HIERARCHICAL SLICED-WASSERSTEIN ON SYMMETRIC POSITIVE DEFINITE MATRICES

Anonymous authors
Paper under double-blind review

ABSTRACT

Comparing distributions of symmetric positive definite (SPD) matrices is a recurring operation in covariance-based learning, but high-dimensional projections can make repeated sliced-Wasserstein evaluations costly. We present SPD hierarchical sliced-Wasserstein (SPDHSW), which computes a small bank of projections in symmetric log-space and mixes their coordinates into many one-dimensional comparisons. Resampling the projection family at each update avoids confining optimization to a fixed subspace. Under Frobenius-uniform sampling, SPDHSW preserves the squared SPDSW population objective. Because the mixed projections share one bank, they are not independent, which prior analysis of hierarchical slicing overlooks; we characterize this dependence through an exact variance decomposition and harmonic analysis, and extend the results to the gradients used in optimization. This establishes separate precision criteria for discrepancy values and training gradients. Experiments on synthetic distributions, hand-action covariances, and EEG adaptation demonstrate improved alignment at a fixed matrix-projection count per update and reduced projection-dominated loop cost relative to large-budget flat slicing under shared optimization settings. Together, the construction and analysis provide a principled approach to allocating projection computation in learning systems that repeatedly compare SPD-valued distributions.

1 INTRODUCTION

Covariance matrices represent multichannel observations as symmetric positive definite (SPD) descriptors. Their representation dimension grows quadratically with the number of channels: a $d \times d$ covariance has $m = d(d+1)/2$ independent entries, so 128-channel EEG yields an 8256-dimensional symmetric matrix space. Comparing distributions in this space matters for EEG adaptation (Barachant et al., 2013; Yger et al., 2017; Lotte et al., 2018; Zanini et al., 2018), visual recognition (Tuzel et al., 2006; Pennec et al., 2006), and SPD representation learning (Huang & Van Gool, 2017; Brooks et al., 2019; Wang et al., 2022), where a learning procedure may evaluate and differentiate a distribution discrepancy at every update.

Sliced-Wasserstein (SW) distances replace a multidimensional optimal-transport problem with one-dimensional comparisons (Rabin et al., 2012; Bonneel et al., 2015; Peyré & Cuturi, 2019). SPDSW performs this comparison along Log-Euclidean geodesic coordinates $t^A(X) = \mathrm{Tr}(A \log X)$ (Bonet et al., 2023). The matrix logarithm maps SPD descriptors to symmetric matrices, and Frobenius-orthonormal coordinates preserve their Log-Euclidean geometry. Even once logarithms are available, L independent directions require $O(Lnm)$ projection work on n descriptors, in addition to sorting. Thus a large slicing budget can remain costly within an optimization loop.

We present SPD hierarchical sliced-Wasserstein (SPDHSW), which separates the number of expensive matrix projections from the number of final one-dimensional comparisons. Following the hierarchical construction of Nguyen et al. (2023b), it computes k projections in symmetric log-space and mixes their coordinates into L slices, reducing projection work to $O(knm + Lkn)$. The intermediate coordinates are shared within one evaluation; both bottleneck directions and mixing weights are drawn afresh at every update. A large final slicing family can therefore be evaluated without maintaining a fixed low-dimensional optimization subspace.

The statistical question is how effectively these mixed slices use the available matrix projections. Because they share a bottleneck, their dependence matters even though the squared population

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

objective is unchanged. We separate the estimator variance into between-bank and within-bank components and identify the degree-two harmonic contribution. Extending the analysis to log-coordinate particle gradients gives a distinct precision criterion for the optimization signal. This distinction is practically relevant: at the same FPHA state, the normalized gradient floor is lower than the scalar floor, while controlled FPHA and EEG experiments show improved alignment from adding mixing at a fixed matrix-projection count and learning rate.

Contributions. We connect the hierarchical construction, its estimation properties, and its behavior in SPD distribution alignment. **(i) Population preservation.** Under the Frobenius-uniform sampler of Bonet et al. (2023), we characterize the mixed-direction law and establish equality of the squared SPDSW and SPDHSW population objectives, with a positive constant-factor relationship for general orders. **(ii) Scalar and gradient precision.** We derive the shared-bank variance decomposition, identify its degree-two harmonic component, and extend it to particle gradients. This accounts for the dependence omitted by an independent-slice Monte Carlo rate and yields separate precision criteria for values and gradients. **(iii) Controlled alignment evidence.** In projection-dominated Log-Euclidean optimization, SPDHSW reaches within 0.23% of a 700-direction flat estimator's final FPHA cost using 50 matrix projections and about 12% of its loop time. On held-out 128-channel EEG subjects, adding mixing at the same matrix-projection count and learning rate consistently improves alignment within the same time, and a smaller improvement persists after each method's budget and rate are selected independently.

2 Related Work

SPD distributions and domain adaptation. Log-Euclidean and affine-invariant geometries provide structured operations on SPD descriptors (Arsigny et al., 2007; Pennec et al., 2006). Riemannian transfer learning, parallel transport, and Procrustes analysis align covariance features across domains (Zanini et al., 2018; Yair et al., 2019; Rodrigues et al., 2019); SPDNet and Riemannian batch normalization learn matrix-valued representations (Huang & Van Gool, 2017; Brooks et al., 2019). Bonet et al. (2023) introduced SPDSW and its symmetric-log-space formulation for M/EEG distribution regression and adaptation; we retain that geometry and analyze a hierarchical estimator for its Frobenius-uniform objective, motivated by the large symmetric dimension and repeated comparisons of the SPD setting.

Sliced transport and manifold geometry. Max-SW optimizes a direction (Deshpande et al., 2019), distributional SW learns a slicing distribution (Nguyen et al., 2021), and energy-based SW weights directions by their projected transport costs (Nguyen & Ho, 2023); generalized, tree-sliced, and Markovian constructions alter the projection family, transport structure, or dependence between directions (Kolouri et al., 2019; Le et al., 2019; Nguyen et al., 2023a). Broader manifold-slicing frameworks include pullback-Euclidean and intrinsic Cartan–Hadamard constructions (Bonet et al., 2025); our Log-Euclidean choice provides linear coordinates in $\mathcal{S}_d(\mathbb{R})$ and efficient matrix products without affine-invariant geodesic projections.

Hierarchical slicing and finite-estimator dependence. Nguyen et al. (2023b) introduced bottleneck mixing, established metricity through the hierarchical Radon transform, and analyzed computation and sample complexity. Their Monte Carlo bound (their Proposition 5) treats all final slices as uncorrelated and so predicts an $L^{-1/2}$ error for the single-bank estimator they use in practice; within a bank the slices share the bottleneck directions, and Theorem 2 gives the exact variance $V_b + V_w/L$, whose bottleneck term does not vanish as L grows (Remark 1; Figure 4). The same statements hold for Euclidean HSW because symmetric log-space is isometric to $\mathbb{R}^m$; for the SPD construction we use the Frobenius-uniform direction law, which distinguishes SPDSW from its spectral implementation (Section 4.1).

Precision of sliced estimates. Statistical analyses of sliced divergences distinguish data-sampling error from directional Monte Carlo error (Nadjahi et al., 2020; Bonet et al., 2023). Our results concern the latter for fixed distributions. Spherical-harmonic control variates reduce integration error by fitting harmonic regression features (Leluc et al., 2024). Here, harmonics characterize the dependence created by bottleneck mixing and lead to value- and gradient-specific precision criteria.

3 Background and Hierarchical Estimation

Symmetric log-space. Let $\mathcal{S}_d(\mathbb{R}) = \{S \in \mathbb{R}^{d\times d} : S^\top = S\}$ and $\mathcal{S}_d^{++}(\mathbb{R}) = \{X \in \mathcal{S}_d(\mathbb{R}) : X \succ 0\}$. The matrix logarithm maps $\mathcal{S}_d^{++}(\mathbb{R})$ bijectively onto $\mathcal{S}_d(\mathbb{R})$, with Log-Euclidean distance $d_{\mathrm{LE}}(X, Z) = \|\log X - \log Z\|_F$ (Bhatia, 2007; Arsigny et al., 2007). Equip $\mathcal{S}_d(\mathbb{R})$ with $\langle S, T\rangle_F = \mathrm{Tr}(ST)$ and let $\iota : \mathcal{S}_d(\mathbb{R}) \to \mathbb{R}^m$, $m = d(d+1)/2$, be a Frobenius-orthonormal coordinate map: it retains diagonal entries and multiplies off-diagonal entries by $\sqrt{2}$. Thus $\|\iota(S)\|_2 = \|S\|_F$. Write $\mathcal{P}_p(\mathcal{S}_d^{++}(\mathbb{R}))$ for measures with finite pth Log-Euclidean moment, $1 \le p < \infty$, and $t_{\#}\mu$ for the push-forward under t. W_p uses the Euclidean ground norm, or the Frobenius norm on $\mathcal{S}_d(\mathbb{R})$.

Let λ_S be the uniform probability law on $\mathbb{S}_F(\mathcal{S}_d(\mathbb{R})) = \{A \in \mathcal{S}_d(\mathbb{R}) : \|A\|_F = 1\}$. Under ι, it is the uniform law σ_{m-1} on $\mathbb{S}^{m-1}$; more generally, σ_{q-1} is uniform on the unit sphere in $\mathbb{R}^q$. For $t^A(X) = \langle A, \log X\rangle_F$, define

$$\mathrm{SPDSW}_p^p(\mu, \nu) = \mathbb{E}_{A\sim\lambda_S} W_p^p\big((t^A)_{\#}\mu, (t^A)_{\#}\nu\big). \tag{1}$$

Writing SymSW_p for slicing in the symmetric matrix space with this same direction law gives the reduction of Bonet et al. (2023):

$$\begin{aligned}\mathrm{SPDSW}_p(\mu, \nu) &= \mathrm{SymSW}_p(\log_{\#}\mu, \log_{\#}\nu)\\ &= \mathrm{SW}_p\big((\iota\circ\log)_{\#}\mu, (\iota\circ\log)_{\#}\nu\big).\end{aligned} \tag{2}$$

The last equality uses Frobenius-uniform directions in m dimensions, not uniform slicing of $\mathrm{vec}(\log X)$ in the ambient $\mathbb{R}^{d^2}$ (Section 4.1). Empirical costs use sorted projections and the weighted quantile formula when counts differ (Appendix A.2).

Hierarchical projections. For integers $k, L \ge 1$, draw

$$A_1, \dots, A_k \overset{\text{i.i.d.}}{\sim} \lambda_S, \quad \psi_1, \dots, \psi_L \overset{\text{i.i.d.}}{\sim} \sigma_{k-1}, \quad A_{1:k} \perp\!\!\!\perp \psi_{1:L}, \tag{3}$$

where $\perp\!\!\!\perp$ denotes independence and $A_{1:k} = (A_1, \dots, A_k)$ is the bottleneck bank. Define

$$\begin{aligned} B_\ell &= \sum_{i=1}^k \psi_{\ell i} A_i, & t^{B_\ell}(X) &= \sum_{i=1}^k \psi_{\ell i}\,\mathrm{Tr}(A_i \log X),\\ Y_\ell &= W_p^p\big((t^{B_\ell})_{\#}\mu, (t^{B_\ell})_{\#}\nu\big), & \widehat{D}_{k,L} &= \frac{1}{L}\sum_{\ell=1}^L Y_\ell.\end{aligned} \tag{4}$$

The population distance is $\mathrm{SPDHSW}_{p,k} = (\mathbb{E}\widehat{D}_{k,L})^{1/p}$. Every B_ℓ is symmetric and defines a Log-Euclidean coordinate. Its generally nonunit length is part of the estimator.

Computation within an update; resampling between updates. For source and target log-coordinate matrices $S_X \in \mathbb{R}^{n_s\times m}$ and $S_Z \in \mathbb{R}^{n_t\times m}$, set $A = [\iota(A_1), \dots, \iota(A_k)]$ and $\Psi = [\psi_1, \dots, \psi_L]$; compute $S_X A, S_Z A$ once in the current evaluation, then form $(S_X A)\Psi, (S_Z A)\Psi$. Both A and Ψ are redrawn at the next update; a fixed-direction control keeps both throughout. For log-coordinate SGD, a fixed bank confines each particle update to its span, and resampling changes this span at every step (Algorithm 1).

4 Statistical and Computational Properties

4.1 Direction law and population preservation

We draw Frobenius-uniform directions using the symmetric-Gaussian construction of Bonet et al. (2023):

$$G_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \qquad A = \frac{G + G^\top}{\|G + G^\top\|_F}. \tag{5}$$

Its coordinates under ι are isotropic Gaussian before normalization, so the law is invariant under the full $O(m)$ action on Frobenius coordinates, which gives the spherical symmetry used in Lemma 1.

Two other conventions appear in prior work. The spectral construction $P\,\mathrm{diag}(\rho)P^\top$, with P Haar on $O(d)$ and $\rho \sim \sigma_{d-1}$, is invariant only under conjugation and has a different second-moment form (Proposition 1); under it the mixed directions are not spherically symmetric, so the angular–radial factorization and Theorem 1 are unavailable. The ambient construction of LogSW and LogHSW, which slices $\mathrm{vec}(\log X) \in \mathbb{R}^{d^2}$ along uniform directions, is $O(m)$-invariant after projection onto the symmetric subspace: its squared population objective is $(m/d^2)\,\mathrm{SPDSW}_2^2$ with an additional radial variance, and rescaling its learning rate by d^2/m removes the accuracy gap between the ambient and symmetric-log-space estimators on BNCI2014-001 (Appendix E.12). Throughout, SPDSW denotes (1), and every SPDSW baseline uses (5).

Lemma 1 (Angular–radial factorization). Under (3), a generic effective direction $B = \sum_i \psi_i A_i$ satisfies $\Pr(B = 0) = 0$ and has the factorization

$$B = RU, \qquad R = \|B\|_F, \qquad U \sim \lambda_S, \qquad U \perp\!\!\!\perp R.$$

Moreover, $\mathbb{E}R^2 = 1$ and $\mathbb{E}R^4 = 1 + \varepsilon_k$, where $\varepsilon_k = 2(k-1)/[m(k+2)]$.

Applying any $Q \in O(m)$ to the whole bank leaves its law unchanged, so $QB \overset{d}{=} B$. Spherical symmetry separates angle and radius. The radial moments follow from $\mathbb{E}\psi\psi^\top = I_k/k$ and the spherical fourth moments; Appendix B.2 gives the details.

Theorem 1 (Population relationship). For $\mu, \nu \in \mathcal{P}_p(\mathcal{S}_d^{++}(\mathbb{R}))$, $1 \le p < \infty$, and every $k \ge 1$,

$$\mathrm{SPDHSW}_{p,k}^p(\mu,\nu) = c_{d,k,p}\mathrm{SPDSW}_p^p(\mu,\nu), \qquad c_{d,k,p} = \mathbb{E}R^p \in (0, k^{p/2}]. \tag{6}$$

For every p, k above, $\mathrm{SPDHSW}_{p,k}$ is a metric because it is a positive constant multiple of SPDSW_p. In particular, $c_{d,k,2} = 1$ for every k, and $c_{d,1,p} = 1$ for every p: the population distances coincide when $p = 2$ or $k = 1$.

Proof sketch. Scaling both projected measures by R multiplies their pth transport cost by R^p, and Lemma 1 makes R independent of the direction; $0 < R \le \sum_i |\psi_i| \le \sqrt{k}$, $\mathbb{E}R^2 = 1$, and $R = 1$ when $k = 1$. □

Normalizing B to $B/\|B\|_F$ gives population equality for every p. Appendix B.4 proves metricity in full, and Appendix B.5 transfers the bounded-support Wasserstein bounds of pullback-Euclidean slicing (Bonnotte, 2013; Bonet et al., 2025) to SPDHSW for every k through (6). Our experiments use the unnormalized squared estimator.

4.2 Shared-bottleneck variance

Fix $p = 2$ and the distribution pair. Let

$$h(u) = W_2^2\big((t^u)_\#\mu, (t^u)_\#\nu\big), \qquad \bar{h} = \mathbb{E}_{\lambda_S} h, \qquad \sigma_h^2 = \mathrm{Var}_{\lambda_S}(h).$$

The bound $h(u) \le W_2^2(\log_\#\mu, \log_\#\nu)$ ensures $h \in L^2(\lambda_S)$. All variance below is over projection randomness, not sampling of the data.

