Controlling the Spectral Power of Muon Through a Volterra State

Abstract

Muon replaces a gradient matrix by a spectral transform of that matrix. We study the case in which the exponent of this transform varies during training. The argument has two logically distinct parts. For any well-posed differentiable Volterra state equation, we derive the exact forward sensitivity, backward adjoint, and first-order optimality condition for the exponent schedule. This part is deterministic. We then state the additional uniform limit theorem needed to identify this state with a high-dimensional training process. That statement is model-dependent and is not implied by the adjoint calculation. We finally place the non-autonomous power-law state used for phase retrieval in this framework and distinguish open-loop control, Hamilton–Jacobi–Bellman control, and entropy-regularized Boltzmann policies.

1 The question

Let G_t be the matrix update used by ordinary gradient descent at time t. If $G_t = U_t \Sigma_t V_t^\top$, regularized Muon with power a applies

$$\Phi_{a,\varepsilon}(G_t) = U_t \psi_{a,\varepsilon}(\Sigma_t) V_t^\top, \qquad \psi_{a,\varepsilon}(s) = s(s^2+\varepsilon^2)^{(a-1)/2}. \tag{1.1}$$

For $\varepsilon = 0$ and $s > 0$, the scalar map is s^a. Thus $a = 1$ is the unpreconditioned update and $a = 0$ is the polar factor. Negative powers require a positive hard-edge regularization.

We seek a measurable schedule

$$a : [0, T] \longrightarrow [a_-, a_+]$$

that minimizes a terminal risk, possibly with a running cost. This raises two different questions:

1. Which schedule is optimal for a given deterministic state equation?
2. Does that state describe the original random training process, uniformly over the schedules being optimized?

The first question is answered below. The second requires a model-specific high-dimensional theorem. Differentiating a proposed DMFT does not prove the DMFT.

The microscopic derivative with respect to the control is explicit:

$$\partial_a \psi_{a,\varepsilon}(s) = \frac{1}{2} \psi_{a,\varepsilon}(s) \log(s^2+\varepsilon^2), \tag{1.2}$$

$$\partial_a \Phi_{a,\varepsilon}(G) = \frac{1}{2} U \Big[\psi_{a,\varepsilon}(\Sigma) \log(\Sigma^2 + \varepsilon^2 I) \Big] V^\top. \tag{1.3}$$

Equation (1.3) is the microscopic source of the control derivative in the state equation.

2 A controlled Volterra state

We first take a finite-dimensional state $X(t) \in \mathbb{R}^m$. This includes a finite discretization of two-time correlation and response kernels. Consider

$$\begin{aligned} \dot{X}(t) &= f(t, X(t), a(t)) \\ &\quad + \int_0^t k(t, s, X(t), X(s), a(t), a(s))\, \mathrm{d}s, \qquad X(0) = X_0. \end{aligned} \tag{2.1}$$

The historical arguments $X(s), a(s)$ encode memory. The present arguments $X(t), a(t)$ allow the current state and control to modify the whole instantaneous memory integral.

Assumption 2.1 (Controlled state)**.** *The functions f and k are measurable in time, continuously differentiable in their state and control arguments, and locally Lipschitz in the state uniformly for $a \in [a_-, a_+]$. Along the trajectories considered, the functions and their first derivatives are bounded by an integrable envelope.*

The usual contraction argument on short intervals, followed by continuation, gives a unique solution on $[0, T]$. The same Volterra–Gronwall estimate controls changes in the schedule.

Proposition 2.2 (Stability in the schedule)**.** *Let $X[a]$ and $X[b]$ solve* (2.1) *on a common bounded set. There is a constant C_T such that*

$$\sup_{t \le T} \|X[a](t) - X[b](t)\| \le C_T \, \|a - b\|_{L^\infty(0,T)} . \tag{2.2}$$

Proof. Subtract the two state equations and set

$$u(t) = \sup_{r \le t} \|X[a](r) - X[b](r)\| .$$

The mean-value theorem gives

$$u(t) \le C \, \|a - b\|_{L^\infty} + C \int_0^t u(s)\, \mathrm{d}s.$$

Gronwall's inequality yields (2.2). On a finite horizon, the nested Volterra integral only changes the constant. □

In a DMFT, X also contains functions on the causal triangle

$$\Delta_T = \{(t, s) : 0 \le s \le t \le T\}.$$

The results below extend to a Banach state space by replacing matrices with bounded Fréchet derivatives and Euclidean products with dual pairings. The finite-dimensional form is retained because it displays every memory term.

