Who Needs Attention Anyway?

Constant-Latency Geometric Corrections for Streaming State-Space Models

Dario Fumarola

Amazon Web Services

Abstract

We study streaming state–space models (SSMs) in nonstationary settings under strict per–step latency. The SSM is frozen; adaptation occurs only at inference in latent space. We endow the latent with a decoder–induced pullback metric and apply a single trust–region correction per step. Concretely, we form a low–rank exponential–moving sketch of the pullback/Gauss–Newton metric, $\mathbf{G} = \lambda I + UU^\top$, from decoder Jacobian products, compute the natural direction $v = -(\lambda I + UU^\top)^{-1}\nabla\ell$ via a Woodbury solve, cap the step by the $\mathbf{G}$–norm trust region, and update with a second–order retraction. Acceptance uses a short frozen rollout of length H. For fixed rank r and horizon H, the per–step cost is $O(d_z r + r^2)$ plus one H–step evaluation, yielding stable mean and p99 latencies.

We provide two concise guarantees: (i) a local Riemannian descent bound under $\mathbf{G}$–smoothness, and (ii) discrete contraction of the composed predict–then–correct map up to $O(\rho^3)$ retraction error. In nonstationary planar navigation and peptide torsion control, the method reduces steps and constraint violations relative to streaming baselines while preserving constant latency (e.g., -46% steps vs. a streaming Transformer and -31% vs. SE(3)+MPC; 2–4× fewer collisions/clashes; p99 ≈ 6 ms and ≈ 4 ms, respectively). The approach is intentionally local and depends on decoder calibration; damping and clipping mitigate misspecification. Implementation details and proofs are deferred to the supplement.

1 Introduction

Streaming state-space models (SSMs) provide constant-memory, linear-time inference, making them well suited to agents under tight latency budgets. When the environment or data distribution drifts, however, a frozen SSM cannot adjust its latent dynamics or the decoder-induced geometry in real time. Attention-based architectures can adapt but forfeit strict latency guarantees because memory and compute scale quadratically.

We seek a principled way to let a frozen SSM adapt at inference while preserving strict latency. We view the latent space as a manifold equipped with a decoder-induced Riemannian metric, obtained as the pullback of the Fisher information (or a Gauss–Newton surrogate) of the decoder likelihood. Each streaming step performs a single correction: after the SSM predicts $z^+ = f_\theta(z, x)$, we take the steepest-descent direction with respect to this metric and update within a small trust region using a retraction.

This yields a constant-latency predict–then–correct procedure that combines the efficiency of linear-time SSMs with local, geometry-aware adaptation. We call the resulting model GeoSSM.

2 Method

We call the procedure GeoSSM. The state update is a predict–then–correct step that treats the SSM latent as a Riemannian manifold with a decoder-induced metric.

Let $z^{+} = f_{\theta}(z, x)$ be the one-step SSM prediction, g_{ϕ} the decoder, and let the instantaneous loss be $\ell(z^{+})$ computed from the fresh observation via $g_{\phi}(z^{+})$ (e.g., negative log-likelihood or a task loss). Write

$$A(z^{+}) \;=\; J_g(z^{+})^{\top}\, W(z^{+})\, J_g(z^{+}),$$

where J_g is the decoder Jacobian (via JVP/VJP) and W is a Gauss–Newton/Fisher weight (e.g., $W = \Sigma^{-1}$ for Gaussian likelihoods). We approximate the pullback metric $A(z^{+})$ by a Nyström factorization of rank r:

$$\mathbf{G}(z^{+}) \;=\; \lambda I{+}UU^{\top}, \qquad \Omega = \big[\, q_j \,\big]_{j=1}^{r}, \quad Y = A(z^{+})\,\Omega, \quad T = \Omega^{\top} Y, \quad U \leftarrow \beta U_{\text{prev}}{+}\sqrt{1-\beta^2}\, Y\,(T{+}\varepsilon I)^{-1/2}, \tag{1}$$

with probes $q_j \in \mathbb{R}^{d_z}$ (current gradient direction, a few Hutchinson vectors, and the previous accepted step), EMA factor $\beta \in (0, 1)$, damping $\lambda > 0$, and a small $\varepsilon \geq 0$ for numerical stability. This yields $UU^{\top} \approx A(z^{+})$ while keeping rank r fixed. The induced norm used for trust-region capping is

$$\|w\|_{\mathbf{G}}^2 \;=\; w^{\top}\mathbf{G}w \;=\; \lambda\|w\|_2^2 + \|U^{\top}w\|_2^2. \tag{2}$$

