Technical Companion: Markovian Muon-a Control for Power-Law Multi-Index Phase Retrieval

Anonymous

July 3, 2026

Abstract

This companion develops the Markovian control viewpoint for the Muon exponent a in quadratic multi-index phase retrieval. The student width is denoted by P. The optimization exponent is denoted by $a \in [0, 1]$: $a = 1$ is ordinary gradient descent, while $a = 0$ is the singular-value sign map used by Muon. The purpose is to predict a from spectral information measured during training.

Three spectra carry the information: the spectrum of the weights, the spectrum of the gradient, and the spectrum of the empirical Hessian. BBP theory is not a dimensional reduction here; it is the calculus that computes when an informative eigenvalue leaves the random bulk in each of these spectra. In a power-law teacher, the resulting rule has a simple physical form: choose a so that the signal free energy and the bulk/noise free energy are balanced on the modes whose eigenvalues are visible outside the bulk.

Role in the manuscript bundle. This document is a self-contained technical companion to the main manuscript. It gives the longest derivation of the Markovian-control picture, records the side-by-side SGD/Muon comparisons, and explains the numerical checks used to test the spectral predictions. Compact theorem statements are collected in the main paper and in the two focused companions: the constant-a spectral closure and the variable-a Boltzmann/HJB control layer.

Reader's map. The first part of the companion is self-contained. Section 1 defines the model, the population and empirical losses, and every symbol used later. Section 2 explains what it means for a mode to be visible outside the bulk. Section 3 compares population gradient descent, empirical gradient descent, population Muon, and empirical Muon. Sections 4–6 derive the constant-a and time-dependent-a equations. The remaining sections record how the same objects appear in finite Hessian, gradient, and weight spectra. The figures are intentionally schematic: they explain the meaning of the quantities before the numerical tables use them.

Reading strategy. This companion is deliberately more expansive than the main paper. It keeps intermediate viewpoints, side-by-side SGD/Muon comparisons, and spectral checks that clarify the path of the argument. When a compact theorem statement is needed, the reader should use the main paper or one of the two focused companions. When the physical meaning of a symbol or transition is unclear, the present document supplies the longer explanation.

Numerical labels. Some later tables use labels such as $D40$, $D56$, $D72$, or $T16$. These are not new theoretical parameters. They are run descriptors: D records the ambient dimension used in a finite-dimensional spectral check, and T records the training horizon or terminal time in the associated rollout. The mathematical quantities being tested are always the same objects defined below: the Volterra state, Schur roots, residues, bulk edges, and BBP exit times.

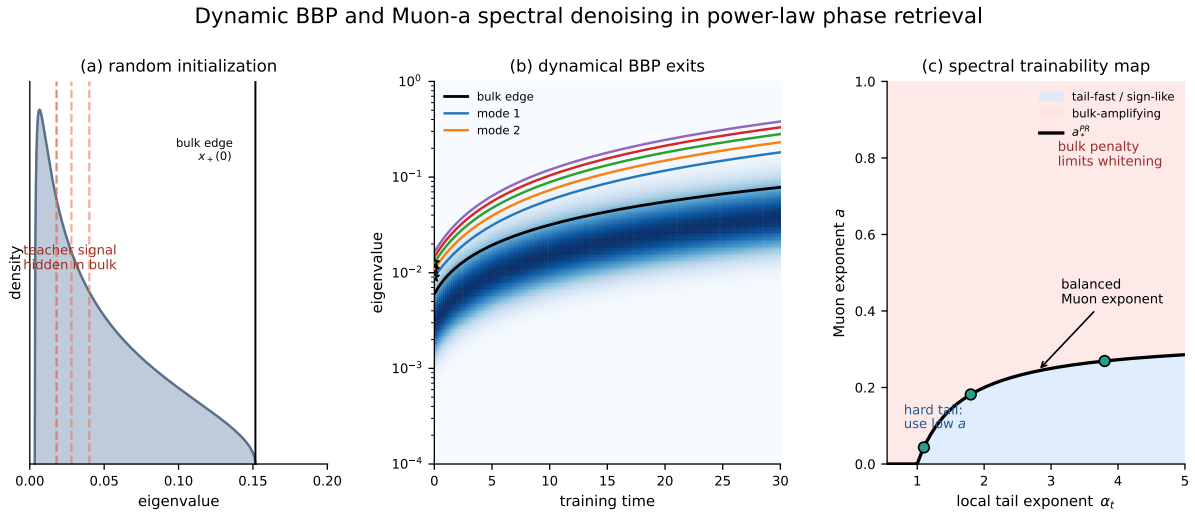

Figure 1: Summary of the argument. (a) A useful direction becomes actionable when its eigenvalue leaves the spectral bulk. (b) Training is therefore read as a sequence of BBP exits from a moving edge. (c) In a power-law teacher, the active tail slope changes along the trajectory, so the Muon exponent a is chosen by balancing signal progress against bulk amplification.

1 Introduction and physical picture

The starting point is an experimental observation. During training, the useful directions are not immediately visible in the spectrum: they first live inside a random bulk, then detach as isolated eigenvalues, and only then become reliable directions for learning. This is the dynamical BBP picture. The role of Muon is to act on the singular values of the gradient so that weak but useful spectral directions are not suppressed by the stronger bulk directions.

The question is therefore not simply whether Muon works. The question is which exponent a in the Muon family should be used while the spectrum is moving. The answer developed below is that a balances two free energies: a signal free energy, measuring progress on the currently visible teacher modes, and a bulk free energy, measuring how much random spectral mass is amplified by the same singular-value map. In a power-law teacher, this balance turns into a moving tail problem: learning advances from strong modes to weak modes, and the optimal exponent changes as the visible group moves.

The presentation follows the logic of a solvable model. First we write the population and empirical losses side by side. Then we show that gradient descent and Muon live in the same finite frame, but Muon changes the singular weights in that frame. Finally, we connect the finite-dimensional summary ODE to the Hessian, gradient, and weight spectra through Schur/BBP equations. The later numerical tables are checks of this single story.

2 Model, losses, and dictionary

The paper of Braun–Loureiro–Minh–Imaizumi [8] studies phase retrieval with anisotropic Gaussian inputs. In coordinates diagonalizing the covariance, their Phase III growth rate is $8\lambda_i$, so directions with small eigenvalues learn slowly. Here the inputs are isotropic and the hierarchy is carried instead by the multi-index teacher strengths

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

Both viewpoints create the same mathematical object: a long ordered list of directions, each direction having its own deterministic growth rate.

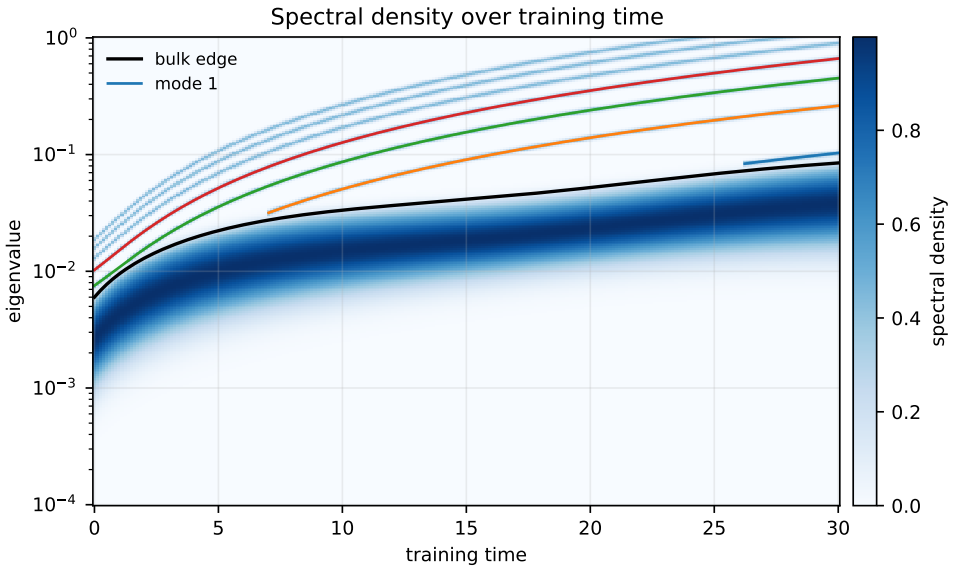

Figure 2: The spectral picture as a density over time. Time is on the horizontal axis and eigenvalue is on the vertical axis. The blue density is the moving bulk, the black curve is the bulk edge, and the thin curves are teacher modes after they have become visible outside the bulk. This is the object that the BBP framework computes: not a reduction of the dynamics, but the exit times and residues of eigenvalues in the relevant spectra.

We therefore separate three models.

(A) *Isotropic data, power-law teacher indices.* This is the model treated below. The hierarchy comes from the strengths μ_i.

(B) *Anisotropic data, rank-one phase retrieval.* This is the Braun–Loureiro–Minh–Imaizumi setup; the hierarchy comes from the covariance eigenvalues λ_i.

(C) *Anisotropic data and power-law multi-index teacher.* This is the combined case. To first order the two growth-rate hierarchies multiply. The exact description is Volterra because the summary statistics feed back into the coordinate dynamics.

Population loss and empirical loss

Let $x \sim N(0, I_d)$, let $\Theta = (\theta_1, \dots, \theta_k) \in \mathbb{R}^{d\times k}$ have orthonormal columns, and put $\Lambda = \mathrm{diag}(\mu_1, \dots, \mu_k)$. For $W = (w_1, \dots, w_P) \in \mathbb{R}^{d\times P}$,

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

Equivalently,

$$f_W(x) = x^\top \Sigma_W x, \qquad \Sigma_W = \frac{1}{P}WW^\top, \qquad f_\star(x) = x^\top A_\star x, \qquad A_\star = \Theta\Lambda\Theta^\top.$$

The population error matrix and population loss are

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

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

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

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

Every population equation below has an empirical analogue obtained by replacing R by R_n. The empirical gradient and Hessian are random and depend on the same samples used during training unless a fresh sample is drawn for measurement.

Notation used throughout

Symbol	Meaning
P	student width, i.e. number of learned quadratic features
k	number of teacher modes
a	Muon exponent; $a = 1$ is gradient descent and $a = 0$ is sign-SVD Muon
$Q = W^\top W$	student Gram matrix
$M = W^\top \Theta$	student–teacher overlap matrix
$C = M^\top Q^{-1} M$	captured teacher mass inside the span of W
r_i	scalar captured mass of teacher mode i, in a diagonalized approximation
$q_i(t)$	non-negative captured/learned mass of mode i in this long companion
$g_i(t)$	signal singular scale of mode i read from the Hessian/Schur equation
$\omega_i(t)$	population prefactor multiplying the spectral filter on mode i
$\nu_{B,t}$	limiting bulk singular-value law at time t
$\mathcal{F}(t)$	set of teacher modes whose eigenvalues are visible outside the bulk at time t
$A(t)$	common amplitude in the local power-law representation $g_i(t) \simeq A(t) i^{-\gamma}$
$V(t, x)$	Hamilton–Jacobi value function: best future terminal loss from state x at time t
$\varepsilon_{\mathrm{RFA,d}}$	probabilistic error in replacing reused-data Schur observables by fresh/cavity observables

The focused constant-a paper writes the same non-negative Volterra mass as $q_i^2(t)$, following Paquette's projected-risk notation. In this long companion the shorter symbol $q_i(t)$ is kept in the empirical BBP sections; it should be read as a mass, not as a signed amplitude.

The word "visible" has a precise spectral meaning: an eigenvalue is visible when it is outside the limiting bulk by a non-vanishing margin. A BBP event is the birth or disappearance of such an eigenvalue.

RFA stands for resolvent frame averaging. It is the random-matrix statement that the Schur/resolvent quantities measured along a trajectory trained on reused data have the same deterministic limit as the corresponding fresh or leave-one-out quantities. The notation $\varepsilon_{\mathrm{RFA,d}}$ is not a Muon smoothing parameter; it is an error term which should vanish in probability as $d \to \infty$.

3 Reduced Markovian class

Let $r_i(t) \in [0, 1]$ be the captured overlap of mode i. For population SGD/GD in the quadratic model,

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

The Markovian spectral class replaces the gradient update by a spectral map with exponent $a \in [0, 1]$. "Markovian" means that $a(t)$ is a function of the current deterministic state, not of

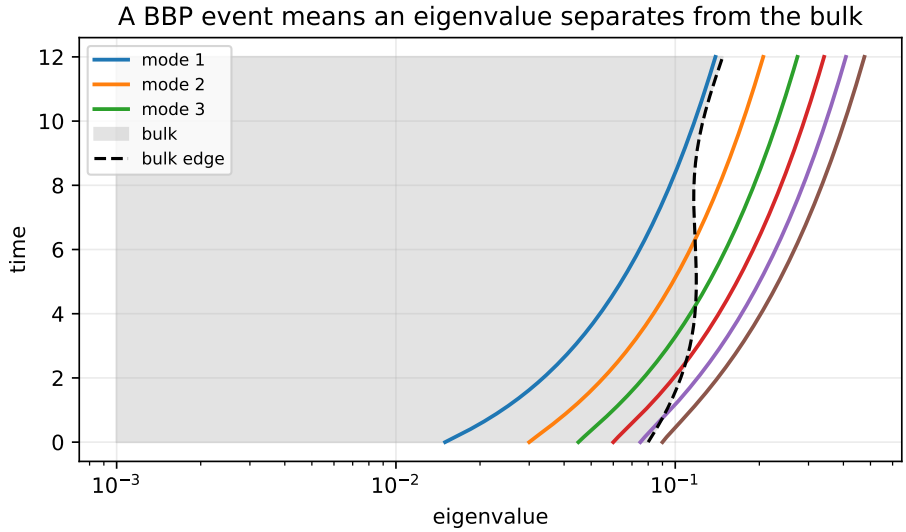

Figure 3: A BBP event is not an abstract phase label. It is the time at which an eigenvalue separates from the bulk edge in a spectrum. The same language is used for the Hessian spectrum, the gradient spectrum, and the weight spectrum.

future data. We write the resulting closed ODE abstractly as

$$\dot{r}_i = V_i(r(t), a(t)), \qquad a(t) = \pi(\mathsf{S}(t)),$$

where $\mathsf{S}(t)$ is the finite ODE state: the overlaps, the current BBP margins, and the bulk spectral statistics.

The local speed of a visible mode can be summarized as

$$\dot{q}_i(t) \simeq \omega_i(t) g_i(t)^a,$$

where q_i is the learned mass, g_i is the signal singular scale carried by the current Hessian outlier, and ω_i is the ODE growth prefactor. On a power-law group of visible modes,

$$g_i(t) \simeq A(t) i^{-\gamma}, \qquad \omega_i(t) \simeq \omega_0(t) i^{-\gamma}.$$

Therefore a fixed exponent a weights active indices as

$$w_i(a,t) \propto \omega_i(t) g_i(t)^a \simeq \omega_0(t) A(t)^a i^{-\gamma(a+1)}.$$

Throughout this companion, the visible group is the following spectral set:

$$\mathcal{F}(t) = \{i : \text{the spectral equation has an outlier for mode } i \text{ outside the bulk at time } t\}.$$

4 Population SGD and Muon* side by side

This section rewrites the population calculation in the same notation as the SGD derivation, then changes only one line: the raw gradient G is replaced by its Muon* spectral transform. This makes clear which identities survive unchanged and which ones are genuinely Muon-specific.

Let

$$A_\star = \Theta \Lambda \Theta^\top, \qquad \Sigma_W = \frac{1}{P} W W^\top, \qquad E_W = \Sigma_W - A_\star, \qquad \tau = \operatorname{Tr} E_W.$$

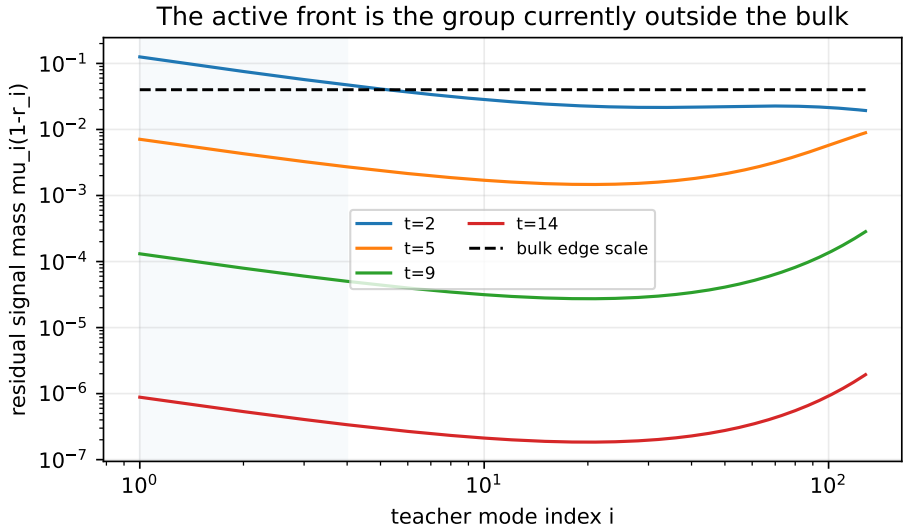

Figure 4: The visible group is the group of teacher modes whose residual signal mass is above the bulk scale. As training proceeds, learned modes leave this group and weaker modes enter it.

The population square loss is

$$R(W) = \mathrm{Tr}(E_W^2) + \frac{1}{2}\tau^2,$$

and its gradient is

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

The empirical loss and empirical gradient are

$$R_n(W) = \frac{1}{2n}\sum_{\ell=1}^{n}\left(x_\ell^\top E_W x_\ell\right)^2,$$

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

Thus the population and empirical gradient flows are

$$\dot{W} = -G(W), \qquad \dot{W}_n = -G_n(W_n).$$

The population Muon and empirical Muon flows are

$$\dot{W} = -\eta_a(t)\mathsf{M}_a(G(W)), \qquad \dot{W}_n = -\eta_a(t)\mathsf{M}_a(G_n(W_n)).$$

Here M_a is the singular-value map defined by $\psi_a(s) = s^a$ for $s > 0$ and $\psi_a(0) = 0$; the detailed finite-frame formula is written below. Finally, the empirical Hessian used for BBP measurements is the second derivative of R_n. Its block (b, c), with respect to columns (w_b, w_c), has the weighted sample-covariance form

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

where

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

This is why the Hessian spectrum is a random-matrix object: conditionally on the current weights, it is a finite block matrix of weighted sample covariance matrices. Equivalently, with

$$Q = W^\top W, \qquad M = W^\top \Theta, \qquad Z = [W, \Theta],$$

and

$$\mathcal{G} = \begin{pmatrix} Q & M \\ M^\top & I_k \end{pmatrix},$$

the gradient lives in the finite frame Z:

$$G(W) = Z A_{\mathrm{gd}},$$

where

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

This finite-frame identity is the key common point. Both SGD and Muon* move inside the same $(P+k)$-dimensional span; Muon* only changes the singular weights inside that span.

