WASSERSTEIN SINGULAR VECTORS AND METRIC DOWKER DUALITY FOR INTERPRETABLE FEATURE GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

The Wasserstein singular vector (WSV) problem asks for a metric on the samples and a metric on the features of a data matrix such that each reproduces the other through optimal transport. We prove existence for the general, unregularized WSV problem, for almost every matrix, and show that degenerate solutions exist exactly when the matrix has a nontrivial lumping. While they determine each other, the WSV sample and feature metrics do not generally agree. We show that their Gromov–Wasserstein distance is bounded by the spread of the features over the samples, and their Gromov–Hausdorff distance by the largest spread and a covering radius. A feature of small spread has a location among the samples, and when the mean spread is small the features together form a metric space that mirrors the sample geometry. This is a global and label-free notion of an interpretable feature geometry. All our mathematical results are machine-verified in Lean 4. Spread is also a quantity a model can be trained on. We regularize WSV spread in autoencoders on geometric toy data and on Gemma-2-2b residual-stream activations along a prompt manifold. As the theory predicts, this allows us to recover the sample geometry on the feature side. An L^1 penalty can have the same effect and we show that also in those cases an interpretable feature geometry is explained by low WSV spread.

1 INTRODUCTION

The geometry of a representation is typically understood as the geometry of the space of samples with the features as coordinates. That neurons/features are directions in activation space is the perspective formalized in the linear representation hypothesis (Park et al., 2024), which provides the generative model for sparse autoencoders (Bricken et al., 2023; Cunningham et al., 2024; Gao et al., 2024; Lieberum et al., 2024). Treating features as directions and organizing those in a sparse

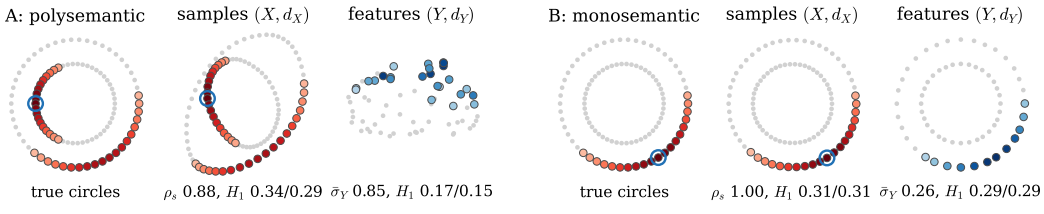

Figure 1: Two exact codes of the same samples, which lie on one of two circles. Left of each triple: the true circles, with the field of one feature, the samples it fires on (red). Middle: the intrinsic (1) sample geometry (X, d_X), the same field in red. Right: the intrinsic feature geometry (Y, d_Y), each feature coloured by its value on one sample (blue; the sample is ringed on the left). Under B every feature fires on one circle, and both sides are the two circles. Under A every feature fires on an arc of each circle. The samples are still the two circles, the features are not. ρ_s: rank correlation of d_X with the true cyclic distance within a circle. H_1: the two longest 1D-persistence bars. $\bar{\sigma}_Y$: spread (Section 3).

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.

dictionary can make them interpretable but says little about how they collectively organize into a geometry. In an alternative perspective, a covariate is encoded by localized receptive fields, each unit responding to a small region of it. Biological neural networks often work in this way: place cells fire at one physical location (O'Keefe & Dostrovsky, 1971), head-direction cells at one heading (Taube et al., 1990), grid cells on a lattice in space (Hafting et al., 2005). In machine learning, sparse coding of natural images yields localized receptive fields (Olshausen & Field, 1996), a unit or dictionary feature is labelled by the inputs it responds to (Bau et al., 2017; Olah et al., 2020; Bricken et al., 2023), and features are matched across dictionaries and models by their activations on shared inputs (Lan et al., 2024; Cao et al., 2026). When features are localized receptive fields it can make sense to represent them as points in a geometric space. Formally, this can be done by normalizing the features to probability distributions on the samples and compare them by optimal transport. This requires a metric on the sample space as the ground cost.

An elegant way to obtain such a metric is to dually treat samples as distribution over the features they activate. Then, the sample and feature metrics can be asked to define each other: two samples are as far apart as their distributions over the features, measured in the feature metric, and two features as far apart as their fields, measured in the sample metric. This is the Wasserstein singular vector (WSV) problem of Huizing et al. (2022). They prove existence for the problem with an additive regularizer $\tau > 0$. Despite much subsequent work on the topic the general unregularized problem remained open (Lins, 2023; Düsterwald et al., 2025; Lin et al., 2025; Auffenberg et al., 2025). We prove that for almost every matrix a positive pair exists, and that degenerate solutions exist exactly when the matrix has a nontrivial lumping, a partition of samples and features through which it factors (Theorems 1 and 2). This is, to our knowledge, the first unregularized existence proof for the general WSV problem.

Small spread of the features' receptive fields makes the feature space reflect the sample space. A field concentrated around one sample has a well defined location, and two such features are about as far apart as the samples they concentrate on. Conversely, a feature spreading over the sample space has no well defined location. This is polysemanticity in geometric terms (Elhage et al., 2022): any point assigned to such a feature lies far from some of the samples it fires on, and two such features can be close while their samples are not. Exact reconstruction does not settle which case one is in. Figure 1 shows two exact codes of the same samples, which lie on one of two circles. Under one code every feature fires on an arc of one circle, and both sides are the two circles. Under the other every feature fires on an arc of each circle, two unrelated meanings. Then the samples are still the two circles, but the features form a different space.

We formalize these observations by bounding the Gromov–Wasserstein distance (Mémoli, 2011) between the row and column sides of the WSV metric by the spread of the features, and the Gromov–Hausdorff distance by the largest spread and a covering radius (Theorem 3). Small spread gives a feature a location among the samples at a cost equal to its radius, zero spread makes the two sides isometric, and small spread on one side forces it on the other (Proposition 1). This can be understood as a metric relaxation of Dowker's theorem (Dowker, 1952), a classical result in algebraic topology, which says that for a binary relation the complex of rows sharing a column and that of columns sharing a row have the same homotopy type. The spread is computed from the matrix and its pair, so the bound is a label-free certificate of a feature space that mirrors the samples, and for a fixed sample metric it is differentiable in the matrix, so a code can be trained for it. We illustrate this by training spread regularized autoencoders and show that so recovered feature geometry can be used for simple steering in Gemma-2-2b.

Our contributions are:

- Existence of the unregularized WSV pair for almost every matrix, and rigidity: degenerate solutions exist exactly for lumpable matrices (Section 2).
- Metric Dowker duality: the two geometries of a matrix agree in Gromov–Wasserstein distance up to the spread and in Gromov–Hausdorff distance up to the largest spread and a covering radius. The spread vanishes exactly at permutation couplings, where the two geometries are isometric, and spread is two-sided (Section 3).
- Spread as a certificate and a training objective. An autoencoder with a hinge on its own spread, recovers the sample geometry on the feature side. Demonstrated on metric toy data and on the residual stream of Gemma-2-2b. The recovered feature ring steers the model's greeting (Section 4).

- All mathematical results machine-verified in Lean 4 and checked with the comparator tool (Section C). Proofs and code are in the supplementary material.

2 WASSERSTEIN SINGULAR VECTORS

Fix finite sets X, Y with $|X|, |Y| \geq 2$ and a joint distribution $\pi \in \mathcal{P}(X \times Y)$ with positive marginals μ_X, μ_Y. Write $\varphi_x = \pi(x, \cdot)/\mu_X(x)$ for the row conditionals and $\psi_y = \pi(\cdot, y)/\mu_Y(y)$ for the column conditionals. Comparing two rows by optimal transport requires a ground metric on the columns, and comparing two columns requires a ground metric on the rows. The fixed points of this circularity are the *self-consistent pairs*: metrics d_X on X and d_Y on Y with

$$\mathcal{W}_2^{d_Y}(\varphi_x, \varphi_{x'}) = \lambda\, d_X(x, x') \quad \text{for all } x, x', \qquad \mathcal{W}_2^{d_X}(\psi_y, \psi_{y'}) = \lambda\, d_Y(y, y') \quad \text{for all } y, y', \tag{1}$$

where $\mathcal{W}_2^d(\mu, \nu) = \left(\min_{\gamma \in \Pi(\mu,\nu)} \langle \gamma, d^2 \rangle\right)^{1/2}$ is the 2-Wasserstein distance over ground cost d^2.

Notation. $\mathcal{P}^+$ denotes couplings with positive marginals. A *pseudometric pair* $d = (d_X, d_Y)$ is *positive* if both are genuine metrics, i.e. no distinct points are at distance zero. The *eigen-map* is $W(\pi, d) = (W_X, W_Y)$ with $W_X(x, x') = \mathcal{W}_2^{d_Y}(\varphi_x, \varphi_{x'})$ and mirror, so (1) reads $W(\pi, d) = \lambda d$. The pair (d, λ) is then an *eigen-pair*. In the *squared coordinate* $v = d^2$ (entrywise) the equations become $T_\pi(v) = \mu v$ with $\mu = \lambda^2$, where $T_\pi(v)$ evaluates optimal transport costs. Most proofs are carried out in this coordinate at no loss of generality (see appendix). Finally, the *spread* of a distribution α on a pseudometric space (Z, d) is

$$\sigma^2(d, \alpha) = \tfrac{1}{2} \sum_{z, z'} \alpha(z)\, \alpha(z')\, d(z, z')^2, \tag{2}$$

the mean squared pairwise distance of the weighted point cloud (Z, α).

Lumpings. Degenerate solutions of (1) come from nontrivial quotients through which the coupling factors.

Definition 1 (Partition pairs, lumpings, unlumpability)**.** A *partition pair* $Q = (\sim_X, \sim_Y)$ is a pair of equivalence relations, one on X and one on Y. The coupling π is *Q-lumpable* if

(1) for all $x \sim_X x'$ and every $\sim_Y$-class $B \subseteq Y$: $\varphi_x(B) = \varphi_{x'}(B)$; and
(2) for all $y \sim_Y y'$ and every $\sim_X$-class $A \subseteq X$: $\psi_y(A) = \psi_{y'}(A)$.

Q is *discrete* if both relations are equality, *indiscrete* if both are total, and *trivial* if it is one of these two. A partition pair for which π is Q-lumpable is a *lumping* of π. We say π is *unlumpable* if every lumping of π is trivial.

A lumping merges rows that place equal conditional mass on every merged block of columns, and mirror. The coupling then factors through the quotient coupling on the classes, and every eigen-pair of the quotient pulls back to an eigen-pair of π that vanishes inside the classes. Unlumpability rules out such hidden quotients. Clause (1) is Kemeny–Snell strong lumpability of the Markov chain $x \to y$ (Kemeny & Snell, 1960, §§6.3–6.4) and clause (2) the same for its Bayes reversal. The two-sided notion appears once before, as the zero locus of a co-clustering cost (Blöchl et al., 2019) but its use as the hypothesis of an existence theorem is new.

Existence.

Theorem 1 (Existence)**.** *If $\pi \in \mathcal{P}^+$ is unlumpable, there exist genuine metrics (d_X, d_Y) and $\lambda \in (0, 1]$ satisfying* (1)*, which can be chosen with (d_X^2, d_Y^2) also a metric pair. Lebesgue-almost-every π is unlumpable.*

A detailed proof is in Section A. Brouwer's fixed point theorem on a compact convex slice of the cone of pseudometric pairs, which the eigen-map preserves, yields a nonzero pseudometric eigen-pair. Every degeneracy of such a pair exhibits a lumping: its zero classes form one, and $\lambda = 0$ forces all rows equal. Under unlumpability both are excluded, and spread contraction gives $\lambda \leq 1$. Nontrivial partition pairs impose polynomial constraints on π, so lumpable couplings form a null set.

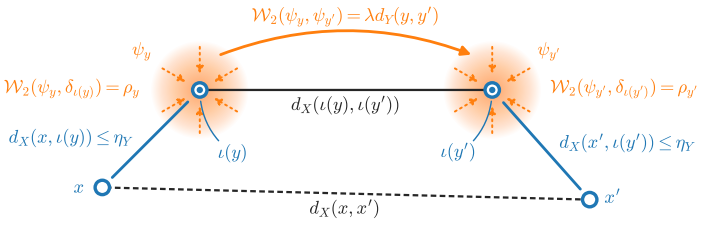

Figure 2: Illustration of the Gromov–Hausdorff bound of Theorem 3. The proof constructs the correspondence $R = \{(x, y) : d_X(x, \iota(y)) \le \eta_Y\}$ and bounds its distortion. Here we have drawn two of its pairs, (x, y) on the left and (x', y') on the right. Every point is a measure on the samples, a sample x as δ_x, a center $\iota(y)$ as $\delta_{\iota(y)}$, and a column y as its distribution ψ_y on the samples, concentrated around its center, and every edge is the transport distance $\mathcal{W}_2^{d_X}$ between its two ends. By the second eigen-equation the transport distance between ψ_y and $\psi_{y'}$ is the feature distance $\lambda d_Y(y, y')$. The distortion of the two pairs compares $d_X(x, x')$ with the OT cost between the two distributions. This distortion can be bounded. Each sample lies within η_Y of its center, so the dashed and the solid chord differ by at most $2\eta_Y$. Each distribution collapses onto its center at cost ρ_y, so the solid chord and the transport differ by at most $\rho_y + \rho_{y'}$. Hence the distortion of R is at most $2\eta_Y + 2\rho^Y_{\max}$, and $d_{\mathrm{GH}}\big((X, d_X), (Y, \lambda d_Y)\big) \le \eta_Y + \rho^Y_{\max}$.

The hypothesis is not merely sufficient. It characterizes exactly the couplings for which no degenerate solution exists at all.

Theorem 2 (Rigidity of the hypothesis)**.** *$\pi \in \mathcal{P}^+$ is unlumpable if and only if every nonzero pseudometric pair solving* (1) *at some eigenvalue is a pair of genuine metrics with $\lambda > 0$.*

One direction is the degeneracy step of the existence proof. For the converse, a nontrivial lumping Q yields a quotient coupling on the classes, and any eigen-pair of the quotient pulls back to an eigen-pair of π that vanishes inside the classes of Q, hence is degenerate. For details see A

Remark 1. A coupling with a nontrivial lumping may still admit a positive eigen-pair alongside its degenerate ones. Thus, Theorem 2 characterizes the absence of *degenerate* solutions and not existence itself. To find a sharply necessary condition for existence is still open.

3 Metric Dowker duality

Dowker's theorem. A relation $R \subseteq X \times Y$ determines two simplicial complexes, the *row complex* $D(R)$, whose faces are the sets of rows that share a column, and its mirror, the *column complex* $S(R)$. Dowker's theorem (Dowker, 1952) says that they have the same homotopy type or in other words, the two sides of a relation carry one shape. The theorem crucially relies on the above complexes being agnostic to how much mass the set of witnesses to a given face carries. This leads to the contractibility conditions, which in turn enable the fiber lemmas that classical proofs like (Björner, 1995) run on. A coupling π determines via WSV two metric spaces instead: the sample geometry (X, d_X) and the feature geometry (Y, d_Y) of Theorem 1. Here the witness set of the column y is its conditional ψ_y, and the two aggregates below are the metric relaxation of what a good cover asks. The radius ρ_y measures how far ψ_y is from being a single point, and the covering radius η_Y whether the conditionals leave a row unwitnessed. Together they provide a bound on the difference between the two sides of a WSV pair.

Locations. Fix a positive eigen-pair (d, λ) of π. For a column y and a point $x \in X$, the cost of collapsing ψ_y onto x is

$$\rho(x, y) := \mathcal{W}_2^{d_X}(\psi_y, \delta_x) = \Big(\textstyle\sum_{x'} \psi_y(x')\, d_X(x', x)^2\Big)^{1/2}, \tag{3}$$

since $\psi_y \otimes \delta_x$ is the only coupling of ψ_y and δ_x. Fix a minimizer $\iota(y) \in \arg\min_x \rho(x, y)$. We call this the *Fréchet center* of y. In general, this is not unique but nothing below depends on the choice.