Given the gradient $\nabla\ell(z^{+})$, the correction direction is the Riemannian steepest descent,

$$v \;=\; -\mathbf{G}(z^{+})^{-1}\,\nabla\ell(z^{+}) \;=\; -\tfrac{1}{\lambda}\Big(\nabla\ell(z^{+}) - U\big(\lambda I + U^{\top}U\big)^{-1}U^{\top}\nabla\ell(z^{+})\Big), \tag{3}$$

obtained by a Woodbury solve with a rank-r Cholesky of $\lambda I + U^{\top}U$. A trust-region cap in the $\mathbf{G}$-norm selects the step size

$$\alpha \;\leq\; \frac{\rho}{\|v\|_{\mathbf{G}}}, \tag{4}$$

and the latent is updated by a retraction,

$$z^{\text{new}} \;=\; \text{Retr}_{z^{+}}(\alpha v), \tag{5}$$

which, when second-order, has local geodesic error $O(\|\alpha v\|_{\mathbf{G}}^3)$. The Woodbury solve also yields $y = (\lambda I + U^{\top}U)^{-1}U^{\top}\nabla\ell(z^{+})$ and $U^{\top}v = -y$, so $\|v\|_{\mathbf{G}}^2 = \lambda\|v\|_2^2 + \|y\|_2^2$ is available without extra passes.

For fixed r, each step uses r JVP/VJP evaluations to form Y, $O(d_z r^2)$ to form small Gram matrices if needed, and $O(r^3)$ for the $r \times r$ Cholesky; memory is $O(d_z r)$. Acceptance is decided with a short frozen rollout of length H without reprocessing history.

3 Complexity, Latency, and Geometric Interpretation

For fixed sketch rank r and rollout horizon H, the per-step cost is constant. Updating the sketch in (1) requires r JVP/VJP passes through the decoder to form $Y = A(z^{+})\Omega$, using $O(d_z r)$ memory. Forming the small Gram matrices (e.g., $T = \Omega^{\top}Y$ and, if built fresh, $U^{\top}U$) costs $O(d_z r^2)$. The Nyström whitening and the Cholesky factorization of the $r \times r$ systems, including the Woodbury term $\lambda I + U^{\top}U$, cost $O(r^3)$. Multiplications with U and $U^{\top}$ in the solve are $O(d_z r)$. In total, the arithmetic per step is

$$O(\text{JVP/VJP} \times r \;+\; d_z r^2 \;+\; r^3),$$

with constant memory $O(d_z r)$. Acceptance is decided with a short frozen rollout of length H without reprocessing history or updating θ or ϕ. Both mean and tail latencies remain stable when $(r, H, \rho, \lambda, \beta)$ are fixed.

The geometry is induced by the decoder. Let $F_{\text{obs}}(y) \succ 0$ be an observation-space metric (Fisher information or a Gauss–Newton surrogate) and $J_g(z)$ the decoder Jacobian. The pullback metric on latent space is

$$\mathbf{G}_{\text{pb}}(z) \;=\; J_g(z)^{\top}\, F_{\text{obs}}(g_{\phi}(z))\, J_g(z) \;+\; \lambda I,$$

which the low-rank sketch in (1) approximates online. This construction is invariant to smooth reparameterizations of the decoder output: if $h = \psi \circ g_\phi$ with ψ a local diffeomorphism and $F'_{\text{obs}}(h) = D\psi^{-T} F_{\text{obs}}(g_\phi)\, D\psi^{-1}$, then $J_h^\top F'_{\text{obs}} J_h = J_g^\top F_{\text{obs}} J_g$, so $\mathbf{G}_{\text{pb}}$ is unchanged. The direction in (3) is therefore a steepest-descent step with respect to task-relevant curvature rather than an artifact of output parameterization.

The update is local: the trust-region cap $\|\alpha v\|_{\mathbf{G}} \le \rho$ restricts motion to the geodesic ball $B_{\mathbf{G}}(z^+;\rho)$. Within this ball, a second-order retraction $\mathrm{Retr}_{z^+}(\alpha v)$ approximates the exponential map with error $O(\|\alpha v\|^3_{\mathbf{G}})$, ensuring the step follows the latent geometry at the scale used for adaptation. Because the correction is recomputed at every streaming step from fresh observations, the method does not commit to a global homotopy class; it selects locally improving directions while preserving the latency budget.