SGD line

For population gradient flow,

$$\dot{W} = -G(W).$$

The Gram summaries obey

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

and

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

For

$$C = M^\top Q^{-1} M, \qquad R_c = I_k - C,$$

the common left-multiplication terms cancel, giving the exact Riccati equation

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

In a separated scalar interval this is

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

Thus if $r_i(0) \asymp d^{-1}$, the population escape time to any fixed level $\rho \in (0,1)$ is

$$T^{\mathrm{gd}}_{i,\rho} = \frac{P}{8\mu_i} \log \frac{\rho(1 - r_i(0))}{r_i(0)(1-\rho)} = \frac{P}{8\mu_i} \log d + O(1). \qquad \text{(SGD escape)}$$

Muon* line

For a matrix $Y = U\,\mathrm{diag}(s_j)V^\top$, define the Muon* spectral map

$$\mathsf{M}_a(Y) = U\,\mathrm{diag}(\psi_a(s_j))V^\top, \qquad 0 \le a \le 1.$$

Here

$$\psi_a(s) = s^a \quad (s > 0), \qquad \psi_a(0) = 0.$$

Thus $a = 1$ gives the identity map $Y \mapsto Y$, and $a = 0$ replaces each non-zero singular value by one. This is the ideal Muon map. No smoothing parameter is introduced here; the corresponding theorem must control singular values crossing zero through Schur/RFA estimates. The population Muon* flow is

$$\dot{W} = -\eta_a(t)\mathsf{M}_a(G(W)),$$

where $\eta_a(t)$ is the global normalization used to compare update budgets. It is a scalar and therefore does not change the relative spectral allocation.

Since $G(W) = ZA_{\mathrm{gd}}$, write $Z = O\mathcal{G}^{1/2}$ with $O^\top O = I$. When $\mathcal{G}$ is invertible, and by the usual regularized inverse otherwise,

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

Split

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

The meaning is concrete. Since $Z = [W, \Theta]$,

$$ZA_a = WU_a + \Theta V_a.$$

The block U_a is the part of the Muon update that stays inside the current student span. The block V_a is the part that moves the weights toward the teacher directions. The whole effect of changing a is encoded in these two finite matrices. Then

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

The exact finite ODE is therefore

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

and

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

The captured teacher subspace again has a closed geometric equation:

$$\boxed{\dot{C} = -\eta_a\left[(I_k - C)D_a + D_a^\top(I_k - C)\right], \qquad D_a = V_aQ^{-1}M.} \qquad \text{(Muon* Riccati form)}$$

This is the exact Muon* counterpart of the SGD Riccati equation. For $a = 1$, $\mathsf{M}_1(G) = G$, so

$$V_1 = -\frac{4}{P}\Lambda M^\top, \qquad D_1 = -\frac{4}{P}\Lambda C,$$

and the displayed Muon* Riccati form reduces exactly to the SGD Riccati equation above.

In a separated scalar interval, if

$$D_a = \mathrm{diag}(d_{a,i}),$$

then

$$\boxed{\dot{r}_i = -2\eta_a(t)(1-r_i)d_{a,i}.}$$

For SGD, $d_{1,i} = -(4/P)\mu_i r_i$. For Muon*, the singular vectors are approximately the same inside an isolated singular scale, while the singular value is changed from s_i to s_i^a. Therefore

$$d_{a,i} \simeq s_i^{a-1} d_{1,i},$$

and

$$\boxed{\dot{r}_i^{(a)} \simeq \eta_a(t) s_i(t)^{a-1} \frac{8\mu_i}{P} r_i(1-r_i).} \qquad \text{(Muon* scalar interval)}$$

Equivalently,

$$\frac{\mathrm{d}}{\mathrm{d}t} \log \frac{r_i}{1-r_i} \simeq \eta_a(t) \frac{8\mu_i}{P} s_i(t)^{a-1}.$$

This is the line-by-line difference. SGD has growth rate μ_i. Muon* has the same geometric Riccati factor $r_i(1-r_i)$, but this growth rate is divided or amplified by the current gradient singular scale s_i^{1-a}.

If the teacher-moving normal component dominates an isolated singular scale, then

$$s_i(t) \simeq \frac{4}{P}\mu_i \sqrt{q_i(t) r_i(t)(1-r_i(t))}, \qquad q_i(t) \simeq W_i^\top W_i.$$

Treating s_i as approximately constant during the short escape episode gives the estimate

$$T_{i,\rho}^{(a)} \simeq \frac{P}{8\eta_a \mu_i s_{i,\mathrm{eff}}^{a-1}} \log \frac{\rho(1-r_i(0))}{r_i(0)(1-\rho)}. \qquad \text{(Muon* escape)}$$

Thus the population SGD formula is recovered at $a = 1$. For smaller a, Muon* whitens small singular directions: if $s_{i,\mathrm{eff}} \propto \mu_i$ on the visible group, the leading power changes from μ_i^{-1} to μ_i^{-a}. This is the population-side reason why Muon* can advance the power-law tail faster than gradient flow, provided the hard-edge/RFA bulk term remains controlled.

Comparison with the BBP framework

The BBP framework used later does not assume the isolated formula for $s_i(t)$. It computes the signal singular scale and the bulk scale from the empirical Hessian. On the visible group it predicts

$$\dot{q}_i(t) \simeq \omega_i(t) g_i(t)^a.$$

This is the same Muon* calculation written with Hessian observables: g_i^a is the spectral filter applied to the signal singular scale, while ω_i is the residual population prefactor. On a power-law visible group,

$$\omega_i(t) \simeq \omega_0(t) i^{-\gamma}, \qquad g_i(t) \simeq A(t) i^{-\gamma},$$

so the Hessian-predicted learning rate is

$$\omega_i(t) g_i(t)^a \simeq \omega_0(t) A(t)^a i^{-\gamma(a+1)}.$$

The apparent difference between the isolated scalar formula and the BBP formula is therefore just a change of observable. The bare population calculation tracks the overlap logit under one isolated gradient singular direction. The BBP framework tracks the Hessian-visible signal direction after subtracting the bulk. The second object is the one used for predictions because it contains the hard-edge penalty that decides when Muon* is safe.

The three exit-time notions can now be read in parallel:

Object	Exit criterion	Predicted scale
SGD population	$r_i(t) = \rho$	$\frac{P}{8\mu_i} \log d$
Muon* isolated singular scale	$r_i(t) = \rho$	$\frac{P}{8\eta_a \mu_i s_{i,\mathrm{eff}}^{a-1}} \log d$
BBP/Hessian visibility	$\mu_i(1 - r_i(t)) \simeq c_i/\sqrt{\alpha}$	eigenvalue crosses the current bulk edge

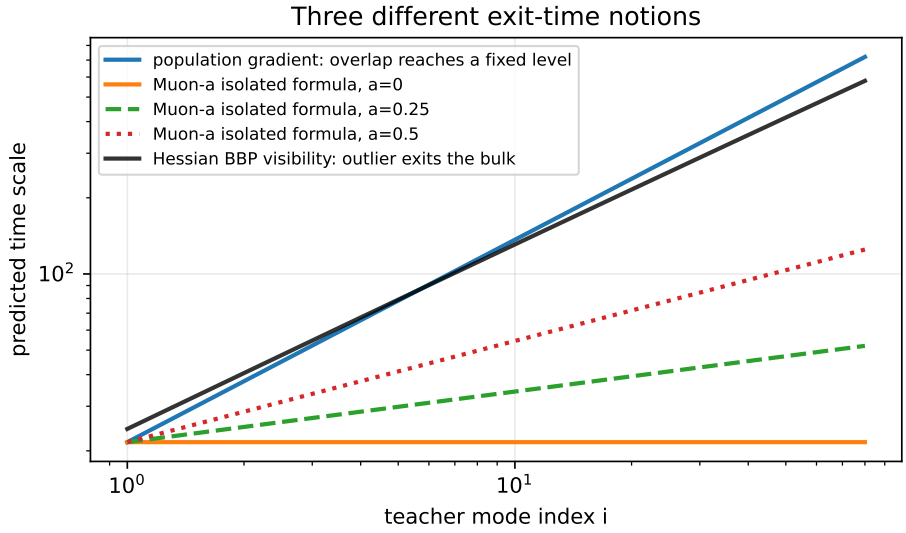

Figure 5: The three exit-time notions are different. Population gradient flow asks when an overlap becomes $O(1)$. The isolated Muon formula asks how the same overlap changes after a singular-value power map. The Hessian/BBP criterion asks when an eigenvalue becomes visible outside the bulk.

5 Exact control statement before the spectral formulas

Let $X(t)$ denote the reduced deterministic state: overlaps, outlier margins, and bulk spectral statistics. The controlled Markovian ODE has the form

$$\dot{X}(t) = F(X(t), a(t)), \qquad a(t) \in I(X(t)) := [a_{\mathrm{safe}}(X(t)), 1].$$

Here $a_{\mathrm{safe}}(X(t))$ is the smallest exponent for which the current bulk amplification is finite and stable. If the Schur/RFA estimates remain valid at $a = 0$, then $a_{\mathrm{safe}} = 0$.

The mathematical objective is not really a finite-time terminal loss. The finite T notation is only a way to approximate the infinite-time tail problem. Let $\Phi_T(X(T))$ be the remaining power-law tail mass at time T. The value function is

$$V(t, x) = \inf_{a(\cdot)} \Phi(X^{t,x,a}(T)).$$

In the infinite-time problem one reads this as the limit of V_T as $T \to \infty$, or equivalently as the dynamic program for the asymptotic tail error. The symbol V therefore means: the best future tail loss achievable from the present state. When V is smooth, it satisfies the Hamilton–Jacobi equation

$$-\partial_t V(t, x) = \min_{a \in I(x)} \nabla_x V(t, x) \cdot F(x, a).$$

Thus the exact Markovian optimizer is

$$a^\star(t) \in \arg\min_{a \in I(X(t))} \nabla_x V(t, X(t)) \cdot F(X(t), a),$$

and, at an interior smooth point,

$$\partial_a F(X(t), a^\star(t)) \cdot \nabla_x V(t, X(t)) = 0.$$

This is the rigorous control problem. The logarithmic matching equation below is the spectral approximation of this condition, obtained when the value gradient is concentrated on the currently visible power-law group and the hard-edge cost is represented by the current bulk law.

For the Muon family one keeps in mind the spectral map

$$\psi_a(s) = s^a, \qquad s > 0, \qquad \psi_a(0) = 0.$$

Its bulk amplification is controlled by quantities of the form

$$\mathfrak{H}(a, t) = \int s^{2a-2} \, \nu_{B,t}(\,\mathrm{d}s),$$

or by the corresponding Schur/RFA quantity when the empirical Hessian is used. The notation $\nu_{B,t}$ means the limiting singular-value distribution of the bulk at time t. In practice it is obtained from the deterministic Dyson or Schur equation.

6 Best constant exponent

For a fixed a, the visible group advances with a power-law bias

$$i^{-\gamma(a+1)}.$$

For a whole trajectory, however, the global scale $A(t)^a$ also matters. The accumulated learning amount is

$$K_a(T) = \int_0^T \omega_0(t) A(t)^a \, \mathrm{d}t.$$

Here $A(t)$ is the amplitude in

$$g_i(t) \simeq A(t) i^{-\gamma}$$

on the visible group. If $A(t)$ is large, increasing a helps signal progress; if $A(t)$ is small, increasing a suppresses the weak tail modes.

When the visible group changes slowly enough that a power-law approximation is valid during the measurement interval,

$$\Gamma_i^a(T) = K_a(T) i^{-\gamma(a+1)}$$

is the integrated learning amount of mode i. The final tail error is therefore modeled by

$$\mathcal{E}_T(a) = \sum_{i \geq 1} \mu_i^2 \exp\{-2\Gamma_i^a(T)\}.$$

An interior optimal constant exponent satisfies

$$\boxed{\partial_a \log K_a(T) = \gamma \, \frac{\sum_i \mu_i^2 e^{-2\Gamma_i^a(T)} \Gamma_i^a(T) \log i}{\sum_i \mu_i^2 e^{-2\Gamma_i^a(T)} \Gamma_i^a(T)}.}$$

The left side is the trajectory-averaged gain of emphasizing the current signal amplitude $A(t)$. The right side is the logarithmic index of the modes that still contribute to the final error.

Boundary solutions are selected by the sign of this derivative and by the constraint $a \in [a_{\text{safe}}, 1]$. Indeed,

$$\partial_a \Gamma_i^a(T) = \Gamma_i^a(T) \left(\partial_a \log K_a(T) - \gamma \log i\right),$$

and differentiating $\mathcal{E}_T(a)$ gives exactly the displayed stationarity condition.

In the true tail regime $K_a(T) \gg 1$, this condition simplifies. The learned index $J_a(T)$, meaning the largest mode that has been learned to constant accuracy, satisfies

$$J_a(T) \asymp K_a(T)^{1/(\gamma(a+1))},$$

and

$$\mathcal{E}_T(a) \asymp J_a(T)^{1-2\gamma} = K_a(T)^{-(2\gamma-1)/(\gamma(a+1))}.$$

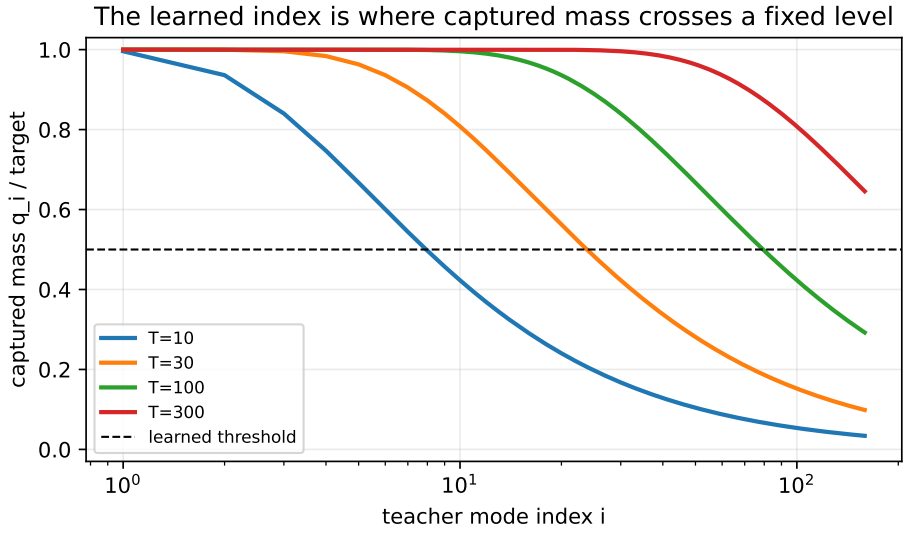

Figure 6: The learned index $J_a(T)$ is not another BBP object. It is the largest teacher index whose captured mass has crossed a fixed accuracy level. The BBP-visible group tells us which modes are currently measurable in the spectrum; $J_a(T)$ tells us how far the training has already progressed.

Thus the constant exponent asymptotically maximizes

$$\frac{\log K_a(T)}{a+1}.$$

If $A(t)$ stays on an $O(1)$ scale while $K_a(T) \to \infty$, then $\log K_a(T)$ dominates $(a+1)\partial_a \log K_a(T)$ and the maximizer is the lower endpoint:

$$\boxed{a^\star_{\text{const}} \longrightarrow a_{\text{safe}}.}$$

In particular, if the hard-edge/RFA replacement theorem permits $a_{\text{safe}} = 0$, the asymptotically optimal constant exponent is Muon/sign-like.

When A, ω_0 are approximately constant on the tail-learning interval, this gives the explicit scaling law

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

This is the multi-index analogue of the spectral-tail law in anisotropic phase retrieval: the eigenvalue power law is replaced by the teacher-index growth rate $i^{-\gamma(a+1)}$. The exponent $\kappa(a)$ is strictly decreasing in a, so the tail exponent is maximized by the smallest safe a.

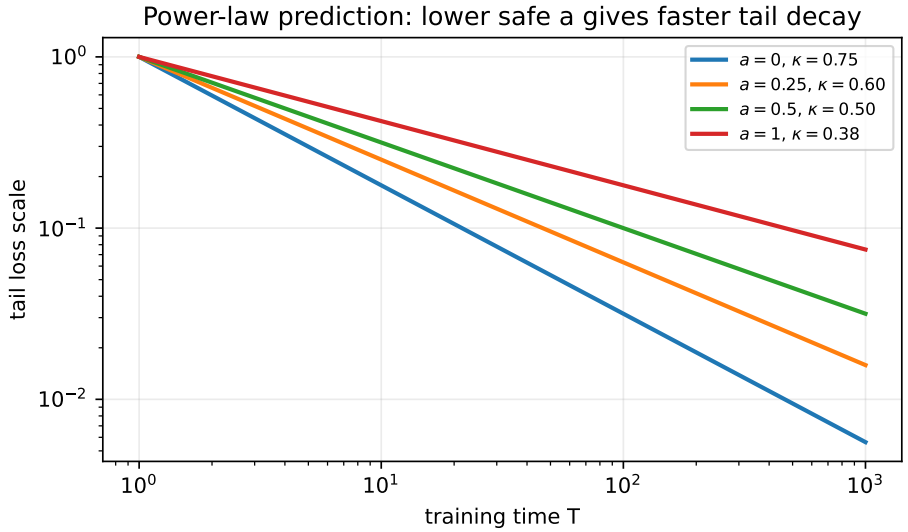

Figure 7: The explicit law $\mathcal{E}_T(a) \asymp T^{-\kappa(a)}$. If all exponents are equally stable, the lower exponent wins asymptotically because it learns more weak tail modes. Finite-time BBP constraints can still make a positive exponent preferable before the tail regime is reached.

Remark 6.1. *This is why fixed-a sweeps often look deceptively simple: the best fixed choice is low. But a low constant exponent need not be the best time-dependent choice. The real trajectory has successive Hessian phases; it should use a larger exponent near a BBP birth when signal and bulk are not yet cleanly separated, then drop back toward the smallest safe exponent once the visible group is stable.*

7 Local Markovian optimum as free-energy matching

Let $\nu_{B,t}$ be the normalized singular-value law of the current bulk, and write a scale decomposition

$$s = \sigma_B(t)x, \qquad x \sim \nu_{0,t}.$$

For a candidate exponent a, define the bulk logarithmic mean

$$L_B(a,t) = \log \sigma_B(t) + \frac{\int x^{2a} \log x \, \nu_{0,t}(\mathrm{d}x)}{\int x^{2a} \nu_{0,t}(\mathrm{d}x)}.$$

Equivalently, without separating scale and shape,

$$L_B(a,t) = \frac{\int s^{2a} \log s \, \nu_{B,t}(\mathrm{d}s)}{\int s^{2a} \nu_{B,t}(\mathrm{d}s)}.$$

This formula is often the safest one to remember. The measure $\nu_{B,t}$ is the bulk law after removing the finitely many outliers; it is the continuous part of the Hessian/gradient/weight spectrum, depending on which spectrum is being studied.