Theorem 2 (Shared-bank variance and its quadratic component). Define $V_b = \mathrm{Var}_A(\mathbb{E}_\psi[Y_1 \mid A])$ and $V_w = \mathbb{E}_A\,\mathrm{Var}_\psi(Y_1 \mid A)$, with $A = A_{1:k}$. Then

$$\begin{aligned} &\mathrm{Cov}(Y_\ell, Y_{\ell'}) = V_b(k) \quad (\ell \ne \ell'), \qquad \mathrm{Var}(\widehat{D}_{k,L}) = V_b(k) + \frac{V_w(k)}{L}, \\ &V_b(k) + V_w(k) = (1+\varepsilon_k)\sigma_h^2 + \varepsilon_k \bar{h}^2. \end{aligned} \tag{7}$$

For $d \ge 2$, write $h = \bar{h} + h_2 + r$, where h_2 is its degree-two spherical-harmonic component and r contains even degrees at least four. With $L^2(\lambda_S)$ norms,

$$V_b(k) = \frac{\|h_2\|^2}{k} + V_{b,\ge 4}(k), \qquad 0 \le V_{b,\ge 4}(k) \le \|r\|^2. \tag{8}$$

Proof sketch. Conditional independence gives the variance decomposition. Since $Y_1 = R^2 h(U)$, Lemma 1 gives the single-slice identity. Define $T_k f = \mathbb{E}_\psi[R^2 f(U) \mid A]$. For each bank $\mathbb{E}_\psi R^2 = 1$,

so weighted Jensen makes T_k an L^2 contraction. Writing $h_2(u) = \iota(u)^\top C\iota(u)$ with $\operatorname{Tr} C = 0$ gives $T_k h_2 = k^{-1}\sum_i h_2(A_i)$. Equivariance under $O(m)$ preserves orthogonality between different harmonic degrees. Their variances therefore add, yielding (8). The full proof is in Appendix B.6. □

The bottleneck determines V_b; increasing L averages only V_w. The $(HL)^{-1/2}$ rate of Nguyen et al. (2023b, Prop. 5) requires uncorrelated slices: averaging H independent banks gives $V_b/H + V_w/(HL)$, which approaches V_b/H as L grows at fixed H (Remark 1). For a globally quadratic directional cost, $r = 0$ and $V_b = \sigma_h^2/k$ exactly. This includes log-space translations and proportional-covariance Gaussian pairs. Such costs cannot gain scalar precision over k independent flat directions by adding finitely many mixtures. For other costs, the higher-degree contribution determines the available precision gain.

4.3 From discrepancy values to training gradients

For empirical log-coordinates $S = (s_i)_{i=1}^{n_s}$ and fixed $T = (t_j)_{j=1}^{n_t}$, let $h_S(u)$ be the directional squared cost and $g_S(u) = \nabla_S h_S(u)$. At differentiable states, with projected optimal coupling π^u,

$$(g_S(u))_i = 2\sum_j \pi^u_{ij}\langle s_i - t_j, u\rangle u. \tag{9}$$

Here s_i, t_j, u are Frobenius coordinate vectors, and π^u has row masses $1/n_s$ and column masses $1/n_t$. This form includes unequal sample counts. The coupling depends on direction, so g_S need not have the same harmonic composition as h_S.

Theorem 3 (Gradient variance of hierarchical projections). Assume $h_S(u)$ is differentiable for almost every direction, $g_S \in L^2(\lambda_S; \mathbb{R}^{n_s\times m})$, and differentiation can be interchanged with directional expectation. Set $g = g_S$, $\bar g = \mathbb{E}g(U)$, and $\sigma_g^2 = \mathbb{E}\|g(U) - \bar g\|_F^2$. With $G_\ell = R_\ell^2 g(U_\ell)$, define $V_b^g = \mathbb{E}\|\mathbb{E}[G_1 \mid A] - \bar g\|_F^2$ and $V_w^g = \mathbb{E}\|G_1 - \mathbb{E}[G_1 \mid A]\|_F^2$. Then $\widehat g_{k,L} = L^{-1}\sum_\ell G_\ell$ is unbiased for $\bar g$ and

$$\begin{aligned} \mathbb{E}\|\widehat g_{k,L} - \bar g\|_F^2 &= V_b^g(k) + \frac{V_w^g(k)}{L}, \\ V_b^g(k) + V_w^g(k) &= (1+\varepsilon_k)\sigma_g^2 + \varepsilon_k\|\bar g\|_F^2. \end{aligned} \tag{10}$$

The decomposition (8) also holds componentwise for g, with the squared L^2 norms summed over gradient coordinates.

The result follows because the sampler is independent of S and $\nabla_S h_S(RU) = R^2 g_S(U)$; Appendix B.7 verifies the empirical differentiability conditions and proves the vector-valued identities, which connects the shared-bank analysis to the gradient used by log-coordinate optimization.

The value and gradient can have different directional variability at the same state: Appendix B.8 gives a Gaussian example with constant directional cost but nonconstant parameter gradient, and the FPHA experiment tests the distinction for empirical particle gradients. Under smoothness assumptions, the variance in (10) is also the noise term in a stochastic-descent bound (Appendix B.9).

4.4 Precision criteria and computational allocation

For either the scalar cost or the gradient, write σ_f^2 for its flat-direction variance and V_b^f, V_w^f for its shared-bank components; gradient variance means trace covariance. The flat average of k directions has mean-squared error σ_f^2/k.

Corollary 1 (Precision at equal matrix-projection count). Hierarchical averaging has strictly smaller mean-squared error than the flat k-direction average if and only if

$$V_b^f(k) < \frac{\sigma_f^2}{k}, \qquad L > \frac{V_w^f(k)}{\sigma_f^2/k - V_b^f(k)}. \tag{11}$$

For $\sigma_f^2 > 0$, the normalized floor $\rho_f = kV_b^f/\sigma_f^2$ quantifies the potential benefit of mixing; its limiting RMS-error gain is $\rho_f^{-1/2}$ when $V_b^f > 0$, and scalar and gradient gains depend on different ρ_f.

Appendix C.2 gives the minimum feasible L under mixing and sorting costs and a cost comparison that includes the flat estimator.

When does the hierarchy reduce computation? With available logarithms, a cost model with projection, mixing, and sorting coefficients c_p, c_m, c_s gives, for $N > k$ flat directions, a break-even symmetric dimension $m_\star = [c_m kL + c_s(L-N)\log n]/[c_p(N-k)]$ above which the hierarchy is cheaper; at $N = L$ this is $(c_m/c_p)kL/(L-k)$, and at $N = k$ mixing adds work that must be earned through a better estimate (Appendix C.1). We therefore target high-dimensional, repeatedly differentiated SPD discrepancies with cached or directly optimized log-coordinates; precision is assessed by Corollary 1.

5 Experiments

We evaluate computational scaling, distribution alignment, downstream classification, and scalar–gradient precision. Unless marked fixed, sliced comparisons use the Frobenius-uniform sampler in (5) and redraw the complete projection family at every update. The training loss is the empirical squared sliced discrepancy. Alignment is scored separately by the full squared Log-Euclidean Wasserstein cost $Q = W^2_{2,\mathrm{LE}}$, reported relative to its initial value Q_0 on High-Gamma and as $0.1Q$ on the Log-Euclidean FPHA pair (Appendix D.3).

5.1 Where the computational savings arise

Figure 1 profiles one post-logarithm estimator call at $n = 600$ and $L = 700$. At $d = 128$, the sums of stage medians are 63.57 ms for flat slicing and 4.66, 10.47, and 12.72 ms for bottleneck counts 20, 50, and 100. Sorting remains close to 1 ms; most of the saving comes from matrix projection (Appendix E.2). Table 1 shows when this saving carries into optimization. Under Log-Euclidean (LEM) updates, particles are optimized directly in log-coordinates and replacing flat slicing reduces a High-Gamma iteration from 29.4 to 5.0 ms. Under affine-invariant (AIM) updates, the particles are SPD matrices, so the matrix logarithm of every source particle is recomputed at each iteration; this alone takes approximately 1679 ms of a 1747 ms iteration and masks most of the estimator-level saving. This comparison locates the benefit in repeated computations with available log-coordinates.

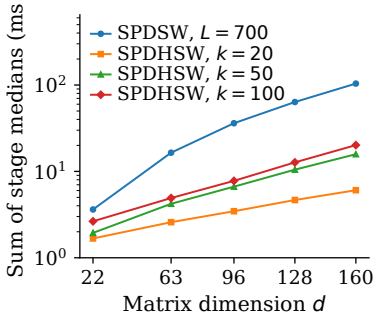


Figure 1: Sum of measured stage medians for one estimator call at $n = 600$, $L = 700$, over matrix dimension (medians of 100 timing repetitions; Appendix E.2).

Table 1: One High-Gamma alignment iteration (ms; $d = 128$, $k = 40$, $L = 500$, subject 1) under affine-invariant (AIM) and Log-Euclidean (LEM) updates. Under LEM the logarithms are computed once outside the loop.

	AIM		LEM	
ms/iter	SPDSW	SPDHSW	SPDSW	SPDHSW
Matrix logarithm	1679.1	1679.5	0	0
Sliced forward	18.8	3.0	18.3	2.8
Backward pass	48.9	39.9	11.1	2.1
Total	1746.75	1722.41	29.42	4.97
SPDSW / SPDHSW	1.01×		5.92×	

5.2 Alignment under controlled projection and time budgets

Fixed FPHA pair. We align 600 source and 575 target covariance descriptors of 63-dimensional hand-joint trajectories ($m = 2016$), following the split of Wang et al. (2022). Source log-coordinates take 500 SGD steps at the shared rate $\eta = 2.5 \times 10^4$, and the independent score $0.1Q$ is evaluated every 25 updates. Figure 2 and Table 2 show that SPDHSW $(50, 700)$ reaches within 0.23% of flat $L = 700$'s final score in 3.61 rather than 30.20 s, or about 12% of its loop time. At the same rate,

flat $L = 50$ takes 3.20 s but its score increases at every recorded interval in all five seeds. Mixing therefore recovers near-large-budget alignment at a fixed matrix-projection count of 50 and a shared learning rate. Affine-invariant and fixed-direction controls are in Appendix E.11; the latter show that resampling, rather than a permanently fixed bank, is what makes the large mixed budget useful.

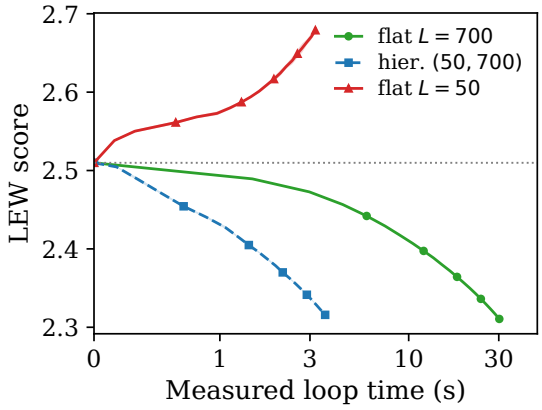


Figure 2: FPHA alignment under Log-Euclidean updates at a shared rate: LEW score (lower is better) against measured loop time, mean of five seeds with ± 1 s.d. bands, evaluated every 25 updates. The dotted line is the initial score; the time axis is linear below 1 s and logarithmic above.

Table 2: Final FPHA LEW scores of the runs in Figure 2 (mean $\pm$ s.d., five seeds) and their 500-update loop time. The original affine-invariant runs are in Appendix E.11.

Method	LEW ↓	Time (s)
No alignment	2.5098	–
SPDSW, $L = 700$	2.3106 ± 0.0003	30.20
SPDSW, $L = 50$	2.6789 ± 0.0034	3.20
SPDHSW, $(50, 700)$	2.3159 ± 0.0008	3.61

Held-out High-Gamma subjects. All 128 channels are retained ($m = 8256$), and each subject's training session is aligned to the unlabeled test session in log-coordinates (Schirrmeister et al., 2017). Configurations are selected on four development subjects and evaluated on the remaining ten with five seeds; each subject–seed pair receives a common measured update-time budget set by reference flat runs, and loop time excludes Q evaluation, initialization, I/O, and classifier fitting (Appendix G).

At 40 matrix projections per update and $\eta = 10^4$, SPDHSW $(40, 500)$ lowers mean Q/Q_0 from 0.6481 to 0.6317 within approximately 13 s (Figure 3; Table 3). Every subject and seed favors the hierarchy (50 of 50 pairs); the mean paired difference is -0.016 with a 95% subject-cluster interval of $[-0.019, -0.014]$. The hierarchy completes fewer updates (3455 versus 3653), so additional iterations do not account for this improvement. Flat $L = 40$ also uses its best rate among the four tested on development subjects. At that rate, larger flat counts $L = 60$ and 100 do not close the development gap (Table 25). The paired difference favors SPDHSW from one quarter of the measured time budget onward.

Separately selected budgets and learning rates. Selecting each family's budget and rate on the development subjects yields flat $L = 500$, $\eta = 10^5$ and SPDHSW $(150, 500)$, $\eta = 3 \times 10^4$. On the same confirmation set, the hierarchy attains mean $Q/Q_0 = 0.6251$ versus 0.6276 within approximately 13 s. This smaller improvement holds in 44 of 50 pairs, with a paired difference of -0.0025 and a subject-cluster interval of $[-0.0040, -0.0001]$. The hierarchy allocates 150 rather than 500 matrix projections to each update and completes about 1400 rather than 500 updates within the budget. The same-count benefit thus persists, in reduced form, after independent selection (full results in Appendix G; classification of these settings in Appendix G.4).

5.3 EEG CLASSIFICATION

Table 4 follows the alignment-then-classification protocol of Bonet et al. (2023) with 500 updates and five seeds: BNCI2014-001 (subjects $\{1, 3, 7, 8, 9\}$) and BNCI2015-001 (all 12 subjects) are cross-subject tasks, and High-Gamma (all 14 subjects) is cross-session. Both methods share the learning rate and final count $L = 500$, and a linear SVM is fitted on the aligned source with source labels only (Appendix D).

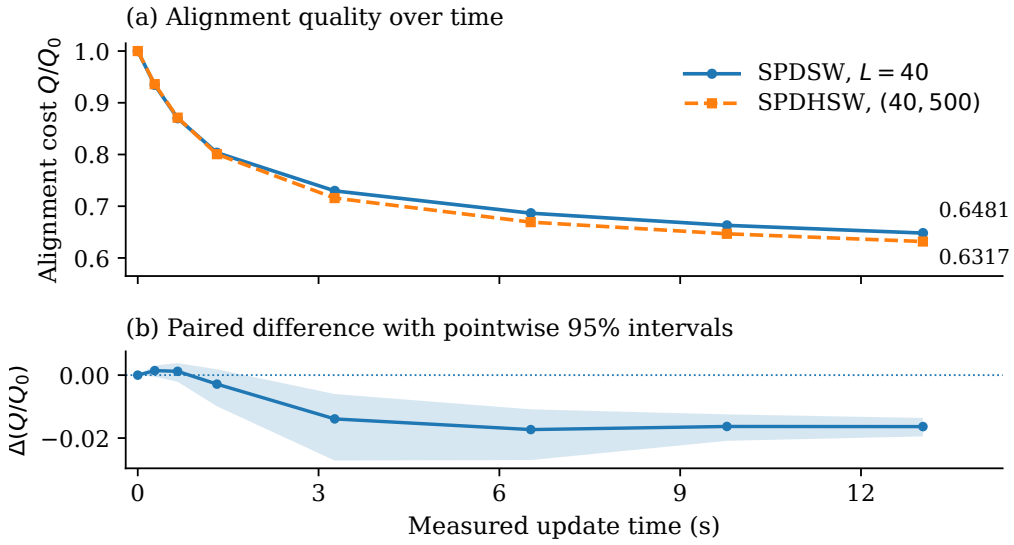


Figure 3: High-Gamma alignment with 40 matrix projections per update and $\eta = 10^4$ (ten confirmation subjects, five seeds). (a) Mean Q/Q_0 against each method's measured update time. (b) Paired SPDHSW-minus-SPDSW difference against mean paired time; shading gives pointwise 95% subject-cluster bootstrap intervals. Lines connect observed checkpoints. Subject-level endpoints are in Appendix G.3.

Table 3: High-Gamma alignment on the ten confirmation subjects under a common measured time budget (means over subjects and five seeds). Updates are those completed within the budget.

Method	Learning rate	$Q/Q_0 \downarrow$	Loop time (s)	Updates
Same matrix-projection count per update and learning rate				
SPDSW, $L = 40$	10^4	0.6481	13.03	3,653
SPDHSW, $(40, 500)$	10^4	0.6317	13.03	3,455
Budgets and learning rates selected on development subjects				
SPDSW, $L = 500$	10^5	0.6276	13.13	504
SPDHSW, $(150, 500)$	3×10^4	0.6251	13.07	1,398

Table 4: EEG classification after alignment (500 updates, $L = 500$, $k = 40$ for SPDHSW; mean $\pm$ s.d. over five seeds). Time ratios pair the same seed and task and cover the alignment loop only. BNCI tasks are cross-subject; High-Gamma is within-subject cross-session.

	BNCI2014 ($d = 22$)	BNCI2015 ($d = 13$)	High-Gamma ($d = 128$)
No alignment	38.29	55.84	77.90
SPDSW, $L = 500$	59.00 ± 0.15	64.56 ± 0.06	79.03 ± 0.21
SPDHSW, $(40, 500)$	58.87 ± 0.09	64.61 ± 0.11	79.31 ± 0.61
Flat/hierarchical loop time	1.66×	1.06×	6.03×

The observed mean accuracies are close across all three datasets. On High-Gamma, SPDHSW obtains 79.31% versus 79.03% with a $6.03\times$ lower alignment-loop time. A separate SPDNet regularization experiment also retains similar observed accuracy, with AFEW's additional regularizer cost falling from 3.87 to 1.54 s per epoch (Appendix E.13).

5.4 Scalar and gradient precision on real SPD data

Theorem 3 distinguishes value and gradient precision at the same state. We test this on the initial FPHA pair: 100 banks with 200 mixtures each estimate the variance components, and 4096 independent flat directions estimate their marginal variances. The gradient includes every source particle. Table 5 normalizes each quantity by its own flat-k variance.