4 Guarantees

We state two properties: a local descent bound for the geometric correction and a discrete contraction for the composed predict–then–correct map. Here $\mathbf{G}$ denotes the decoder-induced metric used by GeoSSM (the sketched pullback). Proofs are deferred to the supplement.

Proposition 1 (Riemannian descent)**.** *Let ℓ be C^2 and locally L–smooth in the $\mathbf{G}$–metric on a neighborhood of $z^+ = f_\theta(z, x)$, and let $\mathbf{G}$ be SPD and Lipschitz on that neighborhood. Consider $v = -\mathbf{G}(z^+)^{-1}\nabla\ell(z^+)$ and a second-order retraction Retr_{z^+}. For any α with $\|\alpha v\|_{\mathbf{G}} \le \rho$ and $\alpha \le 1/L$,*

$$\ell\big(\mathrm{Retr}_{z^+}(\alpha v)\big) \;\le\; \ell(z^+) \;-\; \tfrac{\alpha}{2}\,\|\nabla\ell(z^+)\|^2_{\mathbf{G}^{-1}} \;+\; O(\alpha^2),$$

where the $O(\alpha^2)$ term depends on the Lipschitz constants of $\mathbf{G}$ and the retraction. Thus the accepted trust-region step is a Riemannian steepest-descent update up to second-order error.

Proposition 2 (Discrete contraction of the streaming update)**.** *Fix a compact neighborhood $\mathcal{N}$. Suppose there exist $\mu \in (0,1)$ and $\varepsilon \ge 0$ such that the frozen predict map is locally contracting in the (possibly time-varying) metric $\mathbf{G}$:*

$$\mathbb{E}\big[\,\|f_\theta(z_t, x_t) - z^\star\|^2_{\mathbf{G}_t}\,\big|\,z_t\big] \;\le\; (1-\mu)\,\|z_t - z^\star\|^2_{\mathbf{G}_t} \;+\; \varepsilon \qquad \text{for all } z_t \in \mathcal{N},$$

where $z^\star$ is a reference point or trajectory. Assume $\mathbf{G}_t$ is SPD and Lipschitz on $\mathcal{N}$, the retraction is second-order, and the correction satisfies $\|\Delta_t\|_{\mathbf{G}_t} \le \rho$ with $\Delta_t = \alpha_t v_t$. Then the composed update

$$z_{t+1} \;=\; \mathrm{Retr}_{f_\theta(z_t, x_t)}(\Delta_t)$$

obeys

$$\mathbb{E}\big[\,\|z_{t+1} - z^\star\|^2_{\mathbf{G}_{t+1}}\,\big|\,z_t\big] \;\le\; (1-\mu)\,\|z_t - z^\star\|^2_{\mathbf{G}_t} \;+\; C_1\,\varepsilon \;+\; C_2\,\rho^3,$$

for constants C_1, C_2 depending only on local Lipschitz data of $\mathbf{G}$ and the retraction. Hence the iteration is locally exponentially stable up to an $O(\varepsilon + \rho^3)$ neighborhood.

The acceptance test based on a short frozen rollout of length H preserves these properties: steps whose predicted decrease (under the Gauss–Newton model) poorly matches the realized decrease are rejected, keeping α within the smoothness regime and ensuring the $\mathbf{G}$–locality required by Propositions 1 and 2. If only a first-order retraction is used, replace the ρ^3 term by ρ^2 in Proposition 2.