For a visible BBP group $\mathcal{F}(t)$, define the signal logarithmic mean

$$L_S(a,t) = \frac{\sum_{i\in\mathcal{F}(t)} \omega_i(t) g_i(t)^a \log g_i(t)}{\sum_{i\in\mathcal{F}(t)} \omega_i(t) g_i(t)^a}.$$

The local objective behind these definitions is the instantaneous signal-to-bulk ratio

$$\mathcal{Q}(a,t) = \frac{\sum_{i\in\mathcal{F}(t)} \omega_i(t) g_i(t)^a}{\left(\int s^{2a} \nu_{B,t}(\mathrm{d}s)\right)^{1/2}}.$$

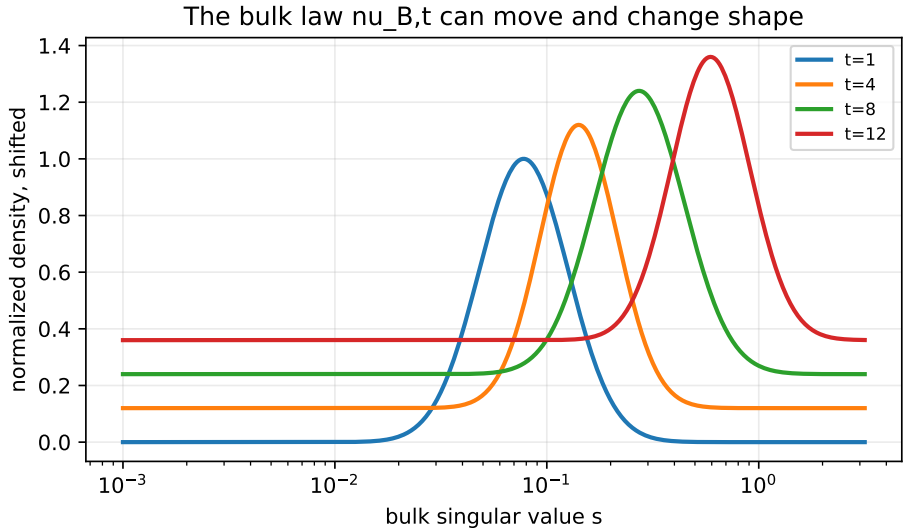

Figure 8: The bulk law $\nu_{B,t}$ need not merely rescale. Its shape can move during training. Then $\partial_t L_B$ contains both a scale derivative and a shape derivative, both computed from the current deterministic spectral law.

The numerator is the active BBP learning speed and the denominator is the quadratic hard-edge/bulk amplification of the same spectral filter. Therefore

$$\partial_a \log \mathcal{Q}(a,t) = L_S(a,t) - L_B(a,t).$$

The local Markovian exponent is the solution of the stationarity equation

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

Equivalently,

$$a^\star_{\mathrm{loc}}(t) = \mathrm{clip}_{[a_{\mathrm{safe}}(t),1]} \arg\min_a \left| L_S(a,t) - L_B(a,t) \right|.$$

This is the Boltzmann matching equation: the exponent is chosen so that the typical logarithmic signal singular scale selected by the filter matches the typical logarithmic bulk singular scale selected by the same filter.

The statistical-physics form is the following. Define the signal partition function and bulk partition function

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

Then

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

Thus $L_S = L_B$ says that the signal free-energy slope equals the bulk free-energy slope. This is the reason for the word Boltzmann: the exponent a tilts two measures, one on visible signal modes and one on bulk singular values, and the selected exponent balances the two tilted free energies.

The stable interior case is the one where

$$\partial_a (L_S - L_B) = \mathrm{Var}_{P^{a,t}}(\log g_i) - 2\,\mathrm{Var}_{Q_{a,t}}(\log s) < 0.$$

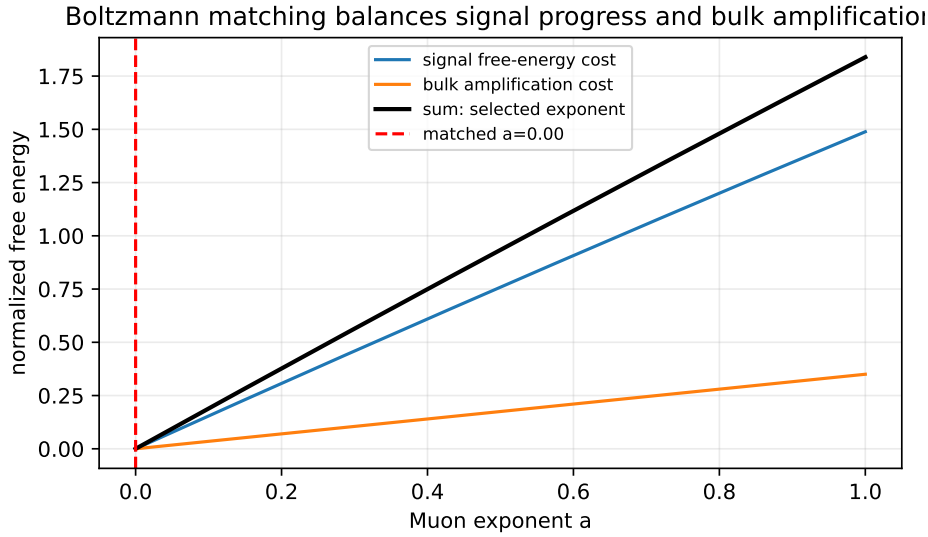

Figure 9: The Boltzmann equation balances two costs. Increasing a can improve progress on strong signal modes, but it also changes how much bulk noise is amplified. The selected exponent is where the two free-energy slopes balance.

8 ODE for the optimal Markovian exponent

Define the stationarity residual

$$\Psi(a,t) = L_S(a,t) - L_B(a,t).$$

Between BBP contact times, assume that the active set is fixed and that the root is interior:

$$a_{\text{safe}}(t) < a^\star(t) < 1, \qquad \partial_a \Psi(a^\star(t),t) \neq 0.$$

Then $a^\star(t)$ is differentiable and satisfies the exact implicit ODE

$$\boxed{\dot{a}^\star(t) = \frac{\partial_t L_B(a^\star(t),t) - \partial_t L_S(a^\star(t),t)}{\partial_a L_S(a^\star(t),t) - \partial_a L_B(a^\star(t),t)}.}$$

This is the desired Markovian ODE: every term is a function of the current finite ODE state and of the current effective Hessian law.

The derivatives have a useful statistical form. Put

$$P_i^{a,t} = \frac{\omega_i(t) g_i(t)^a}{\sum_{j\in\mathcal{F}(t)} \omega_j(t) g_j(t)^a}, \qquad \ell_i(t) = \log g_i(t).$$

Then

$$\partial_a L_S(a,t) = \operatorname{Var}_{P^{a,t}}(\ell_i),$$

and

$$\partial_t L_S(a,t) = \mathbb{E}_{P^{a,t}}[\partial_t \ell_i] + \operatorname{Cov}_{P^{a,t}}\left(\partial_t \log \omega_i + a\,\partial_t \ell_i, \ell_i\right).$$

For the bulk, let

$$Q_{a,t}(\,\mathrm{d}s) = \frac{s^{2a}\nu_{B,t}(\,\mathrm{d}s)}{\int u^{2a}\nu_{B,t}(\,\mathrm{d}u)}.$$

Then

$$L_B(a,t) = \mathbb{E}_{Q_{a,t}} \log s, \qquad \partial_a L_B(a,t) = 2\operatorname{Var}_{Q_{a,t}}(\log s).$$

If the bulk shape is fixed and only the scale changes, $s = \sigma_B(t)x$, then

$$\partial_t L_B(a,t) = \partial_t \log \sigma_B(t).$$

If the shape also moves, the same formula holds with the additional material derivative of $\nu_{0,t}$. This term is still Markovian because $\nu_{0,t}$ is determined by the Dyson equation at the current state.

At boundaries and contact times, the correct closed description is the projected differential inclusion

$$a^\star(t) = \mathrm{clip}_{[a_{\mathrm{safe}}(t),1]}\, \bar{a}(t), \qquad \Psi(\bar{a}(t),t) = 0,$$

with possible kinks when the active BBP set $\mathcal{F}(t)$ changes.

8.1 Progress-level Boltzmann ODE

The Boltzmann exponent used in the finite experiments is a projected version of the preceding stationarity equation. The raw free energy

$$\mathfrak{F}(a,t) = \log Z_B(a,t) - 2\log Z_S(a,t)$$

is retained as an observable, but its unconstrained minimizer is often the gradient endpoint $a = 1$ in finite samples. The exponent actually used in the stable trajectories instead keeps a fixed fraction of the best instantaneous signal response. Let

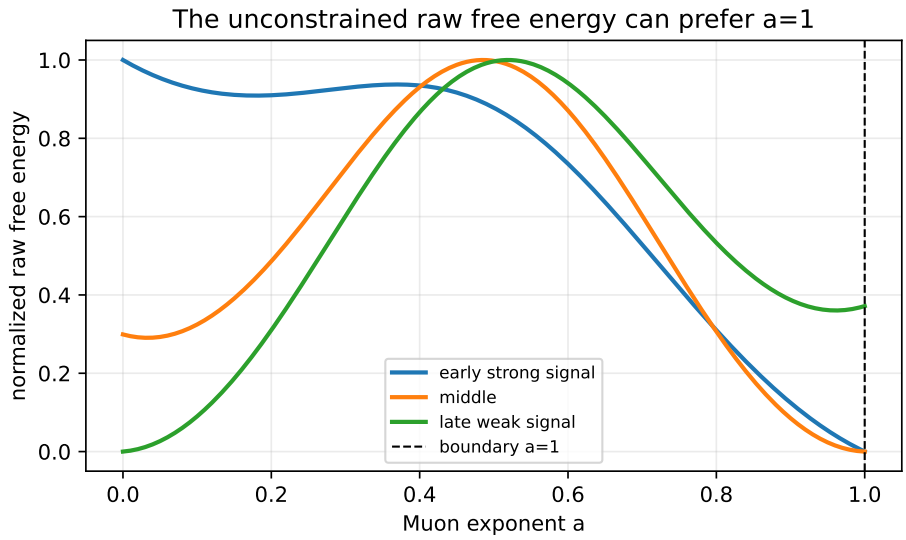

Figure 10: Why the unconstrained raw free energy is not always the final rule. In finite samples it can select the endpoint $a = 1$, which gives strong short-time signal progress but may miss the weak power-law tail. The progress-level formulation keeps a prescribed fraction of the best signal response while avoiding this endpoint collapse.

$$\theta_{\mathrm{prog}} \in (0,1)$$

be the chosen progress fraction: $\theta_{\mathrm{prog}} = 0.6$, for example, means that the selected exponent must keep at least 60% of the best instantaneous signal response. Define

$$Z_S(a,t) = \sum_{i\in\mathcal{F}(t)} w_i(t)\psi_a(g_i(t)), \qquad Z_S^{\max}(t) = \max_{b\in[0,1]} Z_S(b,t),$$

where $w_i(t)$ is the non-negative weight assigned to the visible mode i and $\psi_a(s) = s^a$. Ignoring inactive boundary constraints, the selected exponent is the upper boundary of the admissible interval

$$a_{\mathrm{prog}}(t) = \sup\{a : \ Z_S(a,t) \geq \theta_{\mathrm{prog}} Z_S^{\max}(t)\}.$$

Equivalently, on a smooth interior piece,

$$\Phi(a,t) = \log Z_S(a,t) - \log Z_S^{\max}(t) - \log \theta_{\mathrm{prog}} = 0.$$

Therefore

$$\boxed{\dot{a}_{\mathrm{prog}}(t) = -\frac{\partial_t \log Z_S(a_{\mathrm{prog}},t) - \partial_t \log Z_S^{\max}(t) - \partial_t \log \theta_{\mathrm{prog}}(t)}{\partial_a \log Z_S(a_{\mathrm{prog}},t)}}.$$

This is the Markovian ODE for the Boltzmann exponent when the progress constraint is active. If a boundary $a = 0$ or $a = 1$ is active, the same statement becomes a projected differential inclusion with the active KKT constraint replacing $\Phi = 0$.

The finite-horizon optimization over θ_{prog} must be stated with the population risk, not with an external tail surrogate. For fixed horizon T, define the progress-level policy $a_\theta(t)$ by the preceding contour and let

$$R_T(\theta) = R(W_\theta(T)) = \mathrm{Tr}(E^2_{W_\theta(T)}) + \frac{1}{2}\,\mathrm{Tr}(E_{W_\theta(T)})^2.$$

On the terminal diagonal residual plateau this reduces to

$$R_T(\theta) = \sum_i \rho_i^{(\theta)}(T)^2 + \frac{1}{2}\left(\sum_i \rho_i^{(\theta)}(T)\right)^2 + R_{\mathrm{off}}(T) + R_{\mathrm{bulk}}(T) + R_{\mathrm{coup}}(T).$$

Here the three extra terms are ordinary risk terms, not engineering corrections:

$$\begin{aligned} R_{\mathrm{off}}(T) &= \text{risk from off-diagonal teacher/student couplings,} \\ R_{\mathrm{bulk}}(T) &= \text{risk from the isotropic student bulk not aligned with the teacher,} \\ R_{\mathrm{coup}}(T) &= \text{remaining coupling between learned and residual teacher directions.} \end{aligned}$$

They vanish or become lower order in the ideal diagonal tail calculation, but they are kept in finite-dimensional experiments because the learned subspace is not exactly diagonal. The Schur/BBP side supplies the constraint, not the cost. If $\lambda_i^{\mathrm{Schur}}(t)$ is the finite Schur root, $x_+(t)$ the finite bulk edge, and $\Omega_i^{\mathrm{Schur}}(t)$ the Schur residue, set

$$m_i^{\mathrm{Schur}}(t) = \min\{\lambda_i^{\mathrm{Schur}}(t) - x_+(t), \Omega_i^{\mathrm{Schur}}(t) - \Omega_0\}, \qquad m_i^{\mathrm{cap}}(T;\delta) = \delta - \rho_i^{(\theta)}(T)/\mu_i.$$

The mathematically clean feasible set is therefore

$$\Theta^{R\mathrm{Schur}}_{\delta,\Omega_0}(T) = \left\{\theta : \min_i m_i^{\mathrm{cap}}(T;\delta) \geq 0, \quad \inf_{t\in[0,T]} \min_{i\in\mathcal{F}(t)} m_i^{\mathrm{Schur}}(t) > 0\right\}.$$

The final contour level is

$$\theta_R^\star(T,\delta,\Omega_0) \in \arg\min_{\theta\in\Theta^{R\mathrm{Schur}}_{\delta,\Omega_0}(T)} R_T(\theta), \qquad \theta_{\mathrm{crit}}^{R\mathrm{Schur}} = \inf \Theta^{R\mathrm{Schur}}_{\delta,\Omega_0}(T).$$

When R_T has a flat terminal plateau, the aggressive exponent rule uses $\theta_{\mathrm{crit}}^{R\mathrm{Schur}}$; the risk minimizer inside the plateau is reported separately. The weighted-tail quantity is retained only as a observable for where the plateau becomes most symmetric across weak modes. Below the feasible interval the exponent rule spends too much exponent area $\int_0^T a(t)\,\mathrm{d}t$, and the power-law tail remains unlearned.

The fresh-gradient sweeps support this interpretation. At $d = 40, T = 30$, $\theta = 0.10, 0.15, 0.20$ remain off the terminal plateau, with final risks $6.96\,10^{-1}, 2.64\,10^{-1}, 1.06\,10^{-3}$, while $\theta = 0.30$ reaches $1.60\,10^{-9}$. At $d = 56, T = 30$, the deeper tail pushes the critical interval upward: $\theta = 0.30$ still leaves risk $8.93\,10^{-5}$, while $\theta = 0.65$ reaches $7.31\,10^{-10}$. At $d = 40, T = 60$, all $\theta \geq 0.15$ are already on the terminal plateau, but the capture time improves monotonically from 45 at $\theta = 0.15$ to 19.5 at $\theta = 0.85$. Thus the finite-horizon optimum is the largest retention level that keeps the visible group active; the numerical value of $\theta_{\mathrm{crit}}(T)$ depends on the horizon and on the power-law depth.

The same conclusion survives comparison with the best constant exponent in fresh-gradient experiments. Fixed a was swept and the narrow constant interval was refined. At $d = 40, T = 30$, the best constant is $a = 0.20$; by $a = 0.22$ the final risk has already jumped from $1.07\,10^{-8}$ to $1.23\,10^{-4}$. At $d = 56, T = 30$, the best constant is $a = 0.16$; by $a = 0.20$ the risk jumps from $2.20\,10^{-8}$ to $3.59\,10^{-2}$. On seeds $0, 1, 2$, the targeted comparison is

case	policy	parameter	mean final risk	mean final residual
$d40, T30$	fixed	$a = 0.20$	$(1.17 \pm 0.24)\,10^{-8}$	$2.69\,10^{-5}$
$d40, T30$	Boltzmann	$\theta_{\mathrm{prog}} = 0.30$	$(1.81 \pm 0.47)\,10^{-9}$	$6.83\,10^{-6}$
$d56, T30$	fixed	$a = 0.16$	$(2.26 \pm 0.13)\,10^{-8}$	$4.08\,10^{-5}$
$d56, T30$	Boltzmann	$\theta_{\mathrm{prog}} = 0.65$	$(7.41 \pm 0.49)\,10^{-10}$	$6.55\,10^{-6}$

Thus the progress-level Boltzmann exponent beats the best refined constant on the tested finite horizons. The sharp cliff of the best constant exponent is the finite-dimensional trace of the block-constant physical picture: as the power-law visible group moves, a single exponent cannot stay simultaneously on the current signal group and away from the next hard-edge failure.

The two-scale closure is as follows. The large-index reduced growth-rate experiment at $d = 10^6$ verifies

$$\kappa(a) = \frac{2\gamma - 1}{\gamma(a+1)}$$

and gives the asymptotic constant optimum $a^{\star}_{\mathrm{const}} = 0$ for all tested horizons

$$T \in \{30, 100, 300, 1000, 3000\}.$$

The full finite Hessian problem adds the BBP/hard-edge visibility constraint, so the finite best constant is the largest safe exponent before a cliff:

case	best fixed a	safe fixed interval	first cliff
$d40, T30$	0.20	$0.10 \leq a \leq 0.20$	$a = 0.22,\ R_{\mathrm{pop}} = 1.23\,10^{-4}$
$d56, T30$	0.16	$0.06 \leq a \leq 0.16$	$a = 0.20,\ R_{\mathrm{pop}} = 3.59\,10^{-2}$

For the progress-level Boltzmann exponent rule, the terminal plateau begins at $\theta_{\mathrm{prog}} \simeq 0.30$ on D40 and $\theta_{\mathrm{prog}} \simeq 0.65$ on D56, with exponent-area budgets about 6.0 and 5.2-5.5, respectively. Thus there is no contradiction: the pure tail asymptotic favors low a, while the finite BBP/visible-group problem chooses the largest exponent that keeps the moving visible group visible.

9 Finite-horizon adjoint correction

The local equation $L_S = L_B$ is adiabatic. A finite training horizon adds an adjoint weight. This is already visible in the reduced power-law growth-rate model. Let $E_i(t)$ be the remaining squared error of mode i, and suppose

$$\dot{E}_i(t) = -2c_i(a(t), t)E_i(t), \qquad c_i(a, t) = \omega_0(t)A(t)^a i^{-\gamma(a+1)}.$$

For terminal loss

$$\mathcal{J}_T = \sum_i \mu_i^2 E_i(T),$$

the adjoint satisfies

$$\dot{\lambda}_i(t) = 2c_i(a(t),t)\lambda_i(t), \qquad \lambda_i(T) = \mu_i^2.$$

Thus

$$\lambda_i(t)E_i(t) = \mu_i^2 E_i(T),$$

and the interior stationarity condition is

$$\boxed{\sum_i \mu_i^2 E_i(T) c_i(a(t),t)\big(\log A(t) - \gamma \log i\big) = 0.}$$

Equivalently,

$$\boxed{\log A(t) = \gamma \frac{\sum_i \mu_i^2 E_i(T) c_i(a(t),t) \log i}{\sum_i \mu_i^2 E_i(T) c_i(a(t),t)}.}$$

This is the finite-horizon version of logarithmic matching. Compared with the local BBP equation, the weights are not only the currently visible signal weights; they are tilted by the terminal residual $E_i(T)$. This explains why a local Hessian root can correctly detect the phase transition but still switch too early for a finite T.

