MOSAIC: Closed-Form Riemannian Streaming Updates
on a Low-Rank Covariance Bundle

Andrew Kiruluta
University of California, Berkeley, CA.
kiruluta@berkeley.edu

August 20, 2026

Abstract

We study fixed-memory compression of a non-stationary high-dimensional stream through a moving low-rank Gaussian covariance model. The model state is not globally a Cartesian product of an unframed subspace and an independent positive-definite matrix: the gauge transformation $(U, S) \mapsto (UQ, Q^\top SQ)$ leaves the represented covariance unchanged. We therefore formulate the retained covariance state on the associated quotient bundle $\mathcal{B}_{d,r} = (\mathrm{St}(d, r) \times \mathcal{S}^r_{++})/O(r)$, with an online mean appended in the ambient space. This geometry is established prior art; the contribution of MOSAIC is a streaming specialization that exploits the rank-one structure of each sample update.

For the positive-definite fiber, the instantaneous Gaussian negative log-likelihood $f_c(S) = \frac{1}{2}(\log\det S + c^\top S^{-1}c)$ has affine-invariant gradient $\frac{1}{2}(S - cc^\top)$. Its nominal exponential map appears to require matrix square roots and a matrix exponential, but the one-sample rank-one structure collapses exactly to

$$S_{t+1} = e^{-\eta/2}\left[S_t + \frac{e^{\eta q_t/2} - 1}{q_t} c_t c_t^\top\right], \qquad q_t = c_t^\top S_t^{-1} c_t,$$

an $O(r^2)$ square-root-free recursion with an equally inexpensive inverse update. The effective coefficient on $c_t c_t^\top$ is Mahalanobis-adaptive: it is more conservative than the Euclidean first-order update for low-surprise samples and more aggressive beyond a unique surprise threshold. The exact geodesic preserves positive definiteness for every finite positive step size, whereas the Euclidean truncation is positive definite only for $0 < \eta < 2$.

For the moving subspace, we use the classical rank-one Grassmann exponential but couple it to the sketch through a noise-regularized precision direction $(S_t + \widehat{\sigma}_t^2 I)^{-1} c_t$. We derive the exact Gaussian Fisher metric for horizontal subspace perturbations and show that its formal natural direction is $(S - \sigma^2 I)^{-1} c$ when defined; the singularity at $s_i = \sigma^2$ motivates the stable preconditioner actually used by MOSAIC. A frame-coherent horizontal lift makes the induced fiber coordinate change the identity along each rank-one subspace rotation, avoiding an additional congruence update. On drifting synthetic streams, MOSAIC attains 28% lower distortion than the best tested baseline; at rank 16 it compresses a $50{,}000 \times 256$ stream 2778:1 while landing within 5.8% of the offline optimal distortion. Ablations show that precision preconditioning, rather than replacing a first-order covariance retraction by the exact exponential at small step sizes, is the principal source of the accuracy gain. Implementation repo can be found here:https://github.com/andrew-jeremy/Compressive-Online-Learning

Keywords: online learning, streaming compression, covariance bundle, Grassmann manifold, symmetric positive definite cone, affine-invariant geometry, subspace tracking.

1 Introduction

Compression is not an accessory to machine learning; it is close to a definition of it. A model that ingests a corpus and stores what it found in a space of far lower dimension has performed lossy compression, and the quality of that compression is much of what we mean by the quality of the model. The geometric reading of this is now standard: high-dimensional data do not fill their ambient space but concentrate near a structured subset of much smaller intrinsic dimension, and a useful representation adapts its coordinates to that structure rather than relying on an arbitrary fixed projection (Tenenbaum et al., 2000; Roweis and Saul, 2000; Bengio et al., 2013).

Most such compression is retrospective. The corpus is assembled, the representation is fitted in one or more passes, and the resulting coordinates are then frozen. When the data-generating process drifts, the fitted geometry becomes stale and restoring it can require revisiting data that are no longer available. We ask instead what compression looks like when the representation must be discovered and maintained *prospectively*, from a state whose size never grows with the stream.

The correct global state is a covariance bundle. Let $U \in \mathrm{St}(d, r)$ be an orthonormal frame and let $S \in \mathcal{S}^r_{++}$ describe the retained second moment in that frame. The pair is redundant: for every $Q \in O(r)$,

$$USU^\top = (UQ)(Q^\top SQ)(UQ)^\top. \tag{1}$$

Thus an unframed Grassmann point $[U]$ cannot be paired globally with an independent matrix S without specifying how that matrix transforms when the frame changes. We therefore write the retained covariance state as the equivalence class

$$[U, S] \in \mathcal{B}_{d,r} := \big(\mathrm{St}(d, r) \times \mathcal{S}^r_{++}\big)/O(r), \qquad (U, S) \sim (UQ, Q^\top SQ), \tag{2}$$

and append the online mean,

$$\mathfrak{M}_t = (\mu_t, [U_t, S_t]), \qquad \text{size } O(dr + r^2) \text{ independent of } t. \tag{3}$$

This quotient/associated-bundle construction is not new; closely related representations underlie fixed-rank PSD geometry and low-rank-plus-identity covariance estimation (Bonnabel and Sepulchre, 2010; Massart and Absil, 2020; Bouchard et al., 2021). Our question is algorithmic: what closed forms become available when this geometry is specialized to one-sample streaming Gaussian updates?

Where the novelty is, and is not. The rank-one Grassmann exponential is classical and is already used by streaming subspace methods such as GROUSE (Balzano et al., 2010). The affine-invariant geometry of $\mathcal{S}^r_{++}$, quotient geometries for low-rank PSD matrices, and Riemannian low-rank-plus-identity covariance estimation are also established (Bhatia, 2007; Bonnabel and Sepulchre, 2010; Bouchard et al., 2021). Streaming probabilistic PCA likewise predates this work (Gilman et al., 2025). We therefore make no novelty claim for those constituents. The paper instead isolates three streaming consequences of the particular one-sample likelihood.

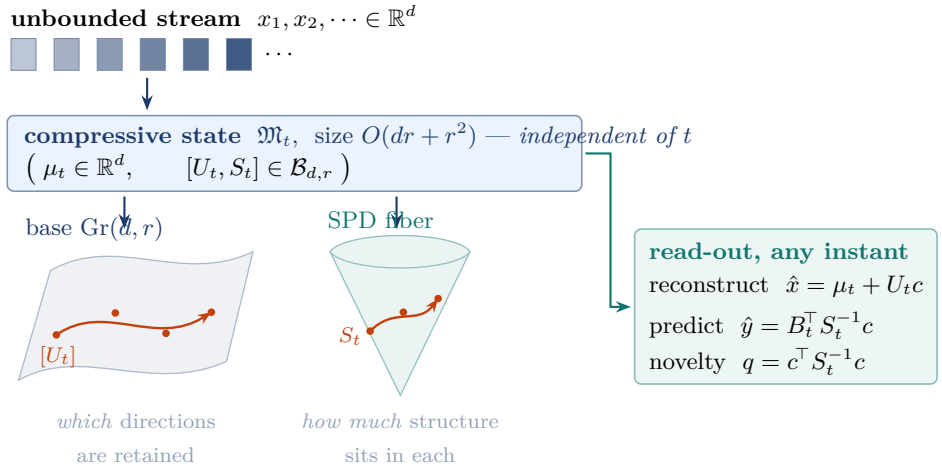


Figure 1: MOSAIC compresses an unbounded stream into a fixed-size state $(\mu_t, [U_t, S_t])$. The Grassmann base records *which* directions are retained and the SPD fiber records *how much* second-order structure is assigned inside the current frame. The pair is gauge-coupled by $(U, S) \sim (UQ, Q^\top SQ)$ rather than being a global Cartesian product. The same state supports reconstruction, prediction, and a Mahalanobis novelty score at any instant.

1. **An exact rank-one affine-invariant likelihood exponential.** For the instantaneous fiber loss, the nominal SPD exponential map collapses to

$$S_{t+1} = e^{-\eta_S/2}\Big[S_t + \frac{e^{\eta_S q_t/2} - 1}{q_t} c_t c_t^\top\Big], \qquad q_t = c_t^\top S_t^{-1} c_t, \tag{4}$$

   with no matrix square root, eigendecomposition, or matrix exponential. The step and its inverse update cost $O(r^2)$. To our knowledge, this exact square-root-free specialization of the one-sample Gaussian AIRM exponential has not previously been used as the core recursion of a fixed-memory streaming covariance tracker.

2. **Frame-coherent coupling with zero coordinate-conversion cost.** For a rank-one Grassmann rotation we choose the horizontal frame lift carried by the same ambient planar rotation. The induced $O(r)$ map on retained coordinates is exactly the identity (Theorem 5.2), so the SPD fiber need not undergo a separate congruence transformation after every subspace move. The novelty claim is not that parallel transport on the Grassmannian is new, but that this gauge choice makes the coupled streaming update separable at rank-one cost.

3. **Noise-regularized precision allocation of angular motion.** The exact Fisher metric of the spiked Gaussian family implies a formal natural coordinate $(S - \sigma^2 I)^{-1} c$ wherever it is defined; it becomes singular when a retained variance meets the noise floor. MOSAIC therefore uses the stable proxy

$$\xi_t = (S_t + \widehat{\sigma}_t^2 I)^{-1} c_t, \tag{5}$$

   which preserves inverse-variance allocation without the Fisher singularity. Our ablation identifies this precision preconditioning as the component that materially changes accuracy.

What the exact SPD geometry buys. The Euclidean first-order covariance update $S_+^{\mathrm{euc}} = (1 - \eta/2)S + (\eta/2)cc^\top$ agrees with (4) to $O(\eta^2)$, so at the tuned small step sizes the

two produce indistinguishable distortion. The exact exponential instead buys structure and a different large-step response. It is positive definite for every finite $\eta > 0$, whereas the Euclidean truncation is positive definite for $0 < \eta < 2$, becomes singular at $\eta = 2$ for $r > 1$, and is indefinite for $\eta > 2$. Moreover, the actual coefficient multiplying $cc^\top$ in the geodesic update,

$$\alpha_\eta(q) = e^{-\eta/2} \frac{e^{\eta q/2} - 1}{q}, \tag{6}$$

is *not* uniformly larger than $\eta/2$. It is smaller for low-surprise samples and larger beyond a unique Mahalanobis-energy threshold (Theorem 5.4). This surprise-adaptive behavior is more informative than the incorrect claim of uniform dominance.

Empirical summary. Under matched memory and per-method tuning, MOSAIC attains 28% lower distortion than the best tested baseline on the principal controlled drifting-stream benchmark. At rank 16 it compresses a $50{,}000 \times 256$ stream 2778:1 while landing within 5.8% of the offline optimal distortion. The advantage widens as the latent spectrum becomes more anisotropic. A primal-to-precision interpolation makes explicit why this can trade subspace-angle fidelity for lower energy-weighted distortion. The revised evaluation also adds four real-data streams and the stronger IncrementalPCA and SHASTA-PCA baselines. Those results are competitive rather than uniformly dominant: MOSAIC consistently improves on GROUSE and Oja in reconstruction, while IncrementalPCA, Frequent Directions, or SHASTA-PCA can win on particular real datasets. This narrows the practical claim to the non-stationary, anisotropic, memory-constrained regime for which the method was designed.

Contributions.

- A globally coherent formulation of the streaming state on the structured covariance bundle $\mathcal{B}_{d,r}$, with explicit acknowledgement of the quotient geometry already established in the literature (Section 3.1).
- The exact square-root-free rank-one AIRM likelihood exponential (4), an $O(r^2)$ inverse update, unconditional SPD preservation, and a corrected surprise-threshold comparison with the Euclidean first-order recursion (Theorems 5.3 and 5.4).
- A Fisher-geometric audit of horizontal subspace motion showing the exact formal natural direction and motivating the implemented noise-regularized precision preconditioner rather than misidentifying it as an exact natural gradient (Section 4.2).
- A frame-coherent rank-one Grassmann lift whose induced fiber coordinate map is the identity, plus empirical ablations identifying which geometric components affect accuracy and which primarily provide guarantees. The revised evaluation adds a tunable subspace–compression trade-off, four real-data streams, downstream classification, and contemporary IncrementalPCA/SHASTA-PCA comparisons (Theorem 5.2 and Section 6).