3 Forward sensitivity

Fix a and a bounded direction h. Let

$$a_\epsilon = a + \epsilon h, \qquad Y(t) = \frac{\mathrm{d}}{\mathrm{d}\epsilon} X[a_\epsilon](t)\bigg|_{\epsilon=0} .$$

All derivatives below are evaluated along $(X[a], a)$. Define

$$A(t) = f_x(t) + \int_0^t k_{x_t}(t, s)\, \mathrm{d}s, \qquad B(t) = f_a(t) + \int_0^t k_{a_t}(t, s)\, \mathrm{d}s, \tag{3.1}$$

$$K_x(t, s) = k_{x_s}(t, s), \qquad K_a(t, s) = k_{a_s}(t, s). \tag{3.2}$$

The labels t and s distinguish present and historical arguments.

Theorem 3.1 (Derivative of the Volterra solution)**.** *Under Assumption 2.1, the solution map $a \mapsto X[a]$ is Fréchet differentiable from $L^\infty(0,T)$ to $C([0,T];\mathbb{R}^m)$. Its derivative in the direction h is the unique solution of*

$$\begin{aligned} \dot{Y}(t) &= A(t)Y(t) + B(t)h(t) \\ &\quad + \int_0^t \{K_x(t,s)Y(s) + K_a(t,s)h(s)\}\,\mathrm{d}s, \qquad Y(0) = 0. \end{aligned} \tag{3.3}$$

Moreover,

$$\sup_{t\le T} \|Y(t)\| \le C_T \,\|h\|_{L^\infty(0,T)}\,. \tag{3.4}$$

Proof. Proposition 2.2 bounds the difference quotient uniformly. Taylor-expand f and k to first order in all state and control arguments. After division by ϵ, the linear terms give (3.3). The remainder is a modulus of continuity of the first derivatives times the bounded difference quotient. A Volterra–Gronwall estimate makes that remainder vanish uniformly. The same estimate applied to (3.3) gives (3.4). □

4 The backward adjoint

Let

$$\mathcal{J}(a) = \Phi(X(T)) + \int_0^T \ell(t, X(t), a(t))\,\mathrm{d}t + \mathcal{R}(a), \tag{4.1}$$

where $\mathcal{R}$ is a differentiable regularizer. Propagating (3.3) separately for every direction h is unnecessary.

Theorem 4.1 (Volterra adjoint)**.** *Suppose Assumption 2.1 holds and $\Phi, \ell, \mathcal{R}$ are continuously differentiable. Let $\lambda : [0,T] \to \mathbb{R}^m$ solve*

$$\begin{aligned} -\dot{\lambda}(t) &= \ell_x(t) + A(t)^\top \lambda(t) + \int_t^T K_x(\tau,t)^\top \lambda(\tau)\,\mathrm{d}\tau, \\ \lambda(T) &= \nabla\Phi(X(T)). \end{aligned} \tag{4.2}$$

Then

$$D\mathcal{J}(a)[h] = \int_0^T g(t)h(t)\,\mathrm{d}t, \tag{4.3}$$

with

$$\begin{aligned} g(t) &= \ell_a(t) + B(t)^\top \lambda(t) + \int_t^T K_a(\tau,t)^\top \lambda(\tau)\,\mathrm{d}\tau \\ &\quad + \nabla\mathcal{R}(a)(t). \end{aligned} \tag{4.4}$$

Proof. Direct differentiation of (4.1) gives

$$\begin{aligned} D\mathcal{J}(a)[h] &= \langle \nabla\Phi(X(T)), Y(T)\rangle + \int_0^T \{\langle \ell_x(t), Y(t)\rangle + \ell_a(t)h(t)\}\,\mathrm{d}t \\ &\quad + D\mathcal{R}(a)[h]. \end{aligned}$$

Multiply (3.3) by $\lambda(t)$ and integrate. Since $Y(0) = 0$,

$$\int_0^T \left\langle \lambda, \dot{Y} \right\rangle \mathrm{d}t = \langle \lambda(T), Y(T)\rangle - \int_0^T \left\langle \dot{\lambda}, Y \right\rangle \mathrm{d}t.$$

Fubini's theorem reverses the memory triangle:

$$\int_0^T \int_0^t \langle \lambda(t), K_x(t,s)Y(s)\rangle \,\mathrm{d}s\,\mathrm{d}t$$

$$= \int_0^T \left\langle \int_s^T K_x(t,s)^\top \lambda(t)\,\mathrm{d}t, Y(s) \right\rangle \mathrm{d}s.$$

The same identity for $K_a(t,s)h(s)$ yields the future-memory term in (4.4). Choosing (4.2) cancels every coefficient of Y and leaves (4.3). □

The integral from t to T is essential. Changing the power at time t changes every future memory integral in which $a(t)$ appears as a historical argument. An ODE adjoint without this term differentiates another model.