Table 5: FPHA same-state precision diagnostic; all variances are normalized by the flat k-direction variance (1). The first two columns are the bottleneck floors for the scalar cost and particle gradient; the last two use 100 direct evaluations at $L = 700$, an RMS gain above 1 meaning smaller gradient fluctuations. Intervals are in Appendix F.

k	Scalar floor ρ_h	Gradient floor ρ_g	Gradient variance ratio	RMS gain
20	1.02	0.58	0.59	$1.30\times$
50	1.04	0.56	0.63	$1.26\times$
100	1.25	0.59	0.73	$1.17\times$

The normalized gradient floor is 0.56–0.59, whereas the scalar floor is near 1 with wider uncertainty; paired intervals for their difference are below zero at all three budgets. At $L = 700$, the observed gradient variance is 0.59–0.73 of flat k, consistent with the decomposition and corresponding to RMS-fluctuation gains of 1.17–$1.30\times$. For $k = 50$, its estimated variance equals that of approximately 79 independent flat directions, quantifying the precision obtained by mixing 50 matrix projections (Appendix F).

The separate gradient-free study on 18 synthetic pairs finds 1.17–$1.43\times$ lower relative RMSE at equal matrix-projection count with 2000 mixtures, but flat sampling is more economical at the 5% scalar-RMSE target for $d = 22, 63$ (Appendix E.1). The paired FPHA diagnosis therefore adds information that scalar error alone cannot provide. It measures a fixed-state gradient effect; its interaction with step-size stability along a full optimization trajectory remains a separate question.

6 Discussion and Conclusion

SPDHSW forms a large family of one-dimensional comparisons from a small bank of symmetric-log-space projections while preserving the squared SPDSW population objective. Our analysis characterizes the dependence introduced by mixing and separates the precision of discrepancy values from that of particle gradients. The experiments connect this construction to measurable benefits: lower projection-dominated loop cost, improved alignment at equal matrix-projection counts and learning rates, and a smaller alignment-cost gain after separate budget-and-rate selection on High-Gamma.

The computational setting is repeated differentiation in high-dimensional symmetric log-space, with logarithms cached or directly optimized. The fixed-state gradient diagnosis complements the alignment experiments; tracking its components along optimization trajectories and comparing alternative direction samplers are natural next steps. Together, the results show how hierarchical mixing makes a limited budget of matrix projections more effective for SPD distribution alignment.

Reproducibility Statement

Section 3 and Appendix A specify the sampling scheme and estimation algorithm. Appendix B provides the proofs, and Appendices D–G describe the experimental protocols and evaluation procedures. We will publicly release the implementation and experimental scripts upon acceptance.

Ethics Statement

This work uses existing EEG and vision benchmark data and involves no new human-subject data collection. The experiments evaluate distribution alignment and classification on these benchmarks, not clinical diagnosis. Applications to EEG require appropriate consent, data privacy, and validation across subjects and recording conditions.

AI Use Statement

Generative AI tools assisted with editing the text and with checking derivations and experimental code. The authors verified all mathematical claims, experimental results, and references, and take full responsibility for the content.

References

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347, 2007.

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.

Rajendra Bhatia. *Positive Definite Matrices*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2007.

Clément Bonet, Benoît Malézieux, Alain Rakotomamonjy, Lucas Drumetz, Thomas Moreau, Matthieu Kowalski, and Nicolas Courty. Sliced-Wasserstein on symmetric positive definite matrices for M/EEG signals. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2777–2805. PMLR, 2023.

Clément Bonet, Lucas Drumetz, and Nicolas Courty. Sliced-Wasserstein distances and flows on Cartan–Hadamard manifolds. *Journal of Machine Learning Research*, 26(32):1–76, 2025.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Nicolas Bonnotte. *Unidimensional and Evolution Methods for Optimal Transportation*. PhD thesis, Université Paris-Sud and Scuola Normale Superiore di Pisa, 2013.

Daniel Brooks, Olivier Schwander, Frédéric Barbaresco, Jean-Yves Schneider, and Matthieu Cord. Riemannian batch normalization for SPD neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G. Schwing. Max-sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10648–10656, 2019.

Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):34–41, 2012.

Josef Faller, Carmen Vidaurre, Teodoro Solis-Escalante, Christa Neuper, and Reinhold Scherer. Autocalibration and recurrent adaptation: Towards a plug and play online ERD-BCI. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(3):313–319, 2012.

Kai-Tai Fang, Samuel Kotz, and Kai Wang Ng. *Symmetric Multivariate and Related Distributions*. Chapman and Hall, 1990.

Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 409–419, 2018.

Zhiwu Huang and Luc Van Gool. A Riemannian network for SPD matrix learning. In *AAAI Conference on Artificial Intelligence*, pp. 2036–2042, 2017.

Vinay Jayaram and Alexandre Barachant. MOABB: Trustworthy algorithm benchmarking for BCIs. *Journal of Neural Engineering*, 15(6):066011, 2018. doi: 10.1088/1741-2552/aadea0.

Max Kochurov, Rasul Karimov, and Serge Kozlukov. Geoopt: Riemannian optimization in PyTorch. *arXiv preprint arXiv:2005.02819*, 2020.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. In *Advances in Neural Information Processing Systems*, volume 32, pp. 261–272, 2019.

Tam Le, Makoto Yamada, Kenji Fukumizu, and Marco Cuturi. Tree-sliced variants of Wasserstein distances. In *Advances in Neural Information Processing Systems*, volume 32, pp. 12283–12294, 2019.

Rémi Leluc, Aymeric Dieuleveut, François Portier, Johan Segers, and Aigerim Zhuman. Sliced-Wasserstein estimation with spherical harmonics as control variates. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27191–27214. PMLR, 2024.

Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update. *Journal of Neural Engineering*, 15(3):031005, 2018.

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Khai Nguyen and Nhat Ho. Energy-based sliced Wasserstein distance. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Khai Nguyen, Nhat Ho, Tung Pham, and Hung Bui. Distributional sliced-Wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2021.

Khai Nguyen, Tongzheng Ren, and Nhat Ho. Markovian sliced Wasserstein distances: Beyond independent projections. In *Advances in Neural Information Processing Systems*, volume 36, 2023a.

Khai Nguyen, Tongzheng Ren, Huy Nguyen, Litu Rout, Tan Nguyen, and Nhat Ho. Hierarchical sliced Wasserstein distance. In *International Conference on Learning Representations*, 2023b.

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

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision (SSVM 2011)*, volume 6667 of *Lecture Notes in Computer Science*, pp. 435–446. Springer, 2012. doi: 10.1007/978-3-642-24785-9_37.

Pedro Luiz Coelho Rodrigues, Christian Jutten, and Marco Congedo. Riemannian Procrustes analysis: Transfer learning for brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 66(8):2390–2401, 2019.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

Elias M. Stein and Guido Weiss. *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton University Press, 1971.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot R. Müller-Putz, et al. Review of the BCI competition IV. *Frontiers in Neuroscience*, 6:55, 2012.

Oncel Tuzel, Fatih Porikli, and Peter Meer. Region covariance: A fast descriptor for detection and classification. In *European Conference on Computer Vision*, pp. 589–600, 2006.

Rui Wang, Xiao-Jun Wu, and Josef Kittler. SymNet: A simple symmetric positive definite manifold deep learning method for image set classification. *IEEE Transactions on Neural Networks and Learning Systems*, 33(5):2208–2222, 2022.

Or Yair, Mirela Ben-Chen, and Ronen Talmon. Parallel transport on the cone manifold of SPD matrices for domain adaptation. *IEEE Transactions on Signal Processing*, 67(7):1797–1811, 2019.

Florian Yger, Maxime Berar, and Fabien Lotte. Riemannian approaches in brain–computer interfaces: A review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1753–1762, 2017.

Paolo Zanini, Marco Congedo, Christian Jutten, Salem Said, and Yannick Berthoumieu. Transfer learning: A Riemannian geometry framework with applications to brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 65(5):1107–1116, 2018.

A ALGORITHM AND NOTATION

Algorithm 1: SPDHSW for two equally weighted empirical measures of size n. Each resampled evaluation draws a new bank and new mixing vectors. Unequal sample counts use the corresponding weighted one-dimensional transport calculation.

Input : $\{X_j\}_{j=1}^n, \{Z_j\}_{j=1}^n \subset \mathcal{S}_d^{++}(\mathbb{R})$; budgets k, L; order p

Output : $\widehat{D}_{k,L} \approx \mathrm{SPDHSW}_{p,k}^p(\mu, \nu)$

$S_j \leftarrow \log X_j, T_j \leftarrow \log Z_j$; `// cached; shared by all sliced methods`

for $i = 1$ **to** k **do**

  $G \sim \mathcal{N}(0,1)^{d\times d}$; $A_i \leftarrow (G+G^\top)/\|G+G^\top\|_F$;

  $u_{ij} \leftarrow \langle A_i, S_j\rangle_F, v_{ij} \leftarrow \langle A_i, T_j\rangle_F$; `// O(nm) each`

Draw $\psi_1, \ldots, \psi_L \overset{\text{i.i.d.}}{\sim} \sigma_{k-1}$;

for $\ell = 1$ **to** L **do**

  $a_{\ell j} \leftarrow \sum_i \psi_{\ell i} u_{ij}, b_{\ell j} \leftarrow \sum_i \psi_{\ell i} v_{ij}$; `// O(kn) each`

  sort $a_{\ell\cdot}, b_{\ell\cdot}$; $Y_\ell \leftarrow \frac{1}{n}\sum_j |a^{\uparrow}_{\ell j} - b^{\uparrow}_{\ell j}|^p$;

return $\widehat{D}_{k,L} = \frac{1}{L}\sum_\ell Y_\ell$;

A.1 FROBENIUS-ORTHONORMAL IDENTIFICATION

The matrices E_{ii} $(i = 1, \ldots, d)$ and $(E_{ij} + E_{ji})/\sqrt{2}$ $(i < j)$ form a Frobenius-orthonormal basis of $\mathcal{S}_d(\mathbb{R})$; the induced coordinate map $\iota : \mathcal{S}_d(\mathbb{R}) \to \mathbb{R}^m$ is a linear isometry, and $\iota(\mathbb{S}_F(\mathcal{S}_d(\mathbb{R}))) = \mathbb{S}^{m-1}$. All statements about $O(m)$ refer to the orthogonal group of $\mathbb{R}^m$ acting on $\mathcal{S}_d(\mathbb{R})$ through ι. In these coordinates $S = G + G^\top$ has independent coordinates $S_{ii} = 2G_{ii} \sim \mathcal{N}(0, 4)$ and $\sqrt{2}S_{ij} = \sqrt{2}(G_{ij} + G_{ji}) \sim \mathcal{N}(0, 4)$, so S is an isotropic Gaussian vector in $\mathbb{R}^m$ and $A = S/\|S\|_F$ is exactly uniform on $\mathbb{S}^{m-1}$, i.e. λ_S-distributed (Bonet et al., 2023).

A.2 ONE-DIMENSIONAL TRANSPORT

For equally weighted samples with equal counts, the one-dimensional transport cost is computed by sorting, as in Algorithm 1. For unequal counts n_s, n_t, we use the empirical quantile formula

$$W_p^p(\alpha, \beta) = \int_0^1 \left|F_\alpha^{-1}(u) - F_\beta^{-1}(u)\right|^p \, du.$$

The empirical quantile functions are piecewise constant, so the integral is evaluated on the union of their cumulative-weight breakpoints. This is the form used for distribution pairs of unequal sizes.

B PROOFS

B.1 PROOF OF PROPOSITION 1 (SECOND MOMENTS OF THE DIRECTION LAWS)

Proposition 1 (Second moments of the direction laws). For $M \in \mathcal{S}_d(\mathbb{R})$, Frobenius-uniform sampling and independent spectral sampling satisfy

$$\mathbb{E}_{\lambda_S}\langle A, M\rangle_F^2 = \frac{\|M\|_F^2}{m}, \qquad \mathbb{E}_{\mathrm{spec}}\langle A, M\rangle_F^2 = \frac{2\|M\|_F^2 + (\operatorname{Tr} M)^2}{d(d+2)}.$$

For $d \geq 2$, the spectral law is invariant under $O(d)$ conjugation but not under the full $O(m)$ action on Frobenius coordinates.

This distinguishes the two sampling conventions discussed by Bonet et al. (2023). The spectral-to-Frobenius ratio is $(d+1)/(d+2)$ for a traceless M and $(d+1)/2$ for $M = I_d/\sqrt{d}$. Thus identical projection formulas need not imply identical sliced objectives.

Spectral construction. Let $A = P\,\mathrm{diag}(\rho)P^\top$ with P Haar on $O(d)$ and $\rho \sim \sigma_{d-1}$ independent. For $M \in \mathcal{S}_d(\mathbb{R})$, $\langle A, M\rangle_F = \sum_i \rho_i (P^\top M P)_{ii}$. Since $\mathbb{E}[\rho_i \rho_j] = \delta_{ij}/d$,

$$\mathbb{E}_\rho \langle A, M\rangle_F^2 = \frac{1}{d}\sum_{i=1}^{d} \big((P^\top M P)_{ii}\big)^2 = \frac{1}{d}\sum_{i=1}^{d} (p_i^\top M p_i)^2,$$

where p_i is the i-th column of P, uniform on $\mathbb{S}^{d-1}$. For p uniform on $\mathbb{S}^{d-1}$ and symmetric M the standard fourth-moment identity $\mathbb{E}[p_a p_b p_c p_e] = (\delta_{ab}\delta_{ce} + \delta_{ac}\delta_{be} + \delta_{ae}\delta_{bc})/(d(d+2))$ gives $\mathbb{E}(p^\top M p)^2 = \big(2\,\mathrm{Tr}(M^2) + (\mathrm{Tr}\,M)^2\big)/(d(d+2)) = \big(2\|M\|_F^2 + (\mathrm{Tr}\,M)^2\big)/(d(d+2))$. Averaging over i leaves this unchanged, which is the first claim.

Frobenius-uniform law. Under ι, A is uniform on $\mathbb{S}^{m-1}$ and $\langle A, M\rangle_F = \langle \iota A, \iota M\rangle$, so $\mathbb{E}\langle A, M\rangle_F^2 = \|\iota M\|^2/m = \|M\|_F^2/m$.

Ratios and invariance. For $\mathrm{Tr}\,M = 0$ the ratio spectral/Frobenius is $\frac{2\|M\|_F^2}{d(d+2)} \cdot \frac{m}{\|M\|_F^2} = \frac{d+1}{d+2}$; for $M = I_d/\sqrt{d}$ ($\|M\|_F = 1$, $\mathrm{Tr}\,M = \sqrt{d}$) it is $\frac{d+2}{d(d+2)} \cdot m = \frac{d+1}{2}$. An $O(m)$-invariant law on $\mathbb{S}_F(\mathcal{S}_d(\mathbb{R}))$ has a second-moment form proportional to $\|M\|_F^2$; the spectral second-moment form is not, for any $d \geq 2$, since it distinguishes the trace direction. Conjugation invariance of the spectral law is immediate from the Haar distribution of P. □

B.2 Proof of Lemma 1 (angular–radial factorization)

Identify $\mathcal{S}_d(\mathbb{R})$ with $\mathbb{R}^m$ through ι and let $Q \in O(m)$. Since λ_S is the uniform measure on $\mathbb{S}^{m-1}$, it is $O(m)$-invariant, so $(QA_1, \ldots, QA_k) \overset{d}{=} (A_1, \ldots, A_k)$ by independence of the A_i. Because $\psi \perp\!\!\!\perp A_{1:k}$ and Q acts linearly, $QB = \sum_i \psi_i (QA_i) \overset{d}{=} \sum_i \psi_i A_i = B$, so the law of B is $O(m)$-invariant, i.e. spherically symmetric. A spherically symmetric random vector in $\mathbb{R}^m$ that is almost surely nonzero factorizes as $B = RU$ with U uniform on $\mathbb{S}^{m-1}$ and $U \perp\!\!\!\perp R$ (Fang et al., 1990); the conditional law of U given $R = r$ is the unique rotation-invariant probability measure on the sphere for almost every r with respect to the law of R, which is the statement $U \sim \lambda_S$, $U \perp\!\!\!\perp R$. Finally $\mathbb{P}(B = 0) = 0$: conditionally on $A_{1:k}$ the linear map $T_A : \psi \mapsto \sum_i \psi_i A_i$ is not identically zero (each $\|A_i\|_F = 1$), so $\ker T_A$ is a proper linear subspace of $\mathbb{R}^k$ and $\{B = 0\} = \{\psi \in \ker T_A\}$ has σ_{k-1}-measure zero; integrating over $A_{1:k}$ gives the claim. □

B.3 Radial moments

With $\|A_i\|_F = 1$, $\|B\|_F^2 = 1 + \sum_{i \neq j} \psi_i \psi_j \langle A_i, A_j\rangle_F$. Since $\mathbb{E}[\psi_i \psi_j] = 0$ for $i \neq j$, $\mathbb{E}\|B\|_F^2 = 1$. For the fourth moment, only index pairs with $\{i', j'\} = \{i, j\}$ survive in $\mathbb{E}[(\sum_{i \neq j} \psi_i \psi_j \langle A_i, A_j\rangle_F)^2]$, giving

$$\begin{aligned}\mathbb{E}\|B\|_F^4 &= 1 + 2\sum_{i \neq j} \mathbb{E}[\psi_i^2 \psi_j^2]\, \mathbb{E}\langle A_i, A_j\rangle_F^2 \\ &= 1 + \frac{2k(k-1)}{k(k+2)m} = 1 + \frac{2(k-1)}{m(k+2)} = 1 + \varepsilon_k,\end{aligned}$$

using $\mathbb{E}[\psi_i^2 \psi_j^2] = 1/(k(k+2))$ for $\psi \sim \sigma_{k-1}$ and $\mathbb{E}\langle A_i, A_j\rangle_F^2 = 1/m$ for independent Frobenius-uniform directions. Hence $\mathrm{Var}(R^2) = \varepsilon_k$, which vanishes as $1/m$: at $k = 50$ it equals 0.1256, 0.00745 and 0.000935 for $d = 5, 22, 63$.