Notation. $\mathrm{Gr}(d, r)$ is the Grassmann manifold of r-dimensional subspaces of $\mathbb{R}^d$, represented by orthonormal $U \in \mathrm{St}(d, r)$. $\mathcal{S}^r_{++}$ is the cone of $r \times r$ symmetric positive definite matrices with affine-invariant metric $g_S(X, Y) = \mathrm{tr}(S^{-1}XS^{-1}Y)$. The bundle equivalence is $(U, S) \sim (UQ, Q^\top SQ)$. We write $z_t = x_t - \mu_{t-1}$, $c_t = U_{t-1}^\top z_t$, $\rho_t = z_t - U_{t-1}c_t$, and $e_t = \|\rho_t\|$. Step sizes for the subspace and SPD updates are η_U and η_S.

Table 1: Scope of the streaming baselines used in the revised evaluation. State scaling is for the direct implementations compared here, with mini-batch size b for IncrementalPCA. SHASTA has capabilities beyond the common fully-observed setting used in our numerical comparison.

Method	Update mode	State scaling	Probabilistic noise model	Missing data	Forgetting/adaptation
Oja	sample	$O(dr)$	no	no	step-size dependent
GROUSE	sample	$O(dr)$	no	yes	step-size dependent
Frequent Directions	sample	$O(dr)$	no	no	no in standard form
IncrementalPCA	mini-batch	$O(dr + bd)$	no	no	no in tested form
SHASTA-PCA	sample	$O(dr^2 + dr)$	yes	yes	yes
MOSAIC	sample	$O(dr + r^2 + d)$	yes	not in this work	yes

2 Related work and novelty boundary

Fixed-rank PSD and structured covariance geometry. The global coupling between a subspace frame and an SPD coordinate matrix is established prior art. Bonnabel and Sepulchre (2010) represent fixed-rank PSD matrices through a quotient involving a Stiefel factor and a positive-definite factor, while Massart and Absil (2020) develop a quotient geometry with simple geodesics for fixed-rank PSD matrices. Particularly close to the present statistical model, Bouchard et al. (2021) study low-rank structured elliptical covariances with low-rank-plus-identity form using a Stiefel–positive-definite quotient geometry. More recently, Marconi (2026) make the associated-bundle representation $\mathrm{St}(n, r) \times_{O(r)} \mathcal{S}^r_{++}$ explicit for fixed-rank covariance geometry. These works rule out any novelty claim based merely on combining a Grassmann subspace with an SPD factor. Our contribution is instead the closed-form *streaming likelihood dynamics* obtained from the one-sample rank-one structure.

Streaming low-rank estimation and probabilistic PCA. Oja's rule (Oja, 1982) and related streaming-PCA methods track a low-dimensional subspace with fixed or slowly varying memory. Representative geometric and recursive approaches include GROUSE (Balzano et al., 2010) and PETRELS (Chi et al., 2013). Frequent Directions (Liberty, 2013; Ghashami et al., 2016) instead maintains a deterministic matrix sketch with covariance-error guarantees, while randomized sketching provides a broader numerical-linear-algebra view (Woodruff, 2014). Streaming probabilistic formulations are also established: SHASTA-PCA jointly tracks a low-rank probabilistic model and heterogeneous noise parameters from streaming data with missing entries (Gilman et al., 2025). Batch-wise incremental PCA is another strong practical reference; the implementation used in our experiments follows the incremental PCA construction of Ross et al. (2008). Accordingly, MOSAIC is not positioned as the first probabilistic or fixed-memory streaming subspace method. The revised experiments compare directly with both IncrementalPCA and the full-observation, single-noise-group specialization of the authors' SHASTA recursion in addition to GROUSE, Oja, and Frequent Directions.

Grassmann geometry. The geometry of $\mathrm{Gr}(d, r)$ and $\mathrm{St}(d, r)$, including rank-one geodesic updates, is classical (Edelman et al., 1998; Absil et al., 2008). The planar rotation used by MOSAIC is the same basic construction exploited by GROUSE, and we make no claim of novelty for it. What differs is the coordinate allocated to that rotation. We derive the exact Fisher metric of the spiked Gaussian model, show why its formal natural direction is singular at the noise floor, and use the bounded precision proxy $(S + \widehat{\sigma}^2 I)^{-1} c$. The resulting update should be

read as covariance-aware preconditioning, not as a new Grassmann exponential.

Geometry of the SPD cone. The affine-invariant metric, geodesic, exponential map, and Hadamard structure of $\mathcal{S}^r_{++}$ are classical (Bhatia, 2007; Pennec et al., 2006). Geodesic convexity in covariance estimation is well developed (Wiesel, 2012; Sra and Hosseini, 2015), and Riemannian covariance methods are widely used in applications such as brain–computer interfaces (Barachant et al., 2013). Euclidean averaging can exhibit the well-known determinant "swelling" effect relative to geometric averaging (Arsigny et al., 2006). The new claim here is narrower: for the one-sample Gaussian loss $f_c(S) = \frac{1}{2}(\log\det S + c^\top S^{-1}c)$, the exact AIRM exponential along $-\operatorname{grad} f_c$ reduces algebraically to a square-root-free rank-one update. This identity makes the exact intrinsic step computationally competitive with its first-order Euclidean truncation.

Riemannian stochastic and online optimization. Riemannian stochastic gradient methods are established (Bonnabel, 2013), as are first-order methods for geodesically convex optimization (Zhang and Sra, 2016). Zhang et al. (2016) study Riemannian SVRG for finite-sum stochastic optimization; that work is not an online-regret theorem. No-regret online learning over Riemannian manifolds was subsequently developed by Wang et al. (2021), with a fuller treatment by Wang et al. (2023). We use those results only to state a standard $O(\sqrt{T})$ corollary for a projected, diminishing-step variant of the SPD subproblem; the constant can depend on sectional-curvature and diameter bounds. The constant-step, unprojected joint algorithm evaluated in this paper is not covered by that theorem.

Continual learning and information geometry. Continual-learning systems address non-stationary streams using mechanisms such as replay, regularization, or architecture growth (Kirkpatrick et al., 2017; Parisi et al., 2019; Lopez-Paz and Ranzato, 2017). MOSAIC instead stores a small structured second-order state rather than exemplars. The Fisher calculation in Section 4.2 is used as a diagnostic of the correct statistical geometry, but the implemented precision rule is deliberately a regularized proxy, not an assertion that the full MOSAIC recursion is exact natural-gradient descent in the sense of Amari (1998).

3 Geometric preliminaries

We first distinguish the global state geometry from the local coordinate calculations used by the algorithm. Grassmann and SPD geometry are classical (Edelman et al., 1998; Absil et al., 2008; Bhatia, 2007); the important structural point is that their coordinates are gauge-coupled.

3.1 The structured covariance bundle

For an orthonormal frame $U \in \mathrm{St}(d, r)$ and $S \in \mathcal{S}^r_{++}$, define the spiked covariance

$$\Sigma(U, S) = USU^\top + \sigma^2(I - UU^\top) = \sigma^2 I + U(S - \sigma^2 I)U^\top. \tag{7}$$

For every $Q \in O(r)$, $\Sigma(UQ, Q^\top SQ) = \Sigma(U, S)$. Hence the natural parameter object is

$$\mathcal{B}_{d,r} = (\mathrm{St}(d, r) \times \mathcal{S}^r_{++})/O(r), \qquad (U, S) \cdot Q = (UQ, Q^\top SQ). \tag{8}$$

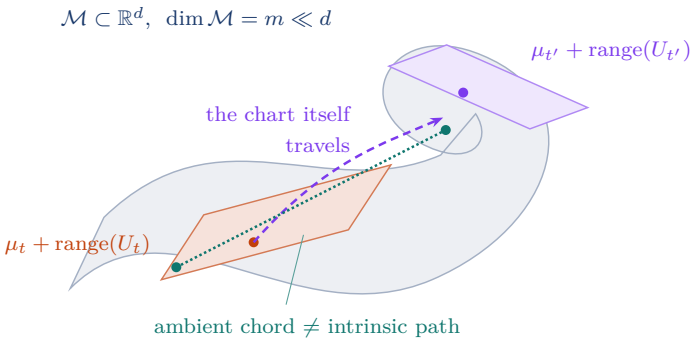


Figure 2: A moving local chart. A rank-r subspace is a local approximation to structured high-dimensional data; under drift the chart must move. MOSAIC tracks a base point, a subspace, and a covariance coordinate attached to the current frame.

Equivalently, $\mathcal{B}_{d,r}$ is an associated bundle over $\mathrm{Gr}(d, r)$ whose SPD fiber contains covariance coordinates in the chosen moving frame. Closely related quotient constructions are standard in fixed-rank PSD and structured covariance geometry (Bonnabel and Sepulchre, 2010; Massart and Absil, 2020; Bouchard et al., 2021; Marconi, 2026). We therefore use the bundle to make the model globally coherent, not as a novelty claim.

A local representative (U, S) remains computationally convenient. The rank-one subspace step moves U horizontally, while the fiber step updates S with the affine-invariant metric. Theorem 5.2 shows that the particular horizontal lift used by MOSAIC induces the identity coordinate map on the fiber for each rank-one rotation.

3.2 Why a chart must move

The premise of low-dimensional representation is that data in $\mathbb{R}^d$ often concentrate near a structured subset of intrinsic dimension much smaller than d. A rank-r linear sketch is therefore a defensible local model, but a single global projection need not remain appropriate under drift. A streaming learner should carry a moving base point, a retained subspace, and a local second-order description. In MOSAIC these are μ_t, $\mathrm{range}(U_t)$, and the fiber coordinate S_t.

3.3 The Grassmann base

$\mathrm{Gr}(d, r)$ is the set of r-dimensional linear subspaces of $\mathbb{R}^d$, realized as $\mathrm{St}(d, r)/O(r)$. A tangent vector at $[U]$ is represented by a horizontal lift $\Delta \in \mathbb{R}^{d\times r}$ with $U^\top \Delta = 0$. Under the canonical metric, $\langle \Delta_1, \Delta_2 \rangle = \mathrm{tr}(\Delta_1^\top \Delta_2)$. Given a thin SVD $\Delta = W\Theta V^\top$, a representative of the Grassmann geodesic is

$$U(\tau) = UV \cos(\Theta\tau)V^\top + W \sin(\Theta\tau)V^\top + U(I - VV^\top). \tag{9}$$

We need only the rank-one case $\Delta = \theta w v^\top$, which reduces to a planar rotation (Theorem 5.1).

3.4 The SPD fiber

On $\mathcal{S}^r_{++}$ we use the affine-invariant metric

$$g_S(X, Y) = \mathrm{tr}(S^{-1}XS^{-1}Y), \qquad X, Y \in T_S\mathcal{S}^r_{++} \cong \mathcal{S}^r. \tag{10}$$

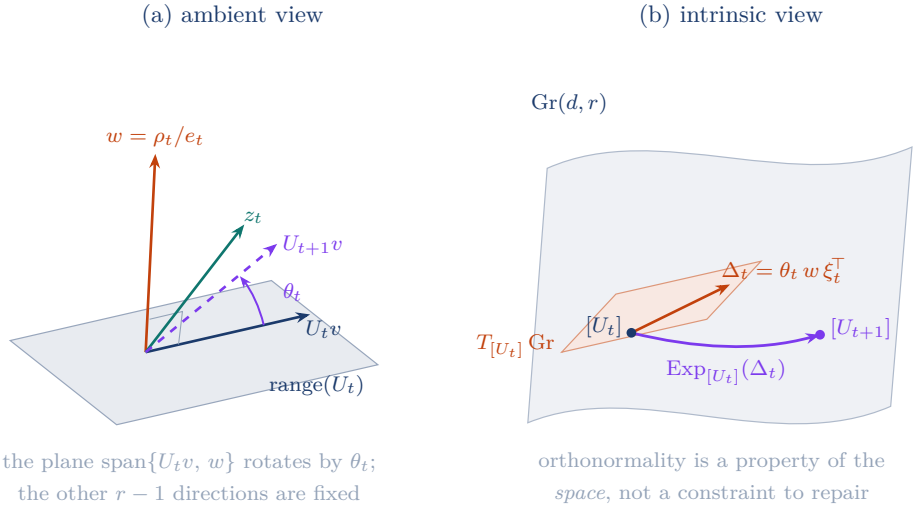


Figure 3: The rank-one Grassmann geodesic. The plane spanned by the retained coordinate direction Uv and residual direction w rotates by θ; the remaining $r-1$ retained directions are fixed.

