GSVD FOR GEOMETRY-GROUNDED DATASET COMPARISON: AN ALIGNMENT ANGLE IS ALL YOU NEED

Arthur Sobrinho
Institute of Computing, UFRJ
arthursfr@ic.ufrj.br

Eduarda Marques
Institute of Computing, UFRJ
eduardasm@ic.ufrj.br

João Paixão
Institute of Computing, UFRJ
jpaixao@ic.ufrj.br

Daniel S. Menasche
Institute of Computing, UFRJ
sadoc@ic.ufrj.br

Heudson Mirandola
Institute of Mathematics, UFRJ
mirandola@im.ufrj.br

ABSTRACT

Geometry-grounded learning asks models to respect structure in the problem domain rather than treating observations as arbitrary vectors. Motivated by this view, we revisit a classical but underused primitive for comparing datasets: *linear relations* between two data matrices, expressed via the co-span constraint $Ax = By = z$ in a shared ambient space.

To operationalize this comparison, we use the generalized singular value decomposition (GSVD) as a joint coordinate system for two subspaces. In particular, we exploit the GSVD form $A = HCU$, $B = HSV$ with $C^\top C + S^\top S = I$, which separates shared versus dataset-specific directions through the diagonal structure of (C, S). From these factors we derive an interpretable *angle score* $\theta(z) \in [0, \pi/2]$ for a sample z, quantifying whether z is explained relatively more by A, more by B, or comparably by both.

The primary role of $\theta(z)$ is as a *per-sample geometric diagnostic*. We illustrate the behavior of the score on MNIST through angle distributions and representative GSVD directions. A binary classifier derived from $\theta(z)$ is presented as an illustrative application of the score as an interpretable diagnostic tool.

1 INTRODUCTION

Comparing datasets is a recurring problem across machine learning and data analysis. It arises when assessing dataset shift between training and deployment Sugiyama et al. (2007), when comparing representations learned by different models (Huang et al., 2007), and when diagnosing similarities and differences between classes or domains (Morcos et al., 2018). While modern practice often compares datasets indirectly (via trained models or embedding distances), such procedures can obscure *why* two datasets appear similar or different.

We take a complementary view: compare datasets through their *geometry*. Many real-world datasets concentrate near low-dimensional structures, exhibit partially shared latent factors, and contain directions that are specific to one domain or class. When datasets live in a common ambient feature space (raw pixels, sensor measurements, or fixed embeddings), their geometric relationship becomes meaningful in its own right.

A relation primitive. Let $A = [a_1, \ldots, a_p] \in \mathbb{R}^{d\times p}$ and $B = [b_1, \ldots, b_q] \in \mathbb{R}^{d\times q}$ denote two dataset matrices (columns are observations in $\mathbb{R}^d$). We characterize similarity through the linear relation

$$Ax = By = z, \tag{1}$$

where $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$ are coefficients and $z \in \mathbb{R}^d$ is a shared ambient vector. We treat (1) as a primitive object: it encodes compatibility in the ambient space without requiring pointwise correspondences between samples or an invertible mapping between domains.

GSVD yields a joint frame. A classical tool for paired subspace analysis is GSVD (Edelman & Wang, 2020). Given A and B with the same row dimension d, one common GSVD form provides

$$A = HCU, \qquad B = HSV, \qquad C^\top C + S^\top S = I, \tag{2}$$

where $H \in \mathbb{R}^{d\times d}$ is invertible (or left-invertible in reduced variants), $U \in \mathbb{R}^{p\times p}$ and $V \in \mathbb{R}^{q\times q}$ are orthogonal, and C and S are diagonal or block-diagonal with nonnegative entries. Intuitively, H defines a shared ambient reference frame, while (C, S) encode how strongly each shared direction contributes to A versus to B.

A single alignment angle. Our goal is to summarize the relative explanatory power of A versus B for a sample z by a minimal, interpretable quantity. We define an *alignment angle* $\theta(z) \in [0, \pi/2]$ derived from the GSVD frame: $\theta(z) \approx 0$ means "more A", $\theta(z) \approx \pi/2$ means "more B", and $\theta(z) \approx \pi/4$ indicates shared structure.

Contributions. We summarize our contributions as follows:

1. We propose linear relations in the co-span form $Ax = By = z$ as a minimal, geometry-grounded primitive for dataset comparison.
2. We use GSVD as a natural joint coordinate system for comparing two subspaces, making shared vs. dataset-specific directions explicit via (C, S).
3. We derive an interpretable angle score $\theta(z)$ that quantifies a per-sample diagnostic of relative dataset alignment and supports a binary classification.
4. We illustrate the geometric behavior of $\theta(z)$ on MNIST through angle distributions and GSVD-derived representative directions.

Organization. The remainder of the paper is organized as follows. Section 2 reviews related work, and Section 3 introduces the background on linear relations and GSVD. Section 4 defines the relative alignment angle $\theta(z)$, explains how to compute it efficiently in the GSVD frame, and describes downstream uses such as classification and diagnostics. Section 5 reports experiments on MNIST, including class-conditional angle distributions and representative GSVD directions. Section 6 discusses additional geometric interpretations, followed by conclusions in Section 7. The appendices provide additional derivations and supporting GSVD properties, including the span-view interpretation and proofs of the main technical results.

2 Related work

Our work sits at the intersection of (i) classical subspace comparisons and multivariate statistics, (ii) instance alignment objectives, and (iii) modern representation-similarity and transfer-learning perspectives.

Subspace geometry and joint decompositions. SVD provides canonical coordinates for a single matrix and underpins least-squares and low-rank approximation (Golub & Van Loan, 2013; Stewart, 1993). To compare two subspaces, principal angles quantify overlap via the singular values of $Q_A^\top Q_B$ for orthonormal bases Q_A, Q_B (Björck & Golub, 1973; Knyazev & Argentati, 2002), and Canonical Correlation Analysis (CCA) identifies maximally correlated projections between variable sets (Hotelling, 1936). GSVD generalizes SVD to matrix pairs (Van Loan, 1976; Paige & Saunders, 1981), producing a joint frame $A = HCU$, $B = HSV$ where the diagonal factors (C, S) expose shared versus dataset-specific directions and encode principal-angle-like structure through (c_i, s_i). We use this joint frame as an interpretable comparative coordinate system and derive from it a per-sample alignment angle $\theta(z)$. Whereas principal angles describe geometry between subspaces, $\theta(z)$ provides a complementary *sample-level diagnostic* indicating whether an observation aligns more strongly with A, B, or the shared structure.

Instance alignment and Procrustes. When sample correspondences are available (or assumed), alignment is often formulated as minimizing an instance-wise discrepancy under a restricted transformation family. Procrustes analysis is the canonical example, aligning two point clouds by an orthogonal (or similarity) transform (Gower, 1975). In many dataset-comparison settings, however,

correspondences are ambiguous or unreliable, motivating our emphasis on relations and subspace geometry. Still, instance alignment is complementary to our approach: the GSVD angles provide a principled way to downweight weakly shared directions before performing alignment (Appendix C).

Representation similarity and dataset distances. Recent work compares learned representations using invariances to re-parameterizations. SVCCA and PWCCA adapt CCA to compare neural representations through truncated subspaces and weighting schemes (Raghu et al., 2017; Morcos et al., 2018), while CKA provides a robust similarity index widely used for layerwise comparisons (Kornblith et al., 2019; Cortes et al., 2012). In parallel, distributional comparators such as MMD (Gretton et al., 2012) and optimal transport (Peyré & Cuturi, 2019) measure dataset discrepancy directly, and FID is a practical feature-statistics score in generative modeling (Heusel et al., 2017). These approaches often output a *single* similarity value (or a layerwise matrix) and prioritize invariance or distribution matching. Our goal is complementary: we extract an explicit shared frame (via GSVD), interpretable representative directions, and a *per-sample* diagnostic $\theta(z)$.

Instance-based transfer learning. A different but related viewpoint is transfer learning by *reweighting* or *selecting* source instances to better match a target domain. Surveyed in Pan & Yang (2010), this line includes boosting-style methods such as TrAdaBoost (Zheng et al., 2020) and covariate-shift / importance-weighting approaches (Sugiyama et al., 2007; Huang et al., 2007). Our angle score $\theta(z)$ plays a conceptually similar role as a *relative compatibility* signal: it can be used to flag samples that are more consistent with one dataset than the other, which is useful for auditing, filtering, or prioritizing instances when comparing or transferring between domains.

3 Background: Linear relations and GSVD

This section presents two equivalent views of linear relations: the *co-span* formulation, which underlies the rest of the paper, and the *span* formulation, developed further in Appendix C. We also introduce GSVD as a joint coordinate framework for expressing relations between datasets, highlighting the block structure of (C, S) that separates shared from dataset-specific directions.

3.1 Linear relations: co-span vs. span

A linear relation between the two subspaces can be formalized as

$$\mathcal{R}_{\text{co-span}} \;=\; \{(x, y) \in \mathbb{R}^p \times \mathbb{R}^q : Ax = By\}. \tag{3}$$

Here, x and y denote coordinate vectors, and the constraint imposes that both sides yield the *same* ambient vector $z := Ax = By \in \mathbb{R}^d$. From a geometric perspective, this encodes the intersection (and approximate intersection) structure between $\mathrm{col}(A)$ and $\mathrm{col}(B)$, which denote the column spaces of A and B, respectively; it can be naturally interpreted as an *alignment of dimensions* by enforcing representability in a shared ambient space. Appendix A summarizes the notation used throughout the paper.