B.4 Proof of Theorem 1 (general orders, normalization, metricity)

By Lemma 1, $t^B(M) = \mathrm{Tr}(RU \log M) = R\,t^U(M)$. Let $s_R(x) = Rx$; then $(t^B)_\# \mu = (s_R)_\# (t^U)_\# \mu$ and the one-dimensional Wasserstein distance is homogeneous, $W_p((s_R)_\#\alpha, (s_R)_\#\beta) = R\,W_p(\alpha, \beta)$ for $R \geq 0$. Taking expectations and using $U \perp\!\!\!\perp R$, $\mathrm{SPDHSW}_{p,k}^p = \mathbb{E}[R^p]\,\mathbb{E}_{U \sim \lambda_S} W_p^p((t^U)_\#\mu, (t^U)_\#\nu) = c_{d,k,p}\mathrm{SPDSW}_p^p$. Finiteness: $\|B\|_F \leq \sum_i |\psi_i| \leq \sqrt{k}$, so $c_{d,k,p} \leq k^{p/2}$; positivity follows from $\mathbb{P}(B = 0) = 0$. For $k = 1$, $\psi_1 \in \{\pm 1\}$ and $R \equiv 1$. The normalized projection equals t^U with $U \sim \lambda_S$, so its population value is SPDSW_p^p

for every p; at $p = 2$, $\mathbb{E}R^2 = 1$ shows the unnormalized estimator carries no scale bias and normalization removes the slice-wise radial factor, whose squared-radius variance is $\mathrm{Var}(R^2) = \varepsilon_k$ of Appendix B.3.

Metricity for every $p \geq 1$ and $k \geq 1$. Since $c_{d,k,p} > 0$ is independent of the pair, $\mathrm{SPDHSW}_{p,k} = c_{d,k,p}^{1/p}\mathrm{SPDSW}_p$ is a positive rescaling of a metric. For completeness, the metric axioms can also be checked directly. Fix $A_{1:k}$ and ψ and write $T = t^{A_{1:k},\psi}$. T is continuous, and $|T(M)| \leq \sum_i |\psi_i||\langle A_i, \log M\rangle_F| \leq \sqrt{k}\|\log M\|_F$, so $T_\# \mu \in \mathcal{P}_p(\mathbb{R})$ whenever $\mu \in \mathcal{P}_p(\mathcal{S}_d^{++}(\mathbb{R}))$ (the finite p-th Log-Euclidean moment about any M_0 bounds $\int \|\log M\|_F^p d\mu$ by the triangle inequality). Nonnegativity and symmetry are inherited from W_p; the triangle inequality follows from the triangle inequality of W_p for each T and Minkowski's inequality in $L^p(\mathbb{S}_F(\mathcal{S}_d(\mathbb{R}))^k \times \mathbb{S}^{k-1})$. For the identity of indiscernibles, $\mathrm{SPDHSW}_{p,k}(\mu,\nu) = 0$ forces $\mathrm{SPDSW}_p(\mu,\nu) = 0$ by the constant-factor relation, hence $(t^A)_\#\mu = (t^A)_\#\nu$ for λ_S-a.e. A. With $\tilde{\mu} = \log_\# \mu$, $\tilde{\nu} = \log_\# \nu$ on $\mathcal{S}_d(\mathbb{R}) \simeq \mathbb{R}^m$, this says the characteristic functions agree on $\{sA : s \in \mathbb{R}\}$ for a.e. $A \in \mathbb{S}_F(\mathcal{S}_d(\mathbb{R}))$; a full-measure subset of $\mathbb{S}_F(\mathcal{S}_d(\mathbb{R}))$ is dense, characteristic functions are continuous, so they agree on all of $\mathbb{R}^m$ and $\tilde{\mu} = \tilde{\nu}$ by injectivity of the Fourier transform of finite measures; log is a bijection, so $\mu = \nu$. □

B.5 Log-Euclidean Wasserstein topology

Corollary 2 (Log-Euclidean Wasserstein topology). Let $\mathcal{Q} \subset \mathcal{P}_p(\mathcal{S}_d^{++}(\mathbb{R}))$ have log push-forwards supported in a common Frobenius ball of radius R_0. For $\mu, \nu \in \mathcal{Q}$,

$$\mathrm{SPDHSW}_{p,k}^p(\mu,\nu) \leq c_{d,k,p} W_p^p(\log_\# \mu, \log_\# \nu),$$
$$W_p(\log_\# \mu, \log_\# \nu) \leq C\,\mathrm{SPDHSW}_{p,k}(\mu,\nu)^{1/[p(m+1)]},$$

for $C = C(d,k,p,R_0) > 0$. Hence SPDHSW, SPDSW, and Log-Euclidean Wasserstein induce the same topology on $\mathcal{Q}$.

Let $\tilde{\mu} = (\iota \circ \log)_\# \mu$ and $\tilde{\nu} = (\iota \circ \log)_\# \nu$. The isometry $\iota \circ \log$ gives

$$W_p(\log_\# \mu, \log_\# \nu) = W_p(\tilde{\mu}, \tilde{\nu}), \qquad \mathrm{SPDSW}_p(\mu,\nu) = \mathrm{SW}_p(\tilde{\mu}, \tilde{\nu}).$$

For the second identity, write $\theta = \iota(A)$. Then $t^A(M) = \langle \theta, (\iota \circ \log)(M)\rangle$ and $\iota_\# \lambda_S = \sigma_{m-1}$.

Upper bound. A projection onto a unit direction is 1-Lipschitz. Consequently,

$$W_p^p\big((\theta^\top)_\#\tilde{\mu}, (\theta^\top)_\#\tilde{\nu}\big) \leq W_p^p(\tilde{\mu}, \tilde{\nu}).$$

Averaging over θ and applying Theorem 1 gives the first bound.

Reverse bound. Both push-forwards are supported in a common ball of radius R_0. The bounded-support sliced-Wasserstein bound (Bonnotte, 2013; Bonet et al., 2025) yields

$$W_p(\tilde{\mu}, \tilde{\nu}) \leq C_0(m,p,R_0)\,\mathrm{SW}_p(\tilde{\mu}, \tilde{\nu})^{1/[p(m+1)]}.$$

Substitute $\mathrm{SW}_p = c_{d,k,p}^{-1/p}\mathrm{SPDHSW}_{p,k}$ and use the isometry. The resulting constant is $C = C_0 c_{d,k,p}^{-1/[p^2(m+1)]} > 0$. The two inequalities give the asserted equivalence of topologies. □

B.6 Proof of Theorem 2 (shared-bottleneck variance)

Conditional dependence. Given $A_{1:k}$, the mixing vectors and hence the slices are independent. For $\ell \neq \ell'$, the conditional covariance is zero, giving

$$\mathrm{Cov}(Y_\ell, Y_{\ell'}) = \mathrm{Cov}\big(\mathbb{E}[Y_\ell \mid A_{1:k}], \mathbb{E}[Y_{\ell'} \mid A_{1:k}]\big) = V_b(k).$$

Therefore,

$$\mathrm{Var}(\widehat{D}_{k,L}) = \frac{L(V_b + V_w) + L(L-1)V_b}{L^2} = V_b(k) + \frac{V_w(k)}{L}.$$

Single-slice variance. By Lemma 1, $Y_1 = R^2 h(U)$ with independent R and U. The radial moments in Appendix B.3 imply

$$\begin{aligned}\mathrm{Var}(Y_1) &= \mathbb{E}R^4\,\mathbb{E}h^2 - (\mathbb{E}R^2)^2(\mathbb{E}h)^2 \\ &= (1+\varepsilon_k)(\sigma_h^2 + \bar{h}^2) - \bar{h}^2 = (1+\varepsilon_k)\sigma_h^2 + \varepsilon_k \bar{h}^2.\end{aligned}$$

The law of total variance gives $\mathrm{Var}(Y_1) = V_b + V_w$. Equivalently,

$$\mathrm{Var}(\widehat{D}_{k,L}) = \frac{\mathrm{Var}(Y_1)}{L} + \left(1 - \frac{1}{L}\right) V_b(k), \qquad \frac{\mathrm{Var}(Y_1)}{L} \le \mathrm{Var}(\widehat{D}_{k,L}) \le \mathrm{Var}(Y_1).$$

Harmonic decomposition. Assume $d \ge 2$ and $\sigma_h^2 > 0$. Since $h(-u) = h(u)$, only even spherical-harmonic degrees appear. Write $h = \bar{h} + h_2 + r$, with r collecting degrees at least four. Orthogonality gives $\sigma_h^2 = \|h_2\|^2 + \|r\|^2$, where norms are in $L^2(\lambda_S)$; define $\alpha = \|h_2\|^2/\sigma_h^2$. Write $h_2(u) = \iota(u)^\top C \iota(u)$ for a symmetric $m \times m$ matrix C with $\mathrm{Tr}\, C = 0$. Define

$$(T_k f)(A_{1:k}) = \mathbb{E}_\psi \left[\|B\|_F^2 f(B/\|B\|_F) \mid A_{1:k} \right].$$

For every bank of unit directions, $\mathbb{E}_\psi[R^2 \mid A] = 1$. Hence $R^2\, d\sigma_{k-1}(\psi)$ is a probability measure conditional on that bank. Weighted Jensen gives

$$\left|\mathbb{E}_\psi[R^2 f(U) \mid A]\right|^2 \le \mathbb{E}_\psi[R^2 |f(U)|^2 \mid A].$$

After integration over the bank, $R \perp\!\!\!\perp U$ and $\mathbb{E}R^2 = 1$ imply

$$\|T_k f\|^2_{L^2(\lambda_S^{\otimes k})} \le \mathbb{E}[R^2 |f(U)|^2] = \|f\|^2_{L^2(\lambda_S)}.$$

Thus T_k is a contraction, has a bounded adjoint, and maps constants to themselves. Moreover, $\mathbb{E}[Y_1 \mid A] = T_k h$ and $V_b = \mathrm{Var}(T_k h)$.

Constant and quadratic components. Since $\mathbb{E}_\psi \|B\|_F^2 = 1$ for every bank, $T_k \bar{h} = \bar{h}$. Also,

$$\|B\|_F^2 h_2(B/\|B\|_F) = \iota(B)^\top C \iota(B) = \sum_{i,j} \psi_i \psi_j a_i^\top C a_j.$$

Here $a_i = \iota(A_i) \in \mathbb{R}^m$. Using $\mathbb{E}[\psi\psi^\top] = I_k/k$ gives

$$T_k h_2 = \frac{1}{k} \sum_{i=1}^k h_2(A_i), \qquad \mathrm{Var}(T_k h_2) = \frac{\|h_2\|^2}{k} = \frac{\alpha \sigma_h^2}{k}.$$

The A_i are independent; mutual orthogonality of the bank is not required.

Orthogonality after averaging. Let $(\rho(Q)f)(u) = f(Q^{-1}u)$ and $(\pi(Q)F)(A_{1:k}) = F(Q^{-1}A_1, \ldots, Q^{-1}A_k)$ for $Q \in O(m)$. Norm preservation and linearity give $T_k \rho(Q) = \pi(Q) T_k$. Since these group actions are unitary, $T_k^* T_k$, acting on $L^2(\lambda_S)$, commutes with $\rho(Q)$. Distinct spherical-harmonic degrees are inequivalent irreducible $O(m)$-modules (Stein & Weiss, 1971, Chapter IV). Schur's lemma therefore yields

$$\langle T_k h_2, T_k r \rangle = \langle h_2, T_k^* T_k r \rangle = 0.$$

Since $\mathbb{E}T_k r = \mathbb{E}R^2\, \mathbb{E}r(U) = 0$, it follows that

$$V_b(k) = \frac{\alpha \sigma_h^2}{k} + V_{b,\ge 4}(k), \qquad V_{b,\ge 4}(k) := \mathrm{Var}(T_k r) \ge 0.$$

Remainder bound. The contraction established above gives

$$V_{b,\ge 4}(k) = \|T_k r\|^2 \le \|r\|^2 = (1 - \alpha)\sigma_h^2.$$

Together with the exact degree-two contribution, this proves (8). In particular, $V_b(k) \le \sigma_h^2$. The weighted Jensen step exploits $\mathbb{E}_\psi R^2 = 1$ for each bank; an unweighted application would give the weaker factor $1 + \varepsilon_k$ on the remainder.

Zero directional variance. If $\sigma_h^2 = 0$, then $h = \bar{h}$ almost everywhere, so $T_k h = \bar{h}$ almost surely. Hence $V_b = 0$ and $V_w = \varepsilon_k \bar{h}^2$; the exact variance identities remain valid without defining α. □

Remark 1 (Quadratic costs and the higher-degree remainder). When h is globally quadratic, $r = 0$ and $V_b(k) = \sigma_h^2/k$ exactly. This includes log-space translations and proportional-covariance Gaussian pairs. For other costs, the upper bound does not establish a decay rate in k for the higher-degree remainder. The parameter α measures the quadratic share of directional variance. The quantity kV_b/σ_h^2 also contains the higher-degree contribution and therefore need not equal α.

The translation case gives an explicit single-bank counterexample to an $(HL)^{-1/2}$ Monte Carlo rate. For log-space Gaussians with a common covariance and means differing by $\Delta \neq 0$, the directional cost is $h(u) = \langle u, \Delta\rangle^2$, whose variance over λ_S is

$$\sigma_h^2 = \|\Delta\|^4\Big(\frac{3}{m(m+2)} - \frac{1}{m^2}\Big) = \frac{2(m-1)\|\Delta\|^4}{m^2(m+2)} > 0.$$

Since h is quadratic, $V_b(k) = \sigma_h^2/k$, so for fixed k, with one sampled bank per estimator evaluation, $\mathrm{Var}(\widehat{D}_{k,L}) = \sigma_h^2/k + V_w(k)/L \to \sigma_h^2/k > 0$ as $L \to \infty$. Conditionally on a realized bank A the variance $\mathrm{Var}_\psi(\widehat{D}_{k,L} \mid A) = \mathrm{Var}_\psi(Y_1 \mid A)/L$ does vanish, but the conditional mean $T_k h(A)$ it converges to varies across banks, so the marginal mean absolute error does not. Averaging H independent banks gives $V_b/H + V_w/(HL)$, which approaches V_b/H as L grows at fixed H.

B.7 Proof of Theorem 3 (gradient estimators)

Directional differentiation for empirical measures. Work in Frobenius coordinates and write $\mu_S = \sum_i a_i \delta_{s_i}$ and $\nu_T = \sum_j b_j \delta_{t_j}$, with fixed positive weights summing to one. Then

$$h_S(u) = \min_{\pi \in \Pi(a,b)} \sum_{i,j} \pi_{ij} \langle s_i - t_j, u\rangle^2.$$

For a direction with no within-measure projection ties, the sorting order is locally constant in S. The monotone one-dimensional coupling therefore has locally fixed weights π_{ij}^u. Differentiating the quadratic cost gives (9). This argument uses the mass-splitting coupling, rather than a permutation, when the sample counts differ.

If the support points within each measure are pairwise distinct, the directions causing projection ties lie in a finite union of lower-dimensional great spheres. These have zero uniform measure. On a bounded neighborhood of S, the gradient obeys

$$\|(g_S(u))_i\|_2 \leq 2a_i \left(\|s_i\|_2 + \max_j \|t_j\|_2\right) \qquad (\|u\|_2 = 1).$$

This is a direction-independent local integrable bound. Differentiation under the expectation is therefore valid at the generic empirical states considered above, and $g_S \in L^2(\lambda_S; \mathbb{R}^{n_s \times m})$. The theorem explicitly assumes differentiability and interchangeability for other states or parameterizations.

Radial scaling and unbiasedness. For fixed sampled $B = RU$, the radius is independent of S and $h_S(B) = R^2 h_S(U)$. Hence $\nabla_S h_S(B) = R^2 g_S(U)$. Lemma 1 yields

$$\mathbb{E}G_1 = \mathbb{E}R^2 \mathbb{E}g_S(U) = \bar{g} = \nabla_S \mathrm{SPDSW}_2^2(\mu_S, \nu_T).$$

This statement concerns the log-coordinate gradient. An affine-invariant Riemannian gradient uses a different state-dependent metric conversion.

Trace-variance decomposition. Let $M_A = \mathbb{E}[G_1 \mid A]$. Conditional on A, the G_ℓ are independent. Expanding the squared norm around M_A and then averaging gives

$$\mathbb{E}\|\widehat{g}_{k,L} - \bar{g}\|_F^2 = \mathbb{E}\|M_A - \bar{g}\|_F^2 + \frac{1}{L}\mathbb{E}\|G_1 - M_A\|_F^2.$$

By radial independence,

$$\begin{aligned}\mathbb{E}\|G_1 - \bar{g}\|_F^2 &= \mathbb{E}R^4 \mathbb{E}\|g(U)\|_F^2 - \|\bar{g}\|_F^2 \\ &= (1+\varepsilon_k)\sigma_g^2 + \varepsilon_k \|\bar{g}\|_F^2.\end{aligned}$$

The law of total variance in the finite-dimensional inner-product space gives the second identity in (10).