Its exponential map, geodesic, and distance are

$$\mathrm{Exp}_S(X) = S^{1/2} \exp(S^{-1/2} X S^{-1/2}) S^{1/2}, \tag{11}$$

$$P \#_t Q = P^{1/2} (P^{-1/2} Q P^{-1/2})^t P^{1/2}, \tag{12}$$

$$\delta(P, Q) = \left\| \log(P^{-1/2} Q P^{-1/2}) \right\|_F. \tag{13}$$

The manifold is complete, simply connected, and nonpositively curved. The boundary is at infinite affine-invariant distance, which underlies the unconditional positive-definiteness of an exact finite geodesic step.

For $P, Q \in \mathcal{S}^r_{++}$,

$$\det\left(\tfrac{1}{2}(P+Q)\right) \geq \det(P \#_{1/2} Q) = \sqrt{\det P \det Q}, \tag{14}$$

with equality iff $P = Q$ (Theorem 5.6). Thus distinct matrices exhibit strict determinant inflation under arithmetic averaging relative to the AIRM midpoint; commutativity alone is not an equality condition.

3.5 Riemannian gradients and preconditioning

For $f : \mathcal{M} \to \mathbb{R}$, the Riemannian gradient satisfies $g_p(\mathrm{grad}\, f(p), v) = df_p(v)$. Changing the metric changes the raised gradient and is the mechanism behind natural-gradient methods (Amari, 1998). This observation must be used carefully: a preconditioner inspired by an SPD sketch is not automatically the Fisher metric of the statistical model. In Section 4.2 we calculate the Gaussian Fisher metric explicitly before introducing MOSAIC's regularized precision proxy.

A retraction agrees with the exponential map to first order and is often cheaper. For the SPD likelihood direction, Theorem 5.5 shows that the standard Euclidean covariance recursion is exactly the first-order truncation of the intrinsic step. The two therefore agree at small step sizes, while only the exact exponential is guaranteed to remain in $\mathcal{S}^r_{++}$ for every finite positive step.

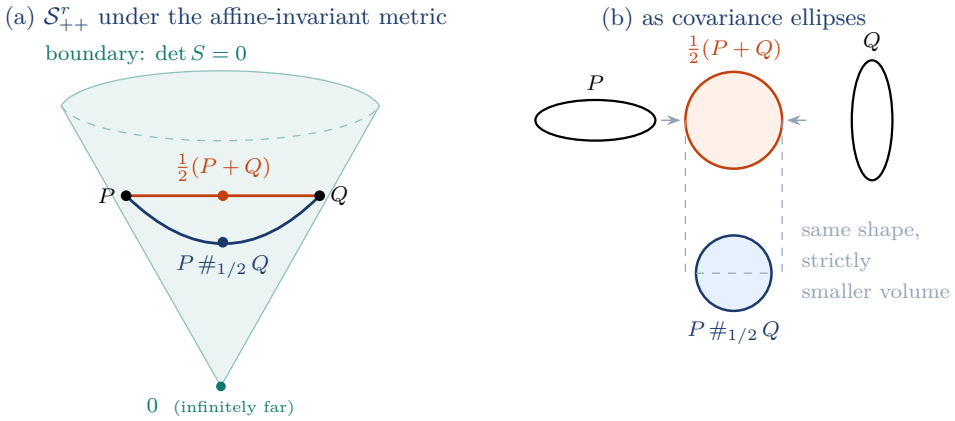


Figure 4: Arithmetic versus affine-invariant geometric averaging on the SPD cone. For distinct P and Q, the arithmetic midpoint has strictly larger determinant than the AIRM midpoint.

4 The MOSAIC scheme

4.1 One likelihood, two different geometric roles

We model the centered stream locally with the spiked covariance

$$\Sigma(U, S) = USU^\top + \sigma^2(I - UU^\top), \tag{15}$$

whose gauge symmetry is described in Section 3.1. Because U is orthonormal, the one-sample negative log-likelihood reduces, up to constants, to

$$f_t(U, S) = \tfrac{1}{2}\left[\log\det S + c^\top S^{-1} c + \frac{e^2}{\sigma^2}\right], \qquad c = U^\top z_t, \quad e = \left\|(I - UU^\top) z_t\right\|. \tag{16}$$

The same likelihood supplies the rank-one residual geometry of the subspace update and the exact AIRM gradient of the SPD update. We do *not*, however, claim that the implemented subspace preconditioner is the exact natural gradient of this statistical model; the Fisher calculation below shows precisely where the two differ.

The residual noise floor is estimated online as $\widehat{\sigma}_t^2 = (1-\beta)\widehat{\sigma}_{t-1}^2 + \beta e_t^2/(d-r)$ with $\beta = 5\times10^{-3}$ throughout.

4.2 The subspace factor: Fisher audit and precision preconditioning

Differentiating (16) in U and projecting onto the horizontal space yields a rank-one canonical Grassmann gradient

$$\operatorname{grad}^{\mathrm{can}}_U f_t = -\rho_t\, \xi_t^{\mathrm{e}\top}, \qquad \xi_t^{\mathrm{e}} = (\sigma^{-2} I - S_t^{-1}) c_t, \tag{17}$$

where $\rho_t = (I - U_t U_t^\top) z_t$. Rank-one structure means that any chosen coordinate vector can be inserted into a closed-form planar Grassmann rotation at $O(dr)$ cost.

What the exact Fisher geometry says. For a horizontal perturbation Δ with $U^\top\Delta = 0$, differentiating (15) gives $\dot{\Sigma} = \Delta A U^\top + U A \Delta^\top$ with $A = S - \sigma^2 I$. The zero-mean Gaussian

Fisher metric for covariance perturbations, $\frac{1}{2}\operatorname{tr}(\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\Sigma}_2)$, therefore restricts to

$$g_F(\Delta_1, \Delta_2) = \sigma^{-2} \operatorname{tr}\left[\Delta_1^\top \Delta_2 \, A S^{-1} A\right]. \tag{18}$$

Where A is invertible, raising the canonical differential with this metric gives the formal Fisher-natural coordinate

$$\xi_t^F \propto (S_t - \sigma^2 I)^{-1} c_t. \tag{19}$$

The formula exposes an identifiability boundary rather than a usable update: if a retained variance equals the residual noise variance, rotations involving that coordinate do not change the covariance to first order and the Fisher metric becomes singular. Near that boundary (19) is numerically unstable.

A two-coordinate picture. The singularity is easiest to see in a basis that diagonalizes S. If $S = \operatorname{diag}(s_1, \ldots, s_r)$, then the Fisher metric weights horizontal motion in coordinate i in proportion to $(s_i - \sigma^2)^2/(\sigma^2 s_i)$. A coordinate whose retained variance is indistinguishable from the residual floor therefore carries essentially no first-order information about subspace rotation, so inverting the metric amplifies that coordinate. For example, with $\sigma^2 = 1$, $S = \operatorname{diag}(9, 1.05)$, and $c = (1, 1)^\top$, the formal Fisher coordinate is proportional to $(1/8, 20)$: the nearly unidentifiable second direction is magnified by a factor of 160 relative to the first. The regularized rule below instead gives $(1/10, 1/2.05)$, a ratio below five. This simple example captures why the exact Fisher calculation is useful diagnostically but is a poor numerical prescription at the noise boundary.

The implemented rule is a stable precision proxy. MOSAIC therefore uses

$$\boxed{\xi_t = (S_t + \widehat{\sigma}_t^2 I)^{-1} c_t} \tag{20}$$

instead of claiming to execute (19). The rule keeps the inverse-variance allocation that motivates the method while replacing the Fisher pole at $s_i = \sigma^2$ by a positive noise-floor regularizer. Directions already carrying large retained variance receive less angular weight; weak directions receive more, but the amplification is capped by $\widehat{\sigma}^{-2}$. This is best described as *noise-regularized precision preconditioning*. The ablation in Table 8 shows that it is the component that materially changes performance.

To expose the compression–subspace trade-off without changing the default algorithm, we also define a diagnostic interpolation

$$\xi_t^{(\lambda)} = \left[(1-\lambda) I + \lambda \, \bar{s}_t (S_t + \widehat{\sigma}_t^2 I)^{-1}\right] c_t, \qquad \bar{s}_t = \frac{\operatorname{tr} S_t}{r} + \widehat{\sigma}_t^2, \quad \lambda \in [0, 1]. \tag{21}$$

The scalar $\bar{s}_t$ only places the primal and precision terms on comparable units; both endpoints are unchanged as directions after normalization. $\lambda = 0$ recovers the primal coordinate used by reconstruction-driven Grassmann tracking, while $\lambda = 1$ is the MOSAIC rule. This interpolation is used only in the trade-off experiment of Section 6.3; the reported MOSAIC results use $\lambda = 1$ throughout.

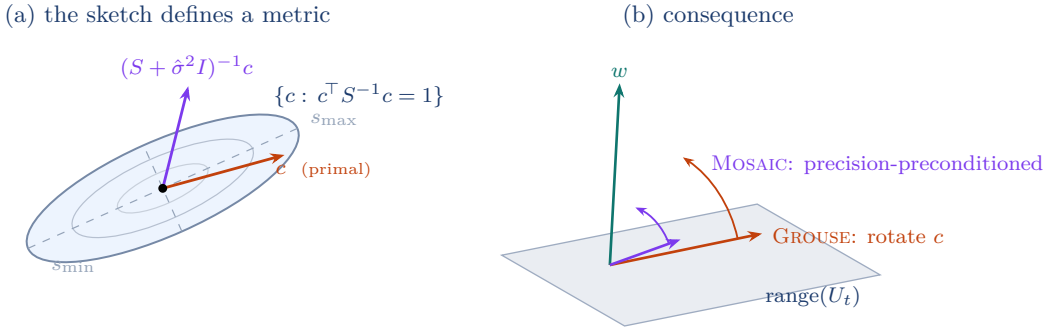


Figure 5: Primal versus precision-preconditioned coordinates. The sketch ellipsoid defines anisotropic retained variance. Multiplication by $(S+\widehat{\sigma}^2 I)^{-1}$ reallocates a fixed angular budget away from already dominant coordinates and toward weak coordinates, while the noise floor prevents unbounded amplification.

Step size. With $v_t = \xi_t / \|\xi_t\|$ and $w_t = \rho_t / e_t$, MOSAIC rotates the plane $\mathrm{span}\{U_t v_t, w_t\}$ by

$$\theta_t = \eta_U \arctan\left(\frac{e_t}{c_t^\top v_t}\right). \tag{22}$$

Because $c_t^\top \xi_t > 0$, the denominator is positive. The angle is bounded by $\eta_U \pi/2$ and coincides with the usual primal-coordinate construction when S_t is isotropic up to scale.

4.3 The SPD fiber: an exact geodesic without square roots

With U fixed, the fiber loss is $f_c(S) = \frac{1}{2}(\log\det S + c^\top S^{-1} c)$. Under the affine-invariant metric,

$$\mathrm{grad}_S\, f_c(S) = \tfrac{1}{2}(S - cc^\top). \tag{23}$$

The intrinsic step $\mathrm{Exp}_S(-\eta_S \,\mathrm{grad}\, f_c)$ appears to require $S^{1/2}$, $S^{-1/2}$, and a matrix exponential. The rank-one sample structure makes all three disappear algebraically (Theorem 5.3):

$$\boxed{S_{t+1} = e^{-\eta_S/2}\left[S_t + \frac{e^{\eta_S q_t/2} - 1}{q_t} c_t c_t^\top\right], \qquad q_t = c_t^\top S_t^{-1} c_t.} \tag{24}$$

The update and the Sherman–Morrison inverse update both cost $O(r^2)$.