Equivalently, one may introduce an explicit shared latent parameter w:

$$\mathcal{R}_{\text{span}} \;=\; \{(x, y) : \exists\, w \text{ s.t. } x = Fw, \;\; y = Gw\}, \tag{4}$$

for suitable linear maps F, G. This view is closer to *instance alignment*, but in many dataset-comparison settings correspondences are unreliable, motivating the relation-first view (3).

3.2 GSVD as a joint geometry

To compare two subspaces simultaneously, GSVD produces a *joint coordinate system*. In one common form, given matrices A and B, GSVD produces the decomposition given by (2). Intuitively, H defines a shared ambient reference frame, while the diagonal factors $C^{d\times p}$ and $S^{d\times q}$ quantify how strongly each shared direction contributes to A and B. In particular, the diagonal (or block-diagonal) factors C and S are ordered so that entries of C decrease and entries of S increase, making "more-A" versus "more-B" directions explicit.

Block structure of C and S. In the GSVD form (2), the diagonal (or block-diagonal) factors C and S can be written by isolating a central diagonal block (the shared subspace) and padding with

identity/zero blocks:

$$C = \begin{bmatrix} I_r & 0 & 0 \\ 0 & \widetilde{C} & 0 \\ 0 & 0 & 0_t \end{bmatrix}, \qquad S = \begin{bmatrix} 0_r & 0 & 0 \\ 0 & \widetilde{S} & 0 \\ 0 & 0 & I_t \end{bmatrix}, \tag{5}$$

where I_r and I_t are identity matrices, 0_r and 0_t are zero matrices, and $\widetilde{C}, \widetilde{S} \in \mathbb{R}^{k\times k}$ are diagonal with strictly positive entries. The sizes satisfy $r + k + t = d$.

The C and S elements are ordered as follows:

$$c_{11} \geq c_{22} \geq \cdots \geq c_{dd},$$
$$s_{11} \leq s_{22} \leq \cdots \leq s_{dd}.$$

Shared vs. specific directions. Directions where a diagonal entry of C dominates correspond to geometry primarily explained by A; directions where S dominates are primarily explained by B; comparable magnitudes indicate shared structure.

4 A RELATIVE ALIGNMENT ANGLE $\theta(z)$ AND ITS USES

This section defines the alignment angle associated with a fixed instance, presents an algorithm for its computation, characterizes the instances that attain its maximum and minimum values, and presents two downstream applications.

4.1 DEFINING THE ALIGNMENT ANGLE

We now turn the relation between two datasets A and B into a single, interpretable score. For a given $z \in \mathrm{col}(A) \cap \mathrm{col}(B)$, i.e., for which the co-span relation $Ax = By = z$ is feasible, define the fiber

$$\mathcal{R}_{\text{co-span}}(z) \;=\; \{(x, y) \in \mathbb{R}^p \times \mathbb{R}^q : Ax = By = z\}. \tag{6}$$

If z does not lie in this intersection, the relation is undefined and the sample is considered *non-relational* with respect to (A, B).

Definition 1 (Alignment angle)**.** The alignment angle of z given matrices A and B is defined as

$$\theta(z) \;:=\; \arctan\left(\frac{\|x\|_2}{\|y\|_2}\right) \in \left[0, \tfrac{\pi}{2}\right], (x, y) \in \mathcal{R}_{\text{co-span}}(z), x \perp \mathrm{Ker}(A), y \perp \mathrm{Ker}(B). \tag{7}$$

Interpretation. $\theta(z) \approx 0$ means z is explained by A with smaller coefficient norm than by B ("more A"); $\theta(z) \approx \pi/2$ means the symmetric situation ("more B"); $\theta(z) \approx \pi/4$ indicates comparable explanatory strength (shared structure).

The constraints $x \perp \mathrm{Ker}(A)$ and $y \perp \mathrm{Ker}(B)$ pick the minimum-ℓ_2 norm coefficients among all pairs satisfying $Ax = By = z$. Hence $\|x\|_2$ and $\|y\|_2$ are canonical "costs" to represent the same z using A versus B, and the ratio $\|x\|_2/\|y\|_2$ compares these costs. Using $\arctan$ maps this positive ratio to a bounded, symmetric score in $[0, \pi/2]$, with equal costs $\|x\|_2 = \|y\|_2$ giving $\theta = \pi/4$.

4.2 COMPUTING $\theta(z)$ IN THE GSVD FRAME

Next, we express the alignment score (1) in terms of the GSVD factors, obtaining a closed-form expression that is easy to compute and geometrically interpretable: H defines a shared ambient basis, while (C, S) weight the relative contributions of A and B along each direction.

Let (H, C, S, U, V) denote the GSVD factors in (2). For an arbitrary sample $z \in \mathbb{R}^d$, compute its coordinates in the shared frame by

$$c(z) \;:=\; \underset{c\in\mathbb{R}^d}{\arg\min}\; \|Hc - z\|_2^2 \qquad \Rightarrow \qquad c(z) = H^\dagger z. \tag{8}$$

Define the GSVD-weighted costs

$$a(z) := \left\|C^\dagger c(z)\right\|_2, \qquad b(z) := \left\|S^\dagger c(z)\right\|_2, \tag{9}$$

where † denotes the Moore–Penrose pseudoinverse (optionally truncated for numerical stability). Because U and V are orthogonal, these costs correspond to the norms of canonical coefficient choices in $\mathcal{R}_{\text{co-span}}(z)$ (up to the same orthogonal transforms).

Theorem 1 (GSVD produces alignment angles)**.**

$$\theta(z) = \arctan\left(\frac{a(z)}{b(z)}\right) = \arctan\left(\frac{\left\|C^\dagger\, c(z)\right\|_2}{\left\|S^\dagger\, c(z)\right\|_2}\right), \tag{10}$$

where $c(z) = H^\dagger z$ and (H, C, S, U, V) are the GSVD factors in (2).

Proof. See Appendix D. □

4.3 Finding Extreme Directions

So far, the pipeline is *forward*: given a sample z, we compute $\theta(z)$ as a relative-alignment diagnostic. We can also ask the inverse question: *which vectors z are maximally A-like or maximally B-like under the GSVD geometry?* Equivalently, we seek

$$z_{\max} \in \arg\max_{z\in\mathcal{Z}}\ \theta(z), \qquad z_{\min} \in \arg\min_{z\in\mathcal{Z}}\ \theta(z), \tag{11}$$

over a constraint set $\mathcal{Z}$ (e.g., $\|z\|_2 = 1$, or z restricted to the span of training data). This viewpoint produces representative "extreme" directions for visualization and diagnostics, complementing the per-sample scoring setting of Section 4.2.

In what follows, we characterize $z_{\max}$ and $z_{\min}$ in terms of the GSVD structure. In particular, $z_{\max}$ and $z_{\min}$ correspond to specific columns of the shared matrix H: namely, h_{r+k} and h_{r+1} respectively. These vectors represent the last and first shared generalized components of the datasets, where "shared" in this instance refers to components associated with indices for which both diagonal matrices C and S have nonzero entries. As such the GSVD provides not only the coordinate system used to calculate (7), but that same basis encodes the solution for optimization problem (11).

Let $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)$ and $\tilde{y} = (\tilde{y}_1, \tilde{y}_2, \tilde{y}_3)$, where the dimensions of the components of $\tilde{x}$ and $\tilde{y}$ correspond to the blocks of C and S given by (5), i.e., $\tilde{x}_1 \in \mathbb{R}^r$, $\tilde{x}_2 \in \mathbb{R}^k$ and $\tilde{x}_3 \in \mathbb{R}^t$. Similarly, $\tilde{y}_1 \in \mathbb{R}^r$, $\tilde{y}_2 \in \mathbb{R}^k$ and $\tilde{y}_3 \in \mathbb{R}^t$.

Theorem 2 (GSVD produces extreme directions)**.** Given datasets A and B, the extreme directions $z_{\max}$ and $z_{\min}$ that maximize and minimize the alignment angle $\theta(z)$, respectively, are unique and given by $z_{\max} = Ax_{\max} = By_{\max}$ and $z_{\min} = Ax_{\min} = By_{\min}$, where $z_{\max} = h_{r+k}$, $z_{\min} = h_{r+1}$ and

$$x_{\max} = U^T \begin{bmatrix} 0 \\ \widetilde{S}\widetilde{C}^{-1} e_k \\ 0 \end{bmatrix}, \quad y_{\max} = V^T \begin{bmatrix} 0 \\ e_k \\ 0 \end{bmatrix}, \quad x_{\min} = U^T \begin{bmatrix} 0 \\ \widetilde{C}\widetilde{S}^{-1} e_1 \\ 0 \end{bmatrix}, \quad y_{\min} = V^T \begin{bmatrix} 0 \\ e_1 \\ 0 \end{bmatrix}.$$

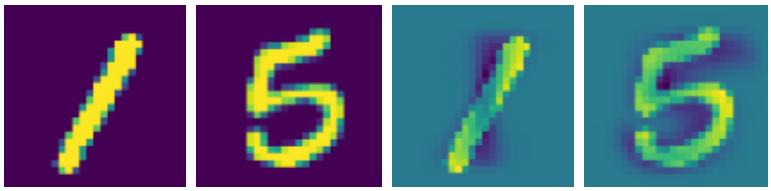
(H, C, S, U, V) are the GSVD factors in (2), e_i is the i-th canonical basis vector, i.e., $(e_i)_j = 1$ if $j = i$ and $(e_i)_j = 0$ otherwise, and h_i is the i-th column of matrix H.