The choice is formalized by a map $\iota : Y \to X$. We call $\rho_y := \min_x \rho(x, y)$ the *Fréchet radius*. This radius and the spread (2) control each other,

$$\sigma(d_X, \psi_y)/\sqrt{2} \;\le\; \rho_y \;\le\; \sqrt{2}\,\sigma(d_X, \psi_y), \tag{4}$$

and both vanish exactly when ψ_y is a Dirac mass. Aggregate the column data as the maximal radius $\rho^Y_{\max} := \max_y \rho_y$, the covering radius $\eta_Y := \max_x \min_y d_X(x, \iota(y))$ of the centers, and the mean squared spread $\bar\sigma^2_Y := \sum_y \mu_Y(y)\,\sigma^2(d_X, \psi_y)$. The row data y^*_x, ρ_x, $\rho^X_{\max}$, η_X, $\bar\sigma_X$ are defined in mirror, with each row φ_x collapsed onto points of (Y, d_Y).

Comparison. The two sides live on different sets, so they are compared with the Gromov–Hausdorff and Gromov–Wasserstein distances (Burago et al., 2001; Mémoli, 2011). The former is the worst distortion $|d_X(x, x') - d_Y(y, y')|$, minimized over correspondences $R \subseteq X \times Y$. The latter is the distortion averaged with $\gamma \otimes \gamma$ for a coupling γ of (μ_X, μ_Y), minimized over couplings.

Theorem 3 (Metric Dowker duality)**.** *For every positive eigen-pair (d, λ) of π, in the column data and in the row data,*

$$d_{\mathrm{GH}}\big((X, d_X),\, (Y, \lambda d_Y)\big) \le \rho^Y_{\max} + \eta_Y, \qquad d_{\mathrm{GW}}\big((X, d_X, \mu_X),\, (Y, \lambda d_Y, \mu_Y)\big) \le \sqrt{2}\,\bar\sigma_Y,$$
$$d_{\mathrm{GH}}\big((X, \lambda d_X),\, (Y, d_Y)\big) \le \rho^X_{\max} + \eta_X, \qquad d_{\mathrm{GW}}\big((X, \lambda d_X, \mu_X),\, (Y, d_Y, \mu_Y)\big) \le \sqrt{2}\,\bar\sigma_X.$$

For a permutation coupling every positive eigen-pair has $\lambda = 1$ and d_Y the pushforward of d_X, and every right-hand side is zero: the duality is an isometry.

The detailed proof together with explanations of all ingredients is in Section B. In Figure 2 we illustrate its mechanism.

Covering. Both bounds rest on the same pairwise estimate, that replacing ψ_y and $\psi_{y'}$ by the point masses at their centers costs at most $\rho_y + \rho_{y'}$. They differ in how this cost is aggregated. The Gromov–Hausdorff distance is a worst case over a correspondence. Since such a correspondence must reach every sample, here through $\{(x, y) : d_X(x, \iota(y)) \le \eta_Y\}$, covering enters additively and the bound is vacuous once the centers leave a part of X unreached. The Gromov–Wasserstein distance is a mean against the free coupling $\gamma(y, x) = \mu_Y(y)\,\psi_y(x)$, namely π transposed, whose sample-side marginal is the mixture $\sum_y \mu_Y(y)\,\psi_y = \mu_X$. Here, covering has moved out of the hypotheses and into the statement. The mean bound is about the geometry where the features put their mass and vice versa.

Every quantity in Theorem 3 is computed from π and d alone. A small radius ρ_y certifies that column y has a location $\iota(y)$ in X, at cost ρ_y. Small radii for all columns and a small covering radius certify that $(Y, \lambda d_Y)$ mirrors (X, d_X) up to $\rho^Y_{\max} + \eta_Y$, and in measure up to $\sqrt{2}\,\bar\sigma_Y$. This is the sense in which a feature geometry of small spread is interpretable.

The column and the row bounds are not independent: spread transfers across the eigen-equation.

Proposition 1 (Spread is two-sided)**.** *For every positive eigen-pair (d, λ) of π and every row x,*

$$\sigma^2(d_Y, \varphi_x) \;\le\; \frac{2}{\lambda^2} \sum_y \varphi_x(y)\,\rho(x, y)^2, \qquad \textit{hence} \qquad \bar\sigma_X \;\le\; \frac{2}{\lambda}\,\bar\sigma_Y,$$

and in mirror $\bar\sigma_Y \le \frac{2}{\lambda}\,\bar\sigma_X$. In particular $\bar\sigma_X = 0$ if and only if $\bar\sigma_Y = 0$.

Small spread of the features over the samples forces small spread of the samples over the features, at exchange rate $2/\lambda$, so a penalty on one side bounds both. This justifies the one sided optimization below.

4 Experiments

Theorem 3 bounds the Gromov–Wasserstein distance between the sample and the feature geometry by the mean spread $\bar\sigma_Y$. We therefore train autoencoders for low spread, ask whether the sample geometry appears on the feature side, and compare with a standard sparsity penalty. On the residual stream of Gemma-2-2b (Section 4.2) we then use the recovered ring to steer.

Table 1: Scores used to evaluate geometric recovery.

relative spread	$\bar{\sigma}_Y/\sigma(d_X, \mu_X)$; 1 uninformative features, 0 point features
ρ_s	Spearman correlation of d_X with the hidden cyclic distance
H_1	longest degree-one persistence bar, as a fraction of the diameter
d_{GW}	Gromov–Wasserstein distance of (Y, d_Y) to the true features
L_0, live, FVU	active latents per sample, latents on ≥ 5 probe samples, unexplained variance

Pair law. Write

$$C(x, x') := \sum_y \mu_Y(y)\, \psi_y(x)\, \psi_y(x') \tag{5}$$

for the probability of two samples drawn independently from one feature. Both marginals are μ_X, and the mean spread is half the second moment of d_X under C, $\bar{\sigma}_Y^2 = \frac{1}{2}\sum_{x,x'} C(x, x')\, d_X(x, x')^2$. For fixed d_X this is differentiable in the coupling.

Model and coupling. $h = \mathrm{ReLU}(W_e(x - b_d) + b_e)$, $\hat{x} = W_d h + b_d$, k latents, unit-norm decoder rows. The coupling is the code itself, $\pi \propto h$ on a fixed probe of samples: $\mu_Y(y)$ is the share of the total activation carried by latent y, ψ_y is the profile of that latent over the probe, $\mu_X(x)$ is the share of the total activation at sample x, and C is the law of two probe samples drawn from one latent.

Objective.

$$\mathrm{rec} + \alpha\|h\|_1 + \beta\left(\left[\bar{\sigma}_{Y,c}^2 - \tau^2\right]_+ + \beta' V\right). \tag{6}$$

The following is a brief explanation of each term and why it is there.

(1) *Reconstruction and sparsity*: $\mathrm{rec} + \alpha\|h\|_1$ is the standard sparse autoencoder (Bricken et al., 2023; Cunningham et al., 2024), and at $\beta = 0$ our baseline. In practice reconstruction also keeps μ_X spread over the data.

(2) *Spread hinge*: $[\bar{\sigma}_{Y,c}^2 - \tau^2]_+$ penalizes the spread of Theorem 3 (floored, see (4)) above τ, measured in a running approximation of the eigen-pair (Section D). Every 200 steps d_X and d_Y are recomputed from the current representation but between these refreshes they are frozen.

Since d_X is intrinsic to the representation and thus learned alongside it some pathologies can arise. Instead of tightening the features they can be shrunk relative to d_X by stretching distances there. Conversely, distances inside a feature may be shrunk. Terms (3) and (4) are designed to prevent these two escapes. Both use a fixed reference distance d_{ref} on the data, the Euclidean distance between centred activations scaled to unit length (Section E).

(3) *Local normalization* (against stretching): Normalized to unit mean over all pairs, d_X can lower the spread of every feature outside one region by stretching that region. In Figure 5 this cuts the ring open. We instead scale d_X so that its mean over all pairs of a sample and one of its ten nearest neighbours in d_{ref} equals the mean of d_{ref} over the same pairs. Then d_X matches the data's scale on average over neighbourhoods, and τ is a spread measured on that scale.

(4) *Floor* (against collapse): A feature also looks tight if d_X shrinks between its samples. We penalize every pair of sample distances below a fraction c of d_{ref}, weighted by how often the pair shares a feature.

$$V := \sum_{x \neq x'} C(x, x')\left[c^2 d_{\mathrm{ref}}(x, x')^2 - d_X(x, x')^2\right]_+. \tag{7}$$

The hinge acts on the *floored spread*

$$\bar{\sigma}_{Y,c}^2 := \bar{\sigma}_Y^2 + \tfrac{1}{2}V \;=\; \tfrac{1}{2}\sum_{x \neq x'} C(x, x')\, \max\left\{d_X(x, x')^2,\; c^2 d_{\mathrm{ref}}(x, x')^2\right\}. \tag{8}$$

It bounds both the spread in d_X and c times the spread in d_{ref} which collapse cannot lower.

(5) *Floor penalty*: The floored hinge removes the reward for collapse. For training it is beneficial to punish collapse directly with $\beta' V$.

In Section D we provide details on training and study ablation of (3) and (4).

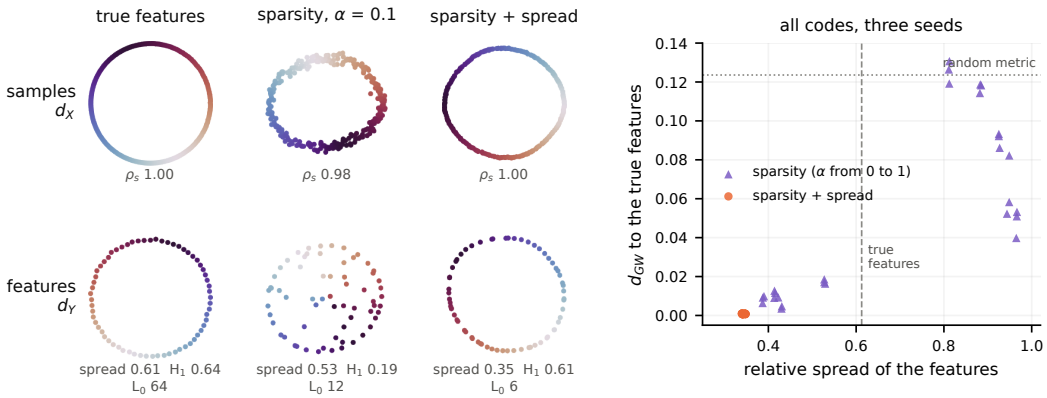

Figure 3: Left: MDS of (X, d_X), coloured by the true angle, and of (Y, d_Y), coloured by the circular mean of a latent's profile. We show the underlying data (left), sparsity regularization at $\alpha = 0.1$ (middle) and sparsity plus spread $\tau = 0.3$ (right). Under each panel ρ_s, or the relative spread, H_1 bar and L_0. Right: every code of both families against its relative spread, the d_{GW} of its feature geometry to the true one.

4.1 A CIRCLE WITH KNOWN FEATURES

The objective is tested on circular and spherical toy data. Here we present the circular one and the spherical one is in Section D. Row i of the data matrix has entries $u_{ij} \propto \exp\big(6\cos(\theta_i - 2\pi j/64)\big)$, $j = 1, \ldots, 64$, normalized to $\sum_j u_{ij} = 1$, so it is the von Mises density centred at $\theta_i = 2\pi i/1440$, evaluated at the 64 reference angles $2\pi j/64$ and normalized, plus Gaussian noise at 0.15 of the entries' standard deviation. Read as a representation, the noiseless matrix has a circle in both the sample and the feature geometry. We train width-64 autoencoders on these vectors, three seeds, once with the sparsity penalty alone and once with the spread penalty added at $\alpha = 0.01$, and ask whether the feature geometry of the learned code reflects the circular geometry Section D). Sparsity alone recovers the sample side, and the feature side only at strong penalties. At every α the sample geometry tracks the hidden angle (Spearman correlation $\rho_s \geq 0.89$), but the sparse representations, $L_0 = 12$ at $\alpha = 0.1$, do not display the circle in their feature geometry. The longest degree-one persistence bar is $H_1 = 0.19$ of the diameter against 0.64 on the underlying data. The Gromov–Wasserstein distance d_{GW} to the underlying geometry is at random level. With the spread penalty the latents tile the ring at every τ and in every seed: six active latents per sample, spread 0.35 relative to the data's, $\rho_s = 1.00$, $H_1 = 0.61$, $d_{\mathrm{GW}} = 0.001$, all 64 latents alive at FVU 0.02. Across the whole experiment, feature geometry is recovered exactly when the spread is low (Figure 3, right); sparsity alone gets there only from $\alpha = 0.2$ on, at higher reconstruction error and with up to half of the latents dead (Table 2).

4.2 GEMMA-2-2B RESIDUALS

We prompt Gemma-2-2b (Gemma Team et al., 2024) with "The time is now {h}:{mm} {am/pm}. Good", once for every minute of the day, and take the residual stream after block 24 at the last token, 1440 vectors in total. At this position the model is about to complete the greeting, *morning*, *afternoon*, *evening* or *night*, so the residual must carry the time of day. The sample geometry should therefore be the circle of the day, up to what the model makes of the number. The sample geomatry of the raw activations already encodes it ($\rho_s = 0.78$). We train codes of width 1024 on these vectors exactly as on the toy, with the same constants and scores, and compare them with the Gemma Scope dictionaries released for this layer (Lieberum et al., 2024).

The circle on both sides. Figure 4. Neither the raw activations nor the released dictionaries have a clean feature ring: their features are spread over almost the whole day (relative spread 0.95–1.00). The raw coordinates do arrange around the day in order, but as a loose, filled cloud whose cycle is weak ($H_1 = 0.20$). The released dictionaries fold it ($H_1 = 0.09$). Sparsity alone at width 1024

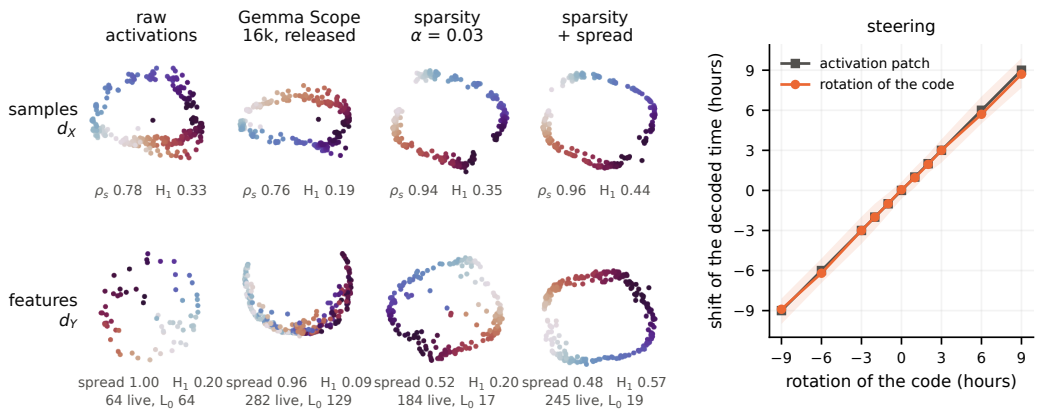

Figure 4: Left: MDS of (X, d_X) over the 240 probe prompts, coloured by the time of day, and of (Y, d_Y) over the live latents, coloured by the circular mean of a latent's profile (colour only), for the raw activations, a released Gemma Scope dictionary, the code trained with the sparsity penalty alone and the code trained with the spread penalty at $\tau = 0.3$; under each panel ρ_s, or the relative spread, H_1 bar, live latents and L_0. Right: shift of the time decoded from the next-token distribution against the rotation of the $\tau = 0.3$ code along its feature ring, median and interquartile range over 120 source times; the activation patch is the ceiling.

lowers the spread to 0.44 and a ring like structure capturing the time becomes visible. It is still contaminated by some features in the middle of the ring. With the spread penalty at $\tau = 0.3$ both sides are clean rings, $\rho_s = 0.96$ on the samples and $H_1 = 0.57$ on the features, with 245 live latents at $L_0 = 19$ and FVU 0.004. The rings are there in every seed for τ from 0.4 down to 0.3, weaken at 0.275, and at 0.25 the ring is cut in two seeds of three (Section E).