The Mahalanobis energy q_t controls the rank-one gain. For comparison with the Euclidean first-order update, the relevant coefficient is not $g(q) = (e^{\eta q/2} - 1)/q$ alone but

$$\alpha_\eta(q) = e^{-\eta/2}\frac{e^{\eta q/2} - 1}{q}, \tag{25}$$

because the outer exponential multiplies the entire update. Consequently the intrinsic recursion is surprise-adaptive rather than uniformly more aggressive: for low q it puts less weight on $cc^\top$ than the Euclidean coefficient $\eta/2$, while for sufficiently high q it puts more (Theorem 5.4). At small η, $\alpha_\eta(q) = \eta/2 + \eta^2(q-2)/8 + O(\eta^3)$, so the crossover is near $q = 2$.

4.4 Frame coherence: a convenient bundle gauge

When U changes, the matrix S is a coordinate in a moving frame and in general transforms by $S \mapsto Q^\top S Q$. For the rank-one Grassmann move, MOSAIC uses the horizontal frame carried by the same ambient planar rotation. The induced coordinate map $Q = U_+^\top \mathcal{P} U$ is exactly I_r (Theorem 5.2). Thus the bundle coupling is respected without an extra congruence transformation or refactorization. This should be understood as a computationally convenient gauge choice inside an established quotient geometry, not as a claim that the bundle or Grassmann transport itself is new.

4.5 Rank allocation, read-out, and cost

Rank. The retained rank need not be fixed. Every $T_{\text{adapt}} = 50$ steps we compare an exponentially-weighted residual energy fraction against a threshold: if the subspace is systematically under-explaining, a new direction is appended along the current residual with initial variance e_t^2 (a rank-inflation limit of geodesics in the larger cone); if the weakest eigendirection of S_t carries less than a $\tau_\downarrow$ fraction of the trace, it is dropped. Growth is information-preserving; shrinkage is the only lossy operation in the scheme, and its cost is exactly the affine-invariant distance between the sketch and its projection.

Read-out. The same state serves inference at any instant, with no separate inference pass. Carrying a cross-moment $B_t \in \mathbb{R}^{r \times p}$ under the same geodesic weighting gives

$$\hat{x}_t = \mu_t + U_t c_t \quad \text{(reconstruction)}, \qquad \hat{y}_t = B_t^\top S_t^{-1} c_t \quad \text{(prediction)}, \qquad q_t = c_t^\top S_t^{-1} c_t \quad \text{(novelty)}. \tag{26}$$

That the novelty score and the update gain are the same quantity is not a coincidence: both ask how surprising the datum is in the metric the model currently holds.

Cost. Per sample, MOSAIC costs $O(dr)$ for the projection and geodesic, $O(r^2)$ for the sketch and its inverse, and $O(r^3)$ for the damped solve in (20); the ambient term dominates whenever $r^2 \lesssim d$, which holds in every configuration we run. Memory is $dr + r^2 + d + rp$ floats, independent of stream length. Measured throughput is 5,771–18,689 samples/s in single-threaded NumPy (Table 9).

5 Theory

Proofs are in Section A. The numerical identities are additionally checked in `code/verify_geometry.py`; the resulting deviations appear in Table 2.

5.1 Exact rank-one geometric identities

Proposition 5.1 (Classical rank-one Grassmann exponential)**.** *Let $U \in \mathrm{St}(d, r)$, let $v \in \mathbb{R}^r$ and $w \in \mathbb{R}^d$ be unit vectors with $U^\top w = 0$, and let $\Delta = \theta w v^\top$. Then a representative of $\mathrm{Exp}_{[U]}(\Delta)$ is*

$$U_+ = U + \left[(\cos\theta - 1) U v + \sin\theta\, w\right] v^\top. \tag{27}$$

It satisfies $U_+^\top U_+ = I_r$ exactly and costs $O(dr)$. This proposition is a specialization of standard Grassmann geometry and is not claimed as novel.

Algorithm 1 MOSAIC: one compressive online step

Input: datum x_t (optional target y_t), step sizes η_U, η_S, η_μ
State: $\mu \in \mathbb{R}^d$, $U \in \mathrm{St}(d,r)$, $S \in \mathcal{S}^r_{++}$, S^{-1}, $\widehat{\sigma}^2$, B

$z \leftarrow x_t - \mu$; $\quad c \leftarrow U^\top z$; $\quad \rho \leftarrow z - Uc$; $\quad e \leftarrow \|\rho\|$ ▷ $O(dr)$
$\widehat{\sigma}^2 \leftarrow (1-\beta)\widehat{\sigma}^2 + \beta\, e^2/(d-r)$ ▷ online noise floor
$\xi \leftarrow (S + \widehat{\sigma}^2 I)^{-1} c$; $\quad v \leftarrow \xi / \|\xi\|$; $\quad w \leftarrow \rho / e$ ▷ precision-preconditioned direction, (20)
$\theta \leftarrow \eta_U \arctan\big(e/(c^\top v)\big)$ ▷ self-limiting angle, (22)
$U \leftarrow U + \big[(\cos\theta - 1)Uv + \sin\theta\, w\big]v^\top$ ▷ Grassmann Exp, exact, $O(dr)$
$c \leftarrow U^\top z$ ▷ frame-coherent coordinates (Theorem 5.2)
$q \leftarrow c^\top S^{-1} c$; $\quad g \leftarrow (e^{\eta_S q/2} - 1)/q$ ▷ Mahalanobis gain
$S \leftarrow e^{-\eta_S/2}\big[S + g\, cc^\top\big]$; $\quad$ update S^{-1} by Sherman–Morrison ▷ SPD Exp, exact, $O(r^2)$
$B \leftarrow e^{-\eta_S/2}\big[B + g\, cy_t^\top\big]$; $\quad \mu \leftarrow \mu + \eta_\mu z$
every T_{adapt} steps: grow or shrink r (Section 4.5)

Proposition 5.2 (Frame-coherent bundle lift)**.** *Let U_+ be given by (27) and let $\mathcal{P}$ be the ambient planar rotation that carries the old frame along the same rank-one geodesic. Then*

$$Q = U_+^\top \mathcal{P} U = I_r. \tag{28}$$

Hence the gauge-consistent fiber transformation $S \mapsto Q^\top S Q$ is the identity for this lift.

Proposition 5.3 (Square-root-free AIRM likelihood exponential)**.** *Let $S \in \mathcal{S}^r_{++}$, $c \in \mathbb{R}^r$, and $f_c(S) = \frac{1}{2}(\log\det S + c^\top S^{-1} c)$. Under the affine-invariant metric, $\mathrm{grad}\, f_c(S) = \frac{1}{2}(S - cc^\top)$. For $\eta > 0$ and $q = c^\top S^{-1} c > 0$,*

$$\mathrm{Exp}_S(-\eta\, \mathrm{grad}\, f_c(S)) = e^{-\eta/2}\left[S + \frac{e^{\eta q/2} - 1}{q} cc^\top\right]. \tag{29}$$

The step costs $O(r^2)$ given S^{-1} and requires no matrix square root, matrix exponential, or eigendecomposition.

Theorem 5.3 is the principal closed-form identity of the paper.

5.2 Surprise gating, positivity, and the Euclidean truncation

Proposition 5.4 (Unconditional positivity and surprise threshold)**.** *Assume $r \geq 2$. Let S_+^{geo} be (29) and let*

$$S_+^{\mathrm{euc}} = (1 - \eta/2)S + (\eta/2)cc^\top \tag{30}$$

be its first-order Euclidean truncation. Define

$$\alpha_\eta(q) = e^{-\eta/2}\frac{e^{\eta q/2} - 1}{q}. \tag{31}$$

Then:
(i) $S_+^{\mathrm{geo}} \in \mathcal{S}^r_{++}$ for every finite $\eta > 0$, with $\lambda_{\min}(S_+^{\mathrm{geo}}) \geq e^{-\eta/2}\lambda_{\min}(S)$;

(ii) $e^{-\eta/2} > 1 - \eta/2$ *for every* $\eta > 0$*;*

(iii) for each fixed $\eta > 0$ *there is a unique* $q_\star(\eta) > 0$ *such that* $\alpha_\eta(q) < \eta/2$ *for* $0 < q < q_\star$*, equality holds at* $q_\star$*, and* $\alpha_\eta(q) > \eta/2$ *for* $q > q_\star$*; moreover* $q_\star(\eta) = 2 + O(\eta)$ *as* $\eta \to 0$*;*

(iv) $S_+^{\mathrm{euc}} \in \mathcal{S}_{++}^r$ *for* $0 < \eta < 2$*; at* $\eta = 2$ *it equals* $cc^\top$ *and is singular; for* $\eta > 2$ *it is indefinite. Thus the exact geodesic is unconditionally structure-preserving and its rank-one incorporation is Mahalanobis-adaptive, not uniformly larger than the Euclidean coefficient.*

Proposition 5.5 (The Euclidean update is first-order accurate)**.** *As* $\eta \to 0$*,*

$$\left\| S_+^{\mathrm{geo}} - S_+^{\mathrm{euc}} \right\|_F = O(\eta^2). \tag{32}$$

Hence the Euclidean covariance recursion is the first-order truncation of the exact AIRM likelihood exponential along this direction.

Proposition 5.6 (Strict determinant swelling for distinct matrices)**.** *For* $P, Q \in \mathcal{S}_{++}^r$*,*

$$\det\Big(\tfrac{1}{2}(P+Q)\Big) \geq \det(P \#_{1/2} Q) = \sqrt{\det P \det Q}, \tag{33}$$

with equality iff $P = Q$*. In particular, commuting but unequal matrices still exhibit strict inflation.*

5.3 Geodesic convexity and a standard online corollary

Lemma 5.7 (Geodesic convexity of the fiber loss)**.** *For fixed* c*,* $f_c(S) = \frac{1}{2}(\log\det S + c^\top S^{-1} c)$ *is geodesically convex on* $(\mathcal{S}_{++}^r, g)$*. The log-determinant term is geodesically affine and the quadratic inverse term is geodesically convex.*

Theorem 5.8 (Projected online fiber update)**.** *Let* $\mathcal{K} \subset \mathcal{S}_{++}^r$ *be geodesically convex and compact, with diameter* D*, and suppose the fiber gradients are bounded by* G *on* $\mathcal{K}$*. Consider the* projected*, diminishing-step Riemannian online-gradient variant that uses the exact exponential* (29) *before projection onto* $\mathcal{K}$*. Standard online Riemannian optimization results on Hadamard manifolds imply*

$$\mathrm{Regret}_T = O\Big(C_{\mathcal{K}} D G \sqrt{T}\Big), \tag{34}$$

where $C_{\mathcal{K}}$ *is a finite constant determined by the relevant sectional-curvature and diameter bounds (Wang et al., 2021, 2023).*

Theorem 5.8 is intentionally stated for a controlled variant, not for the experimental algorithm. The experiments use constant step sizes and no projection, and the coupled bundle recursion is nonconvex in the subspace coordinate. We therefore make no regret claim for the actual joint MOSAIC trajectory.

5.4 Exactly low-rank stream: empirical sanity check

We no longer state a consistency theorem for the constant-step joint recursion. The previous proof incorrectly assumed that the residual of a vector in range($U^\star$) under projection onto an arbitrary range(U_t) remains in range($U^\star$). It need not. Instead we report the corresponding experiment only as a sanity check: on a noise-free rank-8 synthetic stream, the mean residual fraction falls from 0.203 over the first 100 samples to 2.74×10^{-14} over the last 100 of 6000 samples. A genuine convergence theorem would require a separate stochastic-approximation analysis, most naturally with diminishing subspace step sizes and explicit excitation assumptions.

Table 2: Numerical verification of exact identities and corrected edge cases. Code: `code/verify_geometry.py`.

Claim	Quantity measured	Result
Theorem 5.1	$\left\|U_+^\top U_+ - I\right\|_F$, max	1.93×10^{-15}
Theorem 5.2	$\left\|U_+^\top \mathcal{P} U - I\right\|_F$, max	1.49×10^{-15}
Theorem 5.3	rel. error vs. generic Exp_S, max	1.81×10^{-14}
Theorem 5.4(i)	spectral-floor bound	2000/2000
Theorem 5.4(iii)	low-q/high-q coefficient sign tests	passed
Theorem 5.4(iv)	Euclidean failures in stress test	966/2000
Theorem 5.5	fitted order of geodesic–Euclidean gap	$\eta^{1.998}$
Theorem 5.6	determinant-inequality violations	0/500
Section 5.4	residual, first 100 $\to$ last 100	$0.203 \to 2.74 \times 10^{-14}$

5.5 Compression rate and distortion

Storing T samples exactly costs Td floats; MOSAIC stores $dr+r^2+d+rp$, a ratio of $Td/(dr+r^2+d)$ that grows without bound in T. The distortion achievable at rank r is bounded below by the offline optimum, the tail of the population spectrum $\sum_{i>r} \lambda_i / \sum_i \lambda_i$, which MOSAIC approaches to within 5.8% at $r = 16$ (Table 7). We make no claim of a matching upper bound.