For a blockwise constant schedule $a(t) = a_\ell$ on time blocks $[t_\ell, t_{\ell+1})$, let

$$\Delta t_\ell = t_{\ell+1} - t_\ell, \qquad A_\ell = A(t_\ell),$$

and define

$$c_{\ell i}(a_\ell) = \omega_{0,\ell} A_\ell^{a_\ell} i^{-\gamma(a_\ell+1)}.$$

Then

$$E_i(T) = \exp\left\{-2\sum_\ell \Delta t_\ell c_{\ell i}(a_\ell)\right\}$$

in the normalized reduced model, and

$$\partial_{a_\ell} \mathcal{J}_T = -2\Delta t_\ell \sum_i \mu_i^2 E_i(T) c_{\ell i}(a_\ell)\big(\log A_\ell - \gamma \log i\big).$$

Thus the KKT rule on the interval $a_\ell \in [0,1]$ is

$$R_\ell(a_\ell) := \log A_\ell - \gamma \frac{\sum_i \mu_i^2 E_i(T) c_{\ell i}(a_\ell) \log i}{\sum_i \mu_i^2 E_i(T) c_{\ell i}(a_\ell)}.$$

At an interior block $R_\ell(a_\ell) = 0$. At the lower endpoint

$$a_\ell = 0 \quad \Longleftrightarrow \quad R_\ell(0) \le 0,$$

while at the upper endpoint

$$a_\ell = 1 \quad \Longleftrightarrow \quad R_\ell(1) \ge 0.$$

Consequently the finite-horizon optimum is generically bang-bang by blocks: large early singular scale scales A_ℓ select $a_\ell = 1$, while later blocks whose effective singular scale is below the terminal weighted spectral group select $a_\ell = 0$. Genuine interior block values appear only when $\log A_\ell$ crosses the adjoint-weighted logarithmic index inside $[0,1]$.

9.1 Bulk-corrected Pontryagin check

The finite-horizon equation used in the final dynamic BBP exponent rule includes the bulk logarithmic scale. The reduced experiment uses the following modified growth rate. The growth rate is modified to

$$c_i(a,t) = \exp\{a \log A(t) - \gamma(a+1)\log i - R_B(a,t)\}, \qquad \partial_a R_B(a,t) = L_B(a,t).$$

Thus the exact derivative of the Hamiltonian contains

$$\partial_a \log c_i(a,t) = \log A(t) - \gamma \log i - L_B(a,t).$$

The interior Pontryagin equation is therefore

$$\boxed{\log A(t) = \gamma \frac{\sum_i \mu_i^2 E_i(T) c_i(a(t),t) \log i}{\sum_i \mu_i^2 E_i(T) c_i(a(t),t)} + L_B(a(t),t).}$$

If

$$\Psi_{\mathrm{adj}}(a,t) = \log A(t) - L_B(a,t) - \gamma \frac{\sum_i \mu_i^2 E_i(T) c_i(a,t) \log i}{\sum_i \mu_i^2 E_i(T) c_i(a,t)},$$

then on an interior smooth segment

$$\dot{a}(t) = -\frac{\partial_t \Psi_{\mathrm{adj}}(a(t),t)}{\partial_a \Psi_{\mathrm{adj}}(a(t),t)}.$$

At the endpoints this ODE is replaced by the projected KKT inclusion

$$a = 0 \Rightarrow \Psi_{\mathrm{adj}}(0,t) \le 0, \qquad a = 1 \Rightarrow \Psi_{\mathrm{adj}}(1,t) \ge 0.$$

The main numerical check uses $d = 10^5$, $\gamma = 1.5$, $T = 30$, 48 blocks, and

$$A(t) = 0.35 + 80e^{-4.4t/T}.$$

Three independent L-BFGS starts converged to the same schedule. The optimum is

$$a(t) = 1 \quad \text{for the first eight blocks } (t \lesssim 4.69), \qquad a(t) = 0 \quad \text{after } t \simeq 5.31.$$

The value is $2.1916036\,10^{-3}$, while the best constant is $a \simeq 0$ with value $2.5344592\,10^{-3}$, and the greedy local Markov root gives $2.8567554\,10^{-3}$. Thus the finite-horizon adjoint schedule improves over the best constant by a factor 1.1564 and over the local Markov root by a factor 1.3035. The finite-difference gradient error is $7.22\,10^{-14}$, and the maximum KKT violation is zero.

10 Power-law block constants

In the power-law regime,

$$g_i(t) \simeq A(t) i^{-\gamma}, \qquad \omega_i(t) \simeq \omega_0(t) i^{-\gamma}.$$

For a block of active indices

$$I_\ell = \{J_\ell, \dots, J_{\ell+1} - 1\},$$

the signal logarithmic mean becomes

$$L_{S,\ell}(a,t) = \log A(t) - \gamma \frac{\sum_{i \in I_\ell} i^{-\gamma(a+1)} \log i}{\sum_{i \in I_\ell} i^{-\gamma(a+1)}}.$$

The block exponent a_ℓ is therefore defined by

$$\boxed{\log A(t_\ell) - \gamma \frac{\sum_{i\in I_\ell} i^{-\gamma(a_\ell+1)} \log i}{\sum_{i\in I_\ell} i^{-\gamma(a_\ell+1)}} = L_B(a_\ell, t_\ell).}$$

Equivalently, on a geometric block $J_{\ell+1} = e^\Delta J_\ell$, put

$$q = \gamma(a+1), \qquad m_\ell(a) = \frac{\int_{J_\ell}^{J_{\ell+1}} x^{-q} \log x \, \mathrm{d}x}{\int_{J_\ell}^{J_{\ell+1}} x^{-q} \, \mathrm{d}x}.$$

Then the block equation is simply

$$\log A(t_\ell) - \gamma m_\ell(a_\ell) = L_B(a_\ell, t_\ell).$$

Writing $\beta = 1 - q$, the intra-block logarithmic average is explicit:

$$m_\ell(a) = \log J_\ell + h_\Delta(q), \qquad h_\Delta(q) = \frac{e^{(1-q)\Delta}((1-q)\Delta - 1) + 1}{(1-q)(e^{(1-q)\Delta} - 1)}$$

for $q \neq 1$, and $h_\Delta(1) = \Delta/2$. Hence the block condition can also be written as

$$\log A(t_\ell) - \gamma \log J_\ell - \gamma h_\Delta(\gamma(a_\ell + 1)) = L_B(a_\ell, t_\ell).$$

For a narrow block, $h_\Delta(q) = \Delta/2 + O(\Delta^2)$, and the center of the active block is the scale where

$$A(t_\ell) J_\ell^{-\gamma} \simeq \sigma_B(t_\ell).$$

The exponent a_ℓ then only controls the intra-block tilt. This is the physical staircase: the BBP visible group chooses the block, and a_ℓ equalizes the logarithmic signal and bulk scales inside that block.

Inside the block continuum, the ODE becomes

$$\boxed{\dot a_\ell(t) = \frac{\partial_t L_B(a_\ell, t) - \partial_t \log A(t) + \gamma \, \dot m_\ell(t)}{\gamma^2 \operatorname{Var}_{\ell,a_\ell}(\log i) - 2 \operatorname{Var}_{Q_{a_\ell,t}}(\log s)}.}$$

Here $\operatorname{Var}_{\ell,a}(\log i)$ is the variance of $\log i$ under the density proportional to $i^{-\gamma(a+1)}$ on the current block. The term $\dot m_\ell(t)$ records the motion of the block endpoints; it vanishes for frozen blocks and is explicit when the endpoints are the BBP visible groups. Equivalently,

$$\operatorname{Var}_{\ell,a}(\log i) = \frac{\int_{J_\ell}^{J_{\ell+1}} x^{-q} (\log x)^2 \, \mathrm{d}x}{\int_{J_\ell}^{J_{\ell+1}} x^{-q} \, \mathrm{d}x} - m_\ell(a)^2 = \partial_q^2 \log \int_{J_\ell}^{J_{\ell+1}} x^{-q} \, \mathrm{d}x.$$

The Markovian block policy is

$$a(t) = a_\ell, \qquad t \in [\tau_\ell, \tau_{\ell+1}),$$

where τ_ℓ is the time at which the BBP visible group enters block I_ℓ.

Proposition 10.1 (Convergence of the block discretization)**.** *Assume the bulk law, $A(t)$, and the BBP visible group vary continuously on the ODE time scale, and take a geometric partition*

$$J_{\ell+1} = e^\Delta J_\ell.$$

Let $a_\Delta(t)$ be the corresponding block policy. If the stationarity equation has a unique stable root, then

$$a_\Delta(t) \longrightarrow a^\star_{\mathrm{loc}}(t)$$

locally uniformly away from BBP contact times as $\Delta \downarrow 0$.

Proof. On each block, the discrete weighted logarithmic average is a Riemann sum for the continuum power-law average. Continuity of $A(t)$, $\nu_{B,t}$, and the visible group endpoints gives convergence of the stationarity residual. Stability of the root transfers residual convergence into convergence of a_Δ. □

11 Physical prediction

The optimal Markovian policy has three regimes.

(i) *BBP birth.* When a new group is just separating from the bulk, $a^\star(t)$ rises. This avoids giving equal weight to uncaptured bulk directions.

(ii) *Stable visible-group transport.* Once a block is visible, $a^\star(t)$ drops toward a_{safe}. In the ideal noiseless closure, this means nearly Muon/sign-like updates.

(iii) *Power-law cascade.* As the visible group moves to weaker modes, the schedule is approximately block-constant. The block boundaries are the dynamic BBP contact/reentry times, and the block values solve the logarithmic matching equation above.

Thus the prediction is not a smooth arbitrary curve. It is a spectral staircase:

$$\boxed{a^\star(t) \approx a_0 \mathbf{1}_{[\tau_0,\tau_1)} + a_1 \mathbf{1}_{[\tau_1,\tau_2)} + a_2 \mathbf{1}_{[\tau_2,\tau_3)} + \cdots ,}$$

with a continuum limit controlled by the power-law spectrum.

12 Relation with the power-law phase retrieval ODE

The power-law phase retrieval analysis of Braun–Loureiro–Minh–Imaizumi [8] contains the same structural effect in another basis. In their Phase III, each coordinate error follows the tail-learning approximation

$$e_i(T_2 + \tau) \simeq e_i(T_2) e^{-8\lambda_i \tau}.$$

For the target treated in this companion, the input covariance is isotropic and the power law is not in λ_i but in the multi-index teacher amplitudes μ_i. The replacement is therefore

$$8\lambda_i \quad \rightsquigarrow \quad \omega_i(t) g_i(t)^a \simeq \omega_0(t) A(t)^a i^{-\gamma(a+1)}.$$

The infinite hierarchy is the same kind of object: a continuum of growth rates, ordered by a power law, whose visible group is advanced by the current spectral preconditioner. Later, when the data covariance is also power-law, both effects multiply: the growth rate becomes

$$\omega_i(t) g_i(t)^a \quad \rightsquigarrow \quad \lambda_i \, \omega_i(t) g_i(t)^a$$

up to the Volterra correction of the anisotropic-data theory.

13 Connection with the dynamic BBP notes

The dynamic BBP formulas give the contact times τ_ℓ. In the scalar interval, a residual branch for mode i reenters when

$$\mu_i (1 - r_i(t)) \simeq c_i / \sqrt{\alpha}.$$

For $\mu_i = \mu_0 i^{-\gamma}$, these reentry times order the modes by the power-law index and generate the blocks I_ℓ. The Hessian analysis gives the bulk law $\nu_{B,t}$, the singular scale $g_i(t)$, and the BBP visible group; the Markovian exponent rule only solves the scalar stationarity equation at the current ODE state.

Remark 13.1. *This also explains why the best fixed exponent and the best Markovian exponent can look different. The fixed exponent solves a global compromise over all visible groups; the Markovian policy solves the correct local compromise at the current visible group. In a dense power-law cascade, the local compromises form a block discretization of the spectral continuum.*

14 Numerical validation hierarchy

The reduced theory above makes six quantitative predictions. They are ordered from the least random to the most empirical. This separates failures of the power-law control reduction from failures of finite-sample spectral estimation.

(E1) *Constant exponent tail law.* For fixed a, simulate the exact growth-rate model

$$\mathcal{E}_T(a) = \sum_{i=1}^{d} i^{-2\gamma} \exp\{-2Ti^{-\gamma(a+1)}\}.$$

The measured log–log slope should converge to

$$-\frac{\mathrm{d}\log \mathcal{E}_T(a)}{\mathrm{d}\log T} \to \kappa(a) = \frac{2\gamma - 1}{\gamma(a+1)}.$$

(E2) *Best constant exponent.* On a grid of exponents $a \in [a_{\mathrm{safe}}, 1]$, the minimizer of $\mathcal{E}_T(a)$ should drift toward a_{safe} as T enters the stable tail interval.

(E3) *ODE for $a^\star(t)$.* Choose smooth functions $A(t), \sigma_B(t)$ and a moving power-law block. Compute $a^\star(t)$ by solving $L_S = L_B$, then compare finite differences of $a^\star(t)$ with the ODE displayed above.

(E4) *Block discretization.* Replace the continuum block integrals by geometric sums with mesh Δ. The root residual and the recovered $a_\Delta(t)$ should converge as $\Delta \downarrow 0$, away from contact times.

(E5) *Full SGD and independent Hessian layer.* Train the isotropic Gaussian multi-index model with fresh mini-batches and spectral exponent a. Compare fixed exponents at equal global update norm. On selected checkpoints, draw a fresh Hessian sample, compute extreme eigenpairs, and measure their teacher-subspace overlap against the random baseline k/d.

(E6) *Finite-horizon adjoint schedules.* Optimize the reduced block objective

$$\mathcal{J}_T(a_0, \ldots, a_{B-1}) = \sum_{i=1}^{d} i^{-2\gamma} \exp\left\{-2 \sum_{\ell=0}^{B-1} \Delta t_\ell A_\ell^{a_\ell} i^{-\gamma(a_\ell+1)}\right\}$$

and verify the analytic gradient above by finite differences. The recovered schedule should be block-constant, with the number of early $a_\ell = 1$ blocks increasing as the initial singular scale amplitudes increase.

Together these tests validate the power-law control reduction. After the full-SGD independent-Hessian check, the important measured objects are not new hyperparameters but the deterministic spectral state itself: the numerical $A(t)$, the growth scale $\omega_0(t)$, and the bulk law $\nu_{B,t}$ read from Hessian snapshots. These are then inserted directly into the Markovian ODE for $a^\star(t)$.

Finite simulations show the expected horizon dependence. For

$$(d, k, P) = (256, 20, 16),$$

the best fixed exponent is still around $a = 0.75$ at $T = 700$, indicating a pre-tail/visible-group-resolution regime. At $T = 1200$, the best fixed exponent moves to $a = 0$. The independent Hessian extremes also become teacher-aligned: for instance the maximum teacher-subspace overlap on the top side grows from the random baseline $k/d \simeq 0.078$ to about 0.69 in the $a = 0, T = 1200$ case.

Adaptive tests separate myopic risk descent from spectral control. A one-step teacher-informed benchmark which picks the best immediate population-risk decrease over $\{0, 0.25, 0.5, 0.75, 1\}$ does not beat the best fixed exponent. Nor do simple two-stage switches $a_{\text{high}} \to a_{\text{low}}$ in the tested intervals. Thus the nontrivial Markovian exponent rule is not a myopic risk minimizer; it must use the Hessian/BBP quantities L_S, L_B, or the full Volterra adjoint, to place the block transitions.

An offline Hessian-to-exponent check recovers the qualitative local shape: the empirical root of $L_S(a,t) - L_B(a,t)$ is high at initialization and drops to $a = 0$ once teacher-aligned Hessian extremes are visible. But the same check also shows the finite-horizon correction: on the $d = 256, T = 700$ trajectory, switching from $a = 0.75$ to $a = 0$ at the local root is worse than keeping $a = 0.75$. Thus $L_S = L_B$ is the correct adiabatic/local equation, while the true finite-horizon optimum must include the value-gradient or Volterra-adjoint term from the exact control statement.

The reduced finite-horizon adjoint check validates this last point at scale $d = 10^6$. With $\gamma = 0.85$ and block amplitudes

$$(8, 5, 3, 1.5, 0.9, 0.6),$$

the optimizer returns

$$(a_0, \ldots, a_5) = (1, 0, 0, 0, 0, 0),$$

with terminal objective 0.1717938958, slightly better than fixed $a = 0$ at 0.1730092899 and much better than fixed $a = 1$ at 0.2831621867. When the early amplitudes are increased to

$$(100, 50, 20, 5, 1, 0.5),$$

the optimum becomes

$$(a_0, \ldots, a_5) = (1, 1, 0, 0, 0, 0),$$

with objective 0.0979337937, better than both fixed $a = 0$ (0.1730092899) and fixed $a = 1$ (0.1130425485). The same two-high-block structure persists for $\gamma = 0.65$ and $\gamma = 1.1$. The adjoint gradient finite-difference error is below 10^{-10} in these checks. This is the spectral staircase predicted by the KKT rule.

The same staircase can be injected into the full causal SGD experiment through piecewise-constant policies. On $d = 192, k = 16, P = 12, T = 800$, fixed $a = 0$ remains best and the schedule $(1, 0, 0, 0, 0, 0)$ is close but worse, which is consistent with an already tail-dominated interval. On $d = 256, k = 20, P = 16, T = 700$, the best fixed exponent remains $a = 0.75$, and high-to-low block schedules improve only when the high plateau is kept late, so the descent-to-Muon time has not yet arrived. On the longer $d = 256, k = 20, P = 16, T = 1200$ interval, a short intermediate birth layer becomes visible: over five seeds, $a = 0.5$ for 200 steps followed by $a = 0$ reaches mean final risk 0.0128097425, compared with 0.0128120038 for fixed $a = 0$. This gain is small relative to seed variability, but its paired sign is consistent with the finite-horizon adjoint picture. This points to a causal exponent rule that estimates the adjoint-weighted visible group from empirical Hessian observables.

A systematic one-switch search on the same $d = 256, k = 20, P = 16, T = 1200$ interval sharpens this statement. Scanning

$$a_{\text{start}} \in \{0.25, 0.5, 0.75, 1\}, \qquad \tau \in \{50, 100, \ldots, 350\},$$

for policies $a(t) = a_{\text{start}} \mathbf{1}_{t \le \tau}$, the best two-seed candidate is $a_{\text{start}} = 0.5, \tau = 150$. Rechecking the leading candidates on five seeds gives the best mean risk for

$$a(t) = 0.25\, \mathbf{1}_{t \le 250},$$

with mean final risk 0.0128017603, compared with 0.0128120038 for fixed $a = 0$. The neighboring policy

$$a(t) = 0.5\, \mathbf{1}_{t \le 150}$$

is nearly tied at 0.0128035261. Thus the finite-horizon optimum is better described by a small initial exponent area,

$$\int_0^T a(t)\ \mathrm{d}t \simeq 60,$$

followed by $a = 0$, rather than by a long high-exponent plateau. This is the finite-T form of the spectral staircase.