Proof. See Appendix D. □

Remark 1 (Deflation / subsequent directions)**.** A standard way to obtain subsequent maximizers is to augment the optimization problem underlying Theorem 2 with the additional constraint that x be orthogonal to all previously computed maximizers. Let X_{prev} denote the matrix whose columns are the previously obtained maximizers, and impose the constraint $X_{\text{prev}}^\top x = 0$ in the maximization problem. This procedure is commonly referred to as *deflation*, in analogy with extracting successive principal components in PCA. An explicit orthogonality constraint on y is unnecessary, since its orthogonality is induced by that of x. For further details, we refer the reader to (Chu et al., 1997).

4.4 Downstream uses: diagnostics and illustrative classification

Binary classification (relative geometry). Given two domains (e.g., digit "1" vs. digit "5"), we build matrices A and B from their training samples (or low-rank bases thereof). For a test sample z,



(a) (b) (c) (d)

Figure 1: Representative MNIST digit samples for the GSVD pipeline. (a,b) raw digits "1" and "5"; (c,d) corresponding vectors after mean-centering.

compute $\theta(z)$ and predict

$$\hat{\ell}(z) = \begin{cases} A, & \theta(z) \leq \tau, \\ B, & \theta(z) > \tau, \end{cases}$$

with $\tau = \pi/4$ as a symmetry default, or tuned on validation data. The pseudo-code of the classifier is presented in Appendix B (see Algorithm 1).

We emphasize that this rule is presented only as an illustrative consequence of the alignment score. The goal of the method is not to propose a competitive classifier, but to provide a geometrically interpretable diagnostic of how strongly a sample aligns with each dataset.

Diagnostics. Samples whose angles conflict with their nominal labels (e.g., a class-A sample with $\theta(z)$ close to $\pi/2$) are candidates for auditing. The GSVD frame also provides interpretable directions, which can be used to visualize shared versus dataset-specific structure.

Decomposition Complexity. GSVD requires $O(d^3)$ time, which presents a bottleneck for large-scale databases. Since the decomposition is performed as a preprocessing step, however, it does not impact cost at inference time. Broader adoption and use cases of the GSVD in machine learning contexts may motivate the development of more efficient algorithms, leaving reduced decomposition complexity as a promising direction for future work.

Numerical considerations. In practical implementations the diagonal entries of C and S may contain very small values, making the pseudoinverses $C^\dagger$ and $S^\dagger$ sensitive to noise. In our experiments we use a truncated pseudoinverse that ignores singular values below a small tolerance. A systematic analysis of stability with respect to rank truncation, preprocessing, and noise remains an interesting direction for future work.

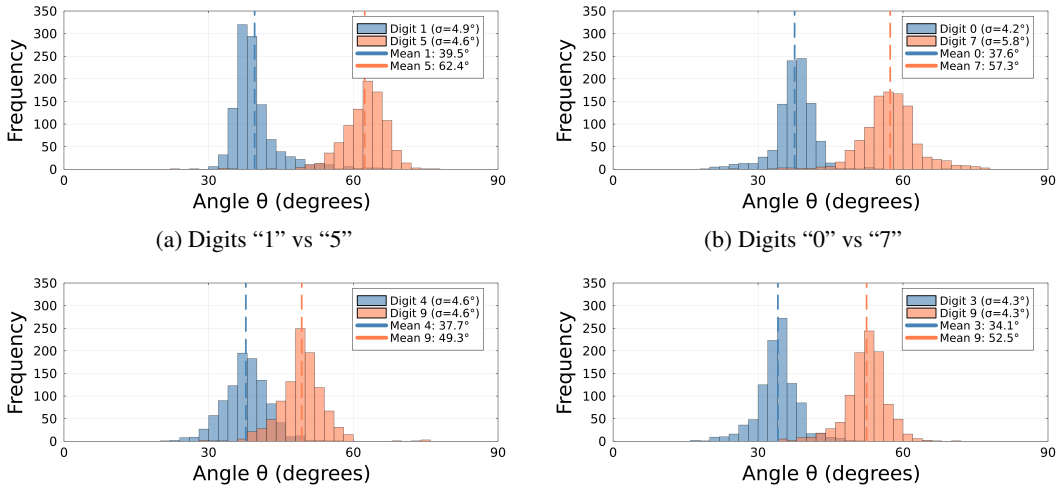
5 Experiments on MNIST

We describe a simple protocol on MNIST (LeCun et al., 1998) to illustrate how θ behaves as an angle score for classification. Additional results on MNIST-Fashion are reported in Appendix E. The purpose of this experiment is not to benchmark classification performance but to illustrate the geometric behavior of the alignment score. MNIST provides a controlled setting in which the induced subspace structure can be visualized and interpreted. The same pipeline can be applied to other representations, including learned feature embeddings, which we leave for future work.

Setup. We choose two digits a and b (e.g., "1" vs. "5"). Each image is vectorized in $\mathbb{R}^{784}$ (from 28×28 pixels) and mean-centered. MNIST consists of 60,000 training images and 10,000 test images. We construct the matrix A by stacking $p = 900$ training images from digit a as columns, and analogously construct B by stacking $q = 800$ training images from digit b, with the columns selected uniformly at random from the training set.

Evaluation. The histograms reported in Figure 2 are obtained by computing $\theta(z)$ for all test samples z belonging to the two selected digits, i.e., for all available instances of digits a and b in the MNIST test set, using 24. We report the empirical class-conditional distributions of $\theta(z)$ as a diagnostic of relative geometric separability.

Interpreting the $\theta(z)$ histograms. If the two digits induce clearly different subspaces, the histograms concentrate on different angle ranges (small angles for the first digit in the pair and



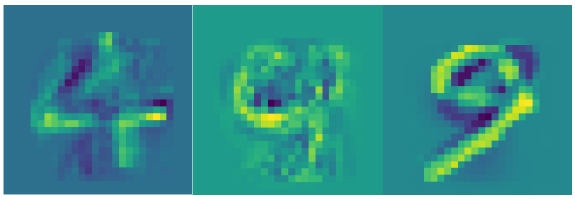
(a) Digits “1” vs “5”

(b) Digits “0” vs “7”

(c) Digits “4” vs “9”

(d) Digits “3” vs “9”

Figure 2: Empirical distributions of the angle $\theta(z)$ on the MNIST test set for four digit pairs. Values closer to 0 indicate stronger alignment with the first digit in the pair, while values closer to $90°$ indicate stronger alignment with the second.



(a) (b) (c)

Figure 3: Representative H component directions reconstructed in the image space as 28x28 images, using the viridis colormap, obtained by the GSVD-based optimization in Section 4.3: (a) representation of a solution from the minimization problem indicating a more “4-like” direction; (c) representation of the solution from the maximization problem indicating a more “9-like” direction; (b) reconstruction of a shared direction that encapsulates the structure of both “4” and “9”.

large angles for the second). Overlap between the histograms indicates directions that are similarly explained by both subspaces, and thus a more ambiguous relative geometry.

For a scalar, threshold-free summary of the separation between the class-conditional θ distributions, we report Fisher–Rao distances between the corresponding histograms in Appendix F (Table 2).

Interpreting the $\theta(z)$ extremes. The extreme columns correspond to directions of the shared GSVD frame H that maximize or minimize the alignment angle $\theta(z)$ as derived in Section 4.3. The columns of H can represent directions ordered by increasing relative-alignment angle $\theta(z)$, ranging from directions that are maximally aligned with A to those that are maximally aligned with B. Intermediate columns capture directions that jointly reflect structural characteristics of both subspaces.

Figure 3 illustrates this behavior for matrices A and B constructed from MNIST images of digits 4 and 9, respectively. The A-aligned extreme exhibits a prototypical “4-like” structure with sharp, edge patterns, while the B-aligned extreme emphasizes rounded contours and a diagonal component. The central image reveals a blended representation, combining elements shared by both digits, and thus illustrates directions along which samples from the two classes are similarly explained.

These visual structures are inherently data-dependent: the appearance of extreme and intermediate directions is determined by the specific instances used to construct A and B. Consequently, they should be interpreted as empirical manifestations of the underlying class geometry rather than canonical templates.

6 DISCUSSION: GEOMETRY, SCALE, AND INFORMATION GEOMETRY

This section connects the alignment angle θ to probabilistic and information-geometric interpretations. While $\theta(z)$ was originally motivated by linear-algebraic considerations, it naturally induces probabilistic structures whose geometry can be studied using tools from information geometry.

From angles to probabilistic geometry. Our goal is to connect the alignment angle θ, originally motivated by linear-algebraic arguments, to probabilistic and information-geometric interpretations. We do so in two complementary steps.

First, we show that each sample-wise angle $\theta(z)$ induces a Bernoulli posterior whose Fisher–Rao geometry is linear in $|\theta(z) - \theta(z')|$. Second, we move from individual samples to distributions over angles and study the Fisher–Rao distance between θ-histograms, showing that this distance is driven primarily by mass near $\theta = \pi/4$, while samples near 0 or $\pi/2$ correspond to confident regions with negligible contribution.

The results in this section are presented using Fisher–Rao geometry. Alternatively, they can also be derived via the Hellinger geometry, as discussed in Appendix G.