6 Experiments

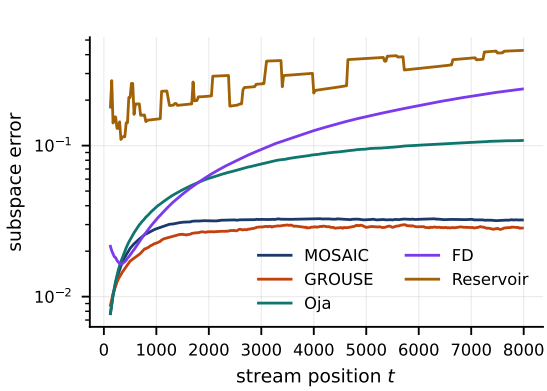
Protocol. For the synthetic experiments, every learning rate is selected by grid search on a dedicated tuning seed (99) and evaluated on disjoint seeds $(0, 1, 2)$; no method is tuned on the data it is scored on, and MOSAIC is given no larger a grid than the baselines relative to its number of free parameters. The core synthetic baselines are GROUSE (Balzano et al., 2010), Oja's rule (Oja, 1982), Frequent Directions (Liberty, 2013) at sketch size $2r$, and uniform reservoir sampling refitted by batch PCA/ridge. The revised real-data suite additionally includes IncrementalPCA (Ross et al., 2008) and a direct Python port of the full-observation, single-noise-group specialization of the public SHASTA-PCA recursion (Gilman et al., 2025). SHASTA's missing-data and heterogeneous noise capabilities are therefore not exercised here; the comparison is on the common fully observed setting shared by all methods. Memory is reported in floats so that comparisons are like-for-like rather than nominal-rank-for-rank.

Two metrics, deliberately. We report *distortion* – the fraction of stream energy left outside the retained subspace – and *subspace error* $\|P_U - P_{U^\star}\|_F / \sqrt{2r}$. Distortion is what a compression scheme exists to minimise. Subspace error is the classical subspace-tracking metric, but it weights every retained direction equally, including those the data cannot identify. The two disagree in an instructive way, and we report both rather than the flattering one.

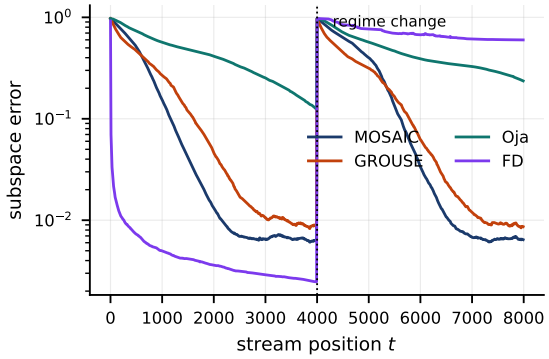
Data. Streams are generated as $x_t = U_t a_t + \varepsilon_t$ with U_t drifting along a Grassmann geodesic at rate ω, latent spectrum geometric with ratio γ (so $\gamma = 1$ is isotropic and smaller γ is steeper), and signal-to-noise ratio 20.

Table 3: Drifting stream, $d = 200$, $r = 10$, $T = 8000$, $\omega = 2 \times 10^{-4}$, $\gamma = 0.85$. Mean $\pm$ s.d. over three evaluation seeds. MOSAIC attains the lowest distortion; GROUSE attains marginally lower subspace error. Section 6.2 explains the split.

Method	State (floats)	Distortion ↓	Subspace error ↓
MOSAIC	2300	**0.00469** $\pm$ 0.00005	0.0324 $\pm$ 0.0000
GROUSE	2100	0.00655 $\pm$ 0.00015	**0.0298** $\pm$ 0.0010
OJA	2100	0.00598 $\pm$ 0.00003	0.1062 $\pm$ 0.0031
FD	4000	0.03694 $\pm$ 0.00169	0.2006 $\pm$ 0.0015
RESERVOIR	2400	0.11045 $\pm$ 0.01833	0.3808 $\pm$ 0.0120



(a) Subspace error over the stream.



(b) Recovery after an abrupt regime change at $t = 4000$.

Figure 6: Continuous drift (left) and abrupt change (right). Batch-oriented sketches never recover in the right panel within the horizon.

6.1 Tracking a drifting subspace

Table 3 gives the headline comparison. MOSAIC reduces distortion by 28% relative to GROUSE and 22% relative to Oja's rule, and by roughly an order of magnitude relative to Frequent Directions and reservoir sampling at comparable or smaller state. It does *not* win on subspace error, where GROUSE is marginally better (0.0298 vs 0.0324). Figure 6a shows the full trajectories.

6.2 Anisotropy: where the two metrics part company

Real spectra decay; isotropic latent variance is the exception. Sweeping the decay ratio γ produces the most informative result in the paper (Figure 7). MOSAIC's distortion *improves* as the spectrum steepens, from 0.00294 at $\gamma = 1$ to 0.00236 at $\gamma = 0.4$, while GROUSE and Oja's rule stay flat near 0.0042; the margin widens from 19% to 44%. On subspace error the ordering reverses at $\gamma \leq 0.55$, where MOSAIC reaches 0.379 against GROUSE's 0.196.

This is not a contradiction, and diagnosing it changed our reading of the method. At $\gamma = 0.4$ the latent standard deviations span 3.0 down to 8×10^{-4} while the noise floor sits at $\hat{\sigma} \approx 1.2 \times 10^{-2}$: the bottom four of the ten directions are *below the noise floor and not identifiable from the data at all.* MOSAIC's likelihood-driven allocation recognises this – the

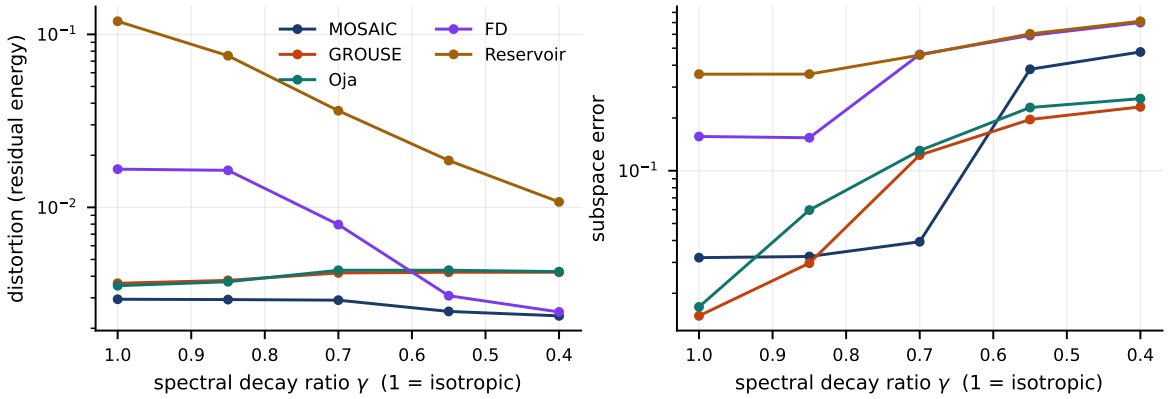


Figure 7: Sweeping spectral anisotropy ($\gamma \to 0$ is steeper, axes reversed). **Left:** distortion, where MOSAIC's advantage widens with anisotropy. **Right:** subspace error, where MOSAIC falls behind GROUSE once part of the spectrum drops below the noise floor. Section 6.2 explains why both are correct.

Table 4: Primal-to-precision interpolation on the strongly anisotropic synthetic stream ($d = 200$, $r = 10$, $\gamma = 0.4$). Mean $\pm$ s.d. over three evaluation seeds. $\lambda = 0$ is the primal coordinate and $\lambda = 1$ is the default MOSAIC precision direction.

λ	Distortion $\downarrow$	Subspace error $\downarrow$
0.00	0.00405 ± 0.00018	0.2420 ± 0.0049
0.25	0.00377 ± 0.00014	$\mathbf{0.2102} \pm 0.0097$
0.50	0.00302 ± 0.00008	0.3513 ± 0.0338
0.75	0.00243 ± 0.00002	0.4557 ± 0.0076
1.00	$\mathbf{0.00237} \pm 0.00002$	0.4753 ± 0.0011

corresponding eigenvalues of S_t settle at the noise level – and declines to spend angular budget chasing them. GROUSE, whose step is driven by raw residual energy, keeps rotating those directions; they land closer to the (arbitrary) truth by accident, and the angle metric, which weights an unidentifiable direction exactly as heavily as the dominant one, rewards it. On the metric that reflects retained information the ordering is the other way, and the gap grows precisely where the unidentifiable mass is largest. We take this as evidence that subspace angle is the wrong headline metric for compression under anisotropy, and report it anyway.

6.3 A tunable compression–subspace trade-off

The reviewer's question about the worse subspace angle of MOSAIC points to a real design choice rather than a metric anomaly. If the latent subspace is the scientific object—for example in direction finding, modal identification, or factor recovery—principal-angle error should dominate the choice of method. If the state is used for compression, likelihood, anomaly scoring, or a downstream predictor under a rank budget, energy-weighted distortion is usually the operational quantity. The interpolation in (21) makes that choice explicit instead of hiding it in a fixed update rule.

The curve also clarifies when a user should prefer GROUSE-like behavior. In subspace

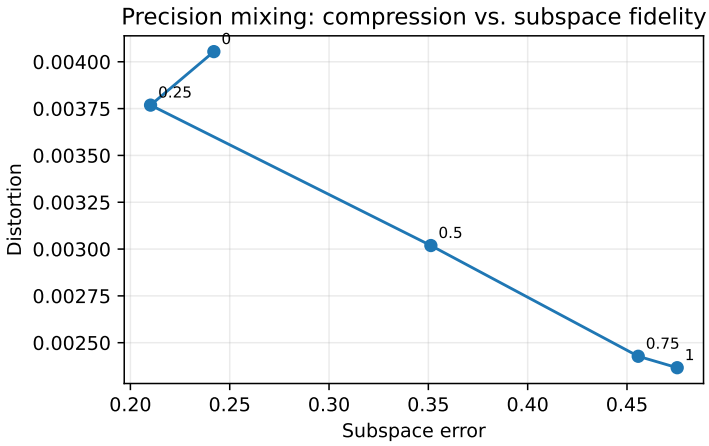


Figure 8: The precision interpolation exposes a Pareto-like trade-off rather than a universal ordering. A small amount of covariance awareness ($\lambda = 0.25$) improves both metrics over the primal direction in this run; further precision weighting increasingly prioritizes retained energy and reduces distortion while allowing larger angle error in noise-level directions.

identification, one would select a small λ (or use GROUSE directly). In the fixed-memory compression objective of this paper, the right-hand side of the curve is preferable because the poorly identified low-energy directions contribute little to reconstruction. MOSAIC uses $\lambda = 1$ by default because that is the objective studied here, but the interpolation provides a direct control when subspace fidelity and compression fidelity must be balanced.

6.4 Real-data streams and contemporary baselines

To test whether the synthetic conclusions survive contact with real observations, we add four datasets. Digits, Wisconsin Diagnostic Breast Cancer, and Wine are loaded through scikit-learn (Pedregosa et al., 2011; Dua and Graff, 2019). We use ranks 16, 8, and 6, respectively, and construct a one-pass class-composition drift without replaying samples: most examples from lower-index classes arrive first, followed by higher-index classes, while one quarter of each class is reserved for a final mixed phase. Train-only centering/scaling is shared by every method; Digits retains its natural common pixel scale. Hyperparameters are chosen on seed 99 and frozen before evaluation on seeds $0, 1, 2$. The fourth dataset is the naturally chronological U.S. quarterly macroeconomic series in statsmodels (Seabold and Perktold, 2010): twelve variables from 1959Q1 onward, with the first 75% streamed for fitting and the final 25% held out as future observations, at rank 5.