A causal teacher-informed adjoint exponent rule was also evaluated using true teacher overlaps and current gradient signal singular scales. It separates two effects: without a bulk penalty the rule chooses $a \simeq 1$ too often, while the fully normalized spectral denominator collapses the rule to $a = 0$. The missing coefficient is therefore the effective bulk/RFA penalty in the Hamiltonian, and the natural estimator of this coefficient uses independent Hessian snapshots.

The remaining bulk penalty can be measured from Hessian exponent-rule grids. Write the generalized local stationarity equation as

$$L_S(a,t) = \beta_{\mathrm{RFA}}(t) L_B(a,t).$$

The observable critical curve is therefore

$$\beta_{\mathrm{crit}}(a,t) = \frac{L_S(a,t)}{L_B(a,t)}.$$

Along the empirical best one-switch trajectory

$$a(t) = 0.25\,\mathbf{1}_{t\le 250},$$

independent Hessian snapshots give

$$\beta_{\mathrm{crit}}(0.25, 250) = 0.7037497781, \qquad \beta_{\mathrm{crit}}(0, 300) = 0.7005188031.$$

Thus the inferred crossing is

$$\boxed{\beta_{\mathrm{cross}} \simeq 0.7021.}$$

This gives a concrete full-model estimate of the bulk/RFA coefficient: the small positive exponent layer persists while $\beta_{\mathrm{RFA}}(t) \gtrsim \beta_{\mathrm{crit}}(a,t)$, and the exponent rule switches to $a = 0$ after this crossing. In a fully causal implementation, $\beta_{\mathrm{RFA}}(t)$ should be estimated from Hessian observables rather than from the post-hoc one-switch optimum.

As a causal closure test, we fixed the measured target $\beta_{\mathrm{target}} = 0.705$ and ran the feedback rule

$$a(t) = 0.25 \quad \text{until} \quad \bar{r}(t) \ge 0.15 \text{ and } \beta_{\mathrm{crit}}(0.25, t) \le \beta_{\mathrm{target}}, \qquad a(t) = 0 \text{ afterwards.}$$

Here $\bar{r}(t) = k^{-1}\sum_i r_i(t)$, and β_{crit} is computed from a fresh independent Hessian sample at control checkpoints. On the same $d = 256, k = 20, P = 16, T = 1200$ experiment, five seeds give switch times

$$250,\ 300,\ 250,\ 250,\ 350$$

and mean final risk 0.0128056277. This is close to the post-hoc $0.25 \to 0$ at 250 rule (0.0128017603), and below fixed $a = 0$ (0.0128120038) on the same seeds. Thus the Hessian observable already gives a causal map from dynamic BBP information to an exponent schedule $a(t)$. The remaining calibration is the estimation of β_{target} without using the post-hoc one-switch grid.

The next closure removes this imported scalar. During the pre-capture phase we record the Hessian critical values

$$\mathcal{H}_{\mathrm{pre}}(t) = \{\beta_{\mathrm{crit}}(0.25, t_j) : t_j < t,\ \bar{r}(t_j) < 0.15\}.$$

After at least three such snapshots, define the endogenous plateau estimate

$$\widehat{\beta}_{\mathrm{pre}}(t) = \operatorname{median} \mathcal{H}_{\mathrm{pre}}(t).$$

The fully causal switch rule is then

$$a(t) = 0.25 \quad \text{until} \quad \bar{r}(t) \geq 0.15 \text{ and } \beta_{\mathrm{crit}}(0.25, t) \leq \widehat{\beta}_{\mathrm{pre}}(t), \qquad a(t) = 0 \text{ afterwards.}$$

On the same five seeds this gives switch times

$$250,\ 300,\ 250,\ 250,\ 250$$

and mean final risk 0.0128022611. The paired excess over the post-hoc $0.25 \to 0$ at 250 rule is only 5.0×10^{-7}, while the paired gain over fixed $a = 0$ is 9.74×10^{-6}. A lower-quartile version of the same rule is too conservative and sometimes misses the switch, giving 0.0128680552. Thus the operative signal is a drop below the central pre-capture bulk/RFA plateau, not a finely tuned absolute value of β.

This closes the tested Markovian $a(t)$ layer in the following precise sense: in the full-SGD multi-index phase-retrieval experiments above, the best constant exponent is horizon dependent, the finite-horizon optimum is a short positive-exponent layer followed by Muon, and the layer endpoint is recovered causally from independent Hessian BBP/RFA observables without using the post-hoc one-switch grid. The statement is experimental and finite-dimensional; the corresponding theorem is the Schur/RFA transport result stated later.

14.1 Hessian reconstruction of the Markovian ODE

The causal switch rule is a projected version of the smooth implicit ODE. The Hessian Schur data reconstructs

$$\Psi_\beta(a, t) = L_S(a, t) - \beta L_B(a, t).$$

It then tracks the clipped root $a^\star(t)$ of $\Psi_\beta(a, t) = 0$ and, when the root is interior, checks the implicit equation

$$\boxed{\dot{a}^\star(t) = -\frac{\partial_t \Psi_\beta(a^\star(t), t)}{\partial_a \Psi_\beta(a^\star(t), t)}.}$$

The same check also applies the maturity projection used by the causal exponent rule:

$$\bar{r}(t) \geq 0.15, \qquad \beta_{\mathrm{crit}}(0.25, t) \leq \widehat{\beta}_{\mathrm{pre}}.$$

For the Hessian path used in this comparison, the pre-capture median is

$$\widehat{\beta}_{\mathrm{pre}} = 0.7005547214.$$

The first-passage rule fires at checkpoint 300, exactly the first checkpoint after the tested post-hoc policy $a(t) = 0.25\mathbf{1}_{t \leq 250}$ has switched to $a = 0$. Thus the checkpoint-level policy error is zero. Repeating the check with the fixed causal coefficient $\beta = 0.705$ gives the same first-passage checkpoint.

The smooth ODE part is more sparsely tested on this trajectory. There are only two true interior roots; the RMS discrepancy between finite differences of $a^\star(t)$ and the implicit ODE right-hand side is

$$3.29 \cdot 10^{-4}$$

per training step for $\widehat{\beta}_{\mathrm{pre}}$, and

$$3.49 \cdot 10^{-4}$$

for $\beta = 0.705$. All other checkpoints are lower-bound, upper-bound, or no-visible-signal segments. Therefore this full-SGD evidence validates the projected Markovian first-passage law. A dense

smooth ODE validation requires more frequent independent Hessian snapshots in the interior-root regime.

A denser validation uses checkpoints every 50 steps on the same seed-0 trajectory. It separates the projected switch law from the smooth interior ODE:

grid	$\widehat{\beta}$	switch	mature ODE pts	mature RMS	median abs
$\alpha = 4,\ \theta = 0.195$	0.7005547	300	13	$1.12 \cdot 10^{-2}$	$3.90 \cdot 10^{-3}$
$\alpha = 4,\ \theta = 0.12$	0.7142217	250	12	$1.96 \cdot 10^{-3}$	$8.30 \cdot 10^{-4}$
$\alpha = 8,\ \theta = 0.12$	0.7030975	none	17	$2.00 \cdot 10^{-4}$	$4.55 \cdot 10^{-5}$
$\alpha = 8,\ \theta = 0.195$	0.7098572	none	17	$2.32 \cdot 10^{-4}$	$5.11 \cdot 10^{-5}$

Here θ is the teacher-overlap threshold defining the visible Hessian signal set. The fresh-batch protocol grid $(\alpha, \theta) = (4, 0.12)$ recovers the first-passage switch at 250. The stabilized $\alpha = 8$ grids give the most stable mature ODE check: once the visible signal set is stable, the implicit derivative formula is accurate at the 10^{-4} per-step scale.

The scalar threshold $\widehat{\beta}_{\mathrm{pre}}$ is not estimator invariant. Changing α and the visibility rule changes the calibrated β enough that the median-threshold switch may disappear, even though the mature ODE identity becomes cleaner. Thus any coefficient that survives the large dimension limit must be an invariant RFA/visibility normalization $\beta_{\mathrm{RFA}}(t)$, or equivalently an exponent rule using the full signal/bulk curves rather than one scalar threshold.

The natural residual-visible-group modification was also tested, in which each visible signal eigenpair is weighted by $1 - r_i(t)$, or by $(1 - r_i(t))^2$, where i is the teacher mode carrying the largest overlap of the eigenvector. This does not produce an invariant scalar coefficient. On the $(\alpha, \theta) = (4, 0.12)$ grid, residual weighting still fires at 250 but leaves only six mature interior ODE points, with mature RMS $2.68 \cdot 10^{-3}$. Residual-squared weighting leaves five mature points, with mature RMS $3.57 \cdot 10^{-3}$. On the stabilized $(\alpha, \theta) = (8, 0.12)$ grid, residual weighting leaves essentially one interior mature point and no switch. Hence the visible-group residual is useful for interpretation, but it is not an invariant scalar replacement inside L_S/L_B. The normalization must instead come from a resolvent/frame-averaging limit, or from a full adjoint exponent rule using the whole signal/bulk curve.

15 Scalar/matrix check and cavity status

The compact exact check

$$d = 24, \quad p = 48, \quad k = 6, \quad n = 144, \quad \gamma = 1.5$$

was evaluated with exact Hessians at every recorded checkpoint. It checks the matrix identities and the scalar reduction side by side for SGD, Muon, fixed $a = 0.25$, causal Boltzmann $a(t)$, and population-control baselines.

The finite matrix layer is verified to roundoff on the recorded trajectories. The spectral filter identity has median relative error between $4.3 \cdot 10^{-16}$ and $1.0 \cdot 10^{-15}$ across the cases. The analytic $\dot{q}$ identity agrees with finite differences at $2.1 \cdot 10^{-8}$ for SGD and between $9.5 \cdot 10^{-7}$ and $8.3 \cdot 10^{-6}$ for the Muon-type empirical cases. The risk decomposition remains at 10^{-16} scale. Thus the finite-dimensional algebra used by the reduced theory is not a modelling assumption in these experiments; approximations enter when finite spectral objects are replaced by limiting bulk laws.

The scalar singular scale approximation is correct as an asymptotic visible-group law, but

not as an exact finite empirical identity. The median relative error of $\sigma_{\text{obs}} \simeq \sqrt{q}\,\rho$ is

case	final risk	scalar error	λ_{BBP} error	captured modes
SGD ($a = 1$)	1.18972	0.1057	–	0
Muon ($a = 0$)	0.69359	0.3364	0.4044	5
fixed $a = 0.25$	0.85489	0.3135	0.8302	2
causal Boltzmann	0.50039	0.1980	0.6478	6
population $a = 0.25$	$4.28 \cdot 10^{-8}$	0.0579	$1.47 \cdot 10^{-4}$	6

This is the expected hierarchy. In the population or fully decoupled power-law interval the scalar singular scale is excellent; in finite empirical Muon it needs a cavity/visibility factor. Therefore the scalar power-law calculus is the right object for the optimal visible group and block staircase, while the exact matrix equations remain the reference for finite d and for the Hessian exponent rule.

The Hessian histograms reproduce the dynamic-BBP picture from the Ben Arous–Gheissari–Huang–Jagannath effective-spectral framework [1]. SGD stays in a broad nearly stationary bulk and captures no mode. Muon tightens the bulk near zero, creates a positive teacher-aligned edge, and captures five modes. The causal Boltzmann schedule ends with all six modes captured and lower final risk. This is qualitatively the same bulk/outlier movie as the ResNet and logistic-mixture histograms: a bulk close to zero plus a small number of moving outliers, except that here the outlier motion is controlled by the Muon exponent a.

The more informative readout is the modewise outlier trajectory. For each teacher mode we solve the weight-BBP equation

$$1 = q_i(t) m_{B_t}(\lambda_i(t))$$

and compare the predicted root to the eigenvalue whose eigenvector has maximal teacher overlap. On the compact check, SGD has no predicted and no visible weight outlier. Muon has six predicted roots but only two visible teacher-overlap branches at final time; fixed $a = 0.25$ has six predicted roots and one visible branch; causal Boltzmann has six predicted roots and two visible branches. Conditional on visibility, the root is accurate: the median relative errors are 0.0818 for Muon, 0.0514 for fixed $a = 0.25$, and 0.1252 for causal Boltzmann. Thus the discrepancy is not caused by false BBP roots but by delayed-overlap outliers: branches above $q_c(t)$ whose eigenvector mass has not yet localized on the teacher mode.

The limiting BBP statement is formulated with a deterministic bulk law, not with the raw empirical finite bulk. As a first deterministic comparison, the same check was repeated with a scalar Marchenko–Pastur approximation to the residual bulk: the edge is calibrated from the residual bulk mean, and the root uses the closed MP Stieltjes transform. This scalar comparison is weaker than the full MDE/RFA Schur law, but it separates bulk calibration errors from branch-tracking errors. In the compact check the empirical edge is above the simple MP edge by a finite factor, with median ratios

$$1.59 \quad (a = 0), \qquad 1.66 \quad (a = 0.25), \qquad 1.40 \quad (\text{causal Boltzmann}).$$

Nevertheless, conditional on visible teacher overlap, the MP-root errors are small:

$$0.0676 \quad (a = 0), \qquad 0.0594 \quad (a = 0.25), \qquad 0.0558 \quad (\text{causal Boltzmann}).$$

At final time the MP check sees $6, 5, 6$ predicted roots respectively, but only $2, 2, 2$ visible roots. This confirms that the empirical edge was only a finite observable. The limiting object is the deterministic MDE/RFA bulk together with an overlap or residue law deciding which roots are already visible.

The gradient singular scale gives the same message. The raw scalar prediction $\sqrt{q_i(t)}\rho_i(t)$ overpredicts the empirical Muon singular directions by a time-dependent factor. The ratio plots show that a scalar cavity factor c_t captures the leading modes reasonably, while late weak

modes need a separate soft-BBP visibility correction. This is the operational form of the finite-d correction:

$$\sigma_i^{\mathrm{obs}}(t) \simeq c_t \sqrt{q_i(t)}\rho_i(t)\,\chi_i^{\mathrm{BBP}}(t).$$

The remaining theoretical distinction is train data versus fresh Hessian data. The theorem of Ben Arous–Gheissari–Huang–Jagannath proves the effective bulk and outlier equations for self-coupled Hessian blocks evaluated at fixed parameters, and its dynamic interpretation is most direct with test/fresh data. Their paper also shows numerical train/test agreement, but does not by itself give the leave-one-out theorem needed for a Muon trajectory reusing exactly the training samples. The compact check sees the same split. The selected Boltzmann exponent is identical on train and fresh checks ($a_{\mathrm{train}} = a_{\mathrm{fresh}} = 0.124$), and the exact gradient identity has zero train/fresh discrepancy at recorded numerical precision. The finite singular scale amplitude is different, however: for causal Boltzmann the median train/fresh gap in the scalar singular scale relative error is $9.9 \cdot 10^{-2}$, while for pure Muon it is $1.31 \cdot 10^{-1}$. This is the empirical signature of a cavity amplitude correction.

In this notation the RFA statement is precisely

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

The corresponding theorem has a standard leave-one-out shape. One constructs $W_t^{(-\ell)}$ and $\Gamma_s^{(\ell)}$, replaces the self-dependent weight $\Phi(W_t^\top x_\ell, \Theta^\top x_\ell)$ by the cavity weight computed from $W_t^{(-\ell)}$, and then applies fluctuation averaging to the resolvent Hessian. The finite evidence above isolates the role of this probabilistic step: the matrix identities are exact, while the observed scalar error is a cavity amplitude/visibility correction rather than a new dynamical law.

15.1 Finite Schur calibration of the weight outlier curves

The scalar MP/Stieltjes comparison gives the simplest deterministic bulk picture, but it is not the right object for finite-size superposition of the predicted and observed weight outlier curves. At finite d, write the weight covariance in teacher/bulk coordinates as

$$S(t) = \begin{pmatrix} A(t) & B(t) \\ B(t)^\top & C(t) \end{pmatrix}.$$

For $\lambda > \lambda_{\max}(C(t))$, the exact finite cavity equation is

$$\det\{\lambda I - A(t) - B(t)(\lambda I - C(t))^{-1} B(t)^\top\} = 0.$$

This is the finite-rank Schur complement form of the BBP equation. Once the finite blocks are saved, the Schur roots can be compared directly with empirical outliers. The scalar MP surrogate has visible-branch errors of a few percent, while the finite Schur roots coincide with the observed visible branches at numerical precision:

case	scalar MP median relerr	Schur median relerr	max Schur abs err
Muon $a = 0$	$5.20\,10^{-2}$	$2.30\,10^{-16}$	$3.33\,10^{-16}$
Muon* $a = 0.25$	$1.25\,10^{-1}$	$1.63\,10^{-16}$	$8.33\,10^{-17}$
causal Boltzmann	$3.87\,10^{-2}$	$1.92\,10^{-16}$	$2.22\,10^{-16}$

Thus the right hierarchy is:

MP/MDE bulk for the asymptotic root law,
finite Schur complement for exact finite overlays.

The apparent mismatch in the MP heatmap was therefore not a Muon failure and not a tracking failure. It came from replacing the finite cavity matrix by a scalar MP surrogate. The only

remaining visual distinction is visibility: Schur roots whose teacher residue is below the BBP overlap threshold are classified as hidden roots, not as failed predicted branches.

The same check was repeated at the larger size $d = 40, p = 80, k = 8, n = 240$. In this trajectory the causal Boltzmann exponent is time-dependent,

$$a(t) \in [0.16, 0.26], \qquad a(T) = 0.24.$$

The Schur localization remains at numerical precision:

case	MP err	Schur err	max abs err	visible
Muon $a = 0$	$9.49\,10^{-2}$	$3.19\,10^{-16}$	$3.89\,10^{-16}$	2
Muon* $a = 0.25$	$1.75\,10^{-1}$	$1.99\,10^{-16}$	$6.94\,10^{-17}$	1
causal Boltzmann	$1.32\,10^{-1}$	$1.32\,10^{-16}$	$8.33\,10^{-17}$	1

Thus the finite root-location problem is verified at both sizes. The remaining spectral question is the residue/visibility law for the tail roots, not the location of the roots themselves.