Harmonic decomposition. Since $h_S(-u) = h_S(u)$, differentiability gives $g_S(-u) = g_S(u)$ almost everywhere. Expand every gradient coordinate into even spherical harmonics, writing $g = \bar{g} + g_2 + r_g$. Apply the scalar operator T_k to each coordinate. The coordinatewise degree-two identities and orthogonality from Appendix B.6 imply

$$V_b^g(k) = \frac{\|g_2\|_{L^2(F)}^2}{k} + V_{b,\geq 4}^g(k), \qquad 0 \leq V_{b,\geq 4}^g(k) \leq \|r_g\|_{L^2(F)}^2,$$

where $\|f\|_{L^2(F)}^2 = \mathbb{E}\|f(U)\|_F^2$. The contraction applies coordinatewise and the sum is finite. No assumption equating the harmonic compositions of h and g is made. □

B.8 A SCALAR–GRADIENT DISTINCTION IN A GAUSSIAN FAMILY

A Gaussian example makes the distinction explicit. Consider log-space Gaussian measures $\mathcal{N}(0, \Sigma)$ and $\mathcal{N}(0, \tau^2 I_m)$, parametrized by the source covariance Σ. Their directional cost is $h_\Sigma(u) = (\sqrt{u^\top \Sigma u} - \tau)^2$. At $\Sigma = s^2 I_m$, with $s, \tau > 0$ and $s \neq \tau$,

$$h_\Sigma(u) = (s - \tau)^2, \qquad \nabla_\Sigma h_\Sigma(u) = \left(1 - \frac{\tau}{s}\right) uu^\top. \tag{12}$$

The scalar directional variance is zero, but the gradient variance is $(1 - \tau/s)^2(1 - 1/m) > 0$ for $m > 1$. Correspondingly, $V_b^h = 0$ while $V_b^g = \sigma_g^2/k$. This analytical example concerns a distribution parameter; Theorem 3 is written for particles, and its same conditional-variance proof applies here. It demonstrates why a low scalar floor alone cannot certify a low gradient floor. The FPHA experiment measures both quantities for particle gradients at the same empirical state.

Calculation. For centered one-dimensional Gaussians with standard deviations a, b, the squared Wasserstein distance is $(a - b)^2$. Thus the projected Gaussians have cost $h_\Sigma(u) = (\sqrt{u^\top \Sigma u} - \tau)^2$, and direct differentiation yields

$$\nabla_\Sigma h_\Sigma(u) = \left(1 - \frac{\tau}{\sqrt{u^\top \Sigma u}}\right) uu^\top.$$

At $\Sigma = s^2 I_m$, write $a = 1 - \tau/s$. Then $\bar{g} = aI_m/m$, $\|uu^\top\|_F^2 = 1$, and

$$\sigma_g^2 = a^2 \mathbb{E}\|uu^\top - I_m/m\|_F^2 = a^2(1 - 1/m).$$

The centered matrix field $a(uu^\top - I_m/m)$ is degree two, so its exact shared-bank contribution is σ_g^2/k. The scalar field is constant and has zero between-bank variance. Its unnormalized single-slice cost can still fluctuate radially, with $V_w^h = \varepsilon_k(s - \tau)^4$. The example distinguishes directional and radial variability and does not require a finite-sample Gaussian approximation.

B.9 CONDITIONAL CONNECTION TO STOCHASTIC DESCENT

Let $F(S) = \mathbb{E}_U h_S(U)$. Suppose its gradient is ℓ-Lipschitz on a convex region containing S and every sampled update $S^+ = S - \eta \widehat{g}_{k,L}$. With the unbiasedness assumptions of Theorem 3, the descent lemma and conditional second-moment decomposition give

$$\begin{aligned}\mathbb{E}[F(S^+) \mid S] &\leq F(S) - \eta \langle \nabla F(S), \mathbb{E}[\widehat{g}_{k,L} \mid S] \rangle + \frac{\ell \eta^2}{2} \mathbb{E}[\|\widehat{g}_{k,L}\|_F^2 \mid S] \\ &= F(S) - \eta \left(1 - \frac{\ell \eta}{2}\right) \|\nabla F(S)\|_F^2 + \frac{\ell \eta^2}{2} \left(V_b^g(k; S) + \frac{V_w^g(k; S)}{L}\right).\end{aligned}$$

This standard conditional bound identifies where gradient trace variance enters an optimization analysis. The empirical sliced cost is piecewise quadratic for fixed sorting orders; uniform smoothness along an entire stochastic trajectory is an additional condition, not established by the fixed-state experiments. Neither the scalar variance nor a single-state gradient estimate determines a globally stable learning rate.

B.10 PROOF OF COROLLARY 1

For a scalar or finite-dimensional vector-valued directional quantity f, both estimators are unbiased. The flat k-direction average has mean-squared error σ_f^2/k. Comparing it with $V_b^f + V_w^f/L$ gives (11) by algebra. If the centered f is purely degree two, $V_b^f = \sigma_f^2/k$, excluding a strict improvement for finite L. When $V_b^f > 0$, taking $L \to \infty$ gives the limiting RMS-error ratio $\sqrt{(\sigma_f^2/k)/V_b^f} = \rho_f^{-1/2}$. □

B.11 LOCAL UPDATE GEOMETRY

The gradient theory characterizes stochastic estimation, not a global convergence rate. A complementary local calculation relates projection geometry to step size. In a fixed sorting chamber, let

$Q_B = L^{-1}\sum_\ell b_\ell b_\ell^\top$, with $b_\ell = \iota(B_\ell)$. For uniform source weights, the source-particle Hessian block is $2Q_B/n_s$. Thus local curvature depends on $\lambda_{\max}(Q_B)$ as well as the loss normalization. Both small-budget flat and hierarchical estimators can concentrate curvature in a low-rank span. Sorting changes and resampling prevent this local calculation from serving as a global stability guarantee.

C Computational Cost and Budget Selection

C.1 Cost components

For two empirical measures with at most n samples each, the costs after computing matrix logarithms are

$$T^{\mathrm{SPDSW}} = O(Ld^2 + Lnm + Ln\log n),$$
$$T^{\mathrm{SPDHSW}} = O(kd^2 + Lk + knm + Lkn + Ln\log n).$$

The terms correspond to direction sampling, matrix projection, hierarchical mixing, and one-dimensional sorting. Frobenius-uniform sampling costs $O(d^2)$ per matrix direction; a spectral construction based on an orthogonal factorization costs $O(d^3)$. Matrix logarithms add $O(nd^3)$ to both methods when they are not already available. Direction storage changes from $O(Lm)$ to $O(km + kL)$.

At equal final budget L, both methods solve L one-dimensional problems. At equal matrix-projection count, SPDSW uses k independent directions while SPDHSW also pays for its L mixtures and sorting operations. Under common unit costs, the projection-work ratio is

$$\frac{knm + Lkn}{Lnm} = \frac{k}{L} + \frac{k}{m}. \tag{13}$$

The ratios are 0.238 for $(k, L, m) = (40, 500, 253)$ and 0.096 for $(50, 700, 2016)$. Actual runtime additionally depends on the relative cost of kernels, sampling, and sorting.

Break-even dimension. With $n = \max(n_s, n_t)$ and available logarithms, a cost model including projection, mixing, and sorting is

$$\begin{aligned} C_{\mathrm{flat}}(N) &= n\big(c_p Nm + c_s N\log n\big), \\ C_{\mathrm{hier}}(k, L) &= n\big(c_p km + c_m kL + c_s L\log n\big), \end{aligned} \tag{14}$$

where the positive coefficients represent operation costs. This is the two-stage cost accounting of Nguyen et al. (2023b) in symmetric dimension m. For $N > k$, hierarchical computation is cheaper in this model when

$$m > m_\star := \frac{c_m kL + c_s(L - N)\log n}{c_p(N - k)}. \tag{15}$$

At $N = L$, both methods sort the same number of columns, and $m_\star = (c_m/c_p)kL/(L - k)$. At $N = k$ with $L \geq k$, mixing adds work; it must earn that work through a better estimate. Equation (15) is a computation threshold; precision is assessed by Corollary 1, and complete optimization by the experiments.

C.2 Precision-constrained choice with a flat alternative

For either scalar or gradient estimation, consider a target mean-squared error $v^\star > 0$. Under known directional variances, the minimum feasible budgets are

$$\begin{aligned} N_{\min}(v^\star) &= \max\left\{1, \left\lceil \frac{\sigma_f^2}{v^\star} \right\rceil\right\}, \\ L_{\min}(k; v^\star) &= \max\left\{1, \left\lceil \frac{V_w^f(k)}{v^\star - V_b^f(k)} \right\rceil\right\}, \qquad V_b^f(k) < v^\star. \end{aligned} \tag{16}$$

For nondecreasing costs in the final-direction count, choose the cheaper of

$$C_{\mathrm{flat}}\big(N_{\min}(v^\star)\big) \quad \text{and} \quad \min_{k: V_b^f(k) < v^\star} C_{\mathrm{hier}}\big(k, L_{\min}(k; v^\star)\big).$$

At the boundary $V_b^f(k) = v^\star$, feasibility requires $V_w^f(k) = 0$, in which case $L = 1$ suffices. The candidates can have arbitrary positive k, L; the condition $L \geq k$ in the synthetic sweep is only that experiment's grid restriction. Mixing is selected only if it is both statistically feasible and computationally preferable to independent directions.

Estimating the components. With J independent banks and M independent mixtures in each bank, let $\bar{Z}_j$ be the within-bank mean and s_j^2 its sample variance (or trace variance). Then

$$\widehat{V}_w = \frac{1}{J}\sum_j s_j^2, \qquad \widehat{V}_b = \frac{1}{J-1}\sum_j \|\bar{Z}_j - \bar{Z}\|^2 - \frac{\widehat{V}_w}{M}.$$

For a scalar, $\|\cdot\|$ is absolute value. For a gradient, it is the Frobenius norm over the full source gradient. The finite-M correction is necessary, and untruncated estimates preserve the diagnostic when sampling noise produces a negative estimate. Independent flat draws estimate σ_f^2; resampling banks and flat blocks propagates both sources of uncertainty.

Pilot validation. The earlier scalar pilot used $J = 100$, $M = 500$, 100,000 flat directions, and 5,000 cluster-bootstrap replicates. For $v_N = \sigma_h^2/N$, upper pilot endpoints U_b, U_w were used in the constraint $U_b + U_w/L \leq v_N$. The hierarchical-only candidate cost was $C(k, L) = c_p km + c_m kL + c_s L \log n$. Validation used 500 fresh draws and a one-sided 95% upper variance bound (Table 19). This pilot tested feasibility within the hierarchical family, not superiority over flat sampling. Indeed, every listed choice has both $k > N$ and $L > N$, so flat N has lower nominal cost under the same positive-coefficient model. The selector above explicitly includes that alternative. In an uncertainty-aware implementation, both families must use the same variance-confidence criterion; an estimated pilot bound is not a deterministic guarantee.

D Experimental Protocols

D.1 BNCI cross-subject adaptation

BNCI2014-001 contains four-class motor-imagery recordings from 22 EEG channels at 250 Hz; we use subjects $\{1, 3, 7, 8, 9\}$ following Bonet et al. (2023). BNCI2015-001 contains two-class recordings from 13 channels at 512 Hz; all 12 subjects are used (Tangermann et al., 2012; Faller et al., 2012). Both datasets are processed through the MOABB `MotorImagery` paradigm (Jayaram & Barachant, 2018): an 8–32 Hz band-pass is applied, trials are epoched from $t_{\min} = 0$ relative to the task interval, and sample covariance matrices are computed from the filtered epochs.

Each subject serves as target in turn, with every other subject treated as a separate source. This gives 20 directed pairs for BNCI2014-001 and 132 for BNCI2015-001. For each pair, only source particles are optimized. The aligned source covariances retain their class labels; a classifier fitted to these source data is evaluated on the target. The no-alignment baseline uses the same classification pipeline without optimizing the source matrices. Accuracy is first averaged over sources for each target, then over targets, and finally summarized across five seeds.

All alignment runs use 500 iterations. SPDSW uses $L = 500$; SPDHSW uses $(k, L) = (40, 500)$. In the resampled regime both bottleneck and mixing directions are redrawn at every iteration; in the fixed regime the directions are retained. Table 4 uses Log-Euclidean updates. Affine-invariant classification results, including full-LEW training, are given in Appendix E.5.

D.2 High-Gamma cross-session adaptation

The High-Gamma dataset contains four-class recordings from 14 subjects (Schirrmeister et al., 2017). All 128 channels are retained, yielding $m = 8256$. Each subject's training session is the source and that subject's test session is the target. Only the source covariance matrices are updated; target log-coordinates are fixed. The two methods use 500 Log-Euclidean gradient steps, learning rate 10^4, and the same projection budgets as the BNCI experiments. The five random seeds are 6398, 3654, 1788, 8515, and 264.

The classifier is a linear support vector machine trained on $\mathrm{vec}(\log X) \in \mathbb{R}^{d^2}$. Its penalty C is selected by three-fold cross-validation on the source data from 10 logarithmically spaced values

between 10^{-2} and 10^{2}. Target class labels are used only for evaluation. All 140 alignment runs and their classifier evaluations are included.

For seed r, accuracy is averaged over the 14 subjects. Table 4 reports the mean and sample standard deviation of these five seed-level means. Runtime ratios are paired by seed and subject; the reported factor is the median over seeds of the subject-median ratios. The alignment-loop timer excludes initial log transforms, final matrix exponentials, and downstream classifier fitting. Runs were performed on a shared NVIDIA RTX A6000. Timing comparisons use matching tasks and seeds; method execution was grouped within each seed.

D.3 Fixed-pair distribution alignment

The fixed-pair experiment has no downstream classifier. BNCI2014-001 session 1 covariances are pooled over subjects $\{1, 3, 7, 8, 9\}$ to form the source distribution, and session 2 covariances from the same subjects form the target. The matrices have size 22×22. This pooled-session comparison is separate from the 20 directed cross-subject classification pairs.

For FPHA (Garcia-Hernando et al., 2018), each sequence is represented by the covariance of the 63-dimensional vector obtained by stacking the 3D positions of 21 hand joints. We follow the train/validation split of Wang et al. (2022), with 600 source and 575 target matrices. All 63 feature dimensions are retained, giving $m = 2016$.

The configurations in Table 20 use 500 affine-invariant gradient updates. Resampled runs draw fresh directions at every iteration; fixed-direction controls retain the initial projection draw. The hierarchical budgets are $(40, 500)$ for BNCI and $(50, 700)$ for FPHA. Flat slicing is evaluated at the same final budget and at $L = k$; intermediate flat budgets and fixed-direction results are reported in Table 20. LEW alignment scores are evaluated independently of the sliced training loss, with the same evaluator and scale within each distribution-pair comparison.

Log-Euclidean FPHA runs. The Log-Euclidean FPHA experiment (Figure 2 and Table 2) uses the same 600 source and 575 target matrices. Source log-coordinates are updated by plain SGD in float64 for 500 updates at the shared rate $\eta = 2.5 \times 10^{4}$, without momentum or weight decay, with symmetrization after every step; the target log-coordinates are fixed. The matrix logarithm is regularized once at initialization (eigenvalue floor 10^{-4}). Directions, and for SPDHSW the mixing vectors, are redrawn at every update, and the effective mixed directions are not renormalized. The alignment score is $0.1\, W_2^2$ between the empirical source and target log-coordinate distributions with uniform masses $1/600$ and $1/575$, computed by exact transport on the squared Frobenius ground cost; no square root or dimension normalization is applied. It is evaluated at initialization and every 25 updates; the final value is the score at exactly update 500. Loop time sums each full update between CUDA synchronizations (direction generation, projection, mixing, one-dimensional transport, backward pass, optimizer step, symmetrization) and excludes data loading, the initial logarithm, evaluation, and checkpointing. The five seeds are 2473, 7490, 7066, 1884, and 929; runs use an NVIDIA RTX 6000 Ada Generation GPU. Reported deviations are sample standard deviations over seeds.

D.4 Gradient-free estimation

The estimand is SPDSW_2^2 for a fixed empirical pair. Samples are generated in symmetric log-space from two rank-eight spiked-covariance families: one uses fixed spike strength and the other dimension-scaled spikes. Source and target covariances are non-proportional. There are three independent seeds per family at $d \in \{22, 63, 160\}$, giving 18 pairs with 600 samples per measure.

The flat budget is $L \in \{20, 40, 50, 100, 200, 500, 700, 1000\}$. The hierarchical grid uses $k \in \{20, 40, 50, 100\}$ and $L \in \{50, 100, 200, 500, 700, 1000, 2000\}$ with $L \geq k$. Additional $L = 1$ draws check single-slice moments. Each configuration uses 200–800 repetitions and 5,000 bootstrap replicates. Comparisons use common random numbers where applicable. A difference is classified as resolved when the paired bootstrap comparison separates the relative RMSEs.

The independent reference averages 200,000 flat directions. Its block standard error is at most 8.1% of the smallest compared RMSE on each pair. The full grid and analysis rules were fixed before measurement. For the matched-work comparison, the primary budget includes projection, mixing,

and sorting; interleaved timing pairs provide an additional wall-clock comparison. Tables 6–9 report the complete per-dimension summaries and pair-level work results.

D.5 Dimension sweep and update parameterization

The dimension sweep uses log-space pairs $S = (G + G^\top)s/2 + cI$, with source parameters $(s, c) = (0.30, 0.00)$ and target parameters $(0.42, 0.15)$. Each measure has 600 samples. Dimensions are $d \in \{22, 63, 96, 128, 160\}$, the final budget is $L = 700$, and $k \in \{20, 50, 100\}$. Matrix logarithms are available before timing. After 20 warm-up calls, we report medians over 100 float64 repetitions, with synchronized timers for sampling, projection, mixing, and sorting.