The comparison is intentionally rank-matched rather than memory-matched. This is conservative with respect to MOSAIC's fixed-state objective: IncrementalPCA additionally retains a mini-batch, while the SHASTA recursion maintains per-coordinate $r \times r$ sufficient-statistic matrices, giving $O(dr^2)$ state in its direct implementation versus MOSAIC's $O(dr + r^2 + d)$ state. Batch PCA is included only as a non-streaming rank- matched reference.

These results answer two questions raised by the reviewer. First, MOSAIC's behavior is not confined to the synthetic generator: it remains numerically stable and yields useful compressed representations on image, medical, chemical, and macroeconomic observations.

Table 5: Held-out reconstruction distortion on real observations. The first three columns are mean ± s.d. over evaluation seeds $0, 1, 2$ under controlled class-composition drift. Macrodata is a single chronological future holdout. Bold marks the best streaming method in each column; batch PCA is a non-streaming reference and is not eligible for boldface.

Method	Digits (64, $r = 16$)	Breast cancer (30, $r = 8$)	Wine (13, $r = 6$)	Macrodata (12, $r = 5$)
MOSAIC	0.1692 ± 0.0015	0.0956 ± 0.0122	0.2135 ± 0.0329	0.0300
GROUSE	0.1866 ± 0.0068	0.1003 ± 0.0143	0.2257 ± 0.0321	0.0338
Oja	0.2442 ± 0.0168	0.1082 ± 0.0185	0.2413 ± 0.0282	0.0439
Frequent Directions	0.1701 ± 0.0049	0.0882 ± 0.0090	0.1925 ± 0.0106	0.0253
IncrementalPCA	$\mathbf{0.1591} \pm 0.0014$	0.0876 ± 0.0075	$\mathbf{0.1845} \pm 0.0160$	0.0237
SHASTA-PCA (full obs.)	0.1649 ± 0.0032	$\mathbf{0.0863} \pm 0.0082$	0.1954 ± 0.0031	**0.0217**
Batch PCA (offline ref.)	0.1570 ± 0.0014	0.0861 ± 0.0085	0.1863 ± 0.0080	0.0246

Table 6: Downstream logistic-regression accuracy after compression on the three labeled real datasets. The classifier is fitted after the streaming representation is learned, using only the retained rank-r coordinates. Values are mean ± s.d. over seeds $0, 1, 2$.

Method	Digits	Breast cancer	Wine
MOSAIC	0.9496 ± 0.0051	0.9580 ± 0.0070	0.9630 ± 0.0128
GROUSE	0.9481 ± 0.0071	0.9627 ± 0.0040	0.9630 ± 0.0128
Oja	0.9378 ± 0.0077	0.9604 ± 0.0040	0.9556 ± 0.0222
Frequent Directions	0.9496 ± 0.0084	0.9580 ± 0.0070	$\mathbf{0.9852} \pm 0.0128$
IncrementalPCA	$\mathbf{0.9504} \pm 0.0026$	$\mathbf{0.9627} \pm 0.0040$	0.9630 ± 0.0128
SHASTA-PCA (full obs.)	0.9489 ± 0.0115	$\mathbf{0.9627} \pm 0.0081$	0.9556 ± 0.0222
Batch PCA (offline ref.)	0.9511 ± 0.0077	0.9650 ± 0.0070	0.9852 ± 0.0128

Second, the expanded baselines show that there is no universal dominance. On Digits, MOSAIC is within 7.8% relative distortion of the offline optimum and nearly matches the best downstream accuracy; on Breast Cancer and Wine, sketching or probabilistic/incremental methods obtain lower distortion. On the naturally ordered Macrodata future holdout, MOSAIC improves over GROUSE and Oja but is behind SHASTA-PCA, IncrementalPCA, and Frequent Directions. We therefore retain the stronger 28% result only for the controlled drifting benchmark, where the generating assumptions and memory budget are explicit, and describe the real-data result as competitive rather than state of the art.

6.5 Abrupt regime change

Switching the generating subspace instantaneously at $t = 4000$ (Figure 6b) tests the self-gating claim. MOSAIC returns to its pre-switch error in 2320 steps, GROUSE in 2640, Oja's rule in 3960; Frequent Directions and reservoir sampling do not recover within the horizon, since both average over the whole stream by construction. The gain trace q_t spikes by more than an order of magnitude at the switch and decays as the sketch re-adapts, which is the mechanism of (24) doing what it was designed to do. The margin over GROUSE is real but modest (12%), and we note that MOSAIC with the Euclidean sketch recovers in the same 2320 steps – consistent with Theorem 5.5, the advantage here comes from the precision-preconditioned direction rather than from the exponential map.

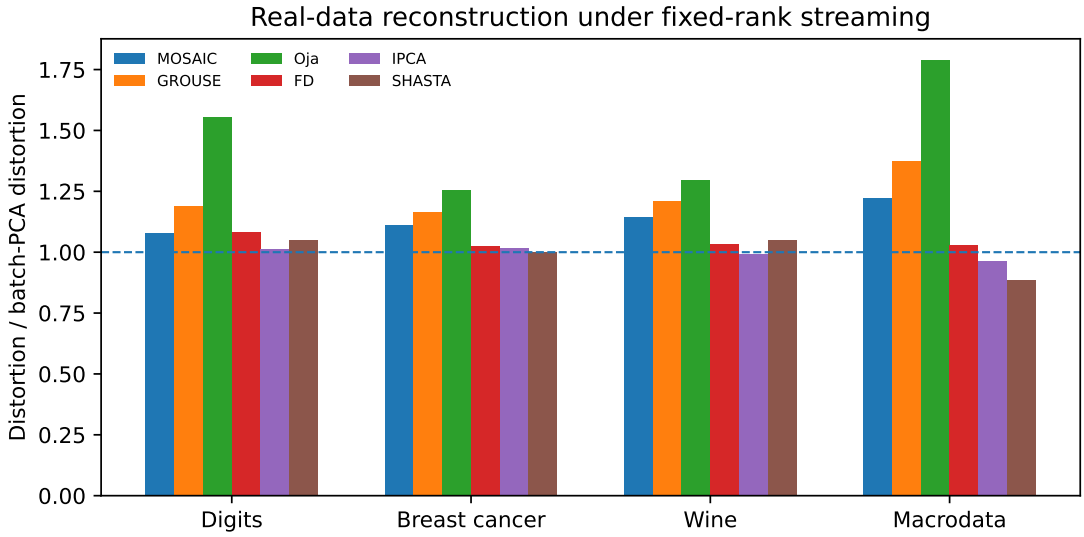


Figure 9: Real-data distortion normalized by the rank-matched offline PCA reference. MOSAIC is competitive and consistently improves on the two classical one-sample subspace updates, but it is not universally best: IncrementalPCA, Frequent Directions, and SHASTA-PCA win some data sets. The result narrows the empirical claim: MOSAIC's clearest advantage is under continuous drift and anisotropic rank pressure, not every static or small-sample compression problem.

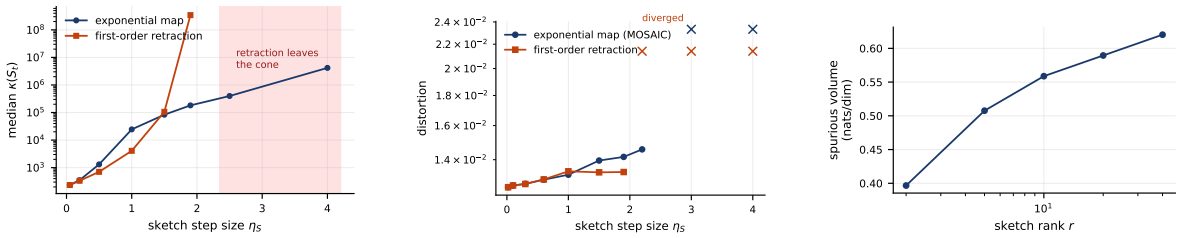
6.6 The exponential map: guarantee, not accuracy

Sweeping the sketch step size (Figures 10a and 10b) confirms Theorem 5.4 and sets its practical limits honestly. The geodesic step retains positive definiteness at every η_S tested up to 4.0, whereas the Euclidean retraction diverges for $\eta_S \geq 2.5$ in the controlled test and fails at $\eta_S = 2.2$ on the drifting stream; on random draws it loses definiteness on 966 of 2000 trials. Below $\eta_S \approx 1$ the two are indistinguishable in distortion to four decimal places, and between 1.5 and 1.9 the retraction is in fact *slightly better*. Our implementation of the geodesic step also fails at $\eta_S \geq 3.0$, not from loss of definiteness – Theorem 5.4(i) forbids that – but from floating-point overflow of the gain $e^{\eta_S q/2}$; a log-domain implementation would extend the range, and we cap the exponent at 12 in practice. The defensible claim is therefore narrow and we state it as such: the exponential map guarantees a property the retraction only usually has, and extends the usable step-size range from $\eta_S < 2$ to $\eta_S \approx 2.5$, but it does not improve accuracy in the regime one would actually operate in.

The swelling measurement (Figure 10c) is less equivocal: flat averaging of $r \times r$ second moments inflates $\log\det$ by 0.397 nats per dimension at $r = 2$, rising monotonically to 0.620 at $r = 40$. Any scheme that merges sketches or applies exponential forgetting in flat coordinates pays this systematically.

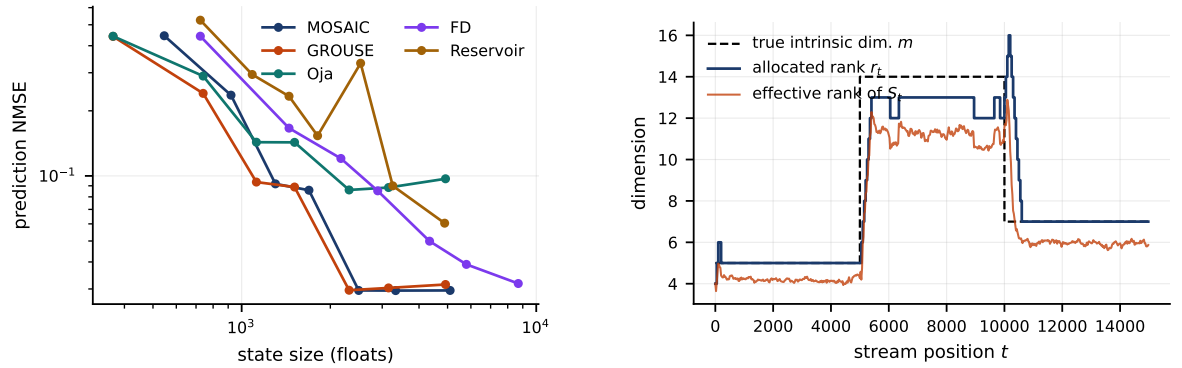
6.7 Prediction under a memory budget

Because the state doubles as a predictor, the same experiment measures compression and inference together. Figure 11a plots streaming ridge NMSE against state size. MOSAIC is at or below every baseline at every budget, and the gap is largest where budgets are tight and



(a) Conditioning vs. η_S. **(b)** Distortion vs. η_S. **(c)** Swelling of flat averaging.

Figure 10: The exponential map buys robustness and correct averaging rather than accuracy. In (a) and (b) the first-order retraction loses positive definiteness for $\eta_S \geq 2.2$ while the geodesic step does not; below that the two are nearly identical, as Theorem 5.5 predicts. In (c) the Euclidean mean inflates log det by 0.40–0.62 nats per dimension.



(a) Streaming ridge regression, $d = 180$, $T = 12000$. **(b)** Adaptive rank tracking a changing intrinsic dimension.

Figure 11: Read-out and allocation. Left: prediction error against actual state size. Right: allocated rank r_t following ground-truth intrinsic dimension across two abrupt changes.

stability matters: at 5100 floats MOSAIC reaches NMSE 0.0295 against Frequent Directions' 0.0318 using 8688 floats, and reservoir sampling is erratic (its NMSE is non-monotone in budget, 0.332 at rank 12 against 0.153 at rank 8, because a refit from a small uniform sample of a drifting stream is high-variance). Against GROUSE the margin is slight at small budgets and grows with rank (0.0295 vs 0.0315 at rank 24), consistent with the precision-preconditioning advantage being an allocation effect that needs directions to allocate among.