Steering. Rotating the spread regularized code by $\Delta\theta$ and decoding the result into the residual moves the time read off its next-token distribution. Over 120 source times and rotations of ± 1 to ± 9 h the decoded time follows the rotation with slope 0.97 ($R^2 = 0.92$), against 1.00 for patching in the other time's activation outright and a scatter of ± 7 h for a random permutation of the latents (Figure 4, right). After a rotation that carries the time into another class, say from a morning time to an afternoon one, we compare the model's output logits at that position for the four greeting words only. In inly 0.56 of the trials the largest is the word of the new class. The four words have very unequal natural frequencies, *morning* at 0.34 against under 0.10 for each of the others, so whenever the state lies between two classes the largest logit falls back on a frequent word. Subtracting each word's log-frequency from its logit before the comparison removes this default, and the new class then wins in 0.82 of the trials (Section E).

5 Related work

Wasserstein singular vectors. Equation (1) is the problem of Huizing et al. (2022), who prove existence for a regularized version, write that the unregularized case is out of reach of the nonlinear Perron–Frobenius theorems of Akian et al. (2016) and ask for its analysis. Lins (2023) proves linear convergence of a damped iteration conditional on an eigenvector existing, and records that at $\tau = 0$ one "may or may not" exist. The follow-ups keep a regularizer or restrict the model to tree-Wasserstein surrogates (Düsterwald et al., 2025; Lin et al., 2025; Auffenberg et al., 2025). The last observes that the fixed-point step of several published proofs invokes Brouwer's theorem on a compact set that is not convex. Our unregularized existence result uses Brouwer's fixed point theorem on a convex slice of the pseudometric cone and is machine-checked. The pair of row and column geometries, each a transport distance with respect to the other, is older than WSV. The coupled geometry program of Coifman & Gavish (2011); Gavish & Coifman (2012) alternates the two sides without a convergence guarantee. Ground metric learning from labels goes back to Cuturi

& Avis (2014); co-optimal transport (Redko et al., 2020) couples the rows and columns of two *given* matrices, with no fixed point.

Eigendistances of Markov chains. Ollivier (2009) discusses the one-sided problem of a metric ρ on the states of a Markov chain P with $\mathcal{W}_p^\rho(P^u, P^v) = (1-\kappa)\,\rho(u,v)$, a metric of constant coarse Ricci curvature κ. Völlering (2017) proves existence under an assumption, shows that algebraic irreducibility of P is equivalent to every eigendistance being a proper metric, and proves uniqueness on finite state spaces. Unlumpability (Definition 1) is the two-sided form of his irreducibility and Theorem 2 of his equivalence. It is tempting to then reduce our problem to his by merging the two sides into the alternating chain on $X \sqcup Y$. This does not work because every step of that chain crosses sides. For a genuine metric this forces $\kappa = 0$, and at $\kappa = 0$ the metric vanishes within X and within Y on every connected component of the support of π. A genuine metric of constant curvature therefore exists only for permutation couplings. His hypotheses fail there in any case, since the bipartite chain is never lazy and his uniqueness argument needs an irreducibility of the optimal coupling kernel that the bipartite case violates in general.

Dowker duality. Dowker's theorem (Dowker, 1952) was put in filtered form by Chazal et al. (2014), who define the Dowker complex of an arbitrary function on two sets and bound their interleaving by the distortion of a pair of correspondences. These results could thus be a possible framework for our duality result but we choose to stay self contained. In any case our contribution is (Chowdhury & Mémoli, 2018) and (Virk, 2021) give functional generalizations of Dowker duality. These are different in spirit to our work as they provide isomorphisms of persistence modules derived from filtrations of the underlying relation. Vaupel & Dunn (2023) extends Dowker duality to compositions of relations and (de Silva et al., 2026) to higher order relations, while (Gaukstad et al., 2026) gives a metric relaxation of Dowker duality but not in a data intrinsic metric like WSV. We believe that all these latter three works could be connected to our framework here in very interesting ways that we would like to explore further in the future in more topology oriented publications.

Interpretability. Sparse autoencoders (Bricken et al., 2023; Cunningham et al., 2024; Gao et al., 2024; Lieberum et al., 2024) produce interpretable dictionaries of features and superposition (Elhage et al., 2022) is the reason a feature may fire in several places. Feature geometry beyond single directions has been observed as circular representations of days and months (Engels et al., 2025) and probed on concept manifolds (Bhalla et al., 2026). Coherence, an objective yielding a bounded row–column interleaving, has been proposed as a topological criterion (Gaukstad et al., 2026) for interpretable feature geometries. Objectives for monosemanticity score each feature separately (Nasiri-Sarvi et al., 2026) and optimal transport between activation distributions has matched features across dictionaries post hoc (Cao et al., 2026). The uniformity term of contrastive learning (Wang & Isola, 2020) maximizes the spread of sample embeddings, the opposite objective to our spread minimization. We are not aware of prior work that scores a feature by the spread of its distribution over the samples in a metric the data determines, or that relates that score to the agreement of the two geometries.

6 Outlook

Theory. Every unlumpable coupling has a positive eigen-pair (Theorem 1), and in every instance we computed it seemed unique up to scale, but we have no general proof. This could also matter for stability. Unlumpability is exactly the absence of degenerate solutions (Theorem 2), yet a lumpable coupling may carry a positive eigen-pair beside its degenerate ones (Remark 1) and which couplings do is not characterized. We leave this as a worthwhile open frontier towards understanding the fascinating WSV construction better.

Application. We believe the dual sample–feature view of a representation and the spread objective can provide novel inductive biases for unsupervised disentanglement. Over the full sample geometry, low spread gives a latent one place, and on a product of factors, say the day and the month, that place is one day at one month, the opposite of disentangled. A day latent fires at one day in every month, so it has one place only once the samples sharing a day are merged. We expect that coupling to carry the geometry of the days, and propose to call a code disentangled when every latent has low spread in the coupling of one factor. A latent with a location in no factor's coupling is polysemantic, the symptom of superposition (Elhage et al., 2022). We intend to pursue this in future work.

AI USE STATEMENT

In this work, generative AI tools (Anthropic's Claude models, through the Claude Code environment) were used for the following tasks with required disclosure: The precise theoretical statements and proofs were developed in collaboration with AI through a close back and forth. In particular the ideas of Section 3 build on intuition from applied topology that the author has acquired over years of prior thinking and research without AI. In all stages human understanding and appropriate presentation in human readable form was a priority. The formalization of the proofs in Lean is purely AI automated. The author is not fully fluent in Lean and did not read the produced scripts in detail. The author approached this experiment carefully and be informed by best practices learned at a workshop about this at ICML 2026 (for example using the comparator tool). Additionally, generative AI tools were used for creating and editing code, creating figures, suggesting experimental parameters, drafting and editing text, brainstorming, identifying and summarizing literature, formatting references. All AI generated text was carefully edited or rewritten by the author. The research questions and the final text are the authors' own. The author takes responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Every mathematical statement of this paper is formalized in Lean 4 over mathlib (de Moura & Ullrich, 2021; The mathlib Community, 2020) and checked with the Lean comparator tool against a file that states it independently of the proofs (Section C). Sections A and B give every proof in human-readable form as well.

The experiments use the publicly released Gemma-2-2b and Gemma Scope dictionaries. Sections D and E state the architecture, the objective and its hyperparameters, the probe, the schedules, the seeds and the readout thresholds. Every number in Section 4 is produced by a recorded script at a fixed seed, and the figures are regenerated from cached runs. The Lean development and all code, including the training and analysis scripts and the cached runs behind the figures, are in the supplementary material.

REFERENCES

Marianne Akian, Stéphane Gaubert, and Roger Nussbaum. Uniqueness of the fixed point of nonexpansive semidifferentiable maps. *Transactions of the American Mathematical Society*, 368(2): 1271–1320, 2016.

Janis Auffenberg, Jonas Bresch, Oleh Melnyk, and Gabriele Steidl. Unsupervised ground metric learning. *arXiv preprint arXiv:2507.13094*, 2025.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: quantifying interpretability of deep visual representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6541–6549, 2017.

Ulrich Bauer. Ripser: efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3):391–423, 2021.

Usha Bhalla, Thomas Fel, Can Rager, Sheridan Feucht, Tal Haklay, Daniel Wurgaft, Siddharth Boppana, Matthew Kowal, Vasudev Shyam, Jack Merullo, Atticus Geiger, and Ekdeep Singh Lubana. Do sparse autoencoders capture concept manifolds? *arXiv preprint arXiv:2604.28119*, 2026.

Anders Björner. Topological methods. In Ronald L. Graham, Martin Grötschel, and László Lovász (eds.), *Handbook of Combinatorics*, volume 2, pp. 1819–1872. Elsevier, 1995.

Clemens Blöchl, Rana Ali Amjad, and Bernhard C. Geiger. Co-clustering via information-theoretic Markov aggregation. *IEEE Transactions on Knowledge and Data Engineering*, 31(4), 2019. arXiv:1801.00584.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec,

Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33 of *Graduate Studies in Mathematics*. American Mathematical Society, 2001.

Tue M. Cao, Nguyen Do, and My T. Thai. Semantic optimal transport for sparse autoencoder feature matching and circuit compression. *arXiv preprint arXiv:2605.28567*, 2026.

Frédéric Chazal, Vin de Silva, and Steve Oudot. Persistence stability for geometric complexes. *Geometriae Dedicata*, 173(1):193–214, 2014.

Samir Chowdhury and Facundo Mémoli. A functorial Dowker theorem and persistent homology of asymmetric networks. *Journal of Applied and Computational Topology*, 2(1-2):115–175, 2018.

Ronald R. Coifman and Matan Gavish. Harmonic analysis of digital data bases. In *Wavelets and Multiscale Analysis: Theory and Applications*, pp. 161–197. Birkhäuser, 2011.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.08600.

Marco Cuturi and David Avis. Ground metric learning. *Journal of Machine Learning Research*, 15: 533–564, 2014.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. *Automated Deduction – CADE 28*, pp. 625–635, 2021.

Vin de Silva, Chad Giusti, Vladimir Itskov, Michael Robinson, Radmila Sazdanovic, Nikola Schonsheck, Melvin Vaupel, and Iris Yoon. Dowker's theorem for higher-order relations. *Journal of Applied and Computational Topology*, 2026. arXiv:2506.10909.

Clifford H. Dowker. Homology groups of relations. *Annals of Mathematics*, 56(1):84–95, 1952.

Kira M. Düsterwald, Samo Hromadka, and Makoto Yamada. Fast unsupervised ground metric learning with tree-Wasserstein distance. In *International Conference on Learning Representations (ICLR)*, 2025.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2405.14860.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78): 1–8, 2021.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Sigurd Gaukstad, Melvin Vaupel, Valdemar Kargård Olsen, Erik Hermansen, and Benjamin Dunn. Learning coherent representations: A topological approach to interpretability. In *International Conference on Machine Learning (ICML)*, 2026. To appear. arXiv:2606.02841.

Matan Gavish and Ronald R. Coifman. Sampling, denoising and compression of matrices by coherent matrix organization. *Applied and Computational Harmonic Analysis*, 33(3):354–369, 2012.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Andrzej Granas and James Dugundji. *Fixed point theory*. Springer, 2003.

Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I. Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436:801–806, 2005.

Geert-Jan Huizing, Laura Cantini, and Gabriel Peyré. Unsupervised ground metric learning using Wasserstein singular vectors. In *International Conference on Machine Learning (ICML)*, volume 162 of *PMLR*, pp. 9429–9443, 2022. arXiv:2102.06278.

John G. Kemeny and J. Laurie Snell. *Finite Markov Chains*. Van Nostrand, 1960.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. Quantifying feature space universality across large language models via sparse autoencoders. *arXiv preprint arXiv:2410.06981*, 2024.

Lean FRO. comparator: a trustworthy judge for Lean proofs. `https://github.com/leanprover/comparator`, 2025.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: open sparse autoencoders everywhere all at once on Gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.

Ya-Wei Eileen Lin, Ronald R. Coifman, Gal Mishne, and Ronen Talmon. Joint hierarchical representation learning of samples and features via informed tree-Wasserstein distance. *arXiv preprint arXiv:2501.03627*, 2025.

Brian Lins. Convergence of iterates in nonlinear Perron–Frobenius theory. *Discrete and Continuous Dynamical Systems – B*, 28, 2023. arXiv:2207.14098.

Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011.

Ali Nasiri-Sarvi, Anh Tien Nguyen, Hassan Rivaz, Dimitris Samaras, and Mahdi S. Hosseini. MonoLoss: a training objective for interpretable monosemantic representations. *arXiv preprint arXiv:2602.12403*, 2026.

John O'Keefe and Jonathan Dostrovsky. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34(1):171–175, 1971.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: an introduction to circuits. *Distill*, 2020. https://distill.pub/2020/circuits/zoom-in.

Yann Ollivier. Ricci curvature of Markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2311.03658.

Ievgen Redko, Titouan Vayer, Rémi Flamary, and Nicolas Courty. CO-optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jeffrey S. Taube, Robert U. Muller, and James B. Ranck. Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *Journal of Neuroscience*, 10(2):420–435, 1990.

The mathlib Community. The Lean mathematical library. *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP)*, pp. 367–381, 2020.

Melvin Vaupel and Benjamin Dunn. The bifiltration of a relation and extended Dowker duality. *arXiv preprint arXiv:2310.11529*, 2023.

Žiga Virk. Rips complexes as nerves and a functorial Dowker-nerve diagram. *Mediterranean Journal of Mathematics*, 18:58, 2021. arXiv:1906.04028.

Florian Völlering. Constant curvature metrics for Markov chains. *arXiv preprint arXiv:1712.02762*, 2017.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, 2020.

A Proofs for Section 2

This appendix proves Theorems 1 and 2 in full. Section A.1 fixes conventions and collects the properties of optimal transport over a ground cost that is not assumed to be a metric. Section A.2 introduces the eigen-map and its squared coordinate as operators on pairs of cost matrices and shows that both preserve the cone of pseudometric pairs. Section A.3 applies Brouwer's fixed point theorem on a slice of that cone and produces a nonzero pseudometric eigen-pair for every coupling, with no hypothesis on π. Section A.4 shows that the zero classes of any pseudometric eigen-pair form a lumping. Section A.5 assembles the proof of Theorem 1 and Section A.6 that of Theorem 2.

Throughout, X and Y are finite sets with $|X| \geq 2$ and $|Y| \geq 2$, and $\pi \in \mathcal{P}^+$, so the marginals μ_X, μ_Y are everywhere positive and the conditionals $\varphi_x = \pi(x,\cdot)/\mu_X(x) \in \mathcal{P}(Y)$ and $\psi_y = \pi(\cdot,y)/\mu_Y(y) \in \mathcal{P}(X)$ are defined for all x and y. Two identities are used repeatedly:

$$\mu_X(x)\,\varphi_x(y) = \pi(x,y) = \mu_Y(y)\,\psi_y(x), \qquad \sum_x \mu_X(x)\,\varphi_x = \mu_Y, \qquad \sum_y \mu_Y(y)\,\psi_y = \mu_X. \tag{9}$$

The size restriction cannot be dropped: for $|X| = 1$ and $|Y| = 2$ every partition pair is trivial, so π is unlumpable, but the only pseudometric on a one-point set is 0 and the second equation of (1) forces $d_Y = 0$.