6.1 PER-SAMPLE FISHER–RAO GEOMETRY INDUCED BY θ

$\theta(z) = \arctan(a(z)/b(z))$. We interpret inverse-squared costs as evidences and adopt the parametric posterior model

$$P(A \mid \theta) = \frac{1/a(z)^2}{1/a(z)^2 + 1/b(z)^2} = \frac{b(z)^2}{a(z)^2 + b(z)^2} = \cos^2\theta, \qquad P(B \mid \theta) = \sin^2\theta, \tag{12}$$

so that $P(A \mid \theta)/P(B \mid \theta) = \cot^2\theta$. This posterior coincides with a Bayes posterior based on likelihoods $p(\theta \mid A)$ and $p(\theta \mid B)$ whenever

$$\frac{p(\theta \mid A)}{p(\theta \mid B)} = \frac{\pi_B}{\pi_A}\cot^2\theta, \tag{13}$$

where $\pi_A = P(A)$ and $\pi_B = P(B)$.

Under the square-root (Bhattacharyya) embedding $\psi(\theta) = (\cos\theta, \sin\theta)$ of the Bernoulli simplex, the Fisher–Rao distance between the posteriors induced by two samples z and z' is

$$d_{\mathrm{FR}}(p(z), p(z')) = 2\,\mathrm{acos}\big(\langle\psi(\theta(z)), \psi(\theta(z'))\rangle\big) = 2\,\mathrm{acos}\big(\cos(\theta(z) - \theta(z'))\big) = 2|\theta(z) - \theta(z')|, \tag{14}$$

since $\theta \in [0, \pi/2]$. Thus, at the sample level, differences in θ correspond exactly to Fisher–Rao distances between induced Bernoulli posteriors.

6.2 FISHER–RAO DISTANCE BETWEEN θ-HISTOGRAMS AND POSTERIOR AMBIGUITY

We now move from individual samples to distributions over angles. Fix a binning of $[0, \pi/2]$ into m bins and let $P = (P_i)$ and $Q = (Q_i)$ be normalized histograms estimating the class-conditional likelihoods $P(\theta \in \text{bin } i \mid A)$ and $P(\theta \in \text{bin } i \mid B)$. Let π_A and π_B denote the class priors and define the mixture weights $m_i = \pi_A P_i + \pi_B Q_i$. For bin index $I = i$, define the binary label random variable $Y = \mathbf{1}\{\text{class} = A\}$, which equals 1 if the sample belongs to class A and 0 otherwise. The posterior distribution of Y given $I = i$ is

$$(Y \mid I = i) \sim \mathrm{Bernoulli}(r_i), \qquad r_i = P(A \mid I = i) = \frac{\pi_A P_i}{\pi_A P_i + \pi_B Q_i}. \tag{15}$$

The Fisher–Rao distance between the histograms, denoted by $d_{\mathrm{FR}}(P, Q)$, and defined as $d_{\mathrm{FR}}(P, Q) = 2\,\mathrm{acos}\big(\sum_{i=1}^m \sqrt{P_i Q_i}\big)$, satisfies

$$d_{\mathrm{FR}}(P, Q) = 2\,\mathrm{acos}\left(\sum_{i=1}^m m_i \sqrt{\frac{r_i(1 - r_i)}{\pi_A \pi_B}}\right) = 2\,\mathrm{acos}\left(\frac{\mathbb{E}_{I\sim m}[\mathrm{Std}(Y|I)]}{\sqrt{\pi_A \pi_B}}\right). \tag{16}$$

Hence, $d_{\mathrm{FR}}(P, Q)$ is a monotone *decreasing* function of the *expected posterior standard deviation* $\mathbb{E}[\mathrm{Std}(Y \mid I)]$: bins with $r_i \approx \frac{1}{2}$ contribute most to the standard deviation, while bins with $r_i \approx 0$

or 1 contribute little. Under the parametric posterior (12) and equal priors, samples with $\theta \approx \pi/4$ concentrate probability mass in posterior bins with $r_i \approx \frac{1}{2}$, whereas samples with $\theta \approx 0$ or $\pi/2$ concentrate probability mass in bins with r_i close to 1 or 0, respectively. This establishes a direct link between the linear-algebraic alignment angle θ, the distance between class-conditional likelihoods, and posterior uncertainty.

Numerical illustration. Figure 2 provides an empirical illustration of the theoretical picture developed above. For digit pairs with clearer geometric separation (e.g., "1" vs. "5", "3" vs. "9" and "0" vs. "7"), the class-conditional θ-histograms concentrate toward opposite extremes, with relatively little mass near $\theta = \pi/4$, indicating confident posterior regions. In contrast, visually similar pairs (e.g., "4" vs. "9") exhibit increased mass around $\theta \approx \pi/4$, reflecting higher posterior ambiguity. This behavior is consistent with the Fisher–Rao analysis: overlap near $\pi/4$ corresponds to bins with $r_i \approx \frac{1}{2}$ that dominate the Bhattacharyya coefficient and reduce histogram-level Fisher–Rao distance.

Together, the results presented in this section complete the connection between θ as introduced in Definition 1 and probabilistic notions of posterior uncertainty through Fisher–Rao geometry.

7 Conclusion

We introduced a geometry-grounded approach to dataset comparison built around a single primitive, the co-span relation $Ax = By = z$, and operationalized it through a GSVD joint frame. The resulting alignment angle $\theta(z)$ provides a minimal, interpretable summary of *relative* explanatory power: small angles indicate that z is represented more economically by A, large angles favor B, and values near $\pi/4$ correspond to shared structure. Beyond yielding a simple illustrative decision rule, the same GSVD factors expose representative directions (including extremes) that act as visual diagnostics of what is shared and what is dataset-specific. Empirically, MNIST digit pairs exhibit characteristic class-conditional θ distributions whose overlap (or lack thereof) matches intuitive notions of separability. The Fisher–Rao distance between θ-histograms complements these plots with a scalar, information-geometric separation metric, linking "dataset similarity" to posterior ambiguity induced by the angle.

This work is limited to two domains and controlled pixel-level datasets. Several extensions are left for future work. First, the framework can be generalized from dataset pairs to multiple domains, for instance by aggregating pairwise angles into a simplex-valued score or using multiway GSVD-style constructions. Second, the robustness of the alignment angle has not yet been characterized; future work should study the sensitivity of θ to preprocessing, rank selection, and the regularization or truncation of $C^\dagger$ and $S^\dagger$, as well as its behavior under noise and partial mismatch when $z \notin \mathrm{col}(A) \cap \mathrm{col}(B)$. Finally, while our experiments focus on raw pixel representations for interpretability, the framework applies more broadly to arbitrary feature spaces. In particular, matrices A and B can be constructed from feature embeddings produced by pretrained models such as convolutional networks or transformers. In such representations the linear subspace assumption may be more appropriate, since modern embeddings are designed to linearize semantic structure. Exploring the behavior of the alignment angle in such representation spaces is therefore a natural direction for future work.

References

Åke Björck and Gene H. Golub. Numerical methods for computing the angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

Moody T. Chu, Robert E. Funderlic, and Gene H. Golub. On a variational formulation of the generalized singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 18 (4):1082–1092, 1997.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13(Mar):795–828, 2012.

Alan Edelman and Yuyang Wang. The GSVD: Where are the ellipses?, matrix trigonometry, and more. *SIAM Journal on Matrix Analysis and Applications*, 41(4):1826–1856, 2020.

Gene Golub and Charles Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4th edition, 2013.

J. C. Gower. Generalized Procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Jiayuan Huang, Arthur Gretton, Karsten M. Borgwardt, Bernhard Schölkopf, and Alexander J. Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, 2007.

Andrew V. Knyazev and Merico E. Argentati. Principal angles between subspaces in an a-based scalar product: algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 23 (6):2008–2040, 2002.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pp. 2278–2324, 1998.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Christopher Paige and Michael Saunders. Towards a generalized singular value decomposition. *SIAM Journal on Numerical Analysis*, 18(3):398–405, 1981.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

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

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *arXiv preprint arXiv:1706.05806*, 2017.

G. W. Stewart. On the early history of the singular value decomposition. *SIAM Review*, 35(4): 551–566, 1993.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

Charles Van Loan. Generalizing the singular value decomposition. *SIAM Journal on Numerical Analysis*, 13(1):76–83, 1976.

Lutao Zheng, Guanjun Liu, Chungang Yan, Changjun Jiang, Mengchu Zhou, and Maozhen Li. Improved TrAdaBoost and its application to transaction fraud detection. *IEEE Transactions on Computational Social Systems*, 7(5):1304–1316, 2020.

A NOTATION

Table 1 summarizes the notation used throughout this work.

Table 1: Notation used throughout the paper.