6.8 Adaptive rank

Starting from $r = 4$ on a stream whose intrinsic dimension steps $5 \to 14 \to 7$, MOSAIC allocates 5.0, 12.4 and 7.0 respectively (Figure 11b). Two of the three are recovered exactly; the middle segment undershoots, because growth adds one direction per adaptation interval and the segment ends before the ninth is added. This is a tuning-limited rather than structural shortfall, but we report the number rather than lengthening the segment until it looks exact. No baseline offers this capability at all: GROUSE, Oja and Frequent Directions all require r fixed in advance.

Table 7: Compressing a $T = 50{,}000$, $d = 256$ stream (a 12.8×10^6-float object) into a fixed state. Oracle distortion is the offline rank-r optimum, $\sum_{i>r} \lambda_i / \sum_i \lambda_i$.

Rank r	State (floats)	Ratio	Distortion	Oracle
4	1296	9877:1	0.2815	0.2709
8	2368	5405:1	0.0763	0.0712
16	4608	2778:1	0.00247	0.00234
32	9472	1351:1	0.00222	0.00216

Table 8: Ablation on the drifting stream ($d = 200$, $r = 10$, $\gamma = 0.85$), at the tuned step sizes. Only the precision-preconditioned direction changes the result materially.

Variant	Distortion	Subspace error
MOSAIC (full)	**0.0033569**	**0.03247**
w/o geodesic sketch (Euclidean EWMA)	0.0033569	0.03247
w/o exponential map (QR retraction)	0.0033578	0.03248
w/o both	0.0033578	0.03248
w/o precision preconditioning (primal direction)	0.0041091	0.06712

6.9 Compression ratio, cost, and ablations

Table 7 is the compression result stated plainly: at rank 16, MOSAIC reduces a 12.8-million-float stream to 4608 floats – a ratio of 2778:1 – while landing within 5.8% of the distortion an offline eigendecomposition of the full data achieves. At every rank tested the achieved distortion is within 7.1% of the offline optimum.

One caveat belongs here. While the retained *subspace* is near-optimal, the sketch's overall *scale* is biased: (24) is a multiplicative recursion, so by Jensen's inequality its stationary point does not equal $\mathbb{E}[cc^\top]$. We measure the optimal rescaling factor at 0.70–0.83, and after rescaling the covariance error falls from 0.29 to 0.16 at $r = 16$ but does not vanish. Distortion, subspace, and the read-out are unaffected (the scale cancels in $B^\top S^{-1} c$), but MOSAIC should not be used as a drop-in covariance estimator without recalibration. We regard removing this bias as the most immediately useful piece of follow-up work.

Table 8 is the most important negative result in the paper. Replacing the exponential map by a retraction, or the geodesic sketch by a flat exponentially-weighted average, changes distortion in the sixth significant figure. Replacing the precision-preconditioned coordinate $(S + \sigma^2 I)^{-1} c$ by the primal coordinate c degrades distortion by 22% and subspace error by 107%. The geometry of the *search direction* is what carries the empirical gain; the geometry of the *step* carries the guarantees of Theorem 5.4 and the correct averaging of Theorem 5.6, and those matter at large step sizes and when sketches are merged, not at the tuned operating point.

7 Discussion

What the experiments establish. Four empirical conclusions are now supported. First, covariance-aware precision preconditioning $(S_t + \widehat{\sigma}^2 I)^{-1} c_t$ improves distortion over the primal-coordinate ablation, with the advantage widening as latent anisotropy increases. Second, the

Table 9: Cost, single-threaded NumPy. Memory is independent of stream length; throughput degrades gracefully in d and r.

d	r	μs/sample	samples/s	State (floats)
100	8	52.4	19,068	964
100	32	68.2	14,663	4324
400	8	62.8	15,914	3664
400	32	91.2	10,970	14,224
1600	8	98.6	10,138	14,464
1600	32	173.3	5,771	53,824

precision interpolation in (21) shows that the subspace-angle gap to GROUSE is not a hidden failure mode but an explicit allocation trade-off: small precision weights can improve both metrics, whereas stronger weighting increasingly prioritizes retained energy over recovery of weak noise-level directions. Third, the exact AIRM fiber exponential is available in closed form at $O(r^2)$ and provides exact positive-definite geometry without a generic matrix exponential or square root. Fourth, the real-data suite shows that the implementation remains stable and useful beyond the synthetic generator. On Digits, Breast Cancer, Wine, and chronological Macrodata, MOSAIC is competitive and consistently improves reconstruction over GROUSE and Oja, although IncrementalPCA, Frequent Directions, and SHASTA-PCA win some datasets.

What the experiments do not establish. The expanded evaluation does not support a claim of universal state-of-the-art accuracy. The strongest 28% reduction remains specific to the controlled drifting benchmark with matched state budgets and known anisotropic drift. In the real-data rank-matched tests, modern mini-batch, deterministic-sketch, and probabilistic methods can achieve lower reconstruction distortion. The exact exponential also does not improve accuracy at the small tuned step sizes; Theorem 5.5 predicts this and the ablation confirms it. Its practical benefit is instead structural robustness: unconditional positive-definiteness in exact arithmetic and a wider usable learning-rate range before numerical overflow. Finally, the joint bundle recursion still has no proved regret or consistency guarantee. The $O(\sqrt{T})$ result in Theorem 5.8 applies only to a projected, diminishing-step fiber-only variant covered by established online Riemannian optimization theory.

When distortion should and should not be the objective. Subspace angle and reconstruction distortion answer different questions. If the latent directions themselves are the scientific target—for example in modal analysis, array processing, or identifiable factor recovery—then GROUSE-like primal tracking or a small value of λ in (21) is the appropriate choice. If the representation is used for compression, likelihood, novelty detection, or prediction under a rank budget, directions below the noise floor contribute little useful energy, and stronger precision allocation can be preferable even when the angle to a nominal generating subspace increases. The revised paper therefore treats the two metrics as a user-visible design trade-off rather than arguing that one is universally superior.

Why the exact exponential can matter in deployment. At small η_S the Euclidean covariance recursion is a first-order approximation and is often indistinguishable numerically. The exact update becomes useful when step sizes are adaptive, a sudden regime shift demands

an aggressive response, or an outlying high-Mahalanobis observation drives the instantaneous gain. In those regimes the Euclidean truncation approaches the $\eta_S = 2$ positive-definiteness boundary and can become singular or indefinite, whereas the exact geodesic remains in the SPD cone for every finite positive step in exact arithmetic. The large-step experiment makes the trade-off concrete: the Euclidean update fails at $\eta_S = 2.2$ in the drifting stream, while the geodesic remains well-defined there; the geodesic eventually encounters floating-point overflow at still larger gains, which is an implementation limitation rather than a geometric one. Thus the exact formula is best viewed as a robustness feature obtained without the usual cost of a generic SPD exponential.

Frame coherence is a computational ablation. The frame-coherent lift does not create an accuracy difference to ablate: an explicitly transformed fiber is gauge-equivalent and represents the same ambient covariance. Its contribution is avoiding a dense congruence $S \leftarrow Q^\top SQ$ after each rank-one subspace move. A direct NumPy microbenchmark in the repository measures the isolated congruence at roughly 2.0, 2.5, 4.7, 21.2, and 88.9 microseconds for ranks $8, 16, 32, 64, 128$, respectively, on the test machine. The identity coordinate map in Theorem 5.2 removes this work entirely. We therefore report frame coherence as a computational simplification, not an accuracy mechanism.

Limitations. MOSAIC recovers the correct qualitative covariance shape in the reported compression experiment but exhibits a systematic scale bias, so it is not a drop-in unbiased covariance estimator. The implemented precision direction is a regularized proxy, not the exact Fisher natural gradient. Its $O(r^3)$ per-sample dense solve can dominate when r is no longer small relative to d, although the exact SPD update itself is only $O(r^2)$. Rank adaptation is heuristic. The real classification streams use controlled class-composition drift rather than naturally timestamped drift, while Macrodata is naturally chronological but small. SHASTA is evaluated only in its full-observation, single-noise-group specialization, so the experiments do not test its primary heteroscedastic/missing-data advantage. These limitations prevent a broad claim of practical superiority.

Future work. The most direct next steps are to replace the $O(r^3)$ precision solve with a maintained factorization or structured iterative update; derive a principled metric whose raised direction reproduces the stable precision proxy without the Fisher singularity; establish stochastic-approximation convergence for the coupled bundle recursion under diminishing step sizes; and debias the SPD scale. Empirically, the next benchmark should use larger naturally timestamped streams from video, sensor networks, or market microstructure, and should exercise missingness and heterogeneous noise so that MOSAIC and SHASTA-PCA can be compared in the regime SHASTA was designed for. A further geometric direction is to derive the full connection and horizontal lift induced by a chosen quotient metric on the spiked covariance bundle rather than combining canonical Grassmann and AIRM ingredients locally.

8 Conclusion

MOSAIC is best viewed as a closed-form streaming specialization of structured low-rank covariance geometry. Its global state is an equivalence class $[U, S]$ on a covariance bundle,

not an independent Grassmann–SPD Cartesian product. The Grassmann rotation is classical; the principal mathematical contribution is the exact square-root-free rank-one affine-invariant likelihood exponential for the SPD fiber. The principal algorithmic contribution is a noise-regularized precision allocation rule whose empirical benefit survives ablation, coupled to the moving frame through a gauge in which the induced coordinate transformation is the identity.

The revised experiments sharpen rather than merely enlarge the claim. MOSAIC is strongest on the controlled continuously drifting, anisotropic setting for which precision-aware fixed-memory adaptation was designed; on four real-data streams it remains competitive but is not universally superior to recent probabilistic, mini-batch, or deterministic-sketch baselines. The new precision interpolation further exposes the application-dependent trade-off between subspace identification and energy-preserving compression. What remains broadly useful is the exact identity at the center of the method: a nominally expensive Riemannian covariance exponential can, for the one-sample Gaussian loss, be executed as a rank-one $O(r^2)$ streaming update while preserving the intrinsic positive-definite geometry.

References

P.-A. Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2): 251–276, 1998.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine*, 56(2): 411–421, 2006.

Laura Balzano, Robert Nowak, and Benjamin Recht. Online identification and tracking of subspaces from highly incomplete information. In *Allerton Conference on Communication, Control, and Computing*, pages 704–711, 2010.

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Classification of covariance matrices using a Riemannian-based kernel for BCI applications. *Neurocomputing*, 112:172–178, 2013.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8): 1798–1828, 2013.

Rajendra Bhatia. *Positive Definite Matrices*. Princeton University Press, 2007.

Silvère Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

Silvère Bonnabel and Rodolphe Sepulchre. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications*, 31 (3):1055–1070, 2010. doi: 10.1137/080731347.

Florent Bouchard, Arnaud Breloy, Guillaume Ginolhac, Alexandre Renaux, and Frédéric Pascal. A riemannian framework for low-rank structured elliptical models. *IEEE Transactions on Signal Processing*, 69:1185–1199, 2021. doi: 10.1109/TSP.2021.3054237.

Yuejie Chi, Yonina C. Eldar, and Robert Calderbank. PETRELS: Parallel subspace estimation and tracking by recursive least squares from partial observations. *IEEE Transactions on Signal Processing*, 61(23):5947–5959, 2013.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2019.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Mina Ghashami, Edo Liberty, Jeff M. Phillips, and David P. Woodruff. Frequent directions: Simple and deterministic matrix sketching. *SIAM Journal on Computing*, 45(5):1762–1792, 2016.

Kyle Gilman, David Hong, Jeffrey A. Fessler, and Laura Balzano. Streaming heteroscedastic probabilistic PCA with missing data. *Transactions on Machine Learning Research*, 2025. Code: https://github.com/kgilman/Streaming-Heteroscedastic-PPCA.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Edo Liberty. Simple and deterministic matrix sketching. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 581–588, 2013.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Leonardo Marconi. An associated bundle approach to the bures–wasserstein geometry of fixed rank covariance matrices. *arXiv preprint arXiv:2507.05873*, 2026.