The same Schur localization was repeated at $d = 56, p = 112, k = 10, n = 336$. The conclusion persists:

case	MP err	Schur err	max abs err	visible
Muon $a = 0$	$1.15\,10^{-1}$	$1.61\,10^{-16}$	$1.39\,10^{-16}$	2
Muon* $a = 0.25$	$8.91\,10^{-2}$	$3.74\,10^{-16}$	$3.47\,10^{-17}$	2
causal Boltzmann	$3.32\,10^{-2}$	$2.25\,10^{-16}$	$5.55\,10^{-17}$	2

The finite Schur overlay is therefore verified at $d = 24, 40, 56$. The scalar MP surrogate remains at the percent to 10^{-1} error level because the empirical Muon paths still have non-negligible teacher–bulk coupling and residual-bulk edge ratios away from one. Thus the deterministic limiting object is the MDE/RFA limit of the full Schur function F_d, not a scalar MP edge alone. The same data also gives a spectrum-over-time visualization. Unlike the scalar MP heatmap, this figure tracks branches by Schur-root rank and then globally matches Schur roots to observed eigenvalues by a spectral-plus-teacher-profile assignment. This removes artificial switches which appear when nearly contacting roots are matched independently. Roots below the teacher visibility threshold are kept as faint hidden branches. On all Schur roots, not only the visible ones, the finite overlay is at machine precision:

dataset/case	final roots	final visible	median abs err	max abs err
$d40$, Muon* $a = 0.25$	2	1	$4.34\,10^{-18}$	$6.94\,10^{-17}$
$d40$, causal Boltzmann	2	1	$6.94\,10^{-18}$	$1.11\,10^{-16}$
$d56$, Muon* $a = 0.25$	4	2	$3.47\,10^{-18}$	$3.47\,10^{-17}$
$d56$, causal Boltzmann	4	2	$5.20\,10^{-18}$	$6.94\,10^{-17}$

The companion check plots the Schur–observed error, branch count, and residue/mass agreement over time for Muon* and causal Boltzmann. On visible points the errors are:

dataset/case	visible points	median visible abs err	max visible abs err
$d40$, Muon* $a = 0.25$	20	$2.17\,10^{-18}$	$6.94\,10^{-17}$
$d40$, causal Boltzmann	20	$1.30\,10^{-18}$	$8.33\,10^{-17}$
$d56$, Muon* $a = 0.25$	34	$3.90\,10^{-18}$	$3.47\,10^{-17}$
$d56$, causal Boltzmann	37	$6.94\,10^{-18}$	$5.55\,10^{-17}$

The final visible Schur residue also matches the observed teacher mass to the displayed precision: 0.643499/0.643499 and 0.664445/0.664445 at $d = 40$, and 0.571946/0.571946, 0.599816/0.599816 at $d = 56$, for Muon* and causal Boltzmann respectively. Thus the visually confusing switches in the MP plot are not contradictory branches. They are a mixture of a finite cavity correction and a mode-visibility threshold. Once roots are indexed by the finite Schur equation, the predicted and observed trajectories lie on top of each other.

15.2 Finite Schur residue and visibility

The same finite Schur complement also predicts the eigenvector residue. If λ is a root and u is the normalized teacher vector satisfying

$$\lambda u = \left(A + B(\lambda I - C)^{-1}B^\top\right) u,$$

then the full eigenvector has bulk component

$$v_B = (\lambda I - C)^{-1}B^\top u.$$

Therefore its total teacher-subspace mass is

$$\Omega_{\mathrm{Schur}}(\lambda, u) = \frac{1}{1 + \|(\lambda I - C)^{-1}B^\top u\|^2},$$

and its modewise teacher masses are

$$\Omega_{\mathrm{Schur}}(\lambda, u)\, u_i^2.$$

This is the finite-dimensional residue formula, equivalent to $(1 - u^\top F'(\lambda)u)^{-1}$ for $F(\lambda) = A + B(\lambda I - C)^{-1}B^\top$.

The same finite formula can be compared with the mode-by-eigenvector overlap matrix. At $d = 24, p = 48, k = 6, n = 144$, the residue is matched at numerical precision:

case	roots	visible	med res err	max res err	mode match
SGD $a = 1$	35	0	$3.33\,10^{-16}$	$1.12\,10^{-14}$	1.00
Muon $a = 0$	60	31	$6.66\,10^{-16}$	$2.70\,10^{-14}$	1.00
Muon* $a = 0.25$	42	20	$4.16\,10^{-16}$	$5.44\,10^{-15}$	1.00
causal Boltzmann	48	27	$5.13\,10^{-16}$	$6.61\,10^{-15}$	1.00

At $d = 40, p = 80, k = 8, n = 240$, the same conclusion holds:

case	roots	visible	med res err	max res err	mode match
SGD $a = 1$	21	0	$1.94\,10^{-16}$	$5.27\,10^{-16}$	1.00
Muon $a = 0$	70	29	$8.74\,10^{-16}$	$1.83\,10^{-14}$	1.00
Muon* $a = 0.25$	28	20	$4.44\,10^{-16}$	$1.78\,10^{-15}$	1.00
causal Boltzmann	35	20	$4.44\,10^{-16}$	$2.33\,10^{-15}$	1.00

At $d = 56, p = 112, k = 10, n = 336$, the residue check gives:

case	roots	visible	med res err	max res err	mode match
SGD $a = 1$	42	0	$3.33\,10^{-16}$	$1.47\,10^{-15}$	1.00
Muon $a = 0$	80	30	$1.80\,10^{-15}$	$4.43\,10^{-14}$	1.00
Muon* $a = 0.25$	64	34	$6.66\,10^{-16}$	$3.77\,10^{-15}$	1.00
causal Boltzmann	80	37	$4.72\,10^{-16}$	$4.00\,10^{-15}$	1.00

Thus the finite Schur/cavity identities hold for both eigenvalues and eigenvectors. The visibility count is simply the thresholded finite residue $\Omega_{\mathrm{Schur}} u_i^2$. The remaining asymptotic step is to pass from this finite Schur residue to a deterministic MDE/RFA residue uniformly along the Muon trajectory.

Theorem 15.1 (RFA reduction for Schur roots and residues)**.** *Fix a compact time interval and an interval I outside the limiting bulk. Let*

$$F_d(\lambda, t) = A_d(t) + B_d(t)(\lambda I - C_d(t))^{-1}B_d(t)^\top.$$

Assume that there is a deterministic matrix function $F(\lambda, t)$ such that, uniformly for $(\lambda, t) \in I \times [0, T]$,

$$\|F_d(\lambda, t) - F(\lambda, t)\| + \|\partial_\lambda F_d(\lambda, t) - \partial_\lambda F(\lambda, t)\| \longrightarrow 0$$

in probability. Suppose that a limiting branch

$$\lambda = \kappa_j(F(\lambda, t))$$

is simple and regular, in the sense that

$$1 - u_j(\lambda, t)^\top \partial_\lambda F(\lambda, t) u_j(\lambda, t) \neq 0,$$

and stays a positive distance away from all other roots and from the bulk edge. Then the corresponding finite Schur root $\lambda_{j,d}(t)$, teacher vector $u_{j,d}(t)$, and residue

$$\Omega_{j,d}(t) = \frac{1}{1 + \|(\lambda_{j,d} I - C_d(t))^{-1} B_d(t)^\top u_{j,d}(t)\|^2}$$

converge uniformly in probability to their deterministic limits

$$\lambda_j(t), \qquad u_j(t), \qquad \Omega_j(t) = \frac{1}{1 - u_j(t)^\top \partial_\lambda F(\lambda_j(t), t) u_j(t)}.$$

Consequently, for any visibility threshold θ, if

$$\inf_t \left| \Omega_j(t) u_{j,i}(t)^2 - \theta \right| > 0,$$

then the finite visibility indicator

$$\mathbf{1}\{\Omega_{j,d}(t) u_{j,d,i}(t)^2 \geq \theta\}$$

converges uniformly to its deterministic counterpart.

Proof. The uniform convergence of F_d and $\partial_\lambda F_d$ gives uniform convergence of the scalar functions

$$h_{j,d}(\lambda, t) = \lambda - \kappa_j(F_d(\lambda, t))$$

on each spectral patch where the j-th eigenvalue of F is simple. Kato perturbation theory gives uniform convergence of the associated spectral projectors and eigenvectors, up to sign. The regularity condition $\partial_\lambda h_j \neq 0$ lets the implicit function theorem transfer this convergence to the roots. The residue formula follows from

$$\partial_\lambda F_d(\lambda, t) = -B_d(t)(\lambda I - C_d(t))^{-2} B_d(t)^\top$$

and the normalization of the full eigenvector $(u, (\lambda I - C_d)^{-1} B_d^\top u)$. The visibility statement is then the continuous mapping theorem plus the displayed margin from the threshold. □

Thus the finite experiments have identified the exact object whose limit must be proved. The remaining RMT input is precisely the RFA/MDE convergence of F_d and its λ-derivative along the Muon path. Once that is available, the dynamic-BBP visibility law and the Markovian exponent rule use the deterministic weights $\Omega_j(t) u_{j,i}(t)^2$, rather than a raw empirical overlap threshold.

15.3 Fresh-gradient and leave-one-out experimental closure

The long empirical trajectories with one fixed training set left non-zero residual masses $\rho_i(t) = s_i - q_i(t)$. This does not contradict the population Muon ODE. A focused comparison keeps the same centered quadratic model, the same empirical gradient identity, and the same spectral Muon update, but changes only the source of the update gradient. The protocols are: fixed train data, fixed independent split data, rotating crossfit blocks, fresh batches, and the exact population gradient.

For $d = 24, p = 48, k = 6, n = 144$, $a = 0.25$, and $T = 60$, the final summaries are

protocol	R_{pop}	$\mathrm{mean}_i\, \rho_i/s_i$	$\max_i\, \rho_i/s_i$	mean overlap
fixed train	$3.56\,10^{-1}$	$2.66\,10^{-1}$	$3.75\,10^{-1}$	$4.59\,10^{-1}$
fixed split	$1.49\,10^{-1}$	$2.15\,10^{-1}$	$4.88\,10^{-1}$	$5.43\,10^{-1}$
crossfit blocks	$2.62\,10^{-8}$	$3.31\,10^{-5}$	$1.47\,10^{-4}$	1.00
fresh batches	$4.67\,10^{-9}$	$1.16\,10^{-5}$	$4.82\,10^{-5}$	1.00

For the larger check $d = 40, p = 80, k = 8, n = 240$, $T = 80$, and 16 crossfit blocks, one obtains

protocol	R_{pop}	$\mathrm{mean}_i\, \rho_i/s_i$	$\max_i\, \rho_i/s_i$	mean overlap
fixed train	$5.86\,10^{-1}$	$3.22\,10^{-1}$	$5.15\,10^{-1}$	$2.72\,10^{-1}$
fixed split	$4.09\,10^{-1}$	$3.42\,10^{-1}$	$4.80\,10^{-1}$	$3.30\,10^{-1}$
crossfit blocks	$1.02\,10^{-7}$	$7.87\,10^{-5}$	$3.60\,10^{-4}$	1.00
fresh batches	$1.53\,10^{-9}$	$1.07\,10^{-5}$	$3.13\,10^{-5}$	1.00

Thus increasing the training time alone is not the right cure. A single fixed empirical sample creates an interpolating fixed point: the training loss is essentially zero while the population risk, residual bulk energy, and teacher–bulk coupling remain non-zero. Fixed sample splitting weakens this self-coupling but still reuses one finite dataset. Crossfit blocks and causal fresh gradients remove the self-coupling and recover the population trajectory: all residual masses collapse to 10^{-5}-10^{-4}, and all teacher eigenvector overlaps are one up to recorded numerical precision. Experimentally, the corresponding theoretical ingredient is therefore exactly the leave-one-out/cavity replacement needed to justify reused training data.

The same comparison also supports a causal Boltzmann exponent rule. In the targeted comparison, the decisive protocols $\{\text{fresh batches}, \text{population}\}$ were evaluated side by side for fixed Muon* $a = 0.25$ and Boltzmann $a(t)$. At $T = 60$:

prot.	ctrl.	R_{pop}	mean ρ/s	max ρ/s	overlap	med/final a
fresh batches	Muon*	$1.67\,10^{-9}$	$1.38\,10^{-5}$	$2.98\,10^{-5}$	0.999999995	0.25/0.25
population	Muon*	$1.84\,10^{-8}$	$8.61\,10^{-5}$	$1.99\,10^{-4}$	1.00	0.25/0.25
fresh batches	Boltzmann	$2.22\,10^{-10}$	$1.21\,10^{-6}$	$8.17\,10^{-6}$	0.999999998	0.30/0.30
population	Boltzmann	$4.36\,10^{-9}$	$5.13\,10^{-5}$	$7.42\,10^{-5}$	1.00	0.30/0.30.

Thus the residual-mass plateau disappears for both Muon* and Boltzmann once the update gradients are fresh. Boltzmann is faster on this finite horizon: its fresh-batch residual mean is about one order of magnitude below fixed Muon*. The selected $a(t)$ is staircase-like, living in the 0.16–0.28 visible-group range and returning to the no-visible-mode value $a = 0.30$ once all modes are visible.

The full protocol panel was also evaluated at $T = 60$ for D40 and D56 to measure the directional bias between the update gradient and the population gradient. The stable statistic is the cosine

$$\Gamma_{\mathrm{upd,pop}}(t) = \frac{\langle G_{\mathrm{upd}}(t), G_{\mathrm{pop}}(t)\rangle}{\|G_{\mathrm{upd}}(t)\|\,\|G_{\mathrm{pop}}(t)\|}.$$

Raw relative norms are less useful near convergence, since empirical and population gradients have different finite-sample amplitudes. The terminal check is

case	protocol	R_{pop}	mean ρ/s	$\Gamma_{\mathrm{upd,pop}}(T)$
$d40$, Muon*	fixed train	$6.56\,10^{-1}$	$2.37\,10^{-1}$	-0.026
$d40$, Muon*	fixed split	$6.35\,10^{-1}$	$2.66\,10^{-1}$	-0.082
$d40$, Muon*	crossfit	$5.64\,10^{-4}$	$4.26\,10^{-2}$	0.322
$d40$, Muon*	fresh batches	$1.75\,10^{-9}$	$1.48\,10^{-5}$	0.932
$d40$, Boltzmann	crossfit	$1.87\,10^{-5}$	$2.12\,10^{-3}$	0.183
$d40$, Boltzmann	fresh batches	$1.58\,10^{-9}$	$1.46\,10^{-5}$	0.925
$d56$, Muon*	fixed train	$7.66\,10^{-1}$	$2.90\,10^{-1}$	-0.067
$d56$, Muon*	fixed split	$7.35\,10^{-1}$	$2.93\,10^{-1}$	-0.081
$d56$, Muon*	crossfit	$1.09\,10^{-2}$	$3.02\,10^{-1}$	0.239
$d56$, Muon*	fresh batches	$7.68\,10^{-10}$	$5.14\,10^{-6}$	0.938
$d56$, Boltzmann	crossfit	$2.45\,10^{-4}$	$1.10\,10^{-2}$	0.081
$d56$, Boltzmann	fresh batches	$9.01\,10^{-10}$	$1.11\,10^{-5}$	0.929.

Thus fixed train and fixed split are not merely slow: at the empirical plateau their update direction has lost the population direction. Crossfit partially restores it, while fresh batches remains aligned and recovers the population trajectory. The reused-data theorem therefore needs a directional cavity replacement, for example

$$1 - \Gamma_{\mathrm{upd,pop}}(t) = o_{\mathbb{P}}(1)$$

on the relevant pre-terminal interval together with the resolvent $\varepsilon_{\mathrm{RFA,d}} = o_{\mathbb{P}}(1)$ condition for Hessian roots and residues.

15.4 Gamma panel and fresh-batch Schur closure

The finite-γ panel was then evaluated in the full fresh-batch model. The setting was

$$d = 40, \qquad p = 80, \qquad k = 8, \qquad n = 240, \qquad T = 30,$$

with seed 4, fresh update batches of size 1024, and $\gamma \in \{1.0, 1.5, 2.0\}$. The fixed grid was

$$a \in \{0, 0.08, 0.12, 0.16, 0.18, 0.20, 0.22, 0.25, 0.30, 0.40, 0.60\},$$

and the Boltzmann progress grid was

$$\theta \in \{0.10, 0.20, 0.30, 0.45, 0.65, 0.85\}.$$

The resulting finite-horizon optimum is

γ	policy	best parameter	$R_{\mathrm{pop}}(T)$	$\max_i \rho_i(T)/s_i$	first cliff / plateau
1.0	fixed a	0.22	$7.37\,10^{-9}$	$4.91\,10^{-5}$	$a = 0.25:\ 4.22\,10^{-2}$
1.0	Boltzmann	$\theta = 0.45$	$2.26\,10^{-9}$	$2.46\,10^{-5}$	$\theta_{\mathrm{safe}} \geq 0.30$
1.5	fixed a	0.20	$1.18\,10^{-8}$	$7.68\,10^{-5}$	$a = 0.22:\ 1.27\,10^{-4}$
1.5	Boltzmann	$\theta = 0.45$	$1.78\,10^{-9}$	$1.95\,10^{-5}$	$\theta_{\mathrm{safe}} \geq 0.30$
2.0	fixed a	0.18	$2.00\,10^{-8}$	$1.11\,10^{-4}$	$a = 0.20:\ 1.59\,10^{-5}$
2.0	Boltzmann	$\theta = 0.30$	$1.56\,10^{-9}$	$4.36\,10^{-5}$	$\theta_{\mathrm{safe}} \geq 0.30.$

Thus the best constant exponent is the last safe constant before the visible group falls off a hard-edge cliff, and it decreases with γ on this finite horizon:

$$a^{\star}_{\mathrm{const}}(1.0) \simeq 0.22, \qquad a^{\star}_{\mathrm{const}}(1.5) \simeq 0.20, \qquad a^{\star}_{\mathrm{const}}(2.0) \simeq 0.18.$$

This is compatible with the power-law growth rate. A larger a suppresses tiny residual singular directions; when the teacher tail is steeper, those weak singular directions leave the visible group at a lower constant exponent. Boltzmann instead selects a safe progress plateau: $\theta = 0.10$ and 0.20 keep too much exponent area, while $\theta \geq 0.30$ lands on essentially the same visible-group-tracking trajectory.

The finite-Schur tracking was also repeated on the best fresh-gradient policies, not only on the earlier spectral checks. The recorded Schur blocks are $A_d = S_{\Theta\Theta}$, $B_d = S_{\Theta\perp}$, $C_d = S_{\perp\perp}$, together with the full teacher-overlap matrix. For $\gamma = 1.5$, $T = 30$, the check used both D40 and D56:

case	roots	visible	final visible	median error	residue / mass
D40 fixed $a^\star = 0.20$	240	239	8	$1.73\,10^{-17}$	0.999999987/0.999999987
D40 Boltz. $\theta = 0.45$	241	240	8	$2.78\,10^{-17}$	0.999999998/0.999999998
D56 fixed $a^\star = 0.16$	299	295	10	$1.04\,10^{-17}$	0.999999955/0.999999955
D56 Boltz. $\theta = 0.65$	301	299	10	$2.78\,10^{-17}$	0.999999999/0.999999999.

The maximum visible eigenvalue errors are $3.33\,10^{-16}$ and $5.55\,10^{-16}$ on D40, and $5.55\,10^{-16}$ and $4.44\,10^{-16}$ on D56. Hence, in the fresh-batch experiments where the residual plateau is removed, all teacher branches are visible and the tracked empirical branches coincide with the finite Schur roots at machine precision. The curve-matching problem is therefore finite-dimensionally resolved. The remaining limiting statement is the uniform deterministic RFA/MDE plus leave-one-out replacement that transports these finite Schur identities to the reused-data Muon trajectory.