Symbol	Meaning
d	Ambient dimension (e.g., 784 for 28×28 MNIST images)
p, q	Number of columns (samples or basis vectors) used to form A and B
$A \in \mathbb{R}^{d\times p}$, $B \in \mathbb{R}^{d\times q}$	Dataset matrices (columns are observations in $\mathbb{R}^d$)
a_i, b_j	Columns of A and B (individual observations)
$\mathrm{col}(A)$, $\mathrm{col}(B)$	Column spaces (subspaces spanned by columns of A and B)
$x \in \mathbb{R}^p$, $y \in \mathbb{R}^q$	Coefficients such that Ax and By represent an ambient vector
$z \in \mathbb{R}^d$	Sample / ambient vector under analysis
$Ax = By = z$	Co-span relation (same z representable by both A and B)
$\mathcal{R}_{\text{co-span}}$	Relation set $\{(x, y) : Ax = By\}$
$\mathcal{R}_{\text{co-span}}(z)$	Fiber $\{(x, y) : Ax = By = z\}$
$\mathrm{Ker}(A)$, $\mathrm{Ker}(B)$	Nullspaces of A and B
$x \perp \mathrm{Ker}(A)$, $y \perp \mathrm{Ker}(B)$	Minimum-ℓ_2 coefficient choice among solutions of $Ax = By = z$
$A = HCU$, $B = HSV$	GSVD form used in the paper
$H \in \mathbb{R}^{d\times d}$	Shared (invertible/left-invertible) ambient frame matrix
$U \in \mathbb{R}^{p\times p}$, $V \in \mathbb{R}^{q\times q}$	Orthogonal factors in the GSVD
C, S	(Block-)diagonal GSVD factors with $C^\top C + S^\top S = I$
r, k, t	Block sizes in (5) with $r + k + t = d$
$\widetilde{C}, \widetilde{S} \in \mathbb{R}^{k\times k}$	Positive diagonal "shared" blocks in (5)
h_i	i-th column of H
$H^\dagger$	Moore–Penrose pseudoinverse of H
$C^\dagger$, $S^\dagger$	Moore–Penrose pseudoinverses (often truncated/regularized in practice)
$c(z) = H^\dagger z$	Coordinates of z in the GSVD ambient frame (least-squares if needed)
$a(z) = \|C^\dagger c(z)\|_2$,	GSVD-weighted representation costs for A
$b(z) = \|S^\dagger c(z)\|_2$	GSVD-weighted representation costs for B
$\theta(z) \in [0, \pi/2]$	Alignment angle $\arctan(\|x\|_2/\|y\|_2) = \arctan(a(z)/b(z))$
τ	Classification threshold on $\theta(z)$ (default $\pi/4$)
e_k	k-th canonical basis vector
$z_{\max}, z_{\min}$	Extreme directions maximizing/minimizing $\theta(z)$ over a constraint set $\mathcal{Z}$
P, Q	Normalized histograms approximating $p(\theta \mid A)$ and $p(\theta \mid B)$
$d_{\mathrm{FR}}(P, Q)$	Fisher–Rao distance between histograms (Eq. (32))

B Angle Score and Minimal Classifier Algorithm

Algorithm 1 computes the GSVD angle score (7) and uses it in for classification as described in Section 4.4.

Matrices A and B are constructed in line 1, and their GSVD is computed in line 2. For each test sample, projection onto the shared basis is performed in line 4, class-specific values are computed in line 5, and the angle score is formed in line 6 before thresholding in line 7.

Algorithm 1 GSVD angle score and a minimal classifier

Require: Two labeled datasets $\mathcal{D}_A, \mathcal{D}_B \subset \mathbb{R}^d$, ranks p, q, threshold τ
1: Build $A \in \mathbb{R}^{d\times p}$ and $B \in \mathbb{R}^{d\times q}$ from training samples or low-rank bases
2: Compute GSVD factors (H, C, S, U, V) such that $A = HCU$ and $B = HSV$
3: **for** each test sample $z \in \mathbb{R}^d$ **do**
4: $\quad c \leftarrow H^\dagger z$ ▷ least-squares projection
5: $\quad a \leftarrow \left\|C^\dagger c\right\|_2, \;\; b \leftarrow \left\|S^\dagger c\right\|_2$
6: $\quad \theta(z) \leftarrow \arctan(a/b)$
7: $\quad$ Predict $\hat{\ell}(z) \leftarrow A$ if $\theta(z) \le \tau$, else B
8: **end for**

Remark 2. Note on Dimensionality Reduction. Exploiting the structure of H and the characterization of its extreme directions discussed in Section 4.3, a practical micro-optimization for improving

classification efficiency without materially degrading performance is to discard directions of H that receive comparable weights from $C^\dagger$ and $S^\dagger$. Intuitively, such directions encode features that are largely shared between the two datasets and therefore contribute little discriminative information when assessing whether a given instance is more strongly aligned with A or with B.

C SPAN VIEW: INSTANCE ALIGNMENT AND A GSVD-BASED RELATIVE PROCRUSTES

This appendix records a complementary interpretation of the same linear relation studied in the main text. While our pipeline is driven by the co-span constraint $Ax = By = z$ (dimension alignment), the same compatibility can be read in a *span* form that emphasizes *instance alignment*.

C.1 SPAN PARAMETERIZATION OF A LINEAR RELATION

A linear relation between two subspaces can be described implicitly (co-span) by

$$\mathcal{R}_{\text{co-span}} = \{(x, y) : Ax = By\},$$

or explicitly (span) by introducing a shared latent parameter w:

$$\mathcal{R}_{\text{span}} = \{(x, y) : \exists\, w \text{ such that } x = Fw,\ y = Gw\}. \tag{17}$$

In this view, paired coordinates (x, y) arise from the *same* latent code w. Geometrically, span emphasizes *instance-level pairing*: one latent point generates a compatible point in each domain.

C.2 RELATIVE PROCRUSTES AS INSTANCE ALIGNMENT

Suppose we have two sets of instance vectors in a shared ambient space (e.g., embeddings)

$$X = [x_1, \dots, x_n] \in \mathbb{R}^{d\times n}, \qquad Y = [y_1, \dots, y_n] \in \mathbb{R}^{d\times n}.$$

A classical alignment primitive is (orthogonal) Procrustes:

$$R^\star \in \operatorname*{arg\,min}_{R^\top R = I} \|RX - Y\|_F^2. \tag{18}$$

In the dataset-comparison setting we often do not want to align in the raw ambient basis, but rather in a *relative geometry* that discounts directions that are not shared and emphasizes directions that are shared.

C.3 GSVD PROVIDES THE NATURAL COORDINATE SYSTEM

Let the two datasets define matrices A and B and consider their GSVD

$$A = HCU, \qquad B = HSV, \qquad C^\top C + S^\top S = I.$$

The shared frame H defines coordinates for any $z \in \mathbb{R}^d$ via $c = H^\dagger z$. Thus instance sets admit GSVD-frame representations

$$\widetilde{X} = H^\dagger X, \qquad \widetilde{Y} = H^\dagger Y.$$

Instance alignment can then be performed in this frame, rather than in the raw ambient basis.

C.4 ANGLE-BASED TRUNCATION FOR RELATIVE ALIGNMENT

The diagonal structure of (C, S) suggests a truncation rule: directions where both factors are non-negligible correspond to shared structure. Let $\mathcal{I}_{\text{shared}}$ denote indices deemed shared (e.g., by thresholding $\min\{c_{ii}, s_{ii}\}$) and define

$$P_{\text{shared}} = \operatorname{diag}(p_1, \dots, p_d), \quad p_i = \mathbf{1}\{i \in \mathcal{I}_{\text{shared}}\}.$$

Then truncated coordinates are $\widetilde{X}_{\text{sh}} = P_{\text{shared}}\widetilde{X}$ and $\widetilde{Y}_{\text{sh}} = P_{\text{shared}}\widetilde{Y}$, and a GSVD-aware Procrustes alignment is

$$R^\star_{\text{sh}} \in \operatorname*{arg\,min}_{R^\top R = I} \|R\widetilde{X}_{\text{sh}} - \widetilde{Y}_{\text{sh}}\|_F^2. \tag{19}$$

This objective focuses alignment on the shared geometry, ignoring weakly shared directions.

D Analytical Properties and Proofs for the GSVD-Based Framework

We begin with some initial definitions. Let $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)$ and $\tilde{y} = (\tilde{y}_1, \tilde{y}_2, \tilde{y}_3)$ where the dimensions correspond to the blocks of C and S,

$$C\tilde{x} = \begin{bmatrix} I_r & 0 & 0 \\ 0 & \widetilde{C} & 0 \\ 0 & 0 & 0_t \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{bmatrix} = \begin{bmatrix} \tilde{x}_1 \\ \widetilde{C}\,\tilde{x}_2 \\ 0 \end{bmatrix},$$

$$S\tilde{y} = \begin{bmatrix} 0_r & 0 & 0 \\ 0 & \widetilde{S} & 0 \\ 0 & 0 & I_t \end{bmatrix} \begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \tilde{y}_3 \end{bmatrix} = \begin{bmatrix} 0 \\ \widetilde{S}\,\tilde{y}_2 \\ \tilde{y}_3 \end{bmatrix},$$

where $\tilde{x}_1 \in \mathbb{R}^r$, $\tilde{x}_2 \in \mathbb{R}^k$ and $\tilde{x}_3 \in \mathbb{R}^t$. Similarly, $\tilde{y}_1 \in \mathbb{R}^r$, $\tilde{y}_2 \in \mathbb{R}^k$ and $\tilde{y}_3 \in \mathbb{R}^t$.

If $C\tilde{x} = S\tilde{y}$, then componentwise

$$\tilde{x}_1 = 0, \qquad \widetilde{C}\,\tilde{x}_2 = \widetilde{S}\,\tilde{y}_2, \qquad \tilde{y}_3 = 0.$$

In what follows, we let

$$\tilde{x} := Ux, \quad \tilde{y} := Vy. \tag{20}$$

Then, our goal in this appendix is twofold. First, we show that

$$Ax = By \Longleftrightarrow C\tilde{x} = S\tilde{y}. \tag{21}$$

with

$$\tilde{x} = \begin{bmatrix} 0 \\ \widetilde{S}w \\ \tilde{x}_3 \end{bmatrix}, \qquad \tilde{y} = \begin{bmatrix} \tilde{y}_1 \\ \widetilde{C}w \\ 0 \end{bmatrix}, \tag{22}$$

where $\tilde{x}_3$ and $\tilde{y}_1$ are free and $w \in \mathbb{R}^k$.