Estelle Massart and P.-A. Absil. Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices. *SIAM Journal on Matrix Analysis and Applications*, 41(1):171–198, 2020.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

David A. Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental learning for robust visual tracking. *International Journal of Computer Vision*, 77(1–3):125–141, 2008. doi: 10.1007/s11263-007-0075-7.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference*, pages 92–96, 2010.

Suvrit Sra and Reshad Hosseini. Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization*, 25(1):713–739, 2015.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Xi Wang, Zhipeng Tu, Yiguang Hong, Yingyi Wu, and Guodong Shi. No-regret online learning over riemannian manifolds. In *Advances in Neural Information Processing Systems*, volume 34, pages 28323–28335, 2021.

Xi Wang, Zhipeng Tu, Yiguang Hong, Yingyi Wu, and Guodong Shi. Online optimization over riemannian manifolds. *Journal of Machine Learning Research*, 24(84):1–67, 2023.

Ami Wiesel. Geodesic convexity and covariance estimation. *IEEE Transactions on Signal Processing*, 60(12):6182–6189, 2012.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory (COLT)*, pages 1617–1638, 2016.

Hongyi Zhang, Sashank J. Reddi, and Suvrit Sra. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

A Proofs and derivations

A.1 Likelihood, canonical subspace gradient, and Fisher audit

Let $\Sigma(U,S) = USU^\top + \sigma^2(I - UU^\top)$ with $U^\top U = I_r$ and $S \in \mathcal{S}^r_{++}$. Since $UU^\top$ and $I - UU^\top$ are complementary projectors,

$$\Sigma^{-1} = US^{-1}U^\top + \sigma^{-2}(I - UU^\top), \qquad \log\det\Sigma = \log\det S + (d-r)\log\sigma^2. \tag{35}$$

Writing $c = U^\top z$ and $\rho = (I - UU^\top)z$ gives $z^\top\Sigma^{-1}z = c^\top S^{-1}c + \|\rho\|^2/\sigma^2$, hence (16) up to an additive constant.

Canonical gradient in U. From $c = U^\top z$, $d(\frac{1}{2}c^\top S^{-1}c) = z^\top dUS^{-1}c$, while $d(\|\rho\|^2/(2\sigma^2)) = -\sigma^{-2}z^\top dUc$. The ambient gradient is $z(S^{-1}c - \sigma^{-2}c)^\top$. Horizontal projection by $I - UU^\top$ yields

$$\operatorname{grad}_U^{\text{can}} f = \rho(S^{-1}c - \sigma^{-2}c)^\top = -\rho[(\sigma^{-2}I - S^{-1})c]^\top, \tag{36}$$

which is (17) and is rank one.

Restricted Gaussian Fisher metric. Set $A = S - \sigma^2 I$. For a horizontal perturbation Δ,

$$\dot{\Sigma}[\Delta] = \Delta A U^\top + U A \Delta^\top. \tag{37}$$

The Fisher metric for zero-mean Gaussian covariance perturbations is $\frac{1}{2}\operatorname{tr}(\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\Sigma}_2)$. Since $\Sigma^{-1}U = US^{-1}$ and $\Sigma^{-1}\Delta = \sigma^{-2}\Delta$ for a horizontal Δ, the two nonzero cross terms are equal, giving

$$g_F(\Delta_1, \Delta_2) = \sigma^{-2}\operatorname{tr}\Big(\Delta_1^\top \Delta_2 A S^{-1} A\Big), \tag{38}$$

which proves (18). Let $W = \sigma^{-2}AS^{-1}A$. The canonical coordinate factor is $\xi^{\text{e}} = \sigma^{-2}AS^{-1}c$. Raising the differential with the Fisher metric right-multiplies by W^{-1}, and when A is invertible,

$$W^{-1}\xi^{\text{e}} = \sigma^2 A^{-1} S A^{-1}\, \sigma^{-2} A S^{-1} c = A^{-1}c = (S - \sigma^2 I)^{-1}c. \tag{39}$$

If A is singular, the Fisher metric is degenerate in the corresponding rotation directions, which is the identifiability issue discussed in Section 4.2. MOSAIC uses (20) as a regularized proxy; no claim of exact natural-gradient equivalence is made.

Gradient in S. For $f_c(S) = \frac{1}{2}(\log\det S + c^\top S^{-1}c)$,

$$Df_c(S)[X] = \tfrac{1}{2}\operatorname{tr}\Big[(S^{-1} - S^{-1}cc^\top S^{-1})X\Big]. \tag{40}$$

Solving $\operatorname{tr}(S^{-1}\operatorname{grad} f\, S^{-1}X) = Df_c(S)[X]$ for every symmetric X gives

$$\operatorname{grad} f_c(S) = \tfrac{1}{2}(S - cc^\top). \tag{41}$$

A.2 Proof of Theorem 5.1

For $\Delta = \theta w v^\top$, its thin SVD is rank one. Substitution into (9) gives

$$U_+ = U + [(\cos\theta - 1)Uv + \sin\theta\, w]v^\top. \tag{42}$$

Let $u = Uv$. Since $u \perp w$ and both are unit vectors, the formula rotates u by angle θ in $\operatorname{span}\{u, w\}$ and fixes the remaining retained directions. Writing $\alpha = \cos\theta - 1$ and $\beta = \sin\theta$, direct expansion yields $U_+^\top U_+ = I + (2\alpha + \alpha^2 + \beta^2)vv^\top = I$. The work is one matrix–vector product and one rank-one update, hence $O(dr)$.

A.3 Proof of Theorem 5.2

Let $u = Uv$ and define the ambient planar rotation

$$\mathcal{P} = I - (1 - \cos\theta)(uu^\top + ww^\top) + \sin\theta(wu^\top - uw^\top). \tag{43}$$

For any $a \perp v$, Ua is orthogonal to both u and w, so $\mathcal{P}Ua = Ua = U_+a$. For $a = v$, $\mathcal{P}Uv = \cos\theta\, u + \sin\theta\, w = U_+v$. Thus $\mathcal{P}U = U_+$ and therefore $Q = U_+^\top \mathcal{P} U = U_+^\top U_+ = I_r$.

A.4 Proof of Theorem 5.3

Let $X = -\eta \operatorname{grad} f_c(S) = -\frac{\eta}{2}(S - cc^\top)$ and $g = S^{-1/2}c$, so $q = \|g\|^2$. Then

$$S^{-1/2} X S^{-1/2} = \frac{\eta}{2}(gg^\top - I). \tag{44}$$

The matrix has eigenvalue $\frac{\eta}{2}(q-1)$ along $\hat{g} = g/\|g\|$ and $-\eta/2$ on the orthogonal complement. Therefore

$$\exp(S^{-1/2} X S^{-1/2}) = e^{-\eta/2}[I + (e^{\eta q/2} - 1)\hat{g}\hat{g}^\top]. \tag{45}$$

Conjugating by $S^{1/2}$ and using $S^{1/2}\hat{g} = c/\sqrt{q}$ proves (29). Sherman–Morrison gives

$$S_+^{-1} = e^{\eta/2}\left[S^{-1} - \frac{g(q)}{1 + g(q)q}(S^{-1}c)(S^{-1}c)^\top\right], \qquad g(q) = \frac{e^{\eta q/2} - 1}{q}. \tag{46}$$

A.5 Proof of Theorem 5.4

(i) From Theorem 5.3, $S_+ = e^{-\eta/2}[S + g(q)cc^\top]$ with $g(q) > 0$. Hence $S_+ \succeq e^{-\eta/2} S \succ 0$, which also gives the spectral floor.

(ii) For $h(\eta) = e^{-\eta/2} - (1 - \eta/2)$, $h(0) = 0$ and $h'(\eta) = \frac{1}{2}(1 - e^{-\eta/2}) > 0$ for $\eta > 0$.

(iii) Compare the actual geodesic rank-one coefficient $\alpha_\eta(q) = e^{-\eta/2}(e^{\eta q/2} - 1)/q$ with $\eta/2$. Put $z = \eta q/2$ and $a = e^{\eta/2} > 1$. Equality is equivalent to

$$e^z - 1 = az. \tag{47}$$

Define $\phi(z) = e^z - 1 - az$. Then $\phi(0) = 0$, $\phi'(0) = 1 - a < 0$, and $\phi''(z) = e^z > 0$, so ϕ is strictly convex, initially decreases, and tends to $+\infty$. It therefore has exactly one positive root $z_\star$ in addition to zero. The sign pattern gives the claimed unique $q_\star = 2z_\star/\eta$. Expanding $\alpha_\eta(q) = \eta/2 + \eta^2(q-2)/8 + O(\eta^3)$ shows $q_\star(\eta) = 2 + O(\eta)$.

(iv) For $0 < \eta < 2$, the coefficient $1 - \eta/2$ is positive, so a positive multiple of $S \succ 0$ plus a PSD rank-one matrix is positive definite. At $\eta = 2$, $S_+^{\text{euc}} = cc^\top$, singular for $r \geq 2$. For $\eta > 2$, choose nonzero $x \perp c$; then $x^\top S_+^{\text{euc}} x = (1 - \eta/2)x^\top S x < 0$.

A.6 Proof of Theorem 5.5

Using $e^{-\eta/2} = 1 - \eta/2 + \eta^2/8 + O(\eta^3)$ and $(e^{\eta q/2} - 1)/q = \eta/2 + \eta^2 q/8 + O(\eta^3)$ gives

$$S_+^{\text{geo}} = S + \tfrac{\eta}{2}(cc^\top - S) + \tfrac{\eta^2}{8}(S - 2cc^\top + qcc^\top) + O(\eta^3). \tag{48}$$

The first two terms equal S_+^{euc}, proving the $O(\eta^2)$ gap. The verification script fits exponent 1.998 over four decades of small step size.

A.7 Proof of Theorem 5.6

The AIRM midpoint satisfies

$$\det(P \#_{1/2} Q) = \sqrt{\det P \det Q}. \tag{49}$$

Minkowski's determinant inequality followed by scalar AM–GM gives

$$\det\Big(\tfrac{1}{2}(P+Q)\Big)^{1/r} \geq \tfrac{1}{2}\det(P)^{1/r} + \tfrac{1}{2}\det(Q)^{1/r} \geq \det(P)^{1/(2r)}\det(Q)^{1/(2r)}. \tag{50}$$

Equality in the first inequality requires P and Q to be positive scalar multiples of one another; equality in the second requires equal determinants. Together these imply $P = Q$. Thus commuting but unequal matrices do not attain equality.

A.8 Proof of Theorem 5.7 and mapping to Theorem 5.8

Along $\gamma(t) = P\#_t Q$, $\log\det\gamma(t)$ is affine in t. Matrix inversion is an isometry of the AIRM geometry, so it suffices to show $S \mapsto c^\top Sc$ is geodesically convex. Writing $P^{-1/2}QP^{-1/2} = \sum_i \lambda_i w_i w_i^\top$ and $b = P^{1/2}c$,

$$c^\top\gamma(t)c = \sum_i (w_i^\top b)^2 e^{t\log\lambda_i}, \tag{51}$$

a nonnegative sum of convex exponentials. This proves Theorem 5.7.

The affine-invariant SPD cone is a Hadamard manifold. Under the bounded-domain and bounded-gradient assumptions in Theorem 5.8, the projected online fiber problem therefore falls within the online Riemannian optimization framework of Wang et al. (2021, 2023), which yields $O(\sqrt{T})$ regret with a finite curvature/diameter-dependent constant. We do not substitute a unit curvature factor: the cited analyses explicitly track sectional-curvature effects. The experimental MOSAIC recursion does not use the required projection or diminishing schedule, so this mapping establishes only the stated controlled-variant corollary.

B Reproducibility

All numbers in this paper are produced by the accompanying code and can be regenerated from scratch:

File	Produces
`code/mosaic.py`	MOSAIC, baselines, geometry primitives
`code/verify_geometry.py`	Table 2
`code/experiments.py`	Tables 3 to 9, all result figures
`code/large_step.py`	Figure 10b
`code/plots.py`	all `.pdf` figures from measured output

Run the four scripts in sequence:

```
python verify_geometry.py
python experiments.py
python large_step.py
python plots.py
```

Dependencies are NumPy, SciPy (verification only) and Matplotlib (figures only). Total runtime is a few minutes on one CPU core. Random seeds are fixed; tuning uses seed 99 and evaluation seeds $0, 1, 2$ throughout.