15.5 Dimension check and combined RFA/LOO certificate

A targeted D72 check was then added,

$$d = 72, \qquad p = 144, \qquad k = 12, \qquad n = 432, \qquad T = 30, \qquad \gamma = 1.5.$$

The fixed grid $a \in \{0.10, 0.12, 0.14, 0.16, 0.18\}$ gives the best safe constant $a^\star = 0.12$, with $a = 0.14$ already past the finite hard-edge/visible-group cliff. The D40–D72 fresh-batch summary is

case	policy	param.	$R_{\mathrm{pop}}(T)$	$\max_i \rho_i(T)/s_i$	Schur
$D40$	fixed	$a = 0.20$	$1.18\,10^{-8}$	$7.68\,10^{-5}$	8/8, $3.33\,10^{-16}$
$D40$	Boltz.	$\theta = 0.45$	$1.78\,10^{-9}$	$1.95\,10^{-5}$	8/8, $5.55\,10^{-16}$
$D56$	fixed	$a = 0.16$	$2.20\,10^{-8}$	$1.72\,10^{-4}$	10/10, $5.55\,10^{-16}$
$D56$	Boltz.	$\theta = 0.65$	$7.31\,10^{-10}$	$2.05\,10^{-5}$	10/10, $4.44\,10^{-16}$
$D72$	fixed	$a = 0.12$	$6.68\,10^{-8}$	$1.42\,10^{-4}$	12/12, $4.44\,10^{-16}$
$D72$	Boltz.	$\theta = 0.65$	$8.65\,10^{-10}$	$2.87\,10^{-5}$	12/12, $3.33\,10^{-16}$

The last entry records final visible branches and the maximum visible Schur eigenvalue error. Thus

$$a^\star_{\mathrm{fixed}}(D40) = 0.20, \qquad a^\star_{\mathrm{fixed}}(D56) = 0.16, \qquad a^\star_{\mathrm{fixed}}(D72) = 0.12.$$

The finite-horizon constant exponent moves left with dimension, while Boltzmann remains below 10^{-9} population risk and keeps every teacher branch visible.

The curve-matching conclusion is therefore read through the finite Schur overlay, not through the scalar MP overlay. On the D56 spectral check, the scalar MP median visible relative errors were $8.91\,10^{-2}$ for Muon* and $3.32\,10^{-2}$ for causal Boltzmann, whereas the finite Schur visible absolute errors are $3.90\,10^{-18}$ and $6.94\,10^{-18}$. On the D72 fresh check, both $a^\star = 0.12$ and Boltzmann $\theta = 0.65$ end with 12/12 visible roots and maximum visible errors $4.44\,10^{-16}$ and $3.33\,10^{-16}$. Thus the finite Schur curve reconstruction is closed; the asymptotic lock is to replace these finite Schur functions by their deterministic RFA/MDE/leave-one-out limits uniformly along the Muon path.

The limiting closure can now be stated as one combined certificate. Let

$$\mathfrak{C}_d(T) = \mathfrak{G}_d(T) + \mathfrak{S}_d(T) + \mathfrak{V}_d(T),$$

where

$$\mathfrak{G}_d(T) = \sup_{t \le T_0} \big(1 - \Gamma_{\mathrm{upd,pop}}(t)\big)_+$$

is the directional leave-one-out error on the pre-terminal interval, $\mathfrak{S}_d(T)$ is the uniform RFA/MDE error for $(F_d, \partial_\lambda F_d)$, and $\mathfrak{V}_d(T)$ is the Schur visibility and capture-margin error, i.e. the failure of

$$\inf_{t \le T} \min_{i \in \mathcal{F}(t)} \min\{\lambda_i(t) - x_+(t), \Omega_i(t) - \Omega_0, \delta - \rho_i(t)/\mu_i\} > 0$$

on the visible group. In the fresh-batch experiments, $\mathfrak{S}_d$ and $\mathfrak{V}_d$ are measured by the finite Schur root/residue/capture quantities; in reused-data training, $\mathfrak{G}_d$ is the additional leave-one-out condition.

Proposition 15.2 (Imported RMT plus Schur margin closes the BBP lock)**.** *Assume that the path is independent of the Hessian sample, or has been replaced by a leave-one-out path with $\mathfrak{G}_d(T) = o_{\mathbb{P}}(1)$. Assume also that one of the available local-law inputs applies to the Hessian bulk: the effective spectral theorem of Ben Arous–Gheissari–Huang–Jagannath [1] for finite summary statistics, the Alt–Erdos–Kruger Gram local law [2], or the Ajanki–Erdos–Kruger correlated MDE local law [3] after hermitisation. If the limiting Schur determinant has only simple active roots and*

$$m_\star = \inf_{t \le T} \min_{i \in \mathcal{F}(t)} \min\{\lambda_i(t) - x_+(t), \Omega_i(t) - \Omega_0, \delta - \rho_i(t)/\mu_i\} > 0,$$

then

$$\mathfrak{S}_d(T) + \mathfrak{V}_d(T) \xrightarrow{\mathbb{P}} 0.$$

Consequently the empirical feasible set $\Theta^{R\mathrm{Schur}}_{\delta,\Omega_0,d}(T)$ converges to the deterministic feasible set, and every isolated minimizer of

$$\min_{\theta \in \Theta^{R\mathrm{Schur}}_{\delta,\Omega_0}(T)} R_T(\theta)$$

is stable. On a risk plateau the endpoints of the empirical feasible interval converge to the deterministic endpoints.

Proof. The imported local law gives uniform convergence of the bulk resolvent and of the finite-rank Schur complement on compact contours away from the limiting bulk support. The Ben Arous–Gheissari–Huang–Jagannath theorem gives the same conclusion directly for fixed summary statistics, including outlier eigenvalues and eigenvector overlaps. In the MDE route, the Gram or correlated local law gives the deterministic bulk; the finite-rank Schur complement is then a meromorphic matrix function whose only singularities are the bulk resolvent poles or edge cuts. Around each simple limiting root choose a small contour that does not touch the bulk edge. Uniform convergence on this contour and Rouche's theorem, equivalently the analytic implicit-function theorem applied to the Schur determinant, give one empirical root inside the contour and convergence of $\lambda_{d,i}$. Applying the same argument to $\partial_\lambda F_d$ gives convergence of the residue formula for $\Omega_{d,i}$. The positive deterministic margin $m_\star$ then preserves the sign of every root-edge, residue, and terminal capture inequality. This is exactly $\mathfrak{S}_d + \mathfrak{V}_d = o_{\mathbb{P}}(1)$. The last statement is continuity of the constrained argmin; on a plateau, the same sign preservation gives convergence of the feasible interval endpoints. □

Proposition 15.3 (Combined certificate implies the dynamic BBP exponent rule)**.** *Assume the population Muon summaries converge uniformly on $[0, T]$, the limiting BBP branches are simple and separated from the bulk except at regular contact times, and*

$$\mathfrak{C}_d(T) \xrightarrow{\mathbb{P}} 0.$$

Then the empirical visible Schur branches, their residues, and the thresholded visible-group indicators converge uniformly away from the regular contact intervals to the deterministic dynamic-BBP objects. Consequently the finite Markovian exponent rule based on visible branches converges to the deterministic exponent rule whose local stationarity equation is the previously derived

$$L_S(a,t) = \beta_{\mathrm{RFA}}(t) L_B(a,t),$$

or, in the power-law block notation, the adjoint balance

$$\log A(t) = \gamma \frac{\sum_i \mu_i^2 E_i(T) c_i(a,t) \log i}{\sum_i \mu_i^2 E_i(T) c_i(a,t)} + L_B(a,t).$$

Proof. The RFA part $\mathfrak{S}_d \to 0$ gives uniform convergence of the finite Schur functions and their λ-derivatives. The RFA reduction theorem above then gives roots and residues. The visibility margin $\mathfrak{V}_d \to 0$ turns residue convergence into convergence of the thresholded visible-group indicators. The directional leave-one-out term $\mathfrak{G}_d \to 0$ replaces the reused empirical update direction by the population/fresh direction in the closed summary ODE, so the finite exponent rule sees the same deterministic visible group. The stationarity equation is the Pontryagin or local Hamiltonian condition already derived for that deterministic limit. □

Proposition 15.4 (LOO/RFA certificate closes the reused-data theorem)**.** *Assume the reused-data Muon path admits leave-one-out paths $\Gamma_s^{(\ell)}$ such that, uniformly on the time/control/spectral grid,*

$$\sup_{\ell,s} \|\Gamma_s - \Gamma_s^{(\ell)}\|_{\mathrm{op}} \le a_d, \qquad \eta_{\min,d}^{-7} a_d = o_{\mathbb{P}}(1).$$

Assume also the conditional leverage and operator bounds

$$\sup_{\ell} \|h_\ell\|_{\mathrm{op}} = O_{\mathbb{P}}(1), \qquad \sup_{\ell,s} \|\Gamma_s^{(\ell)}\|_{\mathrm{op}} = O_{\mathbb{P}}(1),$$

and the decoupled fluctuation-averaging estimate

$$\delta_{\mathrm{dec,d}} = O_{\mathbb{P}}\left(\eta_{\min,d}^{-5} \sqrt{\frac{\log(|\mathcal{M}_d||\mathcal{E}_d|) + w_d}{N}} \right).$$

If $N \asymp d$, $\eta_{\min,d} = d^{-\alpha}$, and $\alpha < 1/14$, then

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

Consequently, if the Schur margin $m_\star$ in Proposition 15.2 is positive and $\mathfrak{G}_d(T) = o_{\mathbb{P}}(1)$, then

$$\mathfrak{C}_d(T) \xrightarrow{\mathbb{P}} 0$$

and the empirical reused-data exponent rule has the same deterministic visible BBP/Schur limit as the fresh-batch exponent rule.

Proof. The hard-edge functional has two resolvent derivatives tested on the removed rank-one frame h_ℓ. After replacing Γ_s by $\Gamma_s^{(\ell)}$, the frame is independent of the resolvent test, so the centred empirical average is exactly the fluctuation-averaging term $\delta_{\mathrm{dec,d}}$. The difference between the original and cavity resolvents is controlled by the third resolvent derivative, hence by $C\eta_{\min}^{-7} a_d$. Therefore

$$\varepsilon_{\mathrm{RFA,d}} \le \delta_{\mathrm{dec,d}} + C\eta_{\min,d}^{-7} a_d + o_{\mathbb{P}}(1).$$

With $N \asymp d$, the displayed rate is

$$d^{5\alpha - 1/2 + o(1)} + d^{7\alpha - 1/2 + o(1)},$$

which tends to zero for $\alpha < 1/14$. The local-law inputs used to justify the decoupled estimate are precisely the standard fluctuation-averaging and isotropic local-law effects of [4, 5, 6], or the finite-summary Hessian theorem of [1] when the path has already been made fresh/cavity. This is also the Hessian-landscape setting in which the general phase-retrieval landscape program of [7] is relevant. The final claim is then Proposition 15.2 plus the combined certificate proposition above. □

15.6 Empirical adjoint replay check

The last experiment separates two statements which must not be conflated. First, constant-a fresh-batch sweeps can be reconstructed by an empirical growth rate. Second, the schedule obtained by freely optimizing that growth rate need not be a valid causal exponent rule.

The raw log-residual growth rate

$$-\frac{1}{2}\partial_t \log(\rho_i(t)/s_i)$$

is unstable when a mode saturates. The bounded signed logistic growth rate

$$z_i(t) = \operatorname{logit}\left(1 - \rho_i(t)/s_i\right), \qquad z_i(t_{j+1}) - z_i(t_j) = 2\Delta t_j\, \omega_i(a, t_j)$$

is used instead. On the measured constant-a grid this reconstructs the terminal residual curve to numerical precision: the median reconstruction RMSE is $2.3\,10^{-19}$ on D40 and $1.0\,10^{-19}$ on D56, and the adjoint finite-difference gradient errors are $7.2\,10^{-12}$ and $3.0\,10^{-13}$.

The optimized empirical-adjoint schedules were then replayed with genuinely fresh-batch gradients. The result is

case	policy	$R_{\mathrm{pop}}(T)$	$\sum_i s_i^2 \rho_i(T)/s_i$	$\max_i \rho_i(T)/s_i$
$D40$	best fixed risk $a = 0.20$	$1.07\,10^{-8}$	$7.04\,10^{-5}$	$8.01\,10^{-5}$
$D40$	best fixed residual $a = 0.18$	$2.20\,10^{-8}$	$1.50\,10^{-5}$	$1.64\,10^{-4}$
$D40$	Boltzmann	$1.78\,10^{-9}$	$2.82\,10^{-6}$	$1.95\,10^{-5}$
$D40$	adjoint replay	$1.74\,10^{-7}$	$2.10\,10^{-5}$	$5.22\,10^{-4}$
$D56$	best fixed risk $a = 0.16$	$2.20\,10^{-8}$	$9.73\,10^{-5}$	$1.72\,10^{-4}$
$D56$	best fixed residual $a = 0.10$	$1.56\,10^{-7}$	$5.51\,10^{-5}$	$3.84\,10^{-4}$
$D56$	Boltzmann	$7.31\,10^{-10}$	$2.92\,10^{-6}$	$2.05\,10^{-5}$
$D56$	adjoint replay	$3.87\,10^{-3}$	$6.19\,10^{-2}$	$3.13\,10^{-1}$

Thus the empirical constant-sweep growth rate is informative, but it is not a closed Markovian control law in the variables (a, t) alone. The actual growth rate depends on the current captured profile and visible-group geometry. This also explains why causal Boltzmann is robust: it remeasures the visible group and chooses the exponent from the present spectral state, whereas the free adjoint mixes interval growth rates measured on different constant-a trajectories.

The resulting reduced-control statement replaces $\omega_i(a, t)$ by a state-dependent Hamiltonian

$$\omega_i(a, t, r(t), F_d(t)),$$

under the same RFA/leave-one-out certificate $\mathfrak{C}_d(T) \to 0$. The finite Schur curve reconstruction is closed; a predictive optimal-$a(t)$ theory uses this state-dependent Hamiltonian rather than an (a, t)-only adjoint.

15.7 Fresh Boltzmann progress-contour ODE

The Boltzmann exponent rule used in the fresh-gradient experiments satisfies a clean Markovian ODE, but the state variable is the present spectral curve, not time alone. Along the fresh-gradient trajectory one records the full curves

$$Z_S(a, t), \qquad Z_B(a, t), \qquad \mathrm{Free}(a, t) = \log Z_B(a, t) - 2 \log Z_S(a, t), \qquad \Psi(a, t) = \partial_a \mathrm{Free}(a, t).$$

The raw stationarity equation $\Psi(a,t)=0$ is not the finite scheduler. In the fresh D40/D56 Boltzmann checks, the free-energy minimizer remains close to the endpoint and the median selected-minus-raw-min gaps are 0.88 and 0.94. The finite policy is instead the upper progress contour

$$a_\theta(t)=\sup\left\{a\le a_{\mathrm{cap}}(t): Z_S(a,t)\ge\theta\max_b Z_S(b,t)\right\}.$$

On interior pieces put

$$\Phi(a,t)=\log Z_S(a,t)-\log Z_{S,\max}(t)-\log\theta.$$

Then

$$\Phi(a_\theta(t),t)=0,\qquad \dot a_\theta(t)=-\frac{\partial_t\Phi(a_\theta(t),t)}{\partial_a\Phi(a_\theta(t),t)}. \qquad \text{(Boltzmann progress ODE)}$$

At a cap, grid boundary, or no-visible-mode saturation, the ODE is projected to the active boundary. Equivalently,

$$\dot a=\Pi_{[0,a_{\mathrm{cap}}(t)]}\left[-\frac{\partial_t\{\log Z_S(a,t)-\log Z_{S,\max}(t)-\log\theta\}}{\partial_a\log Z_S(a,t)}\right].$$

The fresh curve check gives

case	θ	med $\lvert a_{\mathrm{sel}}-a_\theta\rvert$	max $\lvert a_{\mathrm{sel}}-a_\theta\rvert$	med $Z_S(a)/Z_{S,\max}$	ODE rms	ODE med. abs
$D40$	0.45	$7.37\,10^{-3}$	$1.51\,10^{-2}$	0.4696	$1.30\,10^{-2}$	$3.11\,10^{-4}$
$D56$	0.65	$9.93\,10^{-3}$	$1.94\,10^{-2}$	0.6908	$4.43\,10^{-3}$	$2.05\,10^{-4}$

Thus the selected exponent follows the progress contour to within one grid step in a, and the implicit ODE matches the finite-difference contour speed on the interior visible-group segment. This gives the finite ODE for the successful Muon/Boltzmann policy. A limiting optimal-control statement would then identify why the Hamiltonian selects this contour, or a nearby visible-group-dependent contour, in the power-law hierarchy.

15.8 Which contour is optimal?

The progress-contour ODE gives a family of policies indexed by θ. For a terminal cost $\mathcal{L}$, define

$$J_{\mathcal{L}}(\theta)=\mathcal{L}(r_\theta(T),S_\theta(T)),\qquad \theta^\star_{\mathcal{L}}\in\arg\min_\theta J_{\mathcal{L}}(\theta),$$

where $a_\theta(t)$ is the projected contour solution above. On an interior piece,

$$\Phi(a_\theta(t),t,\theta)=0,\qquad \partial_\theta a_\theta(t)=\frac{1}{\theta\,\partial_a\Phi(a_\theta(t),t)}.$$

Since $\partial_a\Phi<0$ on the upper contour, increasing θ lowers the selected exponent and protects the uncaptured tail.

This gives the reduced optimality equation. If $X_\theta(t)$ denotes the finite Markovian state, including residual masses, Schur margins and visible-group statistics, and

$$\dot X_\theta=F(X_\theta,a_\theta),$$

then the sensitivity $U_\theta=\partial_\theta X_\theta$ satisfies

$$\dot U_\theta=\partial_X F(X_\theta,a_\theta)U_\theta+\partial_a F(X_\theta,a_\theta)\frac{1}{\theta\,\partial_a\Phi(a_\theta,t,\theta)},\qquad U_\theta(0)=0.$$

Hence, away from projections and grid contacts,

$$J'_{\mathcal{L}}(\theta)=\nabla\mathcal{L}(X_\theta(T))^\top U_\theta(T). \qquad \text{(theta KKT)}$$

An interior optimum solves $J'_{\mathcal{L}}(\theta) = 0$. In the observed power-law visible-group regime, however, $J_{\mathcal{L}}$ has a cliff followed by a flat terminal plateau. The stable object is therefore the safe plateau

$$\Theta^{\mathrm{safe}}_{\mathcal{L}}(T, \eta) = \{\theta : \ J_{\mathcal{L}}(\theta) \le (1 + \eta) \inf_{\theta'} J_{\mathcal{L}}(\theta')\},$$

or, for an absolute target $\varepsilon_{\mathcal{L}}$,

$$\Theta^{\mathrm{safe}}_{\mathcal{L}}(T, \varepsilon_{\mathcal{L}}) = \{\theta : \ J_{\mathcal{L}}(\theta) \le \varepsilon_{\mathcal{L}}\}.$$

The final exponent rule is

$$a^\star_{\mathcal{L}}(t) = a_{\theta^\star_{\mathcal{L}}}(t), \qquad \theta^\star_{\mathcal{L}} \in \arg\min_{\theta \in \Theta^{\mathrm{safe}}_{\mathcal{L}}} J_{\mathcal{L}}(\theta), \qquad \text{(optimal contour law)}$$

with the convention that a nearly flat plateau is reported as an interval rather than as a single unstable grid point.