A.1 Couplings, pseudometric pairs, and transport over a ground cost

For a finite set Z, $\mathcal{P}(Z)$ is the set of probability vectors on Z, $\delta_z \in \mathcal{P}(Z)$ the Dirac mass at z, and $\alpha(S) = \sum_{z\in S} \alpha(z)$ for $\alpha \in \mathcal{P}(Z)$ and $S \subseteq Z$. For $\mu \in \mathcal{P}(Z)$ and $\nu \in \mathcal{P}(Z')$ a *coupling* is a $\gamma \in \mathcal{P}(Z \times Z')$ with $\sum_{z'} \gamma(z,z') = \mu(z)$ for all z and $\sum_z \gamma(z,z') = \nu(z')$ for all z'. The set of couplings is $\Pi(\mu,\nu)$.

Lemma 1. *$\Pi(\mu,\nu)$ is a nonempty compact convex subset of $\mathbb{R}^{Z\times Z'}$.*

Proof. The product coupling $\mu \otimes \nu : (z,z') \mapsto \mu(z)\,\nu(z')$ lies in $\Pi(\mu,\nu)$. The set is the intersection of the cube $[0,1]^{Z\times Z'}$ with the finitely many affine hyperplanes given by the marginal constraints, hence compact and convex. □

Definition 2 (Pseudometrics and pseudometric pairs)**.** A function $d : Z \times Z \to \mathbb{R}$ is a *pseudometric* if for all $z, z', z'' \in Z$

(i) $d(z,z') \geq 0$;
(ii) $d(z,z) = 0$;
(iii) $d(z,z') = d(z',z)$;
(iv) $d(z,z'') \leq d(z,z') + d(z',z'')$.

It is *positive* if $d(z,z') > 0$ whenever $z \neq z'$, i.e. a genuine metric. Let $V := \mathbb{R}^{X\times X} \times \mathbb{R}^{Y\times Y}$ with elements $d = (d_X, d_Y)$. Order, products, powers and square roots on V are entrywise. $\mathcal{PM} \subseteq V$ is the set of *pseudometric pairs*, the pairs whose components are pseudometrics. A pair in $\mathcal{PM}$ is *positive* if both components are; positive pseudometric pairs are the metric pairs of Theorem 1.

Lemma 2. *$\mathcal{PM}$ is a closed convex cone in V: $0 \in \mathcal{PM}$; if $d, d' \in \mathcal{PM}$ and $s, t \geq 0$ then $s\,d + t\,d' \in \mathcal{PM}$; and $\mathcal{PM}$ is closed.*

Proof. Axioms (i)–(iii) are linear conditions on the entries and are preserved under nonnegative combinations. For (iv), the two triangle inequalities scale by $s, t \geq 0$ and add. Each axiom instance is a non-strict linear inequality or a linear equality in finitely many coordinates of V, so $\mathcal{PM}$ is a finite intersection of closed sets. □

Lemma 3 (Square roots)**.** *If v is a pseudometric on Z then so is $\sqrt{v}$, with the same zero entries. Hence $v \in \mathcal{PM}$ implies $\sqrt{v} \in \mathcal{PM}$, and $\sqrt{v}$ is positive if and only if v is.*

Proof. Axioms (i)–(iii) and the zero entries are immediate. For (iv), the square root is monotone and subadditive on $[0, \infty)$, the latter because $(\sqrt{a}+\sqrt{b})^2 = a+b+2\sqrt{ab} \geq a+b$. So $\sqrt{v(z,z'')} \leq \sqrt{v(z,z') + v(z',z'')} \leq \sqrt{v(z,z')} + \sqrt{v(z',z'')}$. □

The converse fails: three collinear points at mutual distances $1, 1, 2$ form a metric space whose squared distances $1, 1, 4$ violate the triangle inequality. The last clause of Theorem 1 is therefore strictly more than its first.

Definition 3 (Transport over a ground cost)**.** A *ground cost* on Z is any function $c : Z \times Z \to \mathbb{R}$; no axioms are imposed. For $\mu, \nu \in \mathcal{P}(Z)$ set

$$\mathsf{OT}^c(\mu,\nu) := \min_{\gamma \in \Pi(\mu,\nu)} \langle \gamma, c\rangle, \qquad \langle \gamma, c\rangle := \sum_{z,z'} \gamma(z,z')\, c(z,z'), \qquad \mathcal{W}_2^c(\mu,\nu) := \mathsf{OT}^{c^2}(\mu,\nu)^{1/2},$$

where c^2 is the entrywise square. The minimum is attained since $\gamma \mapsto \langle \gamma, c\rangle$ is linear and $\Pi(\mu,\nu)$ is nonempty and compact (Lemma 1); a coupling attaining it is *optimal*. Since $\gamma \geq 0$ and $c^2 \geq 0$, $\mathsf{OT}^{c^2} \geq 0$ and $\mathcal{W}_2^c$ is well defined. For a pseudometric $c = d$ this is the distance of Section 2.

Lemma 4 (Transport lifts the pseudometric axioms)**.** *Let c be a ground cost on Z and $\mu, \nu, \rho \in \mathcal{P}(Z)$.*

(a) If $c \geq 0$ then $\mathsf{OT}^c \geq 0$. If moreover $c(z,z) = 0$ for all z, then $\mathsf{OT}^c(\mu,\mu) = 0$.
(b) If $c(z,z') = c(z',z)$ for all z, z', then $\mathsf{OT}^c(\mu,\nu) = \mathsf{OT}^c(\nu,\mu)$.
(c) If $c(z,z'') \leq c(z,z') + c(z',z'')$ for all z, z', z'', then $\mathsf{OT}^c(\mu,\rho) \leq \mathsf{OT}^c(\mu,\nu) + \mathsf{OT}^c(\nu,\rho)$.
(d) If $c \geq 0$ and $c(z,z'') \leq c(z,z') + c(z',z'')$ for all z, z', z'', then $\mathcal{W}_2^c(\mu,\rho) \leq \mathcal{W}_2^c(\mu,\nu) + \mathcal{W}_2^c(\nu,\rho)$.

Consequently, if d is a pseudometric on Z, then OT^d and $\mathcal{W}_2^d$ are pseudometrics on $\mathcal{P}(Z)$.

Proof. (a) $\langle \gamma, c\rangle \geq 0$ for every $\gamma \geq 0$. The diagonal coupling $\gamma_\Delta := \sum_z \mu(z)\, \delta_{(z,z)}$ lies in $\Pi(\mu,\mu)$ and $\langle \gamma_\Delta, c\rangle = \sum_z \mu(z)\, c(z,z) = 0$.

(b) Transposition $\gamma \mapsto \gamma^{\mathsf{T}}$, $\gamma^{\mathsf{T}}(z,z') := \gamma(z',z)$, is a bijection $\Pi(\mu,\nu) \to \Pi(\nu,\mu)$ that preserves $\langle \cdot, c\rangle$ when c is symmetric.

(c) Gluing. Let $\gamma_1 \in \Pi(\mu,\nu)$ and $\gamma_2 \in \Pi(\nu,\rho)$ be optimal and define $m : Z^3 \to \mathbb{R}$ by

$$m(z,z',z'') := \frac{\gamma_1(z,z')\,\gamma_2(z',z'')}{\nu(z')} \text{ if } \nu(z') > 0, \qquad m(z,z',z'') := 0 \text{ otherwise.}$$

If $\nu(z') = 0$ then $\gamma_1(\cdot, z')$ and $\gamma_2(z', \cdot)$ vanish identically, since $0 \leq \gamma_1(z,z') \leq \nu(z')$ and likewise for γ_2. Summing out z'' gives $\sum_{z''} m(z,z',z'') = \gamma_1(z,z')$, because $\sum_{z''} \gamma_2(z',z'') = \nu(z')$; summing out z gives $\sum_z m(z,z',z'') = \gamma_2(z',z'')$. So $m \geq 0$ has total mass 1, and $\gamma_{13}(z,z'') := \sum_{z'} m(z,z',z'')$ has marginals μ and ρ, i.e. $\gamma_{13} \in \Pi(\mu,\rho)$. Hence

$$\begin{aligned}\mathsf{OT}^c(\mu,\rho) \leq \langle \gamma_{13}, c\rangle &= \sum_{z,z',z''} m(z,z',z'')\, c(z,z'') \\ &\leq \sum_{z,z',z''} m(z,z',z'')\big[c(z,z') + c(z',z'')\big] = \langle \gamma_1, c\rangle + \langle \gamma_2, c\rangle,\end{aligned}$$

using the pointwise triangle inequality and $m \geq 0$ in the second step and the two marginal identities in the third. The right-hand side is $\mathsf{OT}^c(\mu,\nu) + \mathsf{OT}^c(\nu,\rho)$ by optimality.

(d) Take γ_1, γ_2 optimal for the ground cost c^2, glue them into m and γ_{13} as in (c), and write $\|f\|_m := (\sum m f^2)^{1/2}$ for $f : Z^3 \to \mathbb{R}$. Then

$$\begin{aligned}\mathcal{W}_2^c(\mu,\rho) \leq \langle \gamma_{13}, c^2\rangle^{1/2} = \big\|c(z,z'')\big\|_m &\leq \big\|c(z,z') + c(z',z'')\big\|_m \\ &\leq \big\|c(z,z')\big\|_m + \big\|c(z',z'')\big\|_m = \langle \gamma_1, c^2\rangle^{1/2} + \langle \gamma_2, c^2\rangle^{1/2}.\end{aligned}$$

The second inequality uses the pointwise triangle inequality together with $c \geq 0$, which makes squaring monotone; the third is Minkowski's inequality in $L^2(m)$; the last equality uses the marginal identities. The right-hand side is $\mathcal{W}_2^c(\mu,\nu) + \mathcal{W}_2^c(\nu,\rho)$ by optimality.

For the last claim let d be a pseudometric. Then $d^2 \geq 0$ has zero diagonal and is symmetric, so (a) and (b) applied to $c = d^2$ give that $\mathcal{W}_2^d$ is nonnegative, vanishes on the diagonal and is symmetric, and (d) applied to $c = d$ gives the triangle inequality; for OT^d apply (a)–(c) to $c = d$. □

Lemma 5 (Dependence on the ground cost)**.** *Fix $\mu, \nu \in \mathcal{P}(Z)$ and let c, c' be ground costs on Z.*

(a) $\mathsf{OT}^{sc} = s\ \mathsf{OT}^c$ *and* $\mathcal{W}_2^{sc} = s\,\mathcal{W}_2^c$ *for* $s \geq 0$.
(b) $|\,\mathsf{OT}^c(\mu,\nu) - \mathsf{OT}^{c'}(\mu,\nu)| \leq \|c - c'\|_\infty$. *Hence $c \mapsto \mathsf{OT}^c(\mu,\nu)$ and $c \mapsto \mathcal{W}_2^c(\mu,\nu)$ are continuous on $\mathbb{R}^{Z\times Z}$.*
(c) $\mathsf{OT}^c(\delta_z, \delta_{z'}) = c(z,z')$ *and* $\mathcal{W}_2^c(\delta_z,\delta_{z'}) = |c(z,z')|$.

Proof. (a) $\langle \gamma, sc\rangle = s\langle \gamma, c\rangle$ for every γ; pull s out of the minimum. For $\mathcal{W}_2$ apply this to $(sc)^2 = s^2c^2$ and take square roots.

(b) For every $\gamma \in \Pi(\mu,\nu)$, $|\langle \gamma, c\rangle - \langle \gamma, c'\rangle| \leq \sum_{z,z'} \gamma(z,z')\,|c(z,z') - c'(z,z')| \leq \|c - c'\|_\infty$, so the two minima differ by at most $\|c - c'\|_\infty$. The map $c \mapsto \mathcal{W}_2^c(\mu,\nu)$ is the composition of $c \mapsto c^2$, $\mathsf{OT}^{(\cdot)}(\mu,\nu)$ and the square root, each continuous.

(c) The marginal constraints force $\Pi(\delta_z,\delta_{z'}) = \{\delta_{(z,z')}\}$. □

A.2 The eigen-map and its squared coordinate

Definition 4. Define $W_\pi, T_\pi : V \to V$ by

$$W_\pi(d) := \Big((x,x') \mapsto \mathcal{W}_2^{d_Y}(\varphi_x,\varphi_{x'}),\ \ (y,y') \mapsto \mathcal{W}_2^{d_X}(\psi_y,\psi_{y'})\Big),$$
$$T_\pi(v) := \Big((x,x') \mapsto \mathsf{OT}^{v_Y}(\varphi_x,\varphi_{x'}),\ \ (y,y') \mapsto \mathsf{OT}^{v_X}(\psi_y,\psi_{y'})\Big).$$

$W_\pi(d) = W(\pi,d)$ is the eigen-map of Section 2 and T_π its squared coordinate. Both are defined on all of V, because Definition 3 imposes no axioms on the ground cost. By definition, entrywise,

$$W_\pi(d)^2 = T_\pi(d^2), \qquad W_\pi(d) \geq 0. \tag{10}$$

Hence for $d \geq 0$ and $\lambda \geq 0$, $W_\pi(d) = \lambda\, d$ if and only if $T_\pi(d^2) = \lambda^2\, d^2$: both sides of either equation are nonnegative entrywise, so one may square or take square roots.

Proposition 2.

(a) W_π and T_π are continuous on V and positively homogeneous: $W_\pi(s\,d) = s\,W_\pi(d)$ and $T_\pi(s\,v) = s\,T_\pi(v)$ for $s \geq 0$.
(b) Both preserve the cone: if $d \in \mathcal{PM}$ then $W_\pi(d) \in \mathcal{PM}$ and $T_\pi(d) \in \mathcal{PM}$.

Proof. (a) Each entry is $\mathsf{OT}^c(\alpha,\alpha')$ or $\mathcal{W}_2^c(\alpha,\alpha')$ with α,α' fixed and c a component of the argument; apply Lemma 5(a),(b) entrywise.

(b) Let $d \in \mathcal{PM}$ and write $(e_X, e_Y) := T_\pi(d)$. By Lemma 4, OT^{d_Y} is a pseudometric on $\mathcal{P}(Y)$, and $e_X(x,x') = \mathsf{OT}^{d_Y}(\varphi_x,\varphi_{x'})$ is its pullback along $x \mapsto \varphi_x$; the pullback of a pseudometric along any map is a pseudometric. The same holds for e_Y with OT^{d_X}, and for the two components of $W_\pi(d)$ with $\mathcal{W}_2^{d_Y}$ and $\mathcal{W}_2^{d_X}$. □

Nonlinear Perron–Frobenius theory produces eigenvectors in an orthant, and an orthant eigenvector has no reason to satisfy the triangle inequality. Thus clause (b) above is crucial. Because the cone $\mathcal{PM}$ is invariant, the fixed-point argument below runs inside it and the eigen-pair satisfies the pseudometric axioms by construction.

A.3 A PSEUDOMETRIC EIGEN-PAIR EXISTS FOR EVERY COUPLING

The construction is carried out for T_π. (10) with Lemma 3 transfers it to W_π.

Definition 5 (Normalization and slice)**.** Let $\ell : V \to \mathbb{R}$ be the linear functional $\ell(v) := \sum_{x,x'} v_X(x,x') + \sum_{y,y'} v_Y(y,y')$, and

$$\Delta := \{v \in V : v \geq 0,\ \ell(v) = 1\}, \qquad \Delta_{\mathcal{PM}} := \Delta \cap \mathcal{PM}.$$

For $v \geq 0$, $\ell(v) = 0$ if and only if $v = 0$, so $0 \notin \Delta$: the slice excludes the trivial solution $T_\pi(0) = 0$. Every entry of $v \in \Delta$ lies in $[0,1]$, being one nonnegative term of a sum equal to 1.