Second, we show that if $x \perp \mathrm{Ker}(A)$ and $y \perp \mathrm{Ker}(B)$ then each solution (x, y) of $Ax = By$ uniquely characterizes a pair $(\tilde{x}, \tilde{y})$,

$$\tilde{x} = \begin{bmatrix} 0 \\ \widetilde{S}w \\ 0 \end{bmatrix}, \qquad \tilde{y} = \begin{bmatrix} 0 \\ \widetilde{C}w \\ 0 \end{bmatrix}. \tag{23}$$

Lemma 1 (GSVD induces an explicit parameterization of the co-span constraint)**.** Under Definition 2,

$$Ax = By \Longleftrightarrow C\tilde{x} = S\tilde{y}$$

where

$$\begin{bmatrix} \tilde{x}_1 \\ \widetilde{C}\,\tilde{x}_2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \widetilde{S}\,\tilde{y}_2 \\ \tilde{y}_3 \end{bmatrix}$$

and $\tilde{x}_3$ and $\tilde{y}_1$ are free. Moreover, there exists a vector w such that

$$\tilde{x} = \begin{bmatrix} 0 \\ \widetilde{S}\,w \\ \tilde{x}_3 \end{bmatrix}, \qquad \tilde{y} = \begin{bmatrix} \tilde{y}_1 \\ \widetilde{C}\,w \\ 0 \end{bmatrix},$$

with $\tilde{x}_3$ and $\tilde{y}_1$ free.

Proof. The proof proceeds as the following sequence of equivalence steps:

$$\begin{aligned} & Ax = By \\ \Longleftrightarrow \quad & \{ \text{ GSVD definition } \} \\ & HCUx = HSVy. \\ \Longleftrightarrow \quad & \{ \text{ By GSVD definition } \exists \bar{H} \text{ such that } \bar{H}H = I \quad \} \end{aligned}$$

$$
\begin{aligned}
& CUx = SVy \\
\iff & \quad \{ \text{ Change of Variables: } \tilde{x} = Ux \text{ and } \tilde{y} = Vy \ \} \\
& C\tilde{x} = S\tilde{y} \\
\iff & \quad \{ \text{ We can expand the } C, S \text{ matrices by the definition in (3.2) to create: } \} \\
& \begin{bmatrix} I_r & 0 & 0 \\ 0 & \widetilde{C} & 0 \\ 0 & 0 & 0_t \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{bmatrix} = \begin{bmatrix} 0_r & 0 & 0 \\ 0 & \widetilde{S} & 0 \\ 0 & 0 & I_t \end{bmatrix} \begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \tilde{y}_3 \end{bmatrix} \\
\iff & \quad \{ \text{ Algebra } \} \\
& \begin{bmatrix} \tilde{x}_1 \\ \widetilde{C}\,\tilde{x}_2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \widetilde{S}\,\tilde{y}_2 \\ \tilde{y}_3 \end{bmatrix} \text{ with } \tilde{x}_3 \text{ and } \tilde{y}_1 \text{ free.}
\end{aligned}
$$

We now focus on the middle block. For each index i, the relation

$$\widetilde{C}_i\,\tilde{x}_{2i} = \widetilde{S}_i\,\tilde{y}_{2i}$$

holds. Since $\widetilde{C}_i > 0$ and $\widetilde{S}_i > 0$, this equality can be reparameterized by introducing a scalar w_i such that

$$\tilde{x}_{2i} = \widetilde{S}_i\,w_i, \qquad \tilde{y}_{2i} = \widetilde{C}_i\,w_i.$$

Because $\widetilde{C}$ and $\widetilde{S}$ are diagonal, this construction applies entrywise. Collecting the coordinates yields the vector form

$$\tilde{x}_2 = \widetilde{S}\,w, \qquad \tilde{y}_2 = \widetilde{C}\,w, \qquad w \in \mathbb{R}^k.$$

To find the values of $\tilde{x}_1, \tilde{y}_3$ we can look at the equality restriction above.

Finally, the explicit parametrization of $\tilde{x}$ and $\tilde{y}$ is:

$$\tilde{x} = \begin{bmatrix} 0 \\ \widetilde{S}\,w \\ \tilde{x}_3 \end{bmatrix}, \qquad \tilde{y} = \begin{bmatrix} \tilde{y}_1 \\ \widetilde{C}\,w \\ 0 \end{bmatrix},$$

□

Lemma 2 (Nullspace orthogonality fixes the free GSVD blocks)**.** Under Definition 2,

$$
\begin{aligned}
x \perp \mathrm{Ker}(A) \quad &\iff \quad \tilde{x}_3 = 0. \\
y \perp \mathrm{Ker}(B) \quad &\iff \quad \tilde{y}_1 = 0.
\end{aligned}
$$

Proof. We show the claim for A; the proof for B is analogous.

$$
\begin{aligned}
& x \perp \mathrm{Ker}(A) \\
\iff & \quad \{ \text{ GSVD definition } \} \\
& x \perp \mathrm{Ker}(HCU) \\
\iff & \quad \{ \text{ By GSVD definition } \exists \bar{H} \text{ such that } \bar{H}H = I \ \} \\
& x \perp \mathrm{Ker}(CU) \\
\iff & \quad \{ \text{ Perpendicularity to kernel definition } \} \\
& \langle x, k \rangle = 0, \forall k \text{ s.t. } CUk = 0 \\
\iff & \quad \{ \text{ Change of variable: } Uk = \tilde{k}, \text{ so } k = U^T\tilde{k} \ \} \\
& \langle x, U^T\tilde{k} \rangle = 0, \forall \tilde{k} \text{ s.t. } C\tilde{k} = 0 \\
\iff & \quad \{ \text{ Duality of inner product } \} \\
& \langle Ux, \tilde{k} \rangle = 0, \forall \tilde{k} \text{ s.t. } C\tilde{k} = 0 \\
\iff & \quad \{ \ \tilde{x} \text{ definition } \} \\
& \langle \tilde{x}, \tilde{k} \rangle = 0, \forall \tilde{k} \text{ s.t. } C\tilde{k} = 0
\end{aligned}
$$

Expanding the definition of C we have

$$\begin{bmatrix} I_r & 0 & 0 \\ 0 & \widetilde{C} & 0 \\ 0 & 0 & 0_t \end{bmatrix} \begin{bmatrix} \tilde{k}_1 \\ \tilde{k}_2 \\ \tilde{k}_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

which forces $\tilde{k}_1 = 0$ and $\tilde{k}_2 = 0$ while keeping $\tilde{k}_3$ unconstrained. Therefore, to guarantee that $\langle x, k \rangle = 0$, we must enforce $\tilde{x}_3 = 0$ and leave $\tilde{x}_1$ and $\tilde{x}_2$ unconstrained. □

Theorem 1 (GSVD produces alignment angles)**.** Given datasets A and B, and an observation z, the alignment angle between A and B with respect to z (cf. Definition 1) is given by

$$\theta(z) = \arctan\left(\frac{a(z)}{b(z)}\right) = \arctan\left(\frac{\left\|C^\dagger\, c(z)\right\|_2}{\left\|S^\dagger\, c(z)\right\|_2}\right), \tag{24}$$

where

$$c(z) = H^\dagger z, \tag{25}$$

and (H, C, S, U, V) are the GSVD factors in (2).

Proof. For a given vector $z \in \mathbb{R}^d$, we express z in the shared GSVD frame by computing its coordinates with respect to $\mathrm{col}(H)$:

$$\underset{c \in \mathbb{R}^d}{\arg\min}\ \|Hc - z\|_2^2, \qquad \text{so that} \qquad c = H^\dagger z. \tag{26}$$

This least-squares projection provides the natural notion of GSVD-frame coordinates when z is noisy or does not lie exactly in $\mathrm{col}(H)$; when $z \in \mathrm{col}(H)$, the equality $z = Hc$ holds exactly. Under the GSVD parametrization, the co-span constraint $Ax = By = z$ implies the following chain of equivalences:

$$\begin{aligned}
& Ax = By = z \\
\iff\ & \{\ A = HCU,\ B = HSV,\ z = Hc\ \} \\
& (HCU)x = (HSV)y = Hc \\
\iff\ & \{\ \text{By GSVD definition } \exists \bar{H} \text{ such that } \bar{H}H = I\ \} \\
& CUx = SVy = c \\
\iff\ & \{\ \text{Change of variables: } \tilde{x} := Ux,\ \tilde{y} := Vy\ \} \\
& C\tilde{x} = S\tilde{y} = c \\
\iff\ & \{\ \text{By Lemma (1) }\} \\
& \begin{bmatrix} \tilde{x}_1 \\ \widetilde{C}\,\tilde{x}_2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \widetilde{S}\,\tilde{y}_2 \\ \tilde{y}_3 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} \qquad \text{with } \tilde{x}_3 \text{ and } \tilde{y}_1 \text{ free} \\
\iff\ & \{\ \text{The equality tells us that } \tilde{x}_1 = c_1 = 0 \text{ and } \tilde{y}_3 = c_3 = 0\ \} \\
& \begin{bmatrix} 0 \\ \widetilde{C}\,\tilde{x}_2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \widetilde{S}\,\tilde{y}_2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ c_2 \\ 0 \end{bmatrix} \qquad \text{with } \tilde{x}_3 \text{ and } \tilde{y}_1 \text{ free} \\
\iff\ & \{\ \text{Using Lemma 2 }\} \\
& \begin{bmatrix} 0 \\ \widetilde{C}\,\tilde{x}_2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \widetilde{S}\,\tilde{y}_2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ c_2 \\ 0 \end{bmatrix} \\
\iff\ & \{\ \text{Focus on the middle term }\} \\
& \widetilde{C}\,\tilde{x}_2 = \widetilde{S}\,\tilde{y}_2 = c_2 \\
\iff\ & \{\ \text{Isolating expressions }\} \\
& \widetilde{C}\,\tilde{x}_2 = c_2 \qquad \widetilde{S}\,\tilde{y}_2 = c_2 \\
\iff\ & \{\ \text{Solving for } \tilde{x}_2 \text{ and } \tilde{y}_2\ \} \\
& \tilde{x}_2 = \widetilde{C}^{-1} c_2, \quad \tilde{y}_2 = \widetilde{S}^{-1} c_2 \\
\iff\ & \{\ \text{Now consider the ratio expression }\} \\
& \tfrac{\|x\|_2}{\|y\|_2} \\
=\ & \{\ U, V \text{ orthogonal }\}
\end{aligned}$$

$$
\begin{aligned}
& \frac{\|\tilde{x}\|_2}{\|\tilde{y}\|_2} \\
= & \quad \{ \text{ Norm of middle term } \} \\
& \frac{\|\widetilde{C}^{-1} c_2\|_2}{\|\widetilde{S}^{-1} c_2\|_2} \\
= & \quad \{ \text{ Accounting for all blocks } \} \\
& \frac{\|C^{\dagger} c\|_2}{\|\widetilde{S}^{\dagger} c\|_2}
\end{aligned}
$$