5 First-order optimality

Let

$$\mathcal{U} = \{a \in L^\infty(0,T) : a_- \le a(t) \le a_+ \text{ almost everywhere}\}.$$

If $a^\star$ is a local minimizer, its gradient satisfies

$$\int_0^T g^\star(t)\{b(t) - a^\star(t)\}\,\mathrm{d}t \ge 0, \qquad b \in \mathcal{U}. \tag{5.1}$$

Hence, almost everywhere,

$$\begin{cases} g^\star(t) = 0, & a_- < a^\star(t) < a_+, \\ g^\star(t) \ge 0, & a^\star(t) = a_-, \\ g^\star(t) \le 0, & a^\star(t) = a_+. \end{cases} \tag{5.2}$$

Equivalently, for every $\rho > 0$,

$$a^\star(t) = \Pi_{[a_-,a_+]}\{a^\star(t) - \rho g^\star(t)\}. \tag{5.3}$$

These are necessary conditions. They are sufficient only if the reduced functional $a \mapsto \mathcal{J}(a)$ has the required convexity.

For a parametric schedule a_θ,

$$\partial_{\theta_j}\mathcal{J}(a_\theta) = \int_0^T g(t)\partial_{\theta_j} a_\theta(t)\,\mathrm{d}t. \tag{5.4}$$

A sigmoid or a piecewise-constant schedule is therefore a numerical restriction, not a prediction of the Volterra theory. If

$$\mathcal{R}(a) = \frac{\rho}{2}\int_0^T |\dot a(t)|^2\,\mathrm{d}t,$$

then $\nabla\mathcal{R}(a) = -\rho\ddot a$ in the interior, with natural boundary conditions $\dot a(0) = \dot a(T) = 0$ when endpoints are free.

6 The exact discrete adjoint

Suppose the numerical state is the causal recursion

$$X_{n+1} = F_n(X_0,\dots,X_n,a_0,\dots,a_n), \qquad n = 0,\dots,N-1, \tag{6.1}$$

and

$$J_N(a) = \Phi(X_N) + \sum_{n=0}^{N-1} \ell_n(X_n,a_n) + R_N(a).$$

Reverse differentiation gives

$$\lambda_N = \nabla\Phi(X_N), \tag{6.2}$$

$$\lambda_k = \nabla_{X_k} \ell_k + \sum_{n=k}^{N-1} (\partial_{X_k} F_n)^\top \lambda_{n+1}, \tag{6.3}$$

$$g_k = \partial_{a_k} \ell_k + \sum_{n=k}^{N-1} (\partial_{a_k} F_n)^\top \lambda_{n+1} + \partial_{a_k} R_N. \tag{6.4}$$

The sums over $n \geq k$ are the discrete future-memory terms. These identities are exact for the discretized state. Agreement with the continuous adjoint also requires a consistent, stable time discretization.

7 The high-dimensional bridge

Let $X_d[a]$ be observables of a random training process in dimension d, and let $X[a]$ solve the candidate deterministic state equation. Consider

$$\mathcal{U}_L = \{a \in W^{1,\infty}(0,T) : a_- \leq a(t) \leq a_+, \quad \|\dot{a}\|_{L^\infty} \leq L\}. \tag{7.1}$$

This class is compact in the uniform norm.

Theorem 7.1 (Sufficient criterion for uniformity in the schedule)**.** *Assume there are $\epsilon_d, r_d, \delta_d \downarrow 0$ such that:*

1. *for every fixed $a \in \mathcal{U}_L$,*
$$\mathbb{P}\{\|X_d[a] - X[a]\|_{\mathcal{X}_T} > r_d\} \leq \delta_d;$$
2. *with probability tending to one, $X_d[\cdot]$ and $X[\cdot]$ are C_T-Lipschitz on $\mathcal{U}_L$, uniformly in d;*
3. *the covering number satisfies*
$$N(\epsilon_d, \mathcal{U}_L, \|\cdot\|_\infty)\delta_d \longrightarrow 0.$$

Then

$$\sup_{a \in \mathcal{U}_L} \|X_d[a] - X[a]\|_{\mathcal{X}_T} \xrightarrow{\mathbb{P}} 0. \tag{7.2}$$

Proof. Take an ϵ_d-net $\mathcal{N}_d$. A union bound controls the error simultaneously on the net. Given a, choose $\pi_d(a) \in \mathcal{N}_d$ within ϵ_d. Then

$$\begin{aligned} \|X_d[a] - X[a]\| \leq{}& \|X_d[a] - X_d[\pi_d(a)]\| \\ &+ \|X_d[\pi_d(a)] - X[\pi_d(a)]\| \\ &+ \|X[\pi_d(a)] - X[a]\|. \end{aligned}$$

The outer terms are at most $C_T \epsilon_d$ on the stability event and the middle term is at most r_d simultaneously on the net. □

The theorem identifies the missing work. Pointwise convergence for fixed a is insufficient: its probability must pay for the metric entropy of the schedule class, and the microscopic process must be stable in a. Near the hard edge this stability fails without regularization for nonpositive powers. For $\varepsilon > 0$, (1.2) gives

$$|\partial_a \psi_{a,\varepsilon}(s)| \leq \frac{1}{2} |\psi_{a,\varepsilon}(s)| |\log(s^2 + \varepsilon^2)|, \tag{7.3}$$

which is finite on bounded spectral intervals. Sending ε to zero requires a separate small-singular-value estimate.