Lemma 6. *$\Delta_{\mathcal{PM}}$ is nonempty, compact and convex.*

Proof. Convexity: the orthant $\{v \geq 0\}$, the affine hyperplane $\{\ell = 1\}$ and the cone $\mathcal{PM}$ (Lemma 2) are convex. Compactness: all three sets are closed, so $\Delta_{\mathcal{PM}}$ is closed, and it is bounded because every entry lies in $[0,1]$. Nonemptiness: let $\delta = (\delta_X, \delta_Y)$ be the pair of discrete metrics, $\delta_Z(z,z') = 1$ for $z \neq z'$ and 0 on the diagonal. It lies in $\mathcal{PM}$, and $\ell(\delta) = |X|(|X|-1) + |Y|(|Y|-1) > 0$ since $|X|, |Y| \geq 2$, so $\delta/\ell(\delta) \in \Delta_{\mathcal{PM}}$ by Lemma 2. □

The renormalized map $v \mapsto T_\pi(v)/\ell(T_\pi(v))$ is undefined where $T_\pi(v) = 0$, and this happens: for the product coupling $\pi = \mu_X \otimes \mu_Y$ every row equals μ_Y, so $T_\pi(v)_X(x,x') = \mathsf{OT}^{v_Y}(\mu_Y, \mu_Y) = 0$ for every $v \in \mathcal{PM}$ by Lemma 4(a), and mirror. The following damped map has no singularity.

Definition 6 (Averaging map)**.** On the orthant $\{v \in V : v \geq 0\}$ let

$$A(v) := \frac{T_\pi(v) + v}{1 + \ell(T_\pi(v))}.$$

Lemma 7.

(a) A is continuous on the orthant.
(b) $A(\Delta) \subseteq \Delta$.
(c) $A(\Delta_{\mathcal{PM}}) \subseteq \Delta_{\mathcal{PM}}$.
(d) For $v \in \Delta$: $A(v) = v$ if and only if $T_\pi(v) = \mu\, v$ with $\mu := \ell(T_\pi(v))$, and then $\mu \geq 0$.

Proof. (a) For $v \geq 0$ every entry of $T_\pi(v)$ is nonnegative by Lemma 4(a), so the denominator is at least 1. A is a quotient of continuous maps (Proposition 2(a), ℓ linear) with denominator bounded away from 0.

(b) Let $v \in \Delta$. The entries of $A(v)$ are nonnegative, and $\ell(A(v)) = \big(\ell(T_\pi(v)) + \ell(v)\big)/\big(1 + \ell(T_\pi(v))\big) = 1$ by linearity of ℓ and $\ell(v) = 1$.

(c) Let $v \in \Delta_{\mathcal{PM}}$. By Proposition 2(b), $T_\pi(v) \in \mathcal{PM}$, so $A(v)$ is a nonnegative combination of two elements of $\mathcal{PM}$ and lies in $\mathcal{PM}$ by Lemma 2. Combine with (b).

(d) If $A(v) = v$, multiply by the denominator: $T_\pi(v) + v = (1+\mu)\, v$, and subtract v. Conversely, if $T_\pi(v) = \mu\, v$ with $\mu = \ell(T_\pi(v))$, the same identity gives $A(v) = v$. Finally $\mu \geq 0$ because $T_\pi(v) \geq 0$ and ℓ is nonnegative on the orthant. □

For any eigenvector $v \in \Delta$ of T_π the eigenvalue equals $\ell(T_\pi(v))$. Thus, fixed points of A in Δ are exactly the normalized nonnegative eigenvectors of T_π.

Theorem 4 (Brouwer)**.** *Every continuous self-map of a nonempty compact convex subset of a finite-dimensional real vector space has a fixed point (Granas & Dugundji, 2003, §6).*

Theorem 5 (Pseudometric eigen-pair, no hypothesis on π)**.** *For every $\pi \in \mathcal{P}^+$ there is $v \in \Delta_{\mathcal{PM}}$ with $T_\pi(v) = \mu\, v$ and $\mu = \ell(T_\pi(v)) \geq 0$. Consequently $d := \sqrt{v}$ is a nonzero pseudometric pair with*

$$W_\pi(d) = \lambda\, d, \qquad \lambda := \sqrt{\mu} \geq 0,$$

whose square $d^2 = v$ is again a pseudometric pair, and d and d^2 have the same zero entries.

Proof. By Lemma 6 and Lemma 7(a),(c), A restricts to a continuous self-map of the nonempty compact convex set $\Delta_{\mathcal{PM}}$. Theorem 4 gives a fixed point $v \in \Delta_{\mathcal{PM}}$, and Lemma 7(d) converts it into $T_\pi(v) = \mu\, v$ with $\mu \geq 0$. Then $v \in \mathcal{PM}$ and $v \neq 0$ since $\ell(v) = 1$. By Lemma 3, $d = \sqrt{v}$ is a pseudometric pair with the same zero entries as v, in particular nonzero, and by (10), entrywise, $W_\pi(d)^2 = T_\pi(d^2) = T_\pi(v) = \mu\, v = \mu\, d^2$. Taking square roots of nonnegative entries gives $W_\pi(d) = \sqrt{\mu}\, d$. □

A.4 The zero classes of an eigen-pair form a lumping

Read entrywise, $W_\pi(d) = \lambda\, d$ is

$$\mathcal{W}_2^{d_Y}(\varphi_x, \varphi_{x'}) = \lambda\, d_X(x, x'), \qquad \mathcal{W}_2^{d_X}(\psi_y, \psi_{y'}) = \lambda\, d_Y(y, y'). \tag{11}$$

If $\lambda = 0$ every left-hand side vanishes. If $d_X(x, x') = 0$ for some $x \neq x'$ the first vanishes at that pair. In both cases two conditionals are at transport distance zero over a pseudometric ground cost. We first record what this says about the two conditionals, with no hypothesis on π and none on λ.

Lemma 8 (Zero-cost couplings)**.** *Let c be a ground cost on Z and $\mu, \nu \in \mathcal{P}(Z)$. Then $\mathcal{W}_2^c(\mu, \nu) = 0$ if and only if there is $\gamma \in \Pi(\mu, \nu)$ with $\gamma(z, z') = 0$ whenever $c(z, z') \neq 0$.*

Proof. $\mathcal{W}_2^c(\mu, \nu) = 0$ if and only if $\mathsf{OT}^{c^2}(\mu, \nu) = 0$, if and only if some coupling γ (an optimal one) has $\langle \gamma, c^2 \rangle = 0$. This is a sum of nonnegative terms $\gamma(z, z')\, c(z, z')^2$, so it vanishes if and only if every term does, i.e. if and only if $\gamma(z, z') = 0$ wherever $c(z, z') \neq 0$. □

Definition 7 (Zero-closed set)**.** Let c be a ground cost on Z. A subset $C \subseteq Z$ is *zero-closed* for c if $z \in C$ together with $c(z, z') = 0$ or $c(z', z) = 0$ implies $z' \in C$.

Lemma 9 (Trapped transport)**.** *Let c be a ground cost on Z and $\mu, \nu \in \mathcal{P}(Z)$ with $\mathcal{W}_2^c(\mu, \nu) = 0$. Then $\mu(C) = \nu(C)$ for every zero-closed $C \subseteq Z$.*

Proof. Take γ as in Lemma 8. If $z \in C$ and $z' \notin C$ then $c(z, z') \neq 0$, since otherwise $z' \in C$ by zero-closedness, so $\gamma(z, z') = 0$. Hence

$$\mu(C) = \sum_{z \in C} \sum_{z' \in Z} \gamma(z, z') = \sum_{z \in C} \sum_{z' \in C} \gamma(z, z').$$

If $z' \in C$ and $z \notin C$ then zero-closedness at z', second disjunct, gives $c(z, z') \neq 0$, so $\gamma(z, z') = 0$, and

$$\nu(C) = \sum_{z' \in C} \sum_{z \in Z} \gamma(z, z') = \sum_{z' \in C} \sum_{z \in C} \gamma(z, z').$$

The two right-hand sides are the same finite double sum. □

The pseudometric axioms enter through the supply of zero-closed sets.

Proposition 3 (Zero classes)**.** *Let $d \in \mathcal{PM}$ satisfy $W_\pi(d) = \lambda\, d$ for some $\lambda \in \mathbb{R}$. Define*

$$x \sim_X x' \iff d_X(x, x') = 0, \qquad y \sim_Y y' \iff d_Y(y, y') = 0.$$

Then $\sim_X$ and $\sim_Y$ are equivalence relations and $Q_0(d) := (\sim_X, \sim_Y)$ is a lumping of π.

Proof. Equivalence relations. Reflexivity and symmetry of $\sim_X$ are axioms (ii) and (iii) for d_X. Transitivity is the triangle inequality squeezed by nonnegativity: if $d_X(a, b) = d_X(b, c) = 0$ then $0 \leq d_X(a, c) \leq d_X(a, b) + d_X(b, c) = 0$. Likewise for $\sim_Y$.

Every class is zero-closed. For $C \subseteq Y$, Definition 7 asks that $y \in C$ together with $d_Y(y, y') = 0$ or $d_Y(y', y) = 0$ force $y' \in C$. By symmetry both disjuncts say $y \sim_Y y'$, so the condition is that C be closed under $\sim_Y$, which every $\sim_Y$-class is. Likewise on X.

Clause (1). Let $x \sim_X x'$. The first equation of (11) at (x, x') gives $\mathcal{W}_2^{d_Y}(\varphi_x, \varphi_{x'}) = \lambda \cdot 0 = 0$, whatever λ is. For every $\sim_Y$-class B, which is zero-closed for d_Y, Lemma 9 gives $\varphi_x(B) = \varphi_{x'}(B)$. This is clause (1) of Definition 1. *Clause (2)* is the mirror argument with the second equation of (11). □

The proof uses each pseudometric axiom exactly once: the zero diagonal and symmetry make $\sim_X$, $\sim_Y$ reflexive and symmetric, the triangle inequality with nonnegativity makes them transitive, and it is transitivity that makes the classes zero-closed. Without the triangle inequality the zero relation of d_Y need not be transitive, there are no classes, and a vanishing $\mathcal{W}_2$ says nothing about the two conditionals.

A.5 Proof of Theorem 1

Strategy: Two trivial lumpings exist for every π: the discrete pair, for which both clauses of Definition 1 quantify over nothing, and the indiscrete pair, for which every class is a whole side and both clauses read $1 = 1$. Under Proposition 3, $Q_0(d)$ is indiscrete exactly when $d = 0$ and discrete exactly when d is positive. So every nonzero pseudometric eigen-pair that is not positive produces a nontrivial lumping. These non-trivial lumpings are what unlumpability forbids.

Let $\pi \in \mathcal{P}^+$ be unlumpable. Theorem 5 provides a nonzero pseudometric pair d with $W_\pi(d) = \lambda\, d$, $\lambda \geq 0$, such that $d^2 \in \mathcal{PM}$ has the same zero entries as d.

Step 1: positivity. By Proposition 3, $Q_0(d)$ is a lumping of π, hence trivial. It is not indiscrete, since that would mean $d_X \equiv 0$ and $d_Y \equiv 0$. So it is discrete: $d_X(x, x') = 0$ only for $x = x'$, and mirror. Thus d is positive, i.e. d_X and d_Y are genuine metrics, and $d^2 \in \mathcal{PM}$ is positive as well, i.e. (d_X^2, d_Y^2) is a metric pair.

Step 2: $\lambda > 0$. Suppose $\lambda = 0$. The first equation of (11) then reads $\mathcal{W}_2^{d_Y}(\varphi_x, \varphi_{x'}) = 0$ for all x, x'. Since d_Y is a metric, every singleton $\{y\}$ is zero-closed for d_Y, so Lemma 9 gives $\varphi_x(y) = \varphi_{x'}(y)$ for all x, x', y: all rows of π coincide. Let Q be the partition pair with $\sim_X$ total and $\sim_Y$ equality. Clause (1) of Definition 1 quantifies over all x, x' and the singletons $B = \{y\}$, and asks that the rows agree, which they do. Clause (2) is asked only for $y = y'$, where it is trivial. So Q is a lumping. It is nontrivial: $\sim_X$ is not equality because $|X| \geq 2$, and $\sim_Y$ is not total because $|Y| \geq 2$. This contradicts unlumpability, so $\lambda > 0$.

Step 3: $\lambda \leq 1$. The spread (2) is homogeneous of degree two in the cost, $\sigma^2(s\, e, \beta) = s^2 \sigma^2(e, \beta)$, and it is defined for any ground cost e.

Lemma 10 (Spread contraction)**.** *Let c be a ground cost on a finite set W, let $(\alpha_z)_{z \in Z}$ be a family in $\mathcal{P}(W)$, let $\beta \in \mathcal{P}(Z)$, and let $e(z, z') := \mathcal{W}_2^c(\alpha_z, \alpha_{z'})$ be the induced cost on Z. Then*

$$\sigma^2(e, \beta) \;\leq\; \sigma^2\Big(c,\, \textstyle\sum_z \beta(z)\, \alpha_z\Big).$$

Proof. For each pair (z, z') the product coupling $\alpha_z \otimes \alpha_{z'} \in \Pi(\alpha_z, \alpha_{z'})$ bounds the minimum in Definition 3: $e(z, z')^2 \leq \sum_{w,w'} \alpha_z(w)\, \alpha_{z'}(w')\, c(w, w')^2$. Multiply by $\frac{1}{2}\beta(z)\beta(z') \geq 0$, sum over (z, z'), and exchange the order of summation on the right:

$$\sigma^2(e, \beta) \leq \tfrac{1}{2} \sum_{w,w'} c(w, w')^2 \Big(\sum_z \beta(z)\, \alpha_z(w)\Big)\Big(\sum_{z'} \beta(z')\, \alpha_{z'}(w')\Big) = \sigma^2\Big(c, \textstyle\sum_z \beta(z)\, \alpha_z\Big). \quad \square$$

Apply the lemma with $W = Y$, $c = d_Y$, $\alpha_x = \varphi_x$ and $\beta = \mu_X$. The induced cost is $e = \lambda\, d_X$ by the first equation of (11), and the mixture is $\sum_x \mu_X(x)\, \varphi_x = \mu_Y$ by (9), so $\lambda^2\, \sigma^2(d_X, \mu_X) \leq \sigma^2(d_Y, \mu_Y)$. The mirror application, with the second equation, the columns ψ_y, the ground cost d_X and the weights μ_Y, gives $\lambda^2\, \sigma^2(d_Y, \mu_Y) \leq \sigma^2(d_X, \mu_X)$. Composing the two,

$$\lambda^4\, \sigma^2(d_X, \mu_X) \leq \sigma^2(d_X, \mu_X).$$

Since $|X| \geq 2$, μ_X is everywhere positive and d_X is positive off the diagonal, $\sigma^2(d_X, \mu_X) > 0$. Hence $\lambda^4 \leq 1$, and with $\lambda \geq 0$, $\lambda \leq 1$.

Steps 1–3 prove the first sentence of Theorem 1: d_X and d_Y are genuine metrics solving (1) at some $\lambda \in (0, 1]$, and (d_X^2, d_Y^2) is a metric pair.

Step 4: almost every π is unlumpable. Write $n := |X|\,|Y| \geq 4$. The simplex $\mathcal{P}(X \times Y)$ spans the affine hyperplane $H := \{\pi \in \mathbb{R}^{X\times Y} : \sum \pi = 1\}$, and "Lebesgue-almost-every" refers to Lebesgue measure on H, obtained by identifying H with $\mathbb{R}^{n-1}$ through dropping one coordinate. Different choices of the coordinate change the measure by a constant factor, so the null sets are well defined, and the identification takes polynomials in the entries of π to polynomials on $\mathbb{R}^{n-1}$. We use the standard fact that the zero set of a polynomial that does not vanish identically on $\mathbb{R}^m$ is Lebesgue-null.[1]

Positive marginals. The set of $\pi \in H$ with $\pi(x, y) = 0$ is the zero set of the coordinate polynomial $\pi \mapsto \pi(x, y)$, which is not identically zero on H. Off the finite union of these null sets π has full support, hence positive marginals.