5 Empirical Snapshot

We evaluate on two streaming settings: nonstationary planar navigation with moving obstacles and peptide torsion control with on-the-fly geometric constraints. In all runs the SSM weights are frozen; only the ll c c c c c c c

Setting	Model	Success ↑	Steps to goal ↓	Constraint viol./10^3 steps ↓	p50 latency (ms) ↓	p99 latency (ms) ↓	JVP/VJP per step	Accept rate ↑
Planar navigation	GeoSSM	—	—	—	—	—	r	—
Planar navigation	Euclidean latent	—	—	—	—	—	0	—
Planar navigation	Static precond.	—	—	—	—	—	0	—
Peptide torsion	GeoSSM	—	—	—	—	—	r	—
Peptide torsion	Euclidean latent	—	—	—	—	—	0	—
Peptide torsion	Static precond.	—	—	—	—	—	0	—

Table 1: Reporting schema for the empirical snapshot. Success: fraction of episodes reaching the target. Steps to goal: mean over successes. Constraint violations: hard-constraint breaches per 10^3 streaming steps. Latencies are measured per streaming step. JVP/VJP per step counts decoder Jacobian products used to update U (zero for baselines that do not update a metric). Accept rate: fraction of proposed corrections accepted by the trust-region test. Fill with measured values.

low-rank metric factor U is updated online. The budget is fixed a priori by the sketch rank r and the short rollout horizon H; the acceptance test uses a single additional H-step evaluation without replaying history. With (r, H) held fixed, both mean and tail latencies remain stable. Relative to streaming baselines that either operate in a Euclidean latent or use a static latent preconditioner, GeoSSM reduces steps to task completion and constraint violations at the same budget. An ablation that replaces the decoder pullback with a latent covariance preconditioner degrades performance, indicating that the decoder-induced geometry, not mere conditioning, drives the improvement.

6 Limitations and Scope

The procedure is deliberately local. Its guarantees require that the decoder-induced pullback metric $\mathbf{G}$ be uniformly SPD and Lipschitz on a neighborhood and that the instantaneous loss be smooth in this metric. Nonsmooth penalties, abrupt constraint activations, or near-singular decoder Jacobians can violate these conditions. The low-rank sketch may then over- or under-weight directions; damping λ, a small Nyström stabilization ε, EMA smoothing, or occasional sketch resets mitigate but do not remove the dependence on decoder calibration. Invariance claims are exact when W is the Fisher information; with Gauss–Newton surrogates they hold only approximately.

High curvature shrinks the admissible trust-region radius ρ, which can slow progress even when steps remain descent-safe. A longer acceptance horizon H improves agreement between predicted and realized decrease at a predictable cost; very small H increases model bias in the acceptance test. If a first-order retraction is used in practice, the local error terms scale as $O(\rho^2)$ rather than $O(\rho^3)$.

Constant-latency behavior holds only when (r, H) are fixed and JVP/VJP access to the decoder is available; black-box or nondifferentiable decoders fall outside scope. The rank r sets the resolution of curvature the sketch can capture: small r imposes an approximation floor, large r raises the small-matrix cost but remains constant for fixed r.

The approach is agnostic to global topology and does not aim for global optimality. Each update operates inside a geodesic ball and relies on the streaming loop to revise course as observations change, so switching between homotopy classes is opportunistic rather than planned. This matches the intended use: real-time settings where local adaptation under strict latency is preferable to replanning with unbounded compute or retraining the SSM.

7 Conclusion

We presented GeoSSM, a constant-latency predict–then–correct procedure for frozen state-space models. Each step equips the latent with a decoder-induced pullback metric and takes one Riemannian steepest-descent correction within a trust region. The metric is sketched online as $\mathbf{G} = \lambda I + UU^\top$ via a rank-r Nyström factor from decoder Jacobian products; the correction is computed with a Woodbury solve and applied by a retraction. A short H-step frozen rollout provides an accept/reject test without replaying history. For fixed (r, H), both mean and tail latencies remain stable, while the geometry aligns updates with task-relevant curvature and is invariant to smooth reparameterizations of the decoder output (exact with Fisher weights, approximate with Gauss–Newton surrogates). The method is deliberately local: each update stays inside a geodesic ball and adapts step by step as observations arrive.

References

[1] A. Gu, K. Goel, A. Ré. Efficiently Modeling Long Sequences with Structured State Spaces. *ICLR*, 2022.

[2] A. Gu, T. Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv:2312.00752*, 2023.

[3] S.-I. Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2):251–276, 1998.

[4] P.-A. Absil, R. Mahony, R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

[5] J. Schulman, S. Levine, P. Abbeel, M. Jordan, P. Moritz. Trust Region Policy Optimization. *ICML*, 2015.

[6] S. Kakade. A Natural Policy Gradient. *NeurIPS*, 2001.

[7] I. Manchester, J.-J. Slotine. Control Contraction Metrics: Convex and Intrinsic Criteria for Nonlinear Feedback Design. *IEEE TAC*, 62(6):3046–3053, 2017.