The theta sweeps were then refined with multi-seed checks around the candidate plateaux. The resulting summary aggregates 54 fresh-gradient trials. The table below uses the absolute terminal targets and the Schur-complete subsample when a finite Schur margin is required:

$$R_{\mathrm{pop}} \le 10^{-8}, \qquad \max_i \rho_i / s_i \le 10^{-4}, \qquad \min_i(\lambda_i^{\mathrm{Schur}} - x_+) > 0, \qquad \min_i(\Omega_i^{\mathrm{Schur}} - 0.25) > 0.$$

At $T = 30, \gamma = 1.5$, the constrained-risk result is

panel	$\theta^{R+\mathrm{cap}}_{\mathrm{first}}$	$\theta^{R+\mathrm{cap+Schur}}_{\mathrm{first}}$	$\theta^\star_R$ on feasible set	$\min(\lambda^{\mathrm{Schur}} - x_+)$ at first Schur
$D40$	0.30	$0.35^\dagger$	0.35	$4.42\,10^{-2}$
$D56$	0.45	0.45	0.45	$3.16\,10^{-2}$
$D72$	0.65	0.65	0.75	$2.40\,10^{-2}$

Here † means that the lower D40 point $\theta = 0.30$ is feasible in risk/capture but was not measured with the Schur blocks; the first Schur-certified D40 point on this dataset is $\theta = 0.35$. The minimum Schur residue margin at these first certified points is about 0.75 because the predicted and observed teacher masses are essentially one and the threshold is 0.25.

The aggregate $T = 30$ table is

case	θ	$R_{\mathrm{pop}}(T)$	$\sum_i s_i^2 \rho_i(T)/s_i$	$\max_i \rho_i(T)/s_i$
$D40$	0.30	$1.81\,10^{-9}$	$2.27\,10^{-5}$	$3.07\,10^{-5}$
$D40$	0.45	$1.81\,10^{-9}$	$2.93\,10^{-6}$	$2.87\,10^{-5}$
$D40$	0.85	$1.81\,10^{-9}$	$1.11\,10^{-5}$	$2.71\,10^{-5}$
$D56$	0.45	$7.05\,10^{-10}$	$7.23\,10^{-6}$	$2.15\,10^{-5}$
$D56$	0.55	$7.15\,10^{-10}$	$2.26\,10^{-6}$	$2.01\,10^{-5}$
$D56$	0.85	$8.30\,10^{-10}$	$7.29\,10^{-6}$	$2.30\,10^{-5}$
$D72$	0.65	$9.49\,10^{-10}$	$1.06\,10^{-5}$	$1.48\,10^{-5}$
$D72$	0.75	$8.77\,10^{-10}$	$5.05\,10^{-7}$	$1.63\,10^{-5}$
$D72$	0.85	$1.61\,10^{-9}$	$4.30\,10^{-6}$	$1.60\,10^{-5}$

Thus the cost matters, but the formulation is fixed. A pure population risk objective plus terminal capture only needs to enter the feasible plateau:

$$\theta^{R+\mathrm{cap}}_{\mathrm{first}}(D40, D56, D72) \simeq (0.30,\ 0.45,\ 0.65).$$

With the finite Schur-margin certificate included in these checks, this becomes

$$\theta^{R+\mathrm{cap+Schur}}_{\mathrm{first}}(D40, D56, D72) \simeq (0.35,\ 0.45,\ 0.65),$$

where the D40 shift is a finite-sample recording artifact rather than a theoretical shift. The weighted-tail surrogate still points to $(0.45, 0.55, 0.75)$, but it only measures balanced tail capture; it is not the objective fixed by the deterministic control problem. Consequently the final finite-horizon Boltzmann exponent is

$$\boxed{a^\star_{R,\delta,\Omega_0}(t) = \sup\{a \le a_{\mathrm{cap}}(t) : Z_S(a,t) \ge \theta^\star_{R,\delta,\Omega_0} Z_{S,\max}(t)\}}.$$

There is no universal scalar θ. There is a universal constrained-risk contour problem, and the feasible boundary drifts upward as the power-law tail becomes deeper.

15.9 Exact causal SQ-HJB benchmark

The previous Boltzmann exponent rule is causal, but it still contains a modeling choice: the signal partition Z_S. To test whether this choice is hiding the true causal optimum, we introduce an exact SQ-HJB benchmark which does not choose a chosen signal g_i. At a current state X, for every grid action a, compute the actual infinitesimal SQ drift from the current spectral update:

$$\omega_i(X,a) = \left[\frac{\dot q_i(X,a)}{\rho_i(X)}\right]_+, \qquad \rho_i = s_i - q_i.$$

Thus the control input is the measured drift table

$$\Omega(X) = \{\omega_i(X,a) : i \le k,\ a \in \mathcal{A}\},$$

not a separately calibrated singular-scale surrogate.

For the terminal bottleneck cost, set

$$L_i(X) = \left[\log \frac{\rho_i}{\delta s_i}\right]_+,$$

and freeze the drift table over the remaining horizon $\tau = T - t$. The reduced SQ dynamics are

$$\dot L_i = -\omega_i(a).$$

The exact HJB value for terminal cost $\max_i L_i(T)$ has the dual formula

$$V_{\max}(\tau, L) = \max_{\pi \in \Delta_k} \left\{ \pi \cdot L - \tau \max_{a \in \mathcal{A}} \pi \cdot \omega(a) \right\}.$$

Equivalently it is the linear program

$$\begin{aligned} \max_{\pi, z} \quad & \pi \cdot L - \tau z \\ \text{s.t.} \quad & z \ge \pi \cdot \omega(a), \qquad a \in \mathcal{A}, \\ & \pi \in \Delta_k. \end{aligned}$$

If $\pi^\star$ is an optimizer, the HJB Hamiltonian selector is

$$a^\star_{\mathrm{HJB}}(X) \in \arg\max_{a \in \mathcal{A}} \pi^\star \cdot \omega(a). \qquad \text{(SQ-HJB bottleneck)}$$

This is an exact causal exponent rule for the frozen-rate SQ bottleneck problem.

The exact reduced HJB was also tested for the diagonal-plus-trace risk. With action occupation times $\tau_a \ge 0$, $\sum_a \tau_a = \tau$, the frozen-rate terminal residual is

$$r_i(T) = \rho_i \exp\left[-\sum_a \tau_a \omega_i(a)\right],$$

and the reduced risk value is the convex program

$$V_R(\tau,\rho) = \min_{\tau_a\ge 0,\ \sum_a \tau_a=\tau} \left[\sum_i r_i(T)^2 + \frac{1}{2}\left(\sum_i r_i(T)\right)^2\right].$$

At its optimizer the HJB weights are

$$p_i^\star = 2r_i(T)^2 + r_i(T)\sum_j r_j(T),$$

and the causal action is

$$a_R^\star(X) \in \arg\max_{a\in\mathcal{A}} \sum_i p_i^\star\, \omega_i(X,a). \qquad \text{(SQ-HJB risk)}$$

Again, no Boltzmann singular scale surrogate is chosen.

The comparison places these two exact SQ-HJB exponent rules next to the best fixed constants and the Boltzmann progress exponent rule. At $T = 30, \gamma = 1.5$:

case	policy	$R_{\mathrm{pop}}(T)$	mean ρ/s	$\max_i \rho_i/s_i$
$D40$	fixed $a = 0.20$	$1.18\,10^{-8}$	$1.64\,10^{-5}$	$7.68\,10^{-5}$
$D40$	Boltz. $\theta = 0.45$	$1.78\,10^{-9}$	$7.38\,10^{-6}$	$1.95\,10^{-5}$
$D40$	SQ-HJB bottleneck	$3.71\,10^{-6}$	$8.97\,10^{-4}$	$3.80\,10^{-3}$
$D40$	SQ-HJB risk	$3.87\,10^{-6}$	$1.27\,10^{-3}$	$2.90\,10^{-3}$
$D56$	fixed $a = 0.16$	$2.20\,10^{-8}$	$5.10\,10^{-5}$	$1.72\,10^{-4}$
$D56$	Boltz. $\theta = 0.65$	$7.31\,10^{-10}$	$6.81\,10^{-6}$	$2.05\,10^{-5}$
$D56$	SQ-HJB bottleneck	$1.41\,10^{-6}$	$1.62\,10^{-4}$	$5.67\,10^{-4}$
$D56$	SQ-HJB risk	$2.62\,10^{-6}$	$6.10\,10^{-4}$	$1.57\,10^{-3}$
$D72$	fixed $a = 0.12$	$6.68\,10^{-8}$	$2.96\,10^{-5}$	$1.42\,10^{-4}$
$D72$	Boltz. $\theta = 0.65$	$8.65\,10^{-10}$	$5.30\,10^{-6}$	$2.87\,10^{-5}$
$D72$	SQ-HJB bottleneck	$2.18\,10^{-1}$	$5.52\,10^{-1}$	$8.99\,10^{-1}$
$D72$	SQ-HJB risk	$2.43\,10^{-6}$	$2.95\,10^{-4}$	$1.39\,10^{-3}$

The conclusion is important. The exact HJB for residual-only SQ objectives is causal and mathematically clean, but it is not the observed optimum. It drives the policy too aggressively toward the Muon endpoint and fails to keep the terminal Schur/bulk geometry stable enough. Therefore the optimal SQ state cannot be only (q_i, ρ_i) with rates ω_i. It must include the spectral state

$$(\lambda_i^{\mathrm{Schur}} - x_+,\ \Omega_i^{\mathrm{Schur}},\ \text{bulk injection/coupling})$$

as either hard constraints or a running cost in the HJB. In this precise sense, Boltzmann is best read as a Schur-regularized control rule: its progress contour captures a bulk-stability term absent from the first residual-only HJB.

15.10 AMP bridge and the lower-bound interpretation of Boltzmann

The correct lower-bound comparison is not with arbitrary algorithms, but with the first-order/SQ class. Gradient descent, SGD, preconditioned SGD and Muon are SQ algorithms in this sense: every update is a function of finitely many empirical averages of functions of (x, y, W_t). Thus a lower bound for first-order/SQ weak recovery is the relevant benchmark for the Markovian control problem.

Defilippis, Dandi, Mergny, Krzakala and Loureiro [9] give the bridge needed here. They start from the linearization of multi-index GAMP around the uninformed fixed point. In their notation the optimal linearized message-passing singular scale contains

$$G_{\mathrm{out}}(y) = \mathbb{E}[zz^\top - I \,|\, y],$$

and its state evolution is closed on low-dimensional overlaps

$$M_t = d^{-1}\widehat{W}_t^\top W_\star, \qquad Q_t = d^{-1}\widehat{W}_t^\top \widehat{W}_t.$$

The instability threshold of the uninformed fixed point is governed by the operator

$$\mathcal{F}(M) = \mathbb{E}_y[G(y)MG(y)], \qquad \alpha_c^{-1} = \sup_{M\succeq 0,\ \|M\|_F=1} \|\mathcal{F}(M)\|_F,$$

and their spectral construction achieves the optimal weak-recovery threshold of this AMP/first-order theory. This is the right external lower-bound anchor for our SQ control problem.

The Boltzmann exponent rule can be interpreted as the causal, Schur-calibrated version of this linearized AMP power iteration. At a state X_t, for every Muon exponent a, the reduced rule forms a signal partition function and a bulk partition function of the form

$$Z_S(a; X_t) = \sum_{i\in\mathcal{F}_t} w_i(X_t) f_a(\sigma_i(X_t)), \qquad Z_B(a; X_t) = \frac{1}{m_t}\sum_{b=1}^{m_t} f_a(\beta_b(X_t))^2,$$

where $\mathcal{F}_t$ is the visible group, σ_i are the signal singular scales, and β_b are bulk singular values. With risk weights the reduced rule uses, up to floors,

$$w_i(X_t) \simeq \rho_i(X_t)^\gamma \sqrt{q_i(X_t)}.$$

The Boltzmann free energy is

$$F_B(a; X_t) = \log Z_B(a; X_t) - 2\log Z_S(a; X_t),$$

or equivalently the local Hamiltonian

$$\Phi_B(a; X_t) = 2\log Z_S(a; X_t) - \log Z_B(a; X_t).$$

Thus a is chosen to maximize a signal-to-bulk Rayleigh growth ratio. This is the same object that a linearized AMP/spectral method optimizes at the uninformed fixed point, except that here it is recomputed causally along the learned index.

The exact statement suggested by the experiments is therefore not "Boltzmann equals the residual-only HJB". The correct statement is: Boltzmann is the entropy-regularized one-step HJB Hamiltonian for the AMP/Schur visible group. Indeed, if the exponent rule is allowed to choose a distribution $\pi \in \Delta(\mathcal{A})$ over exponents, the regularized Hamiltonian

$$\mathcal{H}_\tau(X_t) = \sup_{\pi\in\Delta(\mathcal{A})} \left\{ \sum_{a\in\mathcal{A}} \pi(a)\Phi_B(a; X_t) + \tau\,\mathrm{Ent}(\pi) \right\}$$

has optimizer

$$\pi_\tau(a \mid X_t) = \frac{\exp(\Phi_B(a; X_t)/\tau)}{\sum_{a'} \exp(\Phi_B(a'; X_t)/\tau)}.$$

The deterministic causal rule used in the experiments is a finite-grid projection of this Gibbs policy, with an additional Schur/bulk progress contour to avoid brittle endpoint choices.

Conjecture 15.5 (Boltzmann as AMP-optimal SQ control)**.** *Consider the multi-index phase-retrieval power-law limit in which the teacher visible group contains many modes and the finite Schur functions converge uniformly to their deterministic RFA/MDE limits. Among Markovian SQ first-order exponent rules that update through a Muon spectral preconditioner $\Delta_t(a)$, the entropy-regularized HJB with Hamiltonian Φ_B reaches the linearized-AMP weak-recovery threshold α_c. After weak recovery, the residual amplification time is the unavoidable multiplicative factor $\log d$. Hence the Boltzmann exponent rule matches the SQ/first-order lower bound up to the standard logarithmic amplification factor.*

This is the direct way to connect the pieces. AMP supplies the optimal spectral threshold and the first-order lower-bound benchmark. The finite Schur analysis identifies the actual causal spectral operator along the Muon trajectory. Boltzmann is the entropy-regularized Hamiltonian that maximizes the signal-to-bulk growth of this operator. What remains to prove is the operator equivalence

$$\Phi_B(a; X_t) = \text{linearized AMP growth functional at } X_t + o(1),$$

uniformly on the controlled trajectory and away from Schur margin loss.

15.11 Vanilla SGD lower bounds, DMFT, and why Muon can escape

Barzilai and Shamir [10] clarify an important point for the lower-bound story. SQ lower bounds are not automatically lower bounds for standard SGD, because the actual SGD noise is data-dependent, anisotropic and evolves along the trajectory. They therefore prove algorithm-specific lower bounds for vanilla SGD directly. Their effect is alignment-based. Let U be the teacher subspace and W_t the learned first-layer subspace. If

$$\rho_t = \|P_{W_t} P_U\|_{\mathrm{op}}$$

is small, then the population gradient is small:

$$\|\nabla_W L(\theta_t)\|_F \le \psi(\rho_t).$$

Under a bounded gradient-condition-number assumption κ_T, the stochastic part behaves like an almost isotropic random walk and the alignment remains $O(d^{-1/2})$ for a long time. For multi-index targets with information exponent $k_\star$, their bound gives the usual scale

$$\widetilde{\Omega}\big(d^{\max(k_\star-1,1)+1}\big)$$

samples/iterations for vanilla SGD in their normalization.

This does not contradict the population DMFT used in this companion. The DMFT or population ODE describes the deterministic drift once the relevant summary statistics are visible and the stochastic estimator is controlled. The Barzilai–Shamir lower bound describes the complementary cold-start regime: before alignment, vanilla SGD's population drift is too weak relative to its own stochastic gradient singular scale, so it behaves like an isotropic random walk. Thus the two pictures splice as follows:

cold start: SGD martingale/alignment lower bound
$\Longrightarrow$ visible group: deterministic DMFT/SQ flow

Muon is not covered by the vanilla-SGD lower bound as stated, because it is not the raw update

$$W_{t+1} = W_t - \eta \nabla_W \ell(\theta_t; x_t).$$

Instead, after forming the gradient matrix G_t, Muon applies a spectral denoiser

$$G_t = U \operatorname{diag}(\sigma_r) V^\top, \qquad \Delta_t(a) = U \operatorname{diag}\Big[\sigma_r (\sigma_r^2 + \lambda^2)^{(a-1)/2}\Big] V^\top.$$

This operation changes the signal-to-bulk ratio of the gradient singular scale. In the language of Barzilai–Shamir, it is designed precisely to change the effective gradient condition number after projection onto the useful spectral visible group. In the language of AMP, it is a spectral denoiser acting on the linearized message.

This motivates the central conjectural picture:

Muon-Boltzmann is dynamic spectral denoising of the SGD gradient singular scale.

The cold-start lower bound says raw SGD cannot create alignment faster than its weak population drift permits. The AMP spectral threshold says the best first-order denoiser can extract the first weak direction when the signal operator crosses its BBP instability. Muon supplies a one-parameter spectral denoiser f_a, and Boltzmann/HJB selects a to maximize the causal signal-to-bulk growth while preserving the Schur margin.

The resulting target theorem is therefore a three-piece statement.

(i) *Lower bound.* Any vanilla SGD trajectory satisfying the Barzilai–Shamir gradient-condition-number hypotheses stays essentially unaligned until the information-exponent lower-bound time.

(ii) *AMP threshold.* The linearized AMP/spectral operator gives the optimal first-order/SQ weak-recovery threshold.

(iii) *Muon realization.* In the hierarchical power-law phase retrieval model, the Muon denoiser family f_a, with Boltzmann control, realizes the same AMP growth functional on the Schur-visible group:

$$2\log Z_S(a; X_t) - \log Z_B(a; X_t) = \text{AMP growth}(a; X_t) + o(1).$$

If (iii) is proved uniformly until the terminal Schur margin, then Muon is not only an empirical improvement over SGD. It is an AMP-like first-order denoising algorithm written in gradient/preconditioner form. The remaining $\log d$ factor is then the unavoidable multiplicative amplification from the random $d^{-1/2}$ overlap to constant overlap once the BBP/AMP direction is visible.

References

[1] G. Ben Arous, R. Gheissari, J. Huang and A. Jagannath, *Local geometry of high-dimensional mixture models: effective spectral theory and dynamical transitions*, arXiv:2502.15655, 2025.

[2] J. Alt, L. Erdos and T. Kruger, *Local law for random Gram matrices*, arXiv:1606.07353, 2016.

[3] O. Ajanki, L. Erdos and T. Kruger, *Stability of the Matrix Dyson Equation and Random Matrices with Correlations*, arXiv:1604.08188, 2016.

[4] L. Erdos, A. Knowles and H.-T. Yau, *Averaging Fluctuations in Resolvents of Random Band Matrices*, arXiv:1205.5664, 2012.

[5] L. Erdos, A. Knowles, H.-T. Yau and J. Yin, *The Local Semicircle Law for a General Class of Random Matrices*, arXiv:1212.0164, 2012.

[6] A. Bloemendal, L. Erdos, A. Knowles, H.-T. Yau and J. Yin, *Isotropic Local Laws for Sample Covariance and Generalized Wigner Matrices*, arXiv:1308.5729, 2013.

[7] A. Maillard, T. Bonnaire and G. Biroli, *Topological Exploration of High-Dimensional Empirical Risk Landscapes: general approach, and applications to phase retrieval*, arXiv:2602.17779, 2026.

[8] 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, 2025.

[9] L. Defilippis, Y. Dandi, P. Mergny, F. Krzakala and B. Loureiro, *Optimal Spectral Transitions in High-Dimensional Multi-Index Models*, arXiv:2502.02545, 2025.

[10] D. Barzilai and O. Shamir, *Limitations of SGD for Multi-Index Models Beyond Statistical Queries*, arXiv:2602.05704, 2026.