A nontrivial lumping exhibits a coincidence. Let $Q = (\sim_X, \sim_Y)$ be a nontrivial lumping of π. We claim there are $x \neq x'$ with $x \sim_X x'$ and a $\sim_Y$-class $B \neq Y$, or the mirror configuration. Then clause (1) gives $\varphi_x(B) = \varphi_{x'}(B)$ on a set B that is nonempty, being a class, and proper. A relation that is not total has a class avoiding some point, hence a proper class, and a total relation relates two distinct points because its side has at least two. If neither relation is total, both sides have proper classes. Q is not discrete, so one relation relates two distinct points, and the proper class on the other side completes the configuration. If one relation is total, say $\sim_X$, then it relates two distinct points of X, and $\sim_Y$ is not total because Q is not indiscrete, so it has a proper class. The case $\sim_Y$ total is the mirror image.

Each coincidence is a null event. Fix $x \neq x'$ and a proper nonempty $S \subseteq Y$. For π with positive marginals, $\varphi_x(S) = \varphi_{x'}(S)$ holds if and only if $p(\pi) = 0$ for the polynomial

$$p(\pi) := \mu_X(x') \sum_{y\in S} \pi(x, y) \;-\; \mu_X(x) \sum_{y\in S} \pi(x', y) = \mu_X(x)\,\mu_X(x')\big[\varphi_x(S) - \varphi_{x'}(S)\big],$$

the first expression being a polynomial in the entries of π since $\mu_X(x) = \sum_y \pi(x, y)$. It does not vanish identically on H: pick $y_0 \in S$ and $y_1 \notin S$ and evaluate at the coupling π_0 with mass $\frac{1}{2}$ on each of (x, y_0) and (x', y_1). Then $\mu_X(x) = \mu_X(x') = \frac{1}{2}$, $\sum_{y\in S} \pi_0(x, y) = \frac{1}{2}$ and $\sum_{y\in S} \pi_0(x', y) = 0$, so $p(\pi_0) = \frac{1}{4}$. Hence $\{p = 0\} \cap H$ is null. The mirror coincidence $\psi_y(T) = \psi_{y'}(T)$, for $y \neq y'$ and a proper nonempty $T \subseteq X$, is null by the same argument.

Finite union. A π with positive marginals that is lumpable carries a nontrivial lumping, hence lies in the union, over the finitely many triples (x, x', S) and (y, y', T), of the null sets of the previous step. Together with the support hyperplanes this is a finite union of null sets. So Lebesgue-almost-every π has positive marginals and is unlumpable, which is the last sentence of Theorem 1. □

A.6 Proof of Theorem 2

Unlumpable implies no degenerate solution. Let π be unlumpable and let $d \neq 0$ be a pseudometric pair with $W_\pi(d) = \lambda\, d$ for some $\lambda \in \mathbb{R}$. Then $\lambda \geq 0$: some off-diagonal entry $d(e)$ is positive, and $\lambda\, d(e) = W_\pi(d)(e) \geq 0$. Steps 1 and 2 of Section A.5 used only that d is a nonzero pseudometric pair with $W_\pi(d) = \lambda\, d$ and $\lambda \geq 0$. They apply verbatim and thus d is a pair of genuine metrics and $\lambda > 0$.

A nontrivial lumping produces a degenerate solution. For a map $q : Z \to \bar{Z}$ of finite sets and $\alpha \in \mathcal{P}(Z)$ the *pushforward* is $q_*\alpha \in \mathcal{P}(\bar{Z})$, $(q_*\alpha)(\bar{z}) := \alpha(q^{-1}(\bar{z}))$. For the quotient map $q_Y : Y \to \bar{Y} := Y/\!\sim_Y$ of a partition pair, the masses of $(q_Y)_*\varphi_x$ are the block masses $\varphi_x(B)$ in clause (1) of Definition 1: clause (1) says that $x \sim_X x'$ implies $(q_Y)_*\varphi_x = (q_Y)_*\varphi_{x'}$, and clause (2) mirror.

Lemma 11 (Pullback)**.** *Let $q : Z \to \bar{Z}$ be a map of finite sets, $\bar{c}$ a ground cost on $\bar{Z}$, $c := \bar{c} \circ (q \times q)$ its pullback, and $\mu, \nu \in \mathcal{P}(Z)$. Then*

$$\mathsf{OT}^c(\mu, \nu) = \mathsf{OT}^{\bar{c}}(q_*\mu, q_*\nu), \qquad \mathcal{W}_2^c(\mu, \nu) = \mathcal{W}_2^{\bar{c}}(q_*\mu, q_*\nu). \tag{12}$$

[1] By induction on m with Fubini's theorem: for $m = 1$ a nonzero polynomial has finitely many roots. In general expand $q = \sum_k c_k(w)\, t^k$ along one variable t, with coefficients c_k polynomial in $w \in \mathbb{R}^{m-1}$, and let N be the set of w at which every c_k vanishes, which is null in $\mathbb{R}^{m-1}$ by induction. The zero set of q lies in $N \times \mathbb{R}$, a null set, together with a set whose slice over each $w \notin N$ is the finite root set of a nonzero one-variable polynomial. Null slices integrate to a null set.

Proof. "$\geq$": for $\gamma \in \Pi(\mu,\nu)$, the pushforward $\bar{\gamma} := (q \times q)_*\gamma$ lies in $\Pi(q_*\mu, q_*\nu)$, and $\langle \bar{\gamma}, \bar{c}\rangle = \langle \gamma, c\rangle$ because c is constant on the fibres of $q \times q$. "$\leq$": for $\bar{\gamma} \in \Pi(q_*\mu, q_*\nu)$ define

$$\gamma(z,z') := \bar{\gamma}(qz, qz')\,\frac{\mu(z)}{(q_*\mu)(qz)} \cdot \frac{\nu(z')}{(q_*\nu)(qz')},$$

with the convention $\gamma(z,z') := 0$ when $\bar{\gamma}(qz,qz') = 0$. This covers every vanishing denominator, since $\bar{\gamma}(\bar{z},\bar{z}') \leq \min\big((q_*\mu)(\bar{z}), (q_*\nu)(\bar{z}')\big)$. Fix z with $\mu(z) > 0$, so $(q_*\mu)(qz) \geq \mu(z) > 0$. For each $\bar{z}'$ with $(q_*\nu)(\bar{z}') > 0$, $\sum_{z':qz'=\bar{z}'} \nu(z')/(q_*\nu)(\bar{z}') = 1$, and for $\bar{z}'$ with $(q_*\nu)(\bar{z}') = 0$ the terms vanish by the convention. So $\sum_{z'} \gamma(z,z') = \mu(z)\,(q_*\mu)(qz)^{-1} \sum_{\bar{z}'} \bar{\gamma}(qz,\bar{z}') = \mu(z)$. If $\mu(z) = 0$ all terms vanish. The second marginal is symmetric, so $\gamma \in \Pi(\mu,\nu)$, and grouping the sum by fibres, $\langle \gamma, c\rangle = \sum_{\bar{z},\bar{z}'} \bar{c}(\bar{z},\bar{z}') \sum_{qz=\bar{z},\, qz'=\bar{z}'} \gamma(z,z') = \langle \bar{\gamma}, \bar{c}\rangle$. For $\mathcal{W}_2$ apply the identity to $\bar{c}^2$, noting $c^2 = \bar{c}^2 \circ (q \times q)$. □

Lemma 12 (Quotient coupling)**.** *Let $Q = (\sim_X, \sim_Y)$ be a lumping of π, with quotient maps $q_X : X \to \bar{X}$ and $q_Y : Y \to \bar{Y}$, and let $\bar{\pi} := (q_X \times q_Y)_*\pi$, i.e. $\bar{\pi}(A,B) = \pi(A \times B)$ for classes A, B. Then $\bar{\pi}$ has positive marginals, and its conditionals are the pushforwards of those of π:*

$$\bar{\varphi}_{q_X(x)} = (q_Y)_*\varphi_x \quad \textit{for all } x, \qquad \bar{\psi}_{q_Y(y)} = (q_X)_*\psi_y \quad \textit{for all } y. \tag{13}$$

Proof. The marginal of $\bar{\pi}$ at a class A is $\bar{\mu}_{\bar{X}}(A) = \pi(A \times Y) = \mu_X(A) > 0$, and mirror. Its row conditional at A evaluated on a class B is

$$\bar{\varphi}_A(B) = \frac{\pi(A \times B)}{\mu_X(A)} = \sum_{x \in A} \frac{\mu_X(x)}{\mu_X(A)}\,\varphi_x(B).$$

By clause (1) of Definition 1, $\varphi_x(B)$ takes a single value as x ranges over A, and the weights sum to 1, so $\bar{\varphi}_A(B) = \varphi_x(B) = \big((q_Y)_*\varphi_x\big)(B)$ for every $x \in A$. The column identity follows from clause (2) in the same way. □

Now let $Q = (\sim_X, \sim_Y)$ be a nontrivial lumping of π. Since Q is not indiscrete, at most one of the two relations is total.

Neither relation is total. Then $|\bar{X}|, |\bar{Y}| \geq 2$, and $\bar{\pi}$ has positive marginals by Lemma 12, so Theorem 5 applies to $\bar{\pi}$ and yields a nonzero pseudometric pair $\bar{d} = (\bar{d}_X, \bar{d}_Y)$ on $(\bar{X}, \bar{Y})$ with $W_{\bar{\pi}}(\bar{d}) = \bar{\lambda}\,\bar{d}$, $\bar{\lambda} \geq 0$. Set

$$d := \big(\bar{d}_X \circ (q_X \times q_X),\ \bar{d}_Y \circ (q_Y \times q_Y)\big).$$

The pseudometric axioms pull back along any map, so $d \in \mathcal{PM}$, and $d \neq 0$ because the quotient maps are surjective. It solves (1) at $\bar{\lambda}$: for all x, x',

$$\begin{aligned} W_\pi(d)_X(x,x') = \mathcal{W}_2^{d_Y}(\varphi_x, \varphi_{x'}) &\overset{(12)}{=} \mathcal{W}_2^{\bar{d}_Y}\big((q_Y)_*\varphi_x, (q_Y)_*\varphi_{x'}\big) \overset{(13)}{=} \mathcal{W}_2^{\bar{d}_Y}\big(\bar{\varphi}_{q_X(x)}, \bar{\varphi}_{q_X(x')}\big) \\ &= \bar{\lambda}\,\bar{d}_X\big(q_X(x), q_X(x')\big) = \bar{\lambda}\,d_X(x,x'), \end{aligned}$$

and mirror for the columns. Since Q is not discrete, some class contains two distinct points, say $x \neq x'$ with $q_X(x) = q_X(x')$ (or mirror), and then $d_X(x,x') = \bar{d}_X(q_X(x), q_X(x)) = 0$. So d is a nonzero pseudometric eigen-pair that is not a pair of genuine metrics, whatever the value of $\bar{\lambda}$.

One relation is total, say $\sim_X$; the case $\sim_Y$ total is the mirror image. Then $\sim_Y$ is not total, so $|\bar{Y}| \geq 2$. No fixed-point theorem is needed. Let $\bar{\delta}$ be the discrete metric on $\bar{Y}$ and set

$$d := \big(0,\ \bar{\delta} \circ (q_Y \times q_Y)\big), \qquad \lambda := 0.$$

This is a pseudometric pair, nonzero because $\sim_Y$ has two distinct classes. On the X side every pair x, x' is related, so clause (1) gives $(q_Y)_*\varphi_x = (q_Y)_*\varphi_{x'}$, and by (12) and Lemma 4(a), $W_\pi(d)_X(x,x') = \mathcal{W}_2^{\bar{\delta}}\big((q_Y)_*\varphi_x, (q_Y)_*\varphi_x\big) = 0$. On the Y side the ground cost is $d_X = 0$, so every entry of $W_\pi(d)_Y$ vanishes. Hence $W_\pi(d) = 0 = 0 \cdot d$: a nonzero pseudometric pair solving (1) at $\lambda = 0$, which is degenerate. □

B PROOFS FOR SECTION 3

This appendix proves Theorem 3 and Proposition 1, together with the two facts used in the "Locations" paragraph of Section 3: the closed form (3) of the collapse cost and the radius–spread comparison (4). Section B.1 defines the two comparison distances. Section B.2 treats transport to a point mass and the Fréchet data of a family of distributions. Section B.3 proves two reconstruction bounds for such a family, one in the worst case and one in the mean. Section B.4 obtains Theorem 3 by reading the eigen-equations as reconstruction instances, and Section B.5 proves Proposition 1.

The conventions of Section A are used. In particular, for a pseudometric d on a finite set Z, $\mathcal{W}_2^d$ is a pseudometric on $\mathcal{P}(Z)$ (Lemma 4) and $\mathcal{W}_2^d(\delta_z, \delta_{z'}) = d(z, z')$ (Lemma 5(c)). Throughout, (d, λ) is a positive eigen-pair of $\pi \in \mathcal{P}^+$: $d = (d_X, d_Y)$ is a pair of genuine metrics with $W_\pi(d) = \lambda\, d$, so that

$$\mathcal{W}_2^{d_Y}(\varphi_x, \varphi_{x'}) = \lambda\, d_X(x, x'), \qquad \mathcal{W}_2^{d_X}(\psi_y, \psi_{y'}) = \lambda\, d_Y(y, y'), \tag{14}$$

and $\lambda \geq 0$ (some entry of d is positive and $W_\pi(d) \geq 0$). Proposition 1 presupposes $\lambda > 0$.

B.1 COMPARISON DISTANCES

Definition 8 (Gromov–Hausdorff distance)**.** A *correspondence* between finite sets X and Y is a subset $R \subseteq X \times Y$ that projects onto both factors: every x occurs in some pair of R and every y does. For pseudometrics d_X on X and d_Y on Y the *distortion* of R is $\mathsf{dis}(R) := \max_{(x,y),(x',y') \in R} |d_X(x, x') - d_Y(y, y')|$, and

$$d_{\mathrm{GH}}\big((X, d_X), (Y, d_Y)\big) := \tfrac{1}{2} \min_R \mathsf{dis}(R),$$

the minimum over all correspondences (Burago et al., 2001, §7.3). The minimum exists since there are finitely many correspondences, and d_{GH} is symmetric in its two arguments since $R \mapsto R^{\mathsf{T}}$ is a bijection between correspondences with the same distortion.

Definition 9 (Gromov–Wasserstein distance)**.** For pseudometrics d_X, d_Y and $\mu_X \in \mathcal{P}(X)$, $\mu_Y \in \mathcal{P}(Y)$,

$$d_{\mathrm{GW}}\big((X, d_X, \mu_X), (Y, d_Y, \mu_Y)\big) := \tfrac{1}{2} \min_{\gamma \in \Pi(\mu_X, \mu_Y)} \sum_{x,y,x',y'} |d_X(x, x') - d_Y(y, y')|\, \gamma(x, y)\, \gamma(x', y'),$$

the order-one Gromov–Wasserstein distance (Mémoli, 2011). The minimum exists since the objective is continuous on the nonempty compact set $\Pi(\mu_X, \mu_Y)$ (Lemma 1), and d_{GW} is symmetric since $\gamma \mapsto \gamma^{\mathsf{T}}$ is a bijection $\Pi(\mu_X, \mu_Y) \to \Pi(\mu_Y, \mu_X)$ preserving the objective.