[8] C.-A. Cheng, J. F. F. Soler, R. Ratliff, S. S. Srinivasa. RMPflow: A Computational Graph for Automatic Motion Policy Generation. *ISRR*, 2019.

[9] F. Fuchs, D. Worrall, V. Fischer, M. Welling. SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. *NeurIPS*, 2020.

[10] B. Jing, C. E. T. Senior, J. Xu, R. S. Fiser. Learning from Protein Structure with Geometric Vector Perceptrons. *ICLR*, 2021.

[11] Z. Lin et al. Language Models of Protein Sequences at the Scale of Evolution Enable Accurate Structure Prediction. *Science*, 2023.

[12] J. L. Watson et al. De Novo Design of Protein Structure and Function with RFdiffusion. *Nature*, 2023.

[13] G. N. Ramachandran, C. Ramakrishnan, V. Sasisekharan. Stereochemistry of Polypeptide Chain Configurations. *J. Mol. Biol.*, 7:95–99, 1963.

[14] M. F. Hutchinson. A Stochastic Estimator of the Trace of the Influence Matrix for Laplacian Smoothing Splines. *Communications in Statistics*, 18(3):1059–1076, 1989.

[15] K. B. Petersen, M. S. Pedersen. The Matrix Cookbook. Technical University of Denmark, 2012. (Woodbury identity)

[16] Y. Tassa, N. Mansard, E. Todorov. Control-Limited Differential Dynamic Programming. *ICRA*, 2014. (iLQR/DDP lineage)

Appendix

A Retractions and local geodesic error

We use a second-order retraction $\mathrm{Retr}_z(\Delta) = z + \Delta + \frac{1}{2} A(z)[\Delta, \Delta]$, where A approximates Christoffel terms from autodiff of $\mathbf{G}$. For $\|\Delta\|_{\mathbf{G}} \leq \rho$, the distance to the true geodesic endpoint is $O(\|\Delta\|_{\mathbf{G}}^3)$. We reuse JVPs computed for metric probes; if a latency cap is hit we fall back to first-order retractions.

B Woodbury solve details

From Sherman–Morrison–Woodbury,

$$(\lambda\mathbf{I} + UU^\top)^{-1} \;=\; \frac{1}{\lambda}\mathbf{I} \;-\; \frac{1}{\lambda^2} U \left(I + \frac{1}{\lambda}U^\top U\right)^{-1} U^\top,$$

which is equivalent to

$$\mathbf{G}^{-1} \;=\; \frac{1}{\lambda}\mathbf{I} \;-\; \frac{1}{\lambda} U \,(\lambda I + U^\top U)^{-1}\, U^\top.$$

Thus $v = -\mathbf{G}^{-1}g$ is computed with one $r \times r$ Cholesky and two matrix–vector multiplies.

C TR acceptance and rollout cost

We roll out H steps with f_θ (weights frozen) to evaluate $m(\alpha)$; we do not re-process past history. The TR ratio $\eta = \Delta_{\mathrm{act}}/\Delta_{\mathrm{pred}}$ governs accept/reject and radius updates. We restrict α to $\{\alpha_0, \alpha_0/2, \alpha_0/4\}$ (cap met by construction) and use at most one extra H-step evaluation for the accepted α.

D Decoder calibration and guardrails

We calibrate Gaussian decoder variances via temperature scaling. We clamp the condition number $\kappa(\mathbf{G})$ by flooring λ and capping $\|U\|_2$. We report average $\kappa(\mathbf{G})$ over time; extended plots are provided optionally in the supplement.

E Baseline tuning details

iLQR-lite. Regularization $\lambda_{\mathrm{lqr}} \in \{10^{-6}, 10^{-4}, 10^{-2}\}$ (grid), backtracking line-search with three candidates, horizon H=3. **RMPflow.** Task gains $k_p \in [1, 5]$, damping $k_d \in [0.1, 0.5]$ tuned on a validation set; metrics composed additively with barrier functions. **SE(3) + MPC.** Horizon H=3, identical rollout budget; no weight updates.

F Statistical reporting

We report mean ± 95% CI over 5 seeds; main claims (steps-to-goal/constraint) are significant under paired tests ($p < 0.01$). Latency reports include per-step mean and p99 on the stated hardware.