For Log-Euclidean updates, write the loss as $\mathcal{L}(X) = \widetilde{\mathcal{L}}(\log X)$ and carry particles as $S = \log X$. The update is

$$S_{t+1} = S_t - \eta \nabla_S \widetilde{\mathcal{L}}(S_t), \qquad X_{t+1} = \exp S_{t+1}.$$

Because $\log$ is a Log-Euclidean isometry, the matrix exponential is only needed when an SPD output is required. In contrast, the affine-invariant implementation applies Riemannian gradient conversion $G \mapsto X\,\mathrm{sym}(G)X$ and a corresponding retraction to the SPD particles (Kochurov et al., 2020). Because the particles are SPD matrices, the source logarithms are recomputed at every iteration and the backward pass runs through them; the logarithm stage, not the backward pass, dominates the AIM iteration in Table 1. The profiling comparison in Table 1 keeps the sliced forward computation unchanged and changes the update parameterization.

D.6 SPD representation-learning experiment

The additional representation-learning experiment uses SPDNet (Huang & Van Gool, 2017): BiMap, ReEig, LogEig, and a linear classifier. A sliced Fisher-ratio regularizer is evaluated on the SPD features after ReEig and before LogEig. For class subsets B_c in a minibatch, let C_B be the classes present, $C_w = \{c : |B_c| \geq 2\}$, and $\widehat{\mu}_c$ the empirical distribution of class-c features. With D equal to SPDSW or SPDHSW at $p = 2$,

$$\mathcal{L}_w = \frac{1}{|C_w|} \sum_{c \in C_w} \frac{2}{|B_c|(|B_c| - 1)} \sum_{\substack{i<j \\ i,j \in B_c}} D(\delta_{Z_i}, \delta_{Z_j})^2,$$

$$\mathcal{L}_b = \frac{2}{|C_B|(|C_B| - 1)} \sum_{c<c'} D(\widehat{\mu}_c, \widehat{\mu}_{c'})^2,$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_w / \mathcal{L}_b.$$

The displayed ratio assumes $|C_B| \geq 2$, $|C_w| > 0$, and $\mathcal{L}_b > 0$. AFEW descriptors are 400×400 covariances of vectorized 20×20 grayscale video frames (Dhall et al., 2012); the BiMap output dimension is 50. Training uses 200 epochs, SGD with learning rate 0.01, and batch size 30. FPHA uses the $600/575$ sequence split above, BiMap output dimension 33, 500 epochs, SGD with learning rate 0.05, and batch size 128. SPDSW uses $L = 1000$ and SPDHSW uses $(k, L) = (100, 1000)$, with fresh directions at every step. Results are averaged over five seeds.

E Additional Experimental Results

E.1 Estimation accuracy at different dimensions

Tables 6–8 compare errors at equal final projection counts and at equal matrix-projection counts. W/L/U denotes lower/higher/unresolved hierarchical RMSE over all tested mixing budgets. The final-budget table distinguishes each method's maximum evaluated budget. Table 9 reports the lowest feasible work at a 5% relative-RMSE target on the evaluated grid.

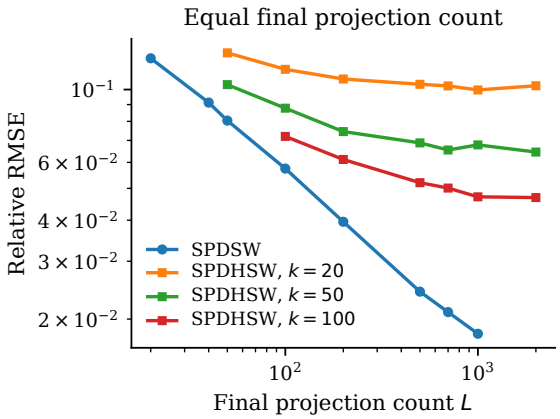


Figure 4: Equal final projection counts on the six $d = 63$ synthetic pairs. Flat SPDSW uses independent directions; each fixed-k hierarchy approaches its between-bank floor as L increases.

Table 6: Gradient-free relative RMSE against the 2×10^5-direction reference at $d = 22$ ($m = 253$; six pairs, two families $\times$ three seeds). (a) Equal final budgets $L = 100, 500$, followed by each method's largest evaluated budget (1000 for SPDSW, 2000 for SPDHSW). (b) Equal matrix-projection count: RMSE uses SPDHSW $(k, 2000)$; W/L/U counts aggregate all tested mixing budgets.

(a) Matched final budget L

	SPDSW	SPDHSW k=20	SPDHSW k=50	SPDHSW k=100
$L = 100$	0.062	0.128	0.092	0.080
$L = 500$	0.028	0.107	0.067	0.054
Maximum tested L	0.020	0.103	0.066	0.046

(b) Matched matrix-projection count

Projections	SPDSW (L=k)	SPDHSW	Ratio	W/L/U
$k = 20$	0.1457	0.1029	1.42	35/0/7
$k = 40$	0.1007	0.0727	1.39	32/4/6
$k = 50$	0.0940	0.0657	1.43	27/3/12
$k = 100$	0.0624	0.0463	1.35	24/7/5

Table 7: Gradient-free relative RMSE against the 2×10^5-direction reference at $d = 63$ ($m = 2016$; six pairs, two families $\times$ three seeds). (a) Equal final budgets $L = 100, 500$, followed by each method's largest evaluated budget (1000 for SPDSW, 2000 for SPDHSW). (b) Equal matrix-projection count: RMSE uses SPDHSW $(k, 2000)$; W/L/U counts aggregate all tested mixing budgets.

(a) Matched final budget L

	SPDSW	SPDHSW k=20	SPDHSW k=50	SPDHSW k=100
$L = 100$	0.057	0.115	0.088	0.072
$L = 500$	0.024	0.104	0.069	0.052
Maximum tested L	0.018	0.103	0.065	0.047

(b) Matched matrix-projection count

Projections	SPDSW (L=k)	SPDHSW	Ratio	W/L/U
$k = 20$	0.1245	0.1027	1.21	30/0/12
$k = 40$	0.0913	0.0744	1.23	25/6/11
$k = 50$	0.0804	0.0645	1.25	24/6/12
$k = 100$	0.0575	0.0469	1.23	23/9/4

Table 8: Gradient-free relative RMSE against the 2×10^5-direction reference at $d = 160$ ($m = 12880$; six pairs, two families $\times$ three seeds). (a) Equal final budgets $L = 100, 500$, followed by each method's largest evaluated budget (1000 for SPDSW, 2000 for SPDHSW). (b) Equal matrix-projection count: RMSE uses SPDHSW $(k, 2000)$; W/L/U counts aggregate all tested mixing budgets.

(a) Matched final budget L

	SPDSW	SPDHSW k=20	SPDHSW k=50	SPDHSW k=100
$L = 100$	0.056	0.120	0.085	0.071
$L = 500$	0.026	0.107	0.069	0.051
Maximum tested L	0.018	0.103	0.067	0.048

(b) Matched matrix-projection count

Projections	SPDSW (L=k)	SPDHSW	Ratio	W/L/U
$k = 20$	0.1289	0.1035	1.25	32/0/10
$k = 40$	0.0903	0.0752	1.20	28/4/10
$k = 50$	0.0814	0.0666	1.22	24/6/12
$k = 100$	0.0558	0.0475	1.17	21/7/8

Table 9: Matched total work at a 5% relative-RMSE target, per pair: projection, mixing, and sorting cost model in Appendix C; ratio is SPDSW work divided by the best feasible SPDHSW work on the evaluated grid ("–": no single-call SPDHSW configuration on the grid reaches the target). Paired wall-clock ratios (10th percentile over interleaved blocks) provide an additional timing comparison.

d	Pair (family, seed)	Work ratio 5%	Wall-clock ratio	Feasible at 1–2%
22	fixed, 0	0.29	0.46	no
22	fixed, 1	0.29	0.46	no
22	fixed, 2	0.29	0.46	no
22	scaled, 0	0.12	0.33	no
22	scaled, 1	–	–	no
22	scaled, 2	0.23	0.39	no
63	fixed, 0	–	–	no
63	fixed, 1	0.29	0.59	no
63	fixed, 2	–	–	no
63	scaled, 0	0.52	0.68	no
63	scaled, 1	–	–	no
63	scaled, 2	0.52	0.69	no
160	fixed, 0	–	–	no
160	fixed, 1	–	–	no
160	fixed, 2	–	–	no
160	scaled, 0	–	–	no
160	scaled, 1	1.44	1.49	no
160	scaled, 2	0.91	1.40	no

E.2 STAGE-LEVEL COMPUTATIONAL COST

Table 10: Stage decomposition of one estimator call (ms; $n = 600$, $L = 700$, float64, medians of 100 repetitions after 20 warm-ups). The total is the sum of the four stage medians.

d	m	Method	k / flat L	sampling	projection	mixing	sorting	total
22	253	SPDSW	700	0.198	2.355	–	1.071	3.624
22	253	SPDHSW	20	0.146	0.173	0.188	1.166	1.673
22	253	SPDHSW	50	0.161	0.246	0.383	1.151	1.942
22	253	SPDHSW	100	0.158	0.696	0.662	1.127	2.644
63	2016	SPDSW	700	0.819	14.430	–	1.223	16.471
63	2016	SPDHSW	20	0.234	0.837	0.253	1.257	2.582
63	2016	SPDHSW	50	0.249	2.299	0.507	1.161	4.215
63	2016	SPDHSW	100	0.239	2.900	0.653	1.125	4.917
96	4656	SPDSW	700	1.603	33.427	–	1.092	36.122
96	4656	SPDHSW	20	0.241	1.721	0.253	1.244	3.460
96	4656	SPDHSW	50	0.262	4.833	0.462	1.100	6.657
96	4656	SPDHSW	100	0.352	5.803	0.576	1.039	7.770
128	8256	SPDSW	700	3.267	59.170	–	1.133	63.570
128	8256	SPDHSW	20	0.233	3.006	0.231	1.194	4.663
128	8256	SPDHSW	50	0.213	8.861	0.388	1.012	10.474
128	8256	SPDHSW	100	0.594	10.759	0.358	1.009	12.720
160	12880	SPDSW	700	5.264	97.559	–	1.097	103.920
160	12880	SPDHSW	20	0.256	4.309	0.228	1.261	6.055
160	12880	SPDHSW	50	0.498	13.601	0.531	1.186	15.815
160	12880	SPDHSW	100	0.806	17.606	0.602	1.136	20.150

Sampling, projection, mixing, and sorting times are given in Table 10. With n and L fixed, matrix-projection time grows strongly with m, whereas the sorting term remains close to 1 ms in this sweep. The measured projection-plus-mixing ratio follows the trend of $k/L + k/m$ but is larger than the unit-cost prediction. This is consistent with different kernel efficiencies at narrow and wide projection outputs.

The per-iteration profile of one High-Gamma alignment step under affine-invariant and Log-Euclidean updates is Table 1 in the main text; the profiling comparison keeps the sliced forward computation unchanged and changes only the update parameterization.

E.3 NORMALIZED HIERARCHICAL PROJECTIONS

We compare normalized and unnormalized projections at $(k, L) = (50, 700)$ for $p \in \{1, 2\}$. The synthetic log-Gaussian pairs have dimensions $d = 5, 22$; the real FPHA pair has $d = 63$. At $p = 1$, the unnormalized comparator is divided by an independently estimated $\mathbb{E}R$ to remove the population scale difference. These constants are 0.9843, 0.9991, and 0.9999, respectively.

Table 11: Relative RMSE reduction from normalizing the effective direction. Positive values favor normalization.

Pair	$\mathrm{Var}(R^2)$	$p = 1$	$p = 2$
Synthetic, $d = 5$	0.1256	5.36%	10.77%
Synthetic, $d = 22$	0.00745	0.57%	0.94%
FPHA, $d = 63$	0.000935	0.04%	0.129%

The benefit is largest at $d = 5$ (Holm-adjusted $p = 0.0012$ for both orders) and small at the larger dimensions. Normalization removes the radial factor from each slice but not dependence through the shared bottleneck. For $p \neq 2$, it also retains the SPDSW population scale exactly.

Table 12: SPDHSW bottleneck ablation at $L = 500$ (accuracy %, mean $\pm$ s.d. over 5 seeds; time in s). The update is affine-invariant. Resampled and fixed directions are compared using the same downstream protocol.

		$k = 20$		$k = 40$		$k = 60$		$k = 80$	
Dataset	Sampling	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time
BNCI2014-001	Resampled	57.06±0.26	3.51	57.07±0.26	3.52	57.07±0.25	3.50	57.16±0.03	3.52
	Fixed	39.75±0.50	3.28	40.58±1.65	3.34	41.50±1.12	3.34	42.66±0.64	3.32
BNCI2015-001	Resampled	64.70±0.10	2.21	64.72±0.04	2.30	64.79±0.10	2.31	64.73±0.05	2.33
	Fixed	57.72±0.21	2.01	58.72±1.00	2.03	60.67±0.80	2.02	62.50±0.63	2.01

E.4 Bottleneck size and resampling

With fixed directions, all final projections remain in the span of the same k bottleneck directions. Accuracy increases with k in this regime, while resampled accuracies vary little over the tested budgets. This is an empirical comparison of finite optimization runs, separately from the population identity.

E.5 BNCI classification under affine-invariant updates

The results below use affine-invariant updates and therefore form a separate comparison from Table 4. Full LEW is an alignment training objective in this table. Per-source accuracies average over the remaining target subjects and then over five seeds.

Table 13: BNCI classification results under the affine-invariant update. Sliced methods resample directions with $L = 500$ and SPDHSW uses $k = 40$. Values are accuracy (%), mean $\pm$ standard deviation over five seeds; full-LEW alignment has no projection-sampling randomness.

Alignment loss	BNCI2014-001	BNCI2015-001
SPDSW	57.16 ± 0.17	64.71 ± 0.08
SPDHSW	57.07 ± 0.26	64.72 ± 0.04
Full LEW	60.16	65.46

Table 14: BNCI2014-001, $k = 40$, $L = 500$, accuracy averaged over the four other target subjects and 5 seeds. Top: resampled. Bottom: fixed (LEW is deterministic).

Src	No Al.	SPDHSW	SPDSW	LogHSW	LogSW
1	41.84	57.44±0.27	57.80±0.36	54.24±0.40	54.20±0.10
3	37.02	63.36±0.36	63.59±0.30	57.88±0.31	57.92±0.49
7	43.06	59.99±0.28	60.04±0.04	55.33±0.32	55.20±0.19
8	37.25	52.14±0.83	52.01±0.40	48.55±0.39	48.46±0.31
9	32.29	52.43±0.26	52.36±0.15	41.70±0.40	41.56±0.12

Src	SPDHSW	SPDSW	LogHSW	LogSW	LEW
1	42.99±2.76	53.71±0.92	43.09±2.53	49.23±1.79	58.85
3	41.30±1.54	59.78±1.06	41.13±1.49	54.17±1.33	63.41
7	45.45±3.19	57.32±0.33	45.40±2.89	53.13±0.55	59.42
8	39.72±2.58	47.02±1.27	39.85±2.26	43.99±1.79	62.46
9	33.43±2.02	45.56±1.90	33.26±1.97	38.86±2.08	56.64

Table 15: BNCI2015-001, $k = 40$, $L = 500$, accuracy averaged over the eleven other target subjects and 5 seeds. Top: resampled. Bottom: fixed (LEW is deterministic).

Src	No Al.	SPDHSW	SPDSW	LogHSW	LogSW
1	50.45	67.26±0.27	67.22±0.09	66.25±0.40	66.15±0.11
2	57.80	69.64±0.32	69.69±0.11	69.65±0.22	70.02±0.15
3	54.26	68.59±0.22	68.53±0.15	68.07±0.13	68.06±0.05
4	58.26	67.73±0.22	67.72±0.22	67.92±0.25	67.85±0.08
5	61.55	68.12±0.24	68.30±0.13	69.43±0.30	69.37±0.06
6	58.85	61.59±0.45	61.51±0.20	63.50±0.24	63.39±0.19
7	51.67	64.21±0.31	64.15±0.06	64.86±0.20	64.89±0.06
8	55.86	64.17±0.49	64.15±0.32	64.65±0.64	64.20±0.14
9	58.05	66.92±0.20	67.05±0.24	67.12±0.34	67.00±0.25
10	59.26	68.55±0.58	68.18±0.62	69.90±0.17	70.08±0.16
11	56.78	57.56±0.22	57.78±0.16	58.18±0.19	58.20±0.15
12	47.29	52.29±0.28	52.28±0.13	50.53±0.34	50.74±0.51

Src	SPDHSW	SPDSW	LogHSW	LogSW	LEW
1	54.70±2.65	66.36±0.60	56.56±4.59	65.39±0.82	69.78
2	60.80±0.93	69.63±0.31	62.17±0.84	69.39±0.38	70.39
3	58.91±2.25	68.46±0.36	58.31±1.61	67.62±0.67	69.56
4	59.01±1.26	68.00±0.33	60.92±1.13	67.77±0.21	69.85
5	60.58±3.99	69.29±0.63	65.35±1.05	69.69±0.40	69.27
6	61.21±0.94	63.70±0.69	60.79±0.84	64.73±0.89	64.81
7	60.14±3.27	64.29±0.69	58.68±2.12	64.54±0.61	65.72
8	57.85±1.08	64.97±1.38	58.36±1.60	64.67±1.02	64.30
9	61.70±2.11	66.91±0.60	62.19±1.64	67.36±0.24	70.74
10	64.86±2.36	69.94±0.79	64.01±1.10	69.61±0.34	63.64
11	56.22±0.44	57.43±1.74	55.34±0.98	57.89±1.37	52.98
12	48.70±1.80	50.61±0.74	50.08±1.08	49.99±0.80	54.48

E.6 Comparison of update geometries

A paired BNCI2014-001 comparison uses the same five subjects, directed source–target pairs, and five seeds for both update geometries. Affine-invariant updates yield 57.06% for SPDSW and 57.07% for SPDHSW, while Log-Euclidean updates yield 59.00% and 58.87%. The pooled change is $+1.87$ percentage points, with a bootstrap 95% interval of $[1.56, 2.17]$. These are separately paired runs rather than a cross-table subtraction.