A correspondence is a hard alignment and the distortion its worst disagreement; a coupling is a soft alignment and the objective of d_{GW} its mean disagreement, weighted by the mass involved. Both definitions are stated for pseudometrics because the index configurations below are in general only pseudometric.

B.2 TRANSPORT TO A POINT MASS AND FRÉCHET DATA

Lemma 13 (Transport to a point mass)**.** *Let d be a pseudometric on a finite set Z, $\alpha \in \mathcal{P}(Z)$ and $z, z' \in Z$. The only coupling of α and δ_z is $\alpha \otimes \delta_z$, so*

$$\mathcal{W}_2^d(\alpha, \delta_z)^2 = \sum_{z'} \alpha(z')\, d(z', z)^2, \qquad \mathcal{W}_2^d(\delta_z, \delta_{z'}) = d(z, z'), \tag{15}$$

and, averaging the first identity against α,

$$\sum_z \alpha(z)\, \mathcal{W}_2^d(\alpha, \delta_z)^2 = \sum_{z,z'} \alpha(z)\, \alpha(z')\, d(z', z)^2 = 2\, \sigma^2(d, \alpha). \tag{16}$$

Proof. A coupling $\gamma \in \Pi(\alpha, \delta_z)$ has second marginal δ_z, so $\gamma(z', w) = 0$ for $w \neq z$ and then $\gamma(z', z) = \alpha(z')$: $\gamma = \alpha \otimes \delta_z$. Its cost for the ground cost d^2 is the stated sum, and no minimization occurs. The second identity is the case $\alpha = \delta_z$ (or Lemma 5(c)). For (16) multiply the first identity by $\alpha(z)$, sum over z, and compare with (2). □

Applied to $Z = X$, $d = d_X$ and $\alpha = \psi_y$, the first identity of (15) is the closed form (3) of $\rho(x, y)$.

Definition 10 (Fréchet data of a family)**.** Let (Z, d_Z) be a finite metric space, I a finite set, and $(a_i)_{i\in I}$ a family in $\mathcal{P}(Z)$. The *index configuration* of the family is the pseudometric $e(i, i') := \mathcal{W}_2^{d_Z}(a_i, a_{i'})$ on I. For $i \in I$ fix a *Fréchet center* and the *Fréchet radius*,

$$z_i^* \in \arg\min_{z\in Z} \mathcal{W}_2^{d_Z}(a_i, \delta_z), \qquad \rho_i := \min_{z\in Z} \mathcal{W}_2^{d_Z}(a_i, \delta_z) = \mathcal{W}_2^{d_Z}(a_i, \delta_{z_i^*}),$$

ties broken arbitrarily, and aggregate them as the *maximal radius* $\rho_{\max} := \max_i \rho_i$ and the *covering radius* $\eta := \max_{z\in Z} \min_{i\in I} d_Z(z, z_i^*)$ of the centers. For $\mu_I \in \mathcal{P}(I)$ the *mixture* is $\mathsf{mix}(a, \mu_I) := \sum_i \mu_I(i)\, a_i \in \mathcal{P}(Z)$ and the *mean squared spread* is $\bar\sigma^2 := \sum_i \mu_I(i)\, \sigma^2(d_Z, a_i)$.

For the column family $(\psi_y)_{y\in Y}$ in (X, d_X) these are the column data of Section 3: $z_y^* = \iota(y)$, ρ_y, $\rho_{\max} = \rho_{\max}^Y$, $\eta = \eta_Y$, and with $\mu_I = \mu_Y$, $\bar\sigma = \bar\sigma_Y$. For the row family $(\varphi_x)_{x\in X}$ in (Y, d_Y) they are the row data y_x^*, ρ_x, $\rho_{\max}^X$, η_X and, with $\mu_I = \mu_X$, $\bar\sigma_X$.

Lemma 14 (Radius and spread)**.** *Let (Z, d) be a finite metric space, $\alpha \in \mathcal{P}(Z)$, and $\rho := \min_z \mathcal{W}_2^d(\alpha, \delta_z)$. Then*

$$\sigma(d, \alpha)/\sqrt{2} \;\le\; \rho \;\le\; \sqrt{2}\,\sigma(d, \alpha),$$

and $\rho = 0$ if and only if $\sigma(d, \alpha) = 0$ if and only if α is a Dirac mass.

Proof. A minimum is at most an average, so by (16) $\rho^2 \le \sum_z \alpha(z)\, \mathcal{W}_2^d(\alpha, \delta_z)^2 = 2\sigma^2(d, \alpha)$. For the other direction fix z. By the triangle inequality for d and $(a + b)^2 \le 2a^2 + 2b^2$,

$$\begin{aligned} 2\sigma^2(d, \alpha) = \sum_{z', z''} \alpha(z')\, \alpha(z'')\, d(z', z'')^2 &\le 2 \sum_{z', z''} \alpha(z')\, \alpha(z'') \big(d(z', z)^2 + d(z, z'')^2\big) \\ &= 4\, \mathcal{W}_2^d(\alpha, \delta_z)^2, \end{aligned}$$

using (15) and $\sum_{z''} \alpha(z'') = 1$ in the last step. Minimizing over z gives $\sigma^2 \le 2\rho^2$. The two inequalities make $\rho = 0$ and $\sigma = 0$ equivalent. If $\alpha = \delta_{z_0}$ then $\mathcal{W}_2^d(\alpha, \delta_{z_0}) = 0$ by (15), so $\rho = 0$. Conversely, if $\rho = 0$ is attained at z, then $\sum_{z'} \alpha(z')\, d(z', z)^2 = 0$. Thus, $\alpha(z') = 0$ whenever $d(z', z) > 0$, i.e. whenever $z' \ne z$ since d is a metric. It then follows that $\alpha = \delta_z$. □

Applied to $Z = X$, $d = d_X$ and $\alpha = \psi_y$, this is (4) and the statement that ρ_y and $\sigma(d_X, \psi_y)$ vanish exactly when ψ_y is a Dirac mass.

Lemma 15 (Pairwise bound)**.** *Let d be a pseudometric on a finite set Z. For $\alpha, \alpha' \in \mathcal{P}(Z)$ and $z, z' \in Z$,*

$$\big|\mathcal{W}_2^d(\alpha, \alpha') - d(z, z')\big| \le \mathcal{W}_2^d(\alpha, \delta_z) + \mathcal{W}_2^d(\alpha', \delta_{z'}).$$

Proof. $\mathcal{W}_2^d$ is a pseudometric on $\mathcal{P}(Z)$ and $\mathcal{W}_2^d(\delta_z, \delta_{z'}) = d(z, z')$. Two chains of triangle inequalities through δ_z and $\delta_{z'}$,

$$\begin{aligned} \mathcal{W}_2^d(\alpha, \alpha') &\le \mathcal{W}_2^d(\alpha, \delta_z) + d(z, z') + \mathcal{W}_2^d(\delta_{z'}, \alpha'), \\ d(z, z') &\le \mathcal{W}_2^d(\delta_z, \alpha) + \mathcal{W}_2^d(\alpha, \alpha') + \mathcal{W}_2^d(\alpha', \delta_{z'}), \end{aligned}$$

bound the difference from above and from below. □

B.3 Reconstruction from a family of distributions

In the setting of Definition 10, the two lemmas below compare the index configuration (I, e) with the ground space (Z, d_Z). Both rest on Lemma 15, aggregated once in the worst case and once in the mean.

Lemma 16 (Gromov–Hausdorff reconstruction)**.** $d_{\mathrm{GH}}\big((I, e), (Z, d_Z)\big) \le \rho_{\max} + \eta$.

Proof. Let $R := \{(i, z) \in I \times Z : d_Z(z, z_i^*) \le \eta\}$. It is a correspondence: $(i, z_i^*) \in R$ for every i, and for every z the minimum in the definition of η is attained at some i with $d_Z(z, z_i^*) \le \eta$. For $(i, z) \in R$, the triangle inequality for $\mathcal{W}_2^{d_Z}$ through $\delta_{z_i^*}$ and (15) give

$$\mathcal{W}_2^{d_Z}(a_i, \delta_z) \le \mathcal{W}_2^{d_Z}(a_i, \delta_{z_i^*}) + \mathcal{W}_2^{d_Z}(\delta_{z_i^*}, \delta_z) = \rho_i + d_Z(z_i^*, z) \le \rho_{\max} + \eta.$$

For $(i,z),(i',z') \in R$, Lemma 15 with $\alpha = a_i, \alpha' = a_{i'}$ gives $|e(i,i') - d_Z(z,z')| \le \mathcal{W}_2^{d_Z}(a_i,\delta_z) + \mathcal{W}_2^{d_Z}(a_{i'},\delta_{z'}) \le 2(\rho_{\max}+\eta)$. So $\mathsf{dis}(R) \le 2(\rho_{\max}+\eta)$, and halving gives the claim. □

Lemma 17 (Gromov–Wasserstein reconstruction)**.** *For every $\mu_I \in \mathcal{P}(I)$,*

$$d_{\mathrm{GW}}\big((I,e,\mu_I),\,(Z,d_Z,\mathsf{mix}(a,\mu_I))\big) \le \sqrt{2}\,\bar{\sigma}.$$

Proof. The natural coupling. $\gamma(i,z) := \mu_I(i)\,a_i(z)$ lies in $\Pi(\mu_I,\mathsf{mix}(a,\mu_I))$: summing over z returns $\mu_I(i)$ since a_i is a probability vector, and summing over i is the definition of the mixture.

Averaging the pairwise bound. Sum Lemma 15, with $\alpha = a_i, \alpha' = a_{i'}$, against $\gamma(i,z)\,\gamma(i',z')$. The two error terms have the same sum, because $\sum_{i',z'}\gamma(i',z') = 1$ and the summand of the second is the first with the primed and unprimed variables exchanged:

$$\sum_{i,z,i',z'} \gamma(i,z)\,\gamma(i',z')\,\big|e(i,i') - d_Z(z,z')\big| \;\le\; 2\sum_{i,z}\gamma(i,z)\,\mathcal{W}_2^{d_Z}(a_i,\delta_z).$$

Two applications of Cauchy–Schwarz. For fixed i, since a_i is a probability vector, Cauchy–Schwarz and (16) give

$$\sum_z a_i(z)\,\mathcal{W}_2^{d_Z}(a_i,\delta_z) \le \Big(\sum_z a_i(z)\,\mathcal{W}_2^{d_Z}(a_i,\delta_z)^2\Big)^{1/2} = \sqrt{2}\,\sigma(d_Z,a_i),$$

and then, since μ_I is a probability vector,

$$\sum_i \mu_I(i)\,\sqrt{2}\,\sigma(d_Z,a_i) \le \sqrt{2}\Big(\sum_i \mu_I(i)\,\sigma^2(d_Z,a_i)\Big)^{1/2} = \sqrt{2}\,\bar{\sigma}.$$

Hence $\sum_{i,z}\gamma(i,z)\,\mathcal{W}_2^{d_Z}(a_i,\delta_z) = \sum_i \mu_I(i)\sum_z a_i(z)\,\mathcal{W}_2^{d_Z}(a_i,\delta_z) \le \sqrt{2}\,\bar{\sigma}$, the objective of Definition 9 at γ is at most $2\sqrt{2}\,\bar{\sigma}$, and halving gives the claim. □

The worst-case bound needs two explicit conditions, covering (η small) and locality ($\rho_{\max}$ small), and no measure. The mean bound needs locality only in the mean, so a few spread-out distributions of small μ_I-mass are tolerated, and the comparison is against the mixture measure, so it is informative about a part of Z only to the extent that the mixture weights it.

B.4 Proof of Theorem 3

Column data. Take $Z = X$ with the metric d_X, $I = Y$, and the family $a_y = \psi_y$. By the second equation of (14), the index configuration is $e(y,y') = \mathcal{W}_2^{d_X}(\psi_y,\psi_{y'}) = \lambda\,d_Y(y,y')$, i.e. $(I,e) = (Y,\lambda d_Y)$, and the Fréchet data of the family are the column data $\iota(y)$, ρ_y, $\rho^Y_{\max}$, η_Y, $\bar{\sigma}_Y$. Lemma 16 and the symmetry of d_{GH} give $d_{\mathrm{GH}}\big((X,d_X),(Y,\lambda d_Y)\big) \le \rho^Y_{\max} + \eta_Y$. With $\mu_I = \mu_Y$, the mixture is $\sum_y \mu_Y(y)\,\psi_y = \mu_X$ by (9), so Lemma 17 and the symmetry of d_{GW} give $d_{\mathrm{GW}}\big((X,d_X,\mu_X),(Y,\lambda d_Y,\mu_Y)\big) \le \sqrt{2}\,\bar{\sigma}_Y$.

Row data. Take $Z = Y$ with the metric d_Y, $I = X$, and the family $a_x = \varphi_x$. By the first equation of (14) the index configuration is $(X,\lambda d_X)$, the Fréchet data are the row data, and with $\mu_I = \mu_X$ the mixture is $\sum_x \mu_X(x)\,\varphi_x = \mu_Y$ by (9). Lemmas 16 and 17 give the two row bounds directly.

Permutation couplings. Let $T : X \to Y$ be a bijection and $\pi = \sum_x \mu_X(x)\,\delta_{(x,T(x))}$. Then $\mu_Y = T_*\mu_X$, $\varphi_x = \delta_{T(x)}$ and $\psi_y = \delta_{T^{-1}(y)}$, so by (15) the eigen-equations (14) read

$$d_Y\big(T(x),T(x')\big) = \lambda\,d_X(x,x'), \qquad d_X\big(T^{-1}(y),T^{-1}(y')\big) = \lambda\,d_Y(y,y').$$

Substituting the first into the second at $y = T(x)$, $y' = T(x')$ gives $d_X(x,x') = \lambda^2 d_X(x,x')$, and $d_X(x,x') > 0$ for some $x \ne x'$ since $|X| \ge 2$. Thus, $\lambda^2 = 1$ and $\lambda = 1$ since $\lambda \ge 0$. The first equation then says $d_Y = T_*d_X$, i.e. $d_Y(y,y') = d_X(T^{-1}(y),T^{-1}(y'))$. Conversely, for every metric d_X the pair (d_X, T_*d_X) solves both equations at $\lambda = 1$ by the same computation, so a permutation coupling has a positive eigen-pair for every d_X. Every conditional is a Dirac mass, so by Lemma 14 every radius and every spread vanishes: $\rho^Y_{\max} = \rho^X_{\max} = 0$ and $\bar{\sigma}_X = \bar{\sigma}_Y = 0$.