8 The phase-retrieval spectral state

For a power-law phase-retrieval teacher, the reduced state used in the companion papers consists of mode energies $q_i^2(t)$ and an aggregate scale $Q(t)$. Conditional on a fresh projected-gradient channel with deterministic one- and two-resolvent coefficients $\delta_i^{(a)}$ and $\nu_i^{(a)}$, the non-autonomous mode equation is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}q_i^2(t) = -2\eta\delta_i^{(a(t))}Q(t)^{(a(t)-1)/2}q_i^2(t) \\ + \eta^2\nu_i^{(a(t))}Q(t)^{a(t)}.\end{aligned} \tag{8.1}$$

The first term is useful spectral drift; the second is transformed-gradient volatility. Once Q and the resolvent coefficients are closed self-consistently, (8.1) is an instance of (2.1).

The projected-risk recursion of Paquette et al. [1] proves one- and two-resolvent formulas for a matrix least-squares model and a general spectral transform. Identifying the same coefficients in nonlinear phase retrieval is an additional channel theorem, not a substitution of notation. Reusing the training samples also requires a dynamic leave-one-out or cavity argument.

For fixed a, integration and summation over modes yield the nonlinear Volterra law in the constant-power companion paper. For variable a, define the raw clock

$$\mathrm{d}\tau = \eta Q(t)^{(a(t)-1)/2}\,\mathrm{d}t. \tag{8.2}$$

The survival action of mode i between u and τ is

$$\mathcal{A}_i(\tau, u) = 2\int_u^\tau \delta_i^{(a(s))}\,\mathrm{d}s. \tag{8.3}$$

Thus the variable-power state remembers the entire schedule. Inserting an instantaneous fixed-power optimum into (8.1) is an adiabatic approximation. The adjoint (4.2) gives the correct first-order calculation for the full memory problem.

9 Adjoint, HJB, and Boltzmann policies

The Volterra adjoint computes the derivative of an open-loop objective with respect to a deterministic schedule. It gives a necessary stationarity condition using one forward and one backward solve.

A Hamilton–Jacobi–Bellman equation gives a globally optimal feedback policy only after the dynamics are Markovian. For a Volterra equation, the Markov state includes the relevant history. The exact value function therefore lives on a history space. A finite-dimensional HJB equation is exact only when the memory has a finite Markov realization; otherwise it is a reduced model.

A Boltzmann policy is exact for an entropy-regularized relaxed control problem. If π_t is a distribution over actions, its drift is

$$\bar{b}(X, \pi_t) = \int b(X, a)\,\pi_t(\mathrm{d}a), \tag{9.1}$$

not $b(X, \int a\,\pi_t(\mathrm{d}a))$ unless b is affine in a. The Gibbs form comes from minimizing a Hamiltonian plus entropy. Replacing the policy by its mean creates a barycentric error controlled by the curvature of $a \mapsto b(X, a)$. The experimental equalization rule is therefore a principled relaxed policy, but it is not automatically the unregularized HJB optimum.

Adding a constant to a value function is an exact gauge symmetry. Adding a term affine in the action is not: it changes the minimizing Hamiltonian unless the induced residual vanishes on the selected support.

10 Status of the argument

The deterministic result is complete. Given a well-posed differentiable controlled Volterra equation, Theorems 3.1 and 4.1 give its exact sensitivity and adjoint, while (5.1) gives the first-order condition for a bounded exponent. The discrete formulas (6.2)–(6.4) are exact for a causal numerical solver.

The identification with high-dimensional Muon training is conditional on a separate probabilistic input. At fixed regularization one needs a pointwise state-evolution or DMFT theorem, quantitative concentration satisfying Theorem 7.1, and stability with respect to the schedule. Pure Muon additionally requires hard-edge control. For nonlinear phase retrieval with sample reuse, the remaining bridge is a dynamic cavity or leave-one-out theorem identifying the same resolvent coefficients along the adapted trajectory.

This separation makes the program reusable. The control calculation need not be repeated for every random model. What must be established model by model is that the random algorithm converges, uniformly over admissible schedules, to the controlled Volterra equation being optimized.

References

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

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

[3] Z. Fan and L. Wang, *High-Dimensional Learning Dynamics of Multi-Pass Stochastic Gradient Descent in Multi-Index Models*, arXiv:2601.21093.