Because the boundary entries of c are constrained to be zero, the resulting system admits multiple matrices that satisfy the final condition. To ensure a unique solution, we therefore employ the Moore–Penrose pseudoinverse. □

Next, we consider the problem of finding extreme directions. Recall that the direction z that maximizes the alignment angle $\theta(z)$ is given by Theorem 2,

$$
\arg\max_{x,y} \arctan\left(\frac{\|x\|_2^2}{\|y\|_2^2}\right) \quad \text{s.t.} \quad Ax = By, \;\; (x,y) \in \mathcal{R}_{\text{co-span}}(z), \;\; x \perp \operatorname{Ker}(A), \;\; y \perp \operatorname{Ker}(B).
$$

Theorem 2 (GSVD produces extreme directions)**.** Given datasets A and B, the extreme directions z_{max} and z_{min} that maximize and minimize the alignment angle $\theta(z)$, respectively, are unique and given by $z_{\text{max}} = Ax_{\text{max}} = By_{\text{max}}$ and $z_{\text{min}} = Ax_{\text{min}} = By_{\text{min}}$, where $z_{\text{max}} = h_{r+k}$, $z_{\text{min}} = h_{r+1}$ and

$$
x_{\text{max}} = U^T \begin{bmatrix} 0 \\ \widetilde{S}\widetilde{C}^{-1} e_k \\ 0 \end{bmatrix}, \quad y_{\text{max}} = V^T \begin{bmatrix} 0 \\ e_k \\ 0 \end{bmatrix}, \quad x_{\text{min}} = U^T \begin{bmatrix} 0 \\ \widetilde{C}\widetilde{S}^{-1} e_1 \\ 0 \end{bmatrix}, \quad y_{\text{min}} = V^T \begin{bmatrix} 0 \\ e_1 \\ 0 \end{bmatrix}.
$$

(H, C, S, U, V) are the GSVD factors in (2), e_i is the i-th canonical basis vector, i.e., $(e_i)_j = 1$ if $j = i$ and $(e_i)_j = 0$ otherwise, and h_i is the i-th column of matrix H.

Proof. Since $\arctan$ is strictly monotonic increasing, we can focus on finding $\arg\max_{x,y} \frac{\|x\|_2^2}{\|y\|_2^2}$ with the same restraints.

By Lemma 1, $Ax = By$ implies $\tilde{x}_1 = 0$, $\tilde{y}_3 = 0$, and $\tilde{x}_2 = \widetilde{S}w$, $\tilde{y}_2 = \widetilde{C}w$ for some w, with $\tilde{x}_3$ and $\tilde{y}_1$ free.

By Lemma 2, the additional constraints $x \perp \operatorname{Ker}(A)$ and $y \perp \operatorname{Ker}(B)$ force $\tilde{x}_3 = 0$ and $\tilde{y}_1 = 0$.

Thus only the middle blocks remain, i.e., $\tilde{x} = (0, \widetilde{S}w, 0)$ and $\tilde{y} = (0, \widetilde{C}w, 0)$.

Since U and V are orthogonal, $\|x\|_2 = \|\tilde{x}\|_2$ and $\|y\|_2 = \|\tilde{y}\|_2$. Therefore

$$
\frac{\|x\|_2^2}{\|y\|_2^2} = \frac{\|\widetilde{S}w\|_2^2}{\|\widetilde{C}w\|_2^2}.
$$

As defined in (3.2) the diagonal entries of the middle matrices are strictly positive, so they have inverses. Now we denote $\tilde{w} = \widetilde{C}w$ and thus $w = \widetilde{C}^{-1}\tilde{w}$:

$$
\frac{\|\widetilde{S}\widetilde{C}^{-1}\tilde{w}\|_2^2}{\|\widetilde{C}\widetilde{C}^{-1}\tilde{w}\|_2^2} = \frac{\|\widetilde{S}\widetilde{C}^{-1}\tilde{w}\|_2^2}{\|\tilde{w}\|_2^2}
$$

The ratio becomes a Rayleigh quotient for the diagonal operator $\widetilde{S}\widetilde{C}^{-1}$, hence is maximized by concentrating all mass on the index of its largest diagonal entry, i.e., $\tilde{w} = e_k$.

Reconstruction of (x, y) follows from $\tilde{x} = Ux$ and $\tilde{y} = Vy$.

Finally, to uncover the desired direction we use substitution on $Ax = By$:

$$
By = HSV(V^{\top}\tilde{y}) = HS\tilde{y} = H \begin{bmatrix} 0 \\ \widetilde{S}e_k \\ 0 \end{bmatrix} = \begin{bmatrix} \widetilde{S}e_k \\ 0 \end{bmatrix} = \tilde{s}_k H \begin{bmatrix} 0 \\ e_k \\ 0 \end{bmatrix} = \tilde{s}_k \cdot h_{r+k}
$$

and

$$Ax = HCU(U^\top \tilde{x}) = HC\tilde{x} = H \begin{bmatrix} 0 \\ \widetilde{C}(\widetilde{S}\widetilde{C}^{-1}e_r) \\ 0 \end{bmatrix} = H \begin{bmatrix} 0 \\ \widetilde{S}e_k \\ 0 \end{bmatrix} = \tilde{c}_k \tilde{s}_k \frac{1}{\tilde{c}_k} H \begin{bmatrix} 0 \\ e_k \\ 0 \end{bmatrix} = \tilde{s}_k \cdot h_{r+k}$$

Note that as $\widetilde{S}, \widetilde{C}, \widetilde{C}^{-1}$ are diagonal and e_k is a canonical basis vector, the operation is simply a multiplication by the scalars $\tilde{s}_k, \tilde{c}_k, \frac{1}{\tilde{c}_k}$.

We can then normalize the vector h_{r+k}. □

Remark 3. Practical notes on numerical stability. In practice, we compute $C^\dagger$ and $S^\dagger$ by inverting diagonal entries above a tolerance (or via truncated SVD / Tikhonov regularization). We also recommend reporting sensitivity of θ to rank choices (p, q) and preprocessing (centering/normalization).

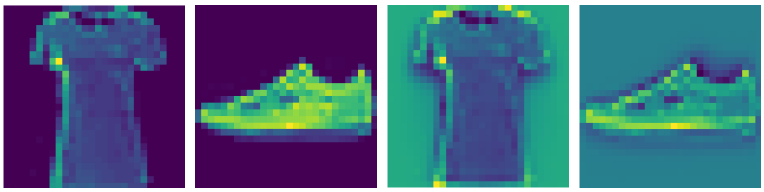
E MNIST FASHION: ANGLE HISTOGRAMS UNDER THE SAME GSVD SETUP

We replicate the MNIST-digit protocol of Section 5 on Fashion-MNIST, keeping the *same preprocessing* (mean centering in the ambient pixel space), the *same GSVD pipeline*, and the *same rank choices* (p, q) and train/test split logic used in the digit experiments. As before, we construct $A \in \mathbb{R}^{d\times p}$ and $B \in \mathbb{R}^{d\times q}$ by stacking (class-specific) training images as columns (after preprocessing), compute the GSVD factors, and evaluate $\theta(z)$ for each test sample z from the two classes.

Figure 4 illustrates representative Fashion-MNIST samples used in our GSVD pipeline, showing both raw images and their corresponding vectors after mean-centering and vectorization.

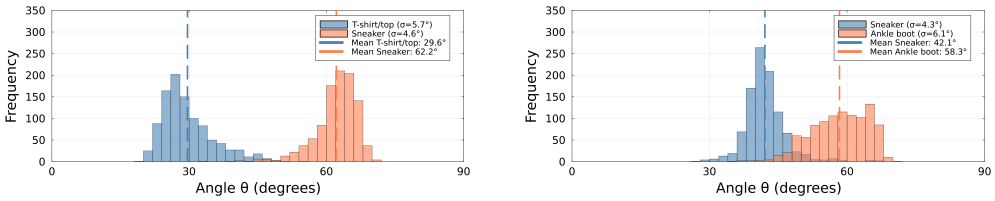
Figure 5 reports empirical histograms of θ for two representative class pairs. In Fig. 5(a) (`T-shirt/top` vs. `Sneaker`), the two distributions are well separated, indicating limited shared structure between upper-body garments and footwear under this linear subspace view. In contrast, Fig. 5(b) (`Sneaker` vs. `Ankle boot`) exhibits substantial overlap, consistent with the fact that both classes correspond to visually related types of shoes and therefore share more dominant directions.