For BNCI2015-001, the Log-Euclidean accuracies are 64.56% and 64.61%. A paired update comparison at seed 6398 gives a change of -0.08 points, with interval $[-0.40, 0.25]$. Thus the benefit of changing update geometry differs between these datasets.

E.7 Common-protocol EEG comparison

The common-protocol experiment uses 500 Log-Euclidean updates, $L = 500$, $k = 40$, and five projection seeds. BNCI2014-001 and BNCI2015-001 are cross-subject tasks; High-Gamma is within-subject cross-session adaptation. Table 4 reports the original common-protocol results. These runs are distinct from the configuration-selection study in Section 5.2, including the hardware used for timing. The fixed-budget ratio is not a tuned-estimator or full-classification-pipeline speed-up. Table 4 in the main text summarizes this common-protocol comparison. The following subsections give its per-seed and target-attainment results.

E.8 High-Gamma: seed-level accuracy and runtime

Each seed contributes all 14 subjects. The $6.03\times$ aggregate time ratio is the median of the seed-wise subject-median ratios, rather than the ratio of the two aggregate mean times. Standard deviations in Table 4 summarize variability across seed-level mean accuracies.

Table 16: High-Gamma cross-session alignment, per seed (14 subjects each). Times are 500-iteration optimization-loop means over subjects (s); ratio is the subject-median SPDSW/SPDHSW runtime ratio; accuracies in %; matched quality: iterations for SPDHSW to reach the SPDSW 500-iteration LEW (subject median) and the estimated time ratio using the measured iteration cost.

	Runtime			Accuracy			Matched quality	
Seed	SPDSW	SPDHSW	Ratio	Unaligned	SPDSW	SPDHSW	Iterations	Est. ratio
6398	31.70	5.21	6.03	77.90	79.38	78.35	614	4.94
3654	31.75	5.23	6.03	77.90	79.06	79.87	612	4.92
1788	32.75	5.23	6.17	77.90	78.97	79.82	616	5.05
8515	31.70	5.21	6.03	77.90	78.84	79.20	611	4.98
264	31.77	5.22	6.02	77.90	78.88	79.33	609	4.99

E.9 High-Gamma: target quality and learning-rate control

For $\mu = n_s^{-1}\sum_i \delta_{X_i}$ and $\nu = n_t^{-1}\sum_j \delta_{Z_j}$, set $a_i = 1/n_s$ and $b_j = 1/n_t$. Let $\Pi(a, b)$ denote nonnegative coupling matrices with row sums a and column sums b. We evaluate the squared Log-Euclidean Wasserstein cost

$$Q(\mu, \nu) = W^2_{2,\mathrm{LE}}(\mu, \nu) = \min_{\pi \in \Pi(a,b)} \sum_{i,j} \pi_{ij} \| \log X_i - \log Z_j \|_F^2.$$

The evaluator solves discrete OT on the squared ground-cost matrix and applies neither a square root nor a display multiplier. It is separate from the sliced training loss.

At the common learning rate 10^4, the target is the SPDSW cost after 500 iterations for each seed–subject pair. SPDHSW is evaluated every 50 iterations through iteration 1500. All 70 pairs reach the target by a recorded checkpoint at or before iteration 1000. Linear interpolation between adjacent checkpoints estimates the crossing iteration $\widehat{e}_{r,s}$.

The corresponding time estimate is

$$\widehat{T}^{\mathrm{HS}}_{r,s} = T^{\mathrm{HS}}_{500,r,s} \frac{\widehat{e}_{r,s}}{500}.$$

Aggregating paired ratios gives $4.98\times$, at a median of approximately 612 iterations. This estimate assumes that the measured iteration cost remains constant. Using the first observed crossing checkpoint instead of interpolation gives approximately $4.66\times$ under the same time model. Full-OT checkpoint evaluation is excluded from the production loop timer.

At method-specific rates 10^5 for SPDSW and 10^4 for SPDHSW, the target is the higher-rate SPDSW cost after 500 iterations. These targets were evaluated from the saved aligned matrices and compared with the same 70 SPDHSW trajectories. No target crossing occurs at any recorded checkpoint through iteration 1500. The median final relative cost excess is 5.33%. The original five-seed control evaluates alignment quality only. The later high-rate classifier check at seed 6398 is reported in Appendix G.

Table 17: High-Gamma target-attainment control. The target is the corresponding SPDSW 500-iteration cost. Attainment is checked at intervals of 50 iterations. Time ratios use the constant-iteration-cost model.

Learning rates (SPDSW / SPDHSW)	Attained by 1500	Crossing iteration	Time ratio
10^4 / 10^4	70/70	≈ 612	$\approx 4.98\times$
10^5 / 10^4	0/70	> 1500	upper bound $\approx 1.994\times$

For the second row, the bound is computed pairwise as $T^{\mathrm{SPDSW}}_{500}/(3T^{\mathrm{SPDHSW}}_{500})$ and summarized by the same nested medians as above. It is an upper bound on checkpoint-defined target-attainment speed-up under the time model, not an observed speed-up. The control compares these specified learning rates; it does not establish a globally optimal rate for either method.

E.10 FPHA VARIANCE AND BUDGET SELECTION

This separate scalar-estimation study uses 600 training and 575 validation matrices without optimization. It is reported separately from the paired scalar–gradient diagnostic in Section 5.4. The point ratios kV_b/σ_h^2 range from 0.74 to 0.93 (Table 18); their intervals are wide. The total-variance identity in Theorem 2 is consistent with the estimates within bootstrap uncertainty.

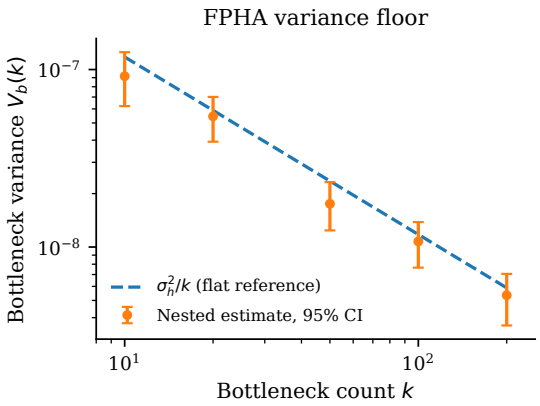


Figure 5: FPHA bottleneck variance and 95% cluster-bootstrap intervals, compared with the k-direction flat variance σ_h^2/k.

Table 18: Nested FPHA variance estimates using 100 bottleneck draws and 500 mixtures per draw. Brackets are 95% cluster-bootstrap intervals; the flat reference moments are $\sigma_h^2 = 1.177 \cdot 10^{-6}$ and $\bar{h} = 8.150 \cdot 10^{-4}$. The identity ratio is $(V_b + V_w)/[(1 + \varepsilon_k)\sigma_h^2 + \varepsilon_k \bar{h}^2]$.

k	V_b [$\times 10^{-8}$]	V_w [$\times 10^{-6}$]	kV_b/σ_h^2	Identity ratio
10	9.17 [6.22, 12.53]	0.97 [0.80, 1.16]	0.78 [0.53, 1.06]	0.900
20	5.45 [3.92, 7.01]	1.14 [1.00, 1.29]	0.93 [0.67, 1.19]	1.015
50	1.75 [1.24, 2.32]	1.10 [1.02, 1.19]	0.74 [0.53, 0.99]	0.951
100	1.07 [0.77, 1.38]	1.14 [1.07, 1.21]	0.91 [0.65, 1.17]	0.976
200	0.53 [0.36, 0.71]	1.19 [1.13, 1.25]	0.91 [0.61, 1.20]	1.015

In direct repeated evaluations, increasing L from 100 to 2000 changes relative RMSE from 0.3376 to 0.2976 at $k = 20$, from 0.2138 to 0.1936 at $k = 50$, and from 0.1786 to 0.1361 at $k = 100$. Flat estimates at $L = k$ give 0.2889, 0.1945, and 0.1275, respectively. Thus the benefit of mixing on these real distributions is smaller than on the anisotropic synthetic pairs.

Table 19: Pilot-based budget selection on FPHA. The target is $v^\star = \sigma_h^2/N$. Predictions use the pilot estimates or their upper confidence endpoints. Observed variance and its upper confidence bound use 500 fresh draws; a target passes when $\mathrm{UCB}_{95} \le v^\star$. All variance columns are scaled by 10^8.

N	$v^\star$	(k, L)	Prediction	Upper prediction	Observed	UCB_{95}	Pass
20	5.89	$(50, 34)$	5.00	5.82	5.90	6.82	no
40	2.94	$(50, 192)$	2.33	2.94	3.09	3.50	no
50	2.35	$(100, 125)$	1.99	2.35	1.91	2.14	yes
80	1.47	$(100, 1354)$	1.16	1.47	1.23	1.38	yes
100	1.18	$(200, 266)$	0.98	1.18	0.98	1.10	yes

Three of five selected budgets meet their targets in independent validation. The two choices at $k = 50$ yield observed variances 0.2% and 4.9% above target. This sensitivity to pilot uncertainty motivates validating a selected budget independently rather than relying only on the pilot point estimate.

E.11 Additional fixed-pair alignment results

Table 20: Fixed-pair alignment with affine-invariant updates: additional flat projection budgets and fixed-direction controls. Scores and times are mean ± standard deviation over five seeds.

Method	Sampling	LEW eval	Runtime (s)
BNCI2014-001 ($d = 22, m = 253$)			
SPDSW ($L = 500$)	Resampled	0.9880 ± 0.0003	38.50 ± 0.02
SPDSW ($L = 50$)	Resampled	0.9882 ± 0.0007	11.19 ± 0.01
SPDSW ($L = 40$)	Resampled	0.9884 ± 0.0008	10.62 ± 0.02
SPDHSW ($k = 40, L = 500$)	Resampled	0.9879 ± 0.0004	10.44 ± 0.25
SPDSW ($L = 500$)	Fixed	1.0187 ± 0.0044	24.20 ± 0.00
SPDSW ($L = 50$)	Fixed	1.1121 ± 0.0059	9.52 ± 0.00
SPDSW ($L = 40$)	Fixed	1.1187 ± 0.0045	9.22 ± 0.01
SPDHSW ($k = 40, L = 500$)	Fixed	1.0967 ± 0.0028	8.98 ± 0.01
FPHA ($d = 63, m = 2016$)			
SPDSW ($L = 700$)	Resampled	2.3707 ± 0.0002	838.16 ± 0.10
SPDSW ($L = 60$)	Resampled	2.4970 ± 0.0024	773.92 ± 0.13
SPDSW ($L = 50$)	Resampled	2.5266 ± 0.0017	769.55 ± 0.06
SPDHSW ($k = 50, L = 700$)	Resampled	2.3738 ± 0.0008	767.58 ± 0.40
SPDSW ($L = 700$)	Fixed	2.5100 ± 0.0010	813.50 ± 0.21
SPDSW ($L = 60$)	Fixed	2.5095 ± 0.0005	771.34 ± 0.16
SPDSW ($L = 50$)	Fixed	2.5094 ± 0.0004	767.30 ± 0.10
SPDHSW ($k = 50, L = 700$)	Fixed	2.5040 ± 0.0007	765.36 ± 0.06

The BNCI experiment pools the same five subjects within each session, as described in Appendix D.3. Its small flat budget already approaches the full-budget score. For FPHA, the fixed-direction configurations converge to similar LEW scores, while resampling separates the large-budget and small-budget estimators.

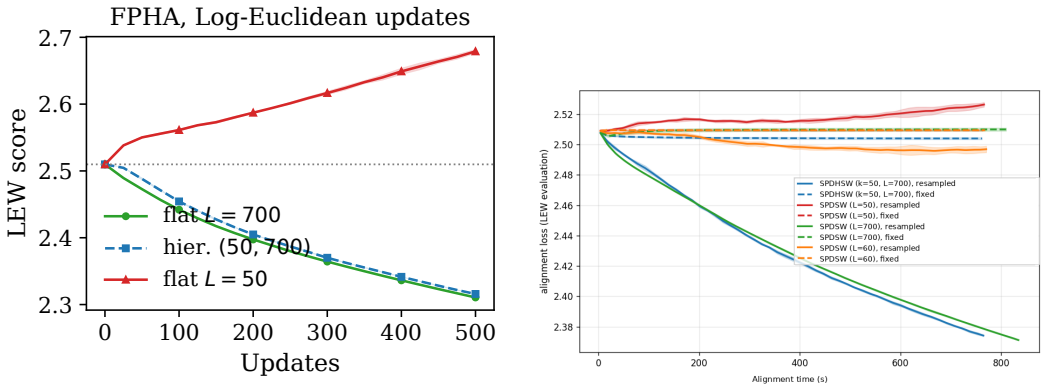


Figure 6: Left: the Log-Euclidean FPHA runs of Figure 2 against the update count (mean of five seeds, ±1 s.d.). Right: the original affine-invariant runs of Table 20 against alignment time; solid lines are resampled and dashed lines fixed directions, bands are ±1 s.d. over five seeds. Under affine-invariant updates the per-iteration logarithm dominates the time axis, so the runtimes nearly coincide and the panel compares quality per update at equal k.

The Log-Euclidean FPHA runs (Appendix D.3) are summarized in Table 2 and shown against updates in Figure 6 (left). At the shared rate 2.5×10^4, SPDHSW $(50, 700)$ ends within 0.23% of flat $L = 700$; the mean 500-update loop times are 3.61 s and 30.20 s (ratio 8.36), and flat $L = 50$ takes 3.20 s. In these runs the score of flat $L = 50$ increases over every recorded 25-update interval in all five seeds, while both other configurations decrease over every interval. An earlier measurement of the same configurations on a shared RTX A6000 gave a seed-median loop-time ratio of 9.07; hardware and aggregation differ, so the two ratios are not comparable. This is a shared-rate comparison; optimizing the learning rate separately for each projection budget remains necessary for a jointly tuned runtime comparison.

E.12 AMBIENT LOG-EUCLIDEAN BASELINES (LOGSW, LOGHSW)

The ambient LogSW baseline projects $\mathrm{vec}(\log M) \in \mathbb{R}^{d^2}$ onto full-matrix Gaussian directions normalized to unit Frobenius norm. LogHSW mixes k such directions with the same normalized-Gaussian mixing law as SPDHSW. Both use $p = 2$ and redraw directions every epoch.

Orthogonal projection of a uniform direction in $\mathbb{R}^{d^2}$ onto the m-dimensional symmetric subspace gives a uniform angle and a random radius. Consequently, the squared ambient population objective is $(m/d^2)\mathrm{SPDSW}_2^2$. The factors are approximately 0.52 at $d = 22$ and 0.54 at $d = 13$: a common nominal learning rate therefore corresponds to a different objective scale.

Table 21 tests the scale explanation on BNCI2014-001 by multiplying the ambient learning rate by d^2/m. LogSW changes from 55.23% at the shared rate to 59.05% at the rescaled rate, against 59.00% for SPDSW. The paired accuracy difference after rescaling is $+0.04$ points (95% interval $[-0.19, +0.28]$); for LogHSW versus SPDHSW it is $+0.07$ ($[-0.05, +0.20]$). The median cumulative log-coordinate displacement ratio is 1.00. Rescaling thus closes the observed ambient-baseline gap without changing the Log-Euclidean representation.

Table 21: Ambient-baseline control on BNCI2014-001 ($k = 40$, $L = 500$, resampled, 5 seeds; accuracy %). "Rescaled" multiplies the LogSW/LogHSW learning rate by the inverse of the population scale factor identified in Appendix E.12, so that the effective step matches SPDSW/SPDHSW.

Method	Update	Shared rate	Rescaled rate
SPDSW	LEM	59.00 ± 0.15	–
LogSW	LEM	55.23 ± 0.07	59.05 ± 0.18
SPDHSW	LEM	58.87 ± 0.09	–
LogHSW	LEM	54.93 ± 0.14	58.93 ± 0.08

E.13 SPD REPRESENTATION LEARNING

Table 22: SPDNet with a sliced Fisher-ratio regularizer (mean $\pm$ s.d. over 5 seeds; sec/epoch includes the regularizer, Δ over the CE-only baseline in parentheses). The regularizer is applied to the SPD features before LogEig.

	AFEW (7 classes, $d_{\mathrm{out}} = 50$)		FPHA (45 classes, $d_{\mathrm{out}} = 33$)	
Loss	Acc. (%)	sec/epoch (Δ)	Acc. (%)	sec/epoch (Δ)
CE only	31.43 ± 1.55	5.83 (–)	83.65 ± 0.87	1.28 (–)
CE + SPDSW Fisher	32.67 ± 1.25	9.70 (+3.87)	83.76 ± 1.26	6.65 (+5.37)
CE + SPDHSW Fisher	33.21 ± 1.16	7.37 (+1.54)	83.79 ± 1.16	6.50 (+5.22)

The two regularizers have similar mean classification accuracies on both datasets. On AFEW, the additional epoch cost over cross-entropy decreases from 3.87 s for SPDSW to 1.54 s for SPDHSW. On FPHA, accuracy differences are small relative to the seed variability. These experiments concern a nonlinear Fisher-ratio objective, whose finite estimates need not be equal even when the individual population discrepancies coincide.

