.
P or p	Number of student directions, depending on the document.
$q_i^2(t)$	Captured mass of teacher mode i.
$\rho_i(t)$	Residual mass of teacher mode i.
$Q(t)$	Total risk scale or aggregate Volterra state.
$F_t(a)$	Instantaneous AMP/spectral denoising free energy.
$A_t(a)$	HJB action or future shadow cost of choosing exponent a.
$b(t, X, a)$	Drift of the reduced Paquette/Volterra state.
c_t	Gauge/intercept in the action; only differences in a matter.
$\varepsilon_{\mathrm{RFA,d}}$	Same-sample Schur/RFA error. This is the remaining RMT transport condition.

9 Logical Spine

The deterministic part of the theory is a chain:

$$\begin{aligned}&\text{finite-frame Muon} \to \text{Paquette Volterra} \to \text{spectral readouts}\\ &\to \text{BBP exit times and overlaps} \to \text{fixed or variable exponent selection.}\end{aligned}$$

The fixed-a branch is:

$$a \text{ fixed} \quad \Longrightarrow \quad Q(t) = F + \eta K * Q^{(a+1)/2} \quad \Longrightarrow \quad \text{learning curve and BBP exits.}$$

The variable-a branch is:

$$F_t(a) + A_t(a) \quad \Longrightarrow \quad \pi_t(\mathrm{d}a) \propto e^{-\beta_t(F_t(a)+A_t(a))}\mathrm{d}a \quad \Longrightarrow \quad a_t = \int a\,\pi_t(\mathrm{d}a),$$

or the corresponding hard-contour/KKT version when the policy is not the smooth Gibbs mean.

10 What Is Closed, Conditional, and Open

The words in the status column have the following precise meanings.

- *Closed finite algebra* means that, once the finite matrices are formed, the statement is an exact identity or an exact Schur complement calculation at the measured state.
- *Closed fresh/frozen* means that the trajectory is treated as fixed before the independent sample or resolvent is drawn. In plain words, the state is produced first and the spectrum is measured with independent data. This is the setting in which Paquette-type formulas are used directly.
- *Closed in the reduced model* means that the reduced state and cost have already been chosen. The resulting policy or ODE is then exact inside that reduced control problem.
- *Conditional* means that the formula is mathematically well-defined and experimentally meaningful, while the larger comparison theorem remains a separate statement.
- *Open RMT input* means that the remaining issue is probabilistic transport from fresh or leave-one-out objects to the reused-data trajectory.

Layer	Status	Meaning
Population and empirical loss algebra	closed finite algebra	The quadratic model, gradients, Hessian blocks, and finite summaries are explicit.
Finite-frame Muon ODE	closed finite algebra	The Muon update reduces to Gram and overlap variables.
Fixed-a Paquette/Volterra translation	closed fresh/frozen	Insert the Muon spectral filter in Paquette's coefficients.
Weight Schur BBP roots	closed finite algebra	The observed-state finite Schur equation predicts branch exits and overlaps.
Hessian Dyson/Schur picture	closed with regular contacts	Detached branches are clean; dense handoff requires smoothing.
Boltzmann/HJB reduced rule	closed in reduced model	Once $F_t + A_t$ is specified, the entropy-regularized policy is exact.
Global SQ optimality of Boltzmann	conditional	Requires a comparison-class theorem beyond the reduced control model.
Same-sample transport	open RMT input	Need $\varepsilon_{\mathrm{RFA,d}} = o_{\mathbb{P}}(1)$ for the reused-data trajectory.

11 Figure Atlas

The main figures to inspect first are:

overview/story figure	main paper, early sections,
learning curves	fixed-a Volterra comparison,
three escape times	population, BBP, and visible exit times,
fresh Schur overlay	weight-spectrum branch matching,
Hessian Schur time spectrum	local geometry and residual branches,
$a(t)$ trajectories	Boltzmann/HJB exponent selection,
theta/Schur comparison	reduced spectral control check.

The important visual rule is that branch identity is read through Schur roots and residues, not through naive eigenvalue ranks. Rank tracking can switch when several branches are close to the same bulk edge.

How to judge a spectral plot

A spectral plot in this bundle is not judged by whether every empirical dot lies exactly on a deterministic curve. It is judged by a hierarchy of evidence.

Visual object	Correct interpretation	Common mistake
Bulk cloud	Continuous spectrum predicted by a Dyson, MP, or Paquette resolvent law. Its edge gives the visibility threshold.	Treating the cloud as noise that can be ignored.
Schur root	Deterministic finite-rank outlier root after the bulk coordinates have been eliminated.	Matching by eigenvalue rank instead of by the Schur branch.
Star or matched marker	Empirical branch matched to the Schur root, usually at a checkpoint.	Requiring the absolute eigenvalue to be perfect when the time of exit is the more stable observable.
Residue or overlap	Teacher mass carried by the outlier eigenvector, computed from the derivative of the Schur resolvent.	Counting any detached eigenvalue as informative without checking residue.
Dense BBP contact	A near-edge handoff where branches are too close to be followed by a raw label; use smoothed margins.	Calling label switches a failure of the theory.

This rule is why the fresh Schur overlays are central. They test the object that the theory actually predicts: roots and residues of a finite Schur problem relative to the bulk edge.

How to judge a learning-curve plot

Learning-curve agreement is read in three windows. The early window tests the initial Paquette drift. The middle window tests the moving front and the Volterra memory kernel. The terminal window is the most delicate because it mixes finite dimension, discretisation, and the final residual floor. A good fit at the beginning and middle is already strong evidence that the reduced state is correct; terminal deviations are compared with spectral checks before changing the theory.

12 Recommended Reading Order

1. Read the introduction and "Summary and outline" of the main paper.
2. Read the model and finite-frame Muon sections of the main paper.
3. Read the constant-a companion if the question is about learning curves, fixed exponents, or the three spectral BBP tests.
4. Read the variable-a companion if the question is about Boltzmann, HJB, AMP projection, or online exponent selection.
5. Return to the RFA section of the main paper only after the deterministic story is clear. It is the probabilistic transport problem, not the source of the optimizer intuition.
6. Use the Markovian companion for long-form intuition and derivations, but cite the main paper or focused companions for the canonical notation.

13 One-Sentence Takeaway

The bundle says that power-law phase retrieval can be read as a moving BBP problem, and that Muon improves learning by reshaping the gradient spectrum so that the visible front of teacher modes is denoised at the right scale.