These results qualitatively mirror the behavior observed on MNIST digits: class pairs with stronger visual or semantic similarity tend to produce overlapping angle distributions, whereas more distinct pairs yield clearer separation. Importantly, no dataset-specific tuning is introduced for Fashion-MNIST, suggesting that the angle score reflects intrinsic geometric relations between class-conditioned subspaces rather than artifacts of hyperparameter choices.



(a) (b) (c) (d)

Figure 4: Representative MNIST Fashion samples for the GSVD pipeline. (a,b) raw "`T-shirt`" and "`Sneakers`"; (c,d) corresponding vectors after mean-centering difference.



(a) (b)

Figure 5: Empirical distributions of the angle θ on the test set for two Fashion-MNIST class pairs under the same GSVD pipeline used for MNIST digits (same preprocessing and the same rank choices (p, q)). (a) `T-shirt/top` (class 0) vs. `Sneaker` (class 7), where A is built from `T-shirt/top` columns and B from `Sneaker` columns; the separation indicates limited shared structure. (b) `Sneaker` (class 7) vs. `Ankle boot` (class 9), which shows larger overlap, consistent with more shared directions between these visually related footwear classes. As in the main text, angles near 0 indicate stronger alignment with A and angles near $90°$ indicate stronger alignment with B.

F GLOBAL SEPARATION VIA FISHER–RAO DISTANCE

Next, we report additional results on the separation via Fisher–Rao distance presented in Section 5.

The per-sample angle $\theta(z)$ characterizes local alignment, while the class-conditional distributions $p(\theta \mid A)$ and $p(\theta \mid B)$ capture *global* geometric separation between datasets. We quantify this separation by measuring the discrepancy between the two θ-histograms with the Fisher–Rao distance on the probability simplex.

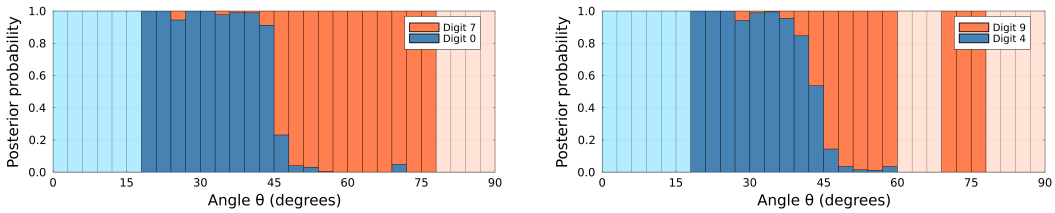
Table 2 reports Fisher–Rao distances (in radians) between class-conditional θ-histograms for several MNIST digit pairs, using 30 uniformly spaced bins over $[0°, 90°]$. Larger Fisher–Rao distances indicate less histogram overlap and thus stronger geometric separability.

The results agree with visual intuition and the underlying angle histograms. Pairs such as $(0, 7)$ and $(1, 5)$ achieve the largest distances, reflecting minimal overlap between their θ distributions and showing that each digit is, on average, represented much more efficiently by its own class-specific subspace than by the other class's. In contrast, $(4, 9)$ yields a much smaller distance, indicating greater overlap and hence higher geometric ambiguity, consistent with the well-known visual similarity between handwritten 4s and 9s.

Table 2: Fisher–Rao distance (in radians) between the class-conditional θ histograms for selected MNIST digit pairs (30 bins over $[0°, 90°]$).

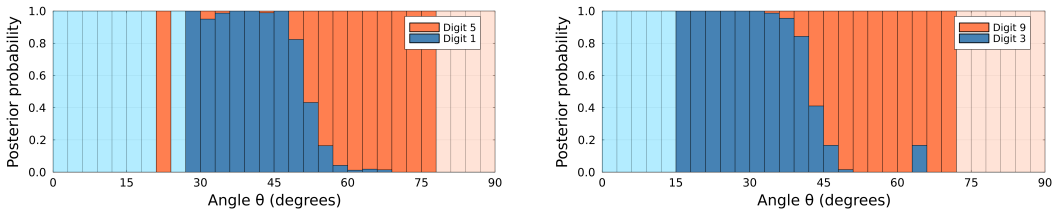
Digit pair (A, B)	Fisher–Rao distance (rad)
$(0, 7)$	2.773465
$(1, 5)$	2.731825
$(3, 9)$	2.665724
$(4, 9)$	2.336575

Figure 6 illustrates the posterior uncertainty described in Section 6.2. Using 1600 images across both digits, the x-axis is discretized into equally spaced angle bins, while the y-axis represents the posterior class probabilities defined in (15). As developed throughout this paper, the posterior distributions reflect the principle that smaller alignment angles are associated with instances from A while larger angles are associated with instances from B intermediate angles capture regions of mutual coherence between the two classes. In panels (a), (b), and (d), this expected structure is clearly observed, with dominant color regions corresponding to their respective angle ranges, even though certain bins are not populated by samples in the test set. In panel (c), an outlier is observed: a single instance from class$B$ is assigned a low alignment angle. Aside from this isolated case, the distribution conforms to the established interpretation.



(a) Digits “0” vs “7”

(b) Digits “4” vs “9”



(c) Digits “1” vs “5”

(d) Digits “3” vs “9”

Figure 6: Posterior distributions across discretized angle bins for four digit pairs in the MNIST test set. Angles near 0 reflect stronger alignment with the first digit in the pair, whereas angles near 90 reflect stronger alignment with the second. Translucent bins indicate angle ranges for which no samples were observed.

G Hellinger vs. Fisher–Rao (Bhattacharyya) geometry

This appendix relates the (squared) Hellinger distance between discrete distributions to the Fisher–Rao distance on the probability simplex. The key bridge is the *Bhattacharyya coefficient*, i.e., the inner product of square-root densities.

Let $P = (P_i)_{i=1}^m$ and $Q = (Q_i)_{i=1}^m$ be probability vectors on m bins, with $P_i \geq 0$, $Q_i \geq 0$, and $\sum_i P_i = \sum_i Q_i = 1$. Define the square-root embedding

$$\psi(P) := \left(\sqrt{P_1}, \ldots, \sqrt{P_m}\right) \in \mathbb{R}^m, \qquad \|\psi(P)\|_2 = 1, \tag{27}$$

and similarly $\psi(Q)$. Recall that the *Bhattacharyya coefficient* is

$$BC(P,Q) := \langle \psi(P), \psi(Q) \rangle = \sum_{i=1}^m \sqrt{P_i Q_i} \in [0,1]. \tag{28}$$

G.1 Hellinger distance as Euclidean distance in the square-root embedding

The (squared) Hellinger distance is

$$H^2(P,Q) := \frac{1}{2} \sum_{i=1}^m \left(\sqrt{P_i} - \sqrt{Q_i}\right)^2 = \frac{1}{2} \|\psi(P) - \psi(Q)\|_2^2. \tag{29}$$

Expanding the squared norm and using $\sum_i P_i = \sum_i Q_i = 1$ gives

$$\begin{aligned} \|\psi(P) - \psi(Q)\|_2^2 &= \|\psi(P)\|_2^2 + \|\psi(Q)\|_2^2 - 2\langle \psi(P), \psi(Q) \rangle \\ &= 1 + 1 - 2\,BC(P,Q) \\ &= 2\big(1 - BC(P,Q)\big). \end{aligned} \tag{30}$$

Plugging into (29) yields the identity

$$H^2(P,Q) = 1 - BC(P,Q) = 1 - \sum_{i=1}^m \sqrt{P_i Q_i}. \tag{31}$$

G.2 Fisher–Rao distance as spherical geodesic distance

Recall from Section 6.2 that under the square-root embedding, the Fisher–Rao distance on the probability simplex admits the closed form

$$d_{\mathrm{FR}}(P,Q) = 2\,\mathrm{acos}\left(\sum_{i=1}^{m}\sqrt{P_i Q_i}\right) = 2\,\mathrm{acos}\big(BC(P,Q)\big). \tag{32}$$

Geometrically, $d_{\mathrm{FR}}(P,Q)$ is twice the angle between the unit vectors $\psi(P)$ and $\psi(Q)$ on the sphere.

Combining (31) and (32) yields a direct conversion between Fisher–Rao and Hellinger distances. For probability vectors P, Q on m bins,

$$d_{\mathrm{FR}}(P,Q) = 2\,\mathrm{acos}\big(1 - H^2(P,Q)\big). \tag{33}$$

Equivalently,

$$H^2(P,Q) = 1 - \cos\left(\frac{d_{\mathrm{FR}}(P,Q)}{2}\right). \tag{34}$$

Indeed, by (31), $BC(P,Q) = 1 - H^2(P,Q)$. Substituting into (32) gives (33). Rearranging yields (34).

Relevance to θ-histograms. When P and Q are the normalized histograms estimating $p(\theta \mid A)$ and $p(\theta \mid B)$, the Bhattacharyya coefficient $BC(P,Q)$ measures histogram overlap in the square-root geometry, the Hellinger distance measures the corresponding Euclidean separation on the sphere (up to the $\frac{1}{\sqrt{2}}$ factor), and Fisher–Rao measures the intrinsic spherical geodesic distance. Eq. (32) shows these are equivalent up to a monotone transform.