F PAIRED OBJECTIVE AND GRADIENT DIAGNOSIS

F.1 DATA, ESTIMANDS, AND NESTED DESIGN

This diagnosis uses one fixed initial pair of FPHA log-covariance distributions, with $n_s = 600$, $n_t = 575$, and $d = 63$. Each sample represents the stacked coordinates of 21 hand joints. No alignment update is taken in this measurement. Both scalar costs and gradients use the same input state, projections, and weighted one-dimensional transport coupling. The quantity g is the gradient with respect to all source log-coordinates, and its variance is the sum of coordinate variances, not the variance of the gradient norm.

For $k = 20, 50, 100$, each of $J = 100$ independent banks has $M = 200$ independent mixtures. We use the finite-M correction in Appendix C.2. Independent 4096 flat directions estimate the scalar and gradient marginal variances. A split-half check requires the ratio of the two half-variance estimates to be at most 1.25 for both quantities; it passes without extending the flat sample. The estimates are

$$\widehat{\sigma}_h^2 = 1.1905 \times 10^{-6}, \qquad \widehat{\sigma}_g^2 = 5.3088 \times 10^{-6}.$$

Their 95% bootstrap intervals are $[1.0610, 1.3318] \times 10^{-6}$ and $[5.0966, 5.5218] \times 10^{-6}$. Bootstrap resampling uses independent banks and independent 64-direction flat blocks, jointly resampling the scalar and gradient quantities. Thus the intervals for $\rho_g - \rho_h$ retain their same-state pairing and denominator uncertainty.

F.2 Finite-mixture gradient fluctuations

Table 23: Check of the predicted gradient trace variance on the fixed FPHA state, in units of 10^{-8}. Predictions use the nested estimates in Theorem 3; observations use 100 independent complete evaluations at $L = 700$. Flat k is the estimated marginal variance divided by k.

k	Flat k	SPDHSW predicted	SPDHSW observed	RMS ratio
20	26.54	16.03	15.74	1.30×
50	10.62	6.66	6.71	1.26×
100	5.31	3.92	3.88	1.17×

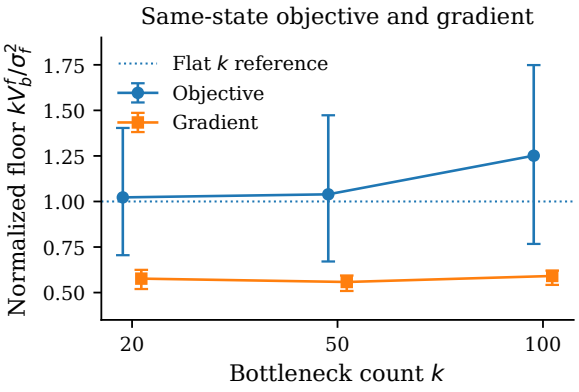


Figure 7: Scalar and gradient normalized bottleneck floors on the same FPHA state, with 95% intervals. The flat-k reference equals one.

An independent set of 100 complete SPDHSW evaluations at $L = 700$ measures trace variance directly. For $k = 20, 50, 100$, the traces are respectively 1.5737×10^{-7}, 6.7127×10^{-8}, and 3.8788×10^{-8}. Dividing the flat reference $\widehat{\sigma}_g^2/k$ by these values and taking square roots yields the RMS-fluctuation ratios in Table 5. The numerator is the estimated variance of an independent flat k-direction average. This calculation does not require a high-precision estimate of the mean gradient.

The normalized between-bank gradient variances are below their scalar counterparts at all three tested k, as shown in Figure 7. The corresponding intervals for $\rho_g - \rho_h$ are $[-0.813, -0.144]$, $[-0.911, -0.128]$, and $[-1.157, -0.190]$. These are fixed-state integration results. They do not measure optimization convergence, establish an HGD mechanism, or show that a hierarchy beats $L = 700$ independent directions at equal final count. The scalar pilot in Table 18 is a separate run and is not pooled with this diagnosis.

F.3 Analytical controls

For a translation test cost with globally quadratic directional value and gradient, the exact normalized floors are both one. An independent 2000-bank control at $k = 20$ gives $\widehat{\rho}_h = 0.984$ with interval

$[0.924, 1.042]$ and $\widehat{\rho}_g = 0.986$ with interval $[0.960, 1.010]$. Tests also verify radial homogeneity, the analytic quadratic conditional mean, and agreement of manual, automatic, and finite-difference gradients on equal- and unequal-count toy problems. These validate the calculations used in the diagnosis without extending its empirical scope to a full training trajectory.

G Configuration-Selected High-Gamma Evaluation

G.1 Shared-rate controls and development protocol

Table 24: High-Gamma common-rate control: 14 subjects, one seed, $\eta = 10^4$, 500 updates. Q/Q_0 is squared alignment cost relative to initialization. Times exclude evaluation and SVM fitting.

Method	Accuracy (%)	$Q/Q_0 \downarrow$	Loop time (s)$\downarrow$
SPDSW, $L = 40$	78.21	0.7761	1.72
SPDSW, $L = 60$	77.63	0.7621	1.89
SPDSW, $L = 500$	78.44	0.7419	12.64
SPDHSW, $(40, 500)$	78.62	0.7642	1.82

All new runs retain 128 channels and align each subject's training-session source to its test-session target. The targets' features are used transductively; their labels do not select the budgets, learning rates, or checkpoints. Optimization uses float64 log-coordinates, plain SGD without momentum or weight decay, Frobenius-uniform directions, and unnormalized mixing. The squared cost Q is the empirical full Log-Euclidean Wasserstein cost defined in Appendix E.9, with no square root or display multiplier.

The shared-rate sweep uses seed 6398, subjects 1–14, 500 updates, and $\eta = 10^4$. Flat counts are $L \in \{20, 40, 60, 100, 150, 200, 500\}$; the hierarchical control is $(k, L) = (40, 500)$. Alignment completes for all 112 runs. Full-subject classifier results are available for the four predefined conditions in Table 24; other partially evaluated configurations are not mixed into those fourteen-subject accuracy means.

Development uses subjects $\{1, 4, 7, 12\}$ and seed 6398. Each family has 24 configurations: flat $L \in \{40, 60, 100, 150, 200, 500\}$ or hierarchical $k \in \{20, 40, 60, 100, 150, 200\}$ with $L = 500$, crossed with $\eta \in \{3000, 10000, 30000, 100000\}$. The primary selection criterion is mean Q/Q_0 at a measured update-time budget across the four subjects. The selected settings are SPDSW $L = 500, \eta = 10^5$ and SPDHSW $(150, 500), \eta = 3 \times 10^4$. Settings are frozen before confirmation. A secondary checkpoint-defined time-to-target criterion selects flat $L = 40, \eta = 10^4$ and SPDHSW $(40, 500), \eta = 10^4$; all four settings and the common-rate flat reference are retained for alignment-quality evaluation.

Table 25 lists every development budget at the common rate $\eta = 10^4$ and the matched update-time budget. A mean is reported only when all four development runs end below their initial cost. At this rate, every flat budget from $L = 40$ to $L = 500$ has a higher mean cost than the hierarchical budgets $(40, 500)$ and $(60, 500)$; the smallest flat budgets perform the most updates but remain above 0.649. For flat $L = 40$ specifically, $\eta = 3 \times 10^3$ gives 0.686, $\eta = 10^4$ gives 0.659, and at $\eta = 3 \times 10^4$ and 10^5 one and four development runs, respectively, end above their initial cost. These development means use one seed and four subjects and were used for selection; the confirmation values in Table 3 are the held-out estimates.

G.2 Confirmation and time budgets

The confirmation subjects are $\{2, 3, 5, 6, 8, 9, 10, 11, 13, 14\}$ and the seeds are $3654, 1788$ and, added afterwards under the identical frozen protocol with alignment only, $2473, 7490, 7066$; a deterministic replay of one original pair reproduced its checkpoint costs exactly before the added seeds were run. A separate budget $B_{s,r}$ for each subject–seed pair is the median of three newly measured 500-update flat reference loops at $L = 500, \eta = 10^4$. Each candidate runs until the time budget and a minimum of 500 updates are both satisfied, with a 20,000-update cap. The reported budget endpoint is the first observed checkpoint at or beyond $B_{s,r}$, not necessarily the final minimum-update horizon.

Table 25: Development grid at the common rate $\eta = 10^4$ (subjects $\{1, 4, 7, 12\}$, seed 6398, matched update-time budget). Mean Q/Q_0 is shown only when all four runs end below their initial cost; the next column counts runs that do not. Mean updates are the completed updates within the budget.

Method	Budget	Mean $Q/Q_0 \downarrow$	Runs above initial	Mean updates
SPDSW	$L = 40$	0.659	0	3745
SPDSW	$L = 60$	0.650	0	3390
SPDSW	$L = 100$	0.668	0	1745
SPDSW	$L = 150$	0.672	0	1480
SPDSW	$L = 200$	0.684	0	1102
SPDSW	$L = 500$	0.735	0	500
SPDHSW	$(20, 500)$	–	1	4838
SPDHSW	$(40, 500)$	0.629	0	3492
SPDHSW	$(60, 500)$	0.633	0	3165
SPDHSW	$(100, 500)$	0.661	0	1648
SPDHSW	$(150, 500)$	0.669	0	1402
SPDHSW	$(200, 500)$	0.683	0	1058

Updates are timed in synchronized blocks of ten, excluding Q evaluation, initial logarithms, matrix exponentials, file I/O, and classifier fitting. New measurements use an NVIDIA RTX 6000 Ada Generation GPU; the earlier five-seed loop times use an RTX A6000 and remain separate.

All 250 confirmation alignments are retained. Table 26 reports alignment costs, measured loop times, and completed updates for five configurations, each on the same fifty subject–seed pairs. For the two primary selections, mean update counts at the budget endpoint are approximately 504 and 1398. The auxiliary two-seed classifier evaluation is reported separately in Appendix G.4, rather than combined with these five-seed alignment columns.

Table 26: Alignment results for all five confirmation configurations on the same ten subjects and five seeds. All columns use all fifty subject–seed pairs. The separate auxiliary classifier results are in Table 27.

Method	Learning rate	$Q/Q_0 \downarrow$	Loop time (s)	Updates
SPDSW, $L = 40$	10^4	0.6481	13.03	3,653
SPDHSW, $(40, 500)$	10^4	0.6317	13.03	3,455
SPDSW, $L = 500$	10^4	0.7442	13.08	502
SPDSW, $L = 500$	10^5	0.6276	13.13	504
SPDHSW, $(150, 500)$	3×10^4	0.6251	13.07	1,398

Quality is observed at time-budget fractions $0, 0.02, 0.05, 0.10, 0.25, 0.50, 0.75, 1$, with block overshoot. The hierarchical primary setting first passes the reference target at the 0.25 checkpoint in all fifty pairs, and the flat one at 0.25 in 48 pairs and at 0.10 in two. Accordingly, we use same-budget quality as the primary comparison rather than interpreting differences between these coarse crossing times as a precise speed-up.

G.3 Paired alignment analysis

The secondary development criterion selected flat $L = 40$ and hierarchical $(40, 500)$ at $\eta = 10^4$ before confirmation. Figure 3 and Table 3 use all ten confirmation subjects and all five seeds, with no further subject or configuration filtering. The mean normalized-cost difference is -0.016359, and per-seed means range from -0.0162 to -0.0165. We first average the paired difference over the five seeds within each subject and then bootstrap the ten subject clusters 50,000 times. The percentile interval is $[-0.019484, -0.013600]$. This descriptive analysis was added to the stored confirmation results; it is not a preregistered hypothesis test and is not adjusted for the development search.

The primary development criterion independently selects flat $L = 500$, $\eta = 10^5$ and hierarchical $(150, 500)$, $\eta = 3 \times 10^4$. Applying the same five-seed, subject-cluster calculation gives a normalized-

cost difference of -0.002482 with interval $[-0.004034, -0.000139]$, favoring the hierarchy in 44 of 50 pairs. Its smaller effect is reported separately from the same-count comparison; neither result is an equivalence test.

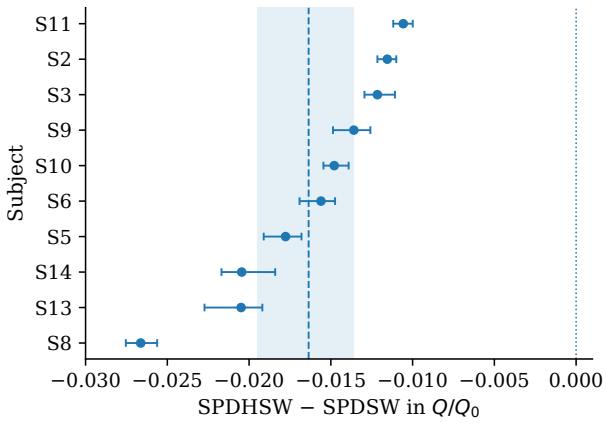


Figure 8: Subject-level endpoints of the same-count High-Gamma comparison (ten subjects, five seeds). Points are within-subject mean SPDHSW-minus-SPDSW differences in Q/Q_0; bars show the seed range. The dashed line and shaded band give the overall mean and its descriptive 95% subject-cluster interval. All fifty paired differences are negative.

The two panels of Figure 3 use only observed checkpoints at the scheduled budget fractions. The upper panel connects each method's mean measured times and mean costs. The lower panel uses the mean paired checkpoint time and the subject-cluster mean difference, with pointwise rather than simultaneous confidence intervals. No unobserved quality value or first-passage time is interpolated. Matching the time budget allows different update counts and includes the small overshoot of a timed ten-update block. Figure 8 preserves the subject-level breakdown of the endpoints.

G.4 Classification of the selected High-Gamma settings

In addition to the complete five-seed classification protocol of Table 4, an auxiliary check evaluates the two development-selected High-Gamma settings on all ten confirmation subjects using seeds 3654 and 1788. These 40 evaluations use the first two confirmation seeds, so Table 27 reports accuracy on its own, separately from the five-seed alignment columns. The evaluation is retained in full and was limited by computation rather than observed accuracy.

Table 27: Auxiliary classification of the development-selected High-Gamma settings: ten subjects, two seeds (3654, 1788), and 20 evaluations per method. Accuracy is averaged over subjects within each seed and then over seeds. These are separate from the complete five-seed classification results in Table 4.

Method	Learning rate	Accuracy (%)
SPDSW, $L = 500$	10^5	81.81
SPDHSW, $(150, 500)$	3×10^4	81.84

Features are $\mathrm{vec}(\log X) \in \mathbb{R}^{16384}$ without an additional scaler. LinearSVC selects C by source-only three-fold cross-validation over ten logarithmically spaced values from 10^{-2} to 10^2, followed by refitting. Both settings use `tol`$= 10^{-4}$, `max_iter`$= 1000$, `dual=auto`, and the projection seed as classifier random state. Target labels are used only after configuration selection. The paired accuracy difference is $+0.031$ percentage points with a descriptive 95% subject-cluster interval of $[-0.906, +1.281]$, computed by averaging the two seeds within each subject before resampling subjects. This interval does not establish equivalence.

Seven flat and ten hierarchical pipelines emit at least one Liblinear convergence warning; the log does not separate candidate fits from final refits. The table records outcomes under the common solver budget, rather than asserting convergence of every inner fit. Classifier cross-validation takes substantially longer than alignment, so alignment-loop savings do not establish an end-to-end training speed-up.

A separate high-rate cache check on all fourteen subjects at seed 6398 gives 81.25% for SPDSW $L = 500$, $\eta = 10^5$, compared with 79.375% and 78.348% for the original common-rate flat and hierarchical caches. This one-seed result is not pooled with either the two-seed check or the five-seed classification protocol.

G.5 Numerical stability

Of the 192 development runs, eight terminate when the loss or log-coordinates become nonfinite: two flat and six hierarchical, all at $\eta = 10^5$. A further 34 completed runs have finite but worsened final cost $Q/Q_0 > 1$. We retain these failures and distinguish numerical completion from convergence. They motivate selecting the step size jointly with projection budget; the experiment does not isolate a unique cause of instability.

Table 28: Development-grid outcomes. The last column counts numerically completed runs whose final cost exceeds initialization.

Method	LR	Runs	Nonfinite termination	Completed with $Q/Q_0 > 1$
SPDHSW	3,000	24	0	0
SPDHSW	10,000	24	0	1
SPDHSW	30,000	24	0	7
SPDHSW	100,000	24	6	12
SPDSW	3,000	24	0	0
SPDSW	10,000	24	0	0
SPDSW	30,000	24	0	2
SPDSW	100,000	24	2	12

Table 29: Development-grid outcomes at $\eta = 10^5$ by projection budget. Each row has four development subjects. Nonfinite termination and finite completion with $Q/Q_0 > 1$ are distinct outcomes.

Method	Budget	Runs	Nonfinite	Finite, $Q/Q_0 > 1$
SPDHSW	$k = 20, L = 500$	4	4	0
SPDHSW	$k = 40, L = 500$	4	1	3
SPDHSW	$k = 60, L = 500$	4	1	3
SPDHSW	$k = 100, L = 500$	4	0	4
SPDHSW	$k = 150, L = 500$	4	0	1
SPDHSW	$k = 200, L = 500$	4	0	1
SPDSW	$L = 40$	4	1	3
SPDSW	$L = 60$	4	1	3
SPDSW	$L = 100$	4	0	4
SPDSW	$L = 150$	4	0	1
SPDSW	$L = 200$	4	0	1
SPDSW	$L = 500$	4	0	0