The center of the column y is the unique minimizer of $x \mapsto \mathcal{W}_2^{d_X}(\delta_{T^{-1}(y)}, \delta_x) = d_X(T^{-1}(y), x)$, namely $\iota(y) = T^{-1}(y)$, so the centers exhaust X and $\eta_Y = \max_x \min_y d_X(x, T^{-1}(y)) = 0$. In mirror $y_x^* = T(x)$ and $\eta_X = 0$. Every right-hand side in Theorem 3 is therefore zero, and T is an isometry $(X, d_X) \to (Y, d_Y)$. □

The deterministic case is also forced from one side alone

Lemma 18 (Point-mass columns)**.** *Suppose every column conditional is a Dirac mass, $\psi_y = \delta_{S(y)}$ for a map $S : Y \to X$, and (d, λ) is a positive eigen-pair of π with $\lambda > 0$. Then S is a bijection and $\pi = \sum_x \mu_X(x)\, \delta_{(x,T(x))}$ with $T := S^{-1}$.*

Proof. By (15) the second equation of (14) reads $\lambda\, d_Y(y, y') = d_X(S(y), S(y'))$. If $S(y) = S(y')$ for some $y \neq y'$ then $d_Y(y, y') = 0$, contradicting positivity of d_Y and thus S is injective. By (9), $\mu_X(x) = \sum_y \mu_Y(y)\, \psi_y(x) = \sum_{y:S(y)=x} \mu_Y(y)$, and $\mu_X(x) > 0$ forces some y with $S(y) = x$ and thus S is surjective. Finally $\pi(x, y) = \mu_Y(y)\, \psi_y(x) = \mu_Y(y)\, \mathbf{1}[x = S(y)]$, which is the stated coupling with $\mu_X(x) = \mu_Y(T(x))$. □

Remark 2 (Removing the scale)**.** Theorem 3 compares (X, d_X) with the rescaled $(Y, \lambda d_Y)$, which is where (14) places the exact identity. Since $\lambda \leq 1$ (Section A.5, Step 3), the identity correspondence gives $d_{\mathrm{GH}}\big((Y, d_Y), (Y, \lambda d_Y)\big) \leq \frac{1}{2}(1-\lambda) \max_{y,y'} d_Y(y, y')$, and the diagonal coupling gives $d_{\mathrm{GW}}\big((Y, d_Y, \mu_Y), (Y, \lambda d_Y, \mu_Y)\big) \leq \frac{1}{2}(1-\lambda) \sum_{y,y'} \mu_Y(y)\, \mu_Y(y')\, d_Y(y, y')$. The triangle inequalities for d_{GH} (Burago et al., 2001, §7.3) and d_{GW} (Mémoli, 2011, Theorem 5.1) then bound the unscaled distances between (X, d_X) and (Y, d_Y) by the column bounds of Theorem 3 plus these terms, and in mirror for the row bounds. At $\lambda = 1$ the corrections vanish.

B.5 Proof of Proposition 1

Let $\lambda > 0$ and fix a row x. For columns y, y' the second equation of (14) and the triangle inequality for $\mathcal{W}_2^{d_X}$ through δ_x give

$$\lambda\, d_Y(y, y') = \mathcal{W}_2^{d_X}(\psi_y, \psi_{y'}) \leq \mathcal{W}_2^{d_X}(\psi_y, \delta_x) + \mathcal{W}_2^{d_X}(\delta_x, \psi_{y'}) = \rho(x, y) + \rho(x, y'). \tag{17}$$

Square, use $(a+b)^2 \leq 2a^2 + 2b^2$, and average against $\varphi_x(y)\, \varphi_x(y')$:

$$\begin{aligned} 2\lambda^2\, \sigma^2(d_Y, \varphi_x) &= \lambda^2 \sum_{y,y'} \varphi_x(y)\, \varphi_x(y')\, d_Y(y, y')^2 \\ &\leq \sum_{y,y'} \varphi_x(y)\, \varphi_x(y') \big(2\rho(x, y)^2 + 2\rho(x, y')^2\big) = 4 \sum_y \varphi_x(y)\, \rho(x, y)^2, \end{aligned}$$

using $\sum_{y'} \varphi_x(y') = 1$ in the last step. Dividing by $2\lambda^2$ gives the pointwise bound of Proposition 1.

For the mean bound, average the pointwise bound against μ_X, use $\mu_X(x)\, \varphi_x(y) = \pi(x, y) = \mu_Y(y)\, \psi_y(x)$ from (9) to exchange the order of conditioning, and then (16) with $\alpha = \psi_y$, $d = d_X$, recalling $\rho(x, y) = \mathcal{W}_2^{d_X}(\psi_y, \delta_x)$:

$$\begin{aligned} \bar{\sigma}_X^2 = \sum_x \mu_X(x)\, \sigma^2(d_Y, \varphi_x) &\leq \frac{2}{\lambda^2} \sum_{x,y} \pi(x, y)\, \rho(x, y)^2 = \frac{2}{\lambda^2} \sum_y \mu_Y(y) \sum_x \psi_y(x)\, \rho(x, y)^2 \\ &= \frac{2}{\lambda^2} \sum_y \mu_Y(y)\, 2\sigma^2(d_X, \psi_y) = \frac{4}{\lambda^2}\, \bar{\sigma}_Y^2. \end{aligned}$$

Taking square roots, $\bar{\sigma}_X \leq \frac{2}{\lambda}\, \bar{\sigma}_Y$. The mirror bound $\bar{\sigma}_Y \leq \frac{2}{\lambda}\, \bar{\sigma}_X$ is the same argument with the first equation of (14), the row cross cost $\mathcal{W}_2^{d_Y}(\varphi_x, \delta_y)$ in place of $\rho(x, y)$, the average against $\psi_y \otimes \psi_y$ and then against μ_Y, and (16) with $\alpha = \varphi_x$, $d = d_Y$. If $\bar{\sigma}_Y = 0$ then $\bar{\sigma}_X \leq 0$, hence $\bar{\sigma}_X = 0$, and in mirror. □

C Formal verification

Every mathematical statement of this paper is formalized and proved in Lean 4 (de Moura & Ullrich, 2021) over mathlib (The mathlib Community, 2020). This includes the numbered theorems,

propositions, lemmas and remarks, the claims made inside definitions, for instance that a minimum is attained, and the mathematical facts stated in the text, from the pair law of Section 4 to the alternating chain of Section 5. Brouwer's fixed point theorem, which is not in mathlib, is proved from singular homology, and the triangle inequalities for d_{GH} and d_{GW} that Remark 2 cites are proved as well. The proofs contain no gaps and use only the three standard axioms of Lean, propositional extensionality, choice and quotient soundness.

A formal proof certifies its formal statement, so what a reader has to trust is that each formal statement says what the printed one says. We therefore keep the statements apart from the proofs. One file states all 58 results, written from the text of this paper and without proofs. The comparator tool of the Lean developers (Lean FRO, 2025) checks the proofs against this file. It rebuilds both, checks that every proved theorem has exactly the type of its statement in the file and uses only the standard axioms, and replays every proof and everything it depends on through the Lean kernel. Changing a single statement in the file, for instance $\lambda \le 1$ into $\lambda < 1$ in Remark 2, makes the check fail.

To trust the results a reader therefore compares this file with the paper, together with the definitions it uses, such as the transport distance, the spread and the two comparison distances, and need not read the proofs. This also makes the results independent of how the proofs were produced. The file, the Lean development and the configuration of the check are in the supplementary material. The written proofs in Sections A and B are for the reader and are not themselves checked. The empirical results of Section 4 are outside the formalization.

D Details for Section 4

Training. Encoder and decoder are initialized as one Gaussian matrix and its transpose, biases at zero. We use 4000 full-batch Adam steps at rate $2 \cdot 10^{-3}$, gradient norm clipped at 1. The reconstruction loss is the mean squared error over the mean squared norm of a sample. Decoder rows are renormalized after each step. The penalty is off for 500 steps and ramped in over 800. Constants for both experiments: $\beta = 0.3$, $\beta' = 10$, $c = 0.7$, $\alpha = 0.01$. The codes trained with the sparsity penalty alone are (6) at $\beta = 0$. We use dead-latent resampling, re-injecting random directions into the frozen metric.

The running eigen-pair. The metric is computed on a fixed probe, every sixth sample (240 of 1440). Every 200 steps the eigen-pair of the probe coupling is refreshed by two alternations warm-started from the previous d_Y. Each d_X is a batched entropic transport between the rows over the k features (regularization 0.1 of the mean cost, 60 Sinkhorn iterations). Each d_Y is the pseudometric between the Fréchet centres of the columns in d_X, $y \mapsto \arg\min_x \sum_{x'} \psi_y(x') d_X(x, x')^2$. This is the point-mass reconstruction of Section B.2. By Lemmas 14 and 15 it differs from the transport distance between two features by at most $\sqrt{2}\,(\sigma(d_X, \psi_y) + \sigma(d_X, \psi_{y'}))$, and the hinge drives the mean of these spreads down. The metrics are then normalized, d_X by one factor that sets its mean over the pairs of each sample and its ten nearest neighbours in d_{ref} to the data's, d_Y to unit mean, and frozen. Between refreshes the spread and V are computed on a random sub-batch of 120 probe samples with the feature masses frozen. Gradients pass through the coupling. A width-256 run on Gemma with exact transports between features reproduces the width-1024 result ($\rho_s = 0.88$, $H_1 = 0.34$ and 0.37 at $\tau = 0.3$, period position), so the Fréchet-centre metric is not what produces it.

Local normalization and floor. Figure 5. Under a global unit-mean normalization a segment has a larger mean distance than a circle at the same local scale, so cutting the circle open lowers the spread of every feature away from the cut at no cost, and a trained code does that (first panel); the floor alone still straightens (second). The local normalization alone is met on average, by collapsing one half of the circle onto a few latents and stretching the other (third; learned local scale 0.5–0.7 of the data's on the collapsed half, 2.2 on the other). With both the code tiles and the local scale tracks the data's around the clock, 0.8–1.2 (fourth). Before training, on five cached width-64 codes (the sparsity code at $\alpha = 0.03$, which tiles the ring, the same code with the spread hinge added under the global normalization and no floor, at $\tau = 0.6$, where it keeps the ring, and at $\tau = 0.4$, where it folds the ring onto itself, a code trained on the spread alone, which collapses the ring onto a few latents at $\lambda = 1$, and the code at $\alpha = 0$ with every latent active), (6) ranks the tiled code first at every τ. The spread in the code's own metric under a global normalization ranks the collapsed or the folded code

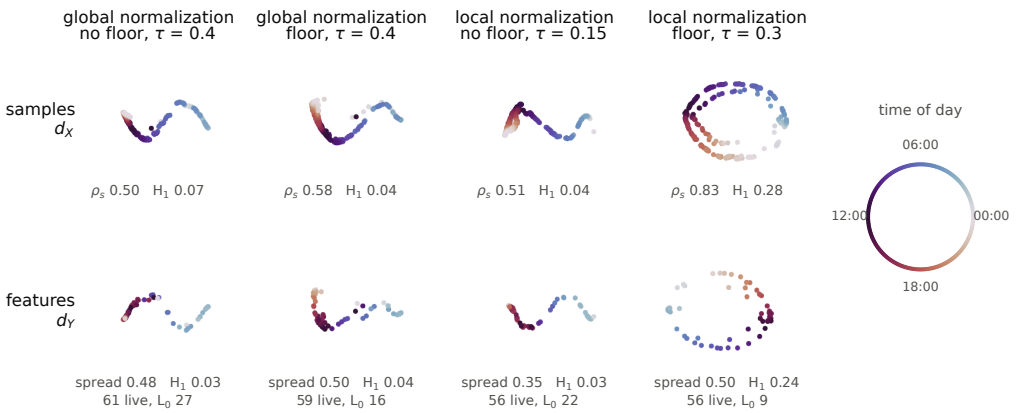

Figure 5: Four codes trained with the spread penalty on the residual at the period of "The time is now t.", block 18, width 64, $\alpha = 0.03$, seed 0. Each column is one code. Top: MDS of (X, d_X) over the 240 probe prompts, coloured by the time of day (clock at right). Bottom: MDS of (Y, d_Y) over the live latents, coloured by the circular mean of a latent's profile (colour only). Under each panel ρ_s and the H_1 bar, resp. the relative spread, H_1 bar, live latents and L_0. With d_X at unit mean over all pairs (global normalization, first two columns, $\tau = 0.4$) the code opens the ring into an arc, with or without the floor: at unit mean an arc has a smaller local scale than a ring, so the spread of every feature away from the cut falls at no cost. With the local normalization but no floor (third, $\tau = 0.15$) the code shrinks d_X on one half of the day, collapsing it onto a few latents, and stretches the other half, so the local scale matches the data's only on average. With both (fourth, $\tau = 0.3$) the latents tile the day and both sides are rings.

first, and the spread in any of eight fixed data metrics ranks the collapsed code first, since a feature concentrated on a few nearby samples has a small spread in any fixed metric.

Readout. A latent is deemed alive if it is active on at least five probe samples. The eigen-pair of the readout coupling is computed by power iteration of T_π in the squared coordinate. Row costs under the column metric, normalized to maximum one, then column costs under that row metric. λ is the fourth root of the product of the two normalizers. Transport is entropic (60 Sinkhorn iterations, regularization 0.02 on the max-normalized squared metric), 300 iterations or a residual below 10^{-8}. Feature-side scores are taken on the rank-transformed metrics (pairwise distances replaced by their ranks, scaled to $[0, 1]$), as the Spearman correlation does on the sample side. In our runs the eigen-metrics of the cyclic toy coupling are concave in the cyclic distance, and the raw d_{GW} measures that profile (the true features score 0.083 against a random metric's 0.092). H_1 is the longest bar of the Vietoris–Rips diagram in degree one (Bauer, 2021). The score d_{GW} is the square-loss Gromov–Wasserstein distance (Mémoli, 2011; Flamary et al., 2021) with the coupling marginals as masses, not the order-one distance of Definition 9 that Theorem 3 bounds. The ideal circle on 64 points scores $H_1 = 0.66$ and $d_{\mathrm{GW}} = 0$; a random metric 0.23 and 0.12.

The toy circle. Table 2, medians over seeds. The codes trained with the spread penalty are the same code at every τ: 64 tiles covering the circle have spread 0.17 in the local normalization of the loss, so the hinge is active only below $\tau = 0.17$ and then moves it by 0.002; the floor penalty alone puts the code there. Unimodal von Mises features at 64 regular positions span relative spread 0.99 to 0.38 for concentrations 0.25 to 40 and score $H_1 = 0.64$–0.65 throughout: a feature-side circle exists at every spread, and the threshold in Figure 3 is a property of the trained codes, not of the bound.

The toy sphere. The same experiment on a 2-manifold (Figure 6 and Table 3). Given: 1440 vectors in $\mathbb{R}^{64}$, one per point of a Fibonacci lattice on S^2 (spacing 5°), the von Mises–Fisher bump $u(x)_j \propto \exp(10\,\langle x, c_j\rangle)$ at 64 lattice centres c_j, normalized to sum one, plus the same noise. The true features read at $\lambda = 0.961$, relative spread 0.65, $\rho_s = 1.00$ against the great-circle distance, and its feature geometry is the ideal sphere on 64 points, $H_2 = 0.61$, $H_1 = 0.06$ (a random metric 0.24

Table 2: Toy circle, width 64, medians over three seeds. FVU: reconstruction error; live: latents entering the coupling; $\bar{\sigma}_Y$: relative feature spread; H_1, d_{GW}: longest bar of d_Y and its Gromov–Wasserstein distance to the true features (rank metrics); ρ_s: Spearman correlation of d_X with the true cyclic distance.

code	weight	FVU	live	λ	$\bar{\sigma}_Y$	H_1	d_{GW}	ρ_s
true features	—	—	64	0.961	0.61	0.64	0.000	1.00
random metric	—	—	64	—	—	0.23	0.124	—
spread, $\alpha = 0.01$	$\tau \in [0.2, 0.5]$	0.020	64	0.998	0.35	0.61	0.001	1.00
	$\tau \in [0.025, 0.15]$	0.022	64	0.998	0.34	0.61	0.001	1.00
sparsity	$\alpha = 0$	0.000	64	0.603	0.97	0.17	0.051	0.89
	0.003	0.000	64	0.714	0.93	0.11	0.092	0.94
	0.01	0.001	64	0.788	0.88	0.20	0.118	0.94
	0.03	0.004	64	0.859	0.81	0.09	0.126	0.97
	0.1	0.017	64	0.983	0.53	0.19	0.017	0.98
	0.2	0.023	64	0.994	0.43	0.49	0.005	0.98
	0.3	0.026	58	0.994	0.41	0.57	0.012	0.94
	1	0.051	30	0.996	0.39	0.57	0.009	0.96

Table 3: Toy sphere, width 64, medians over three seeds. Columns as in Table 2, with H_2, H_1 the longest bars of d_Y in degrees two and one and ρ_s against the great-circle distance.

code	weight	FVU	live	λ	$\bar{\sigma}_Y$	H_2	H_1	d_{GW}	ρ_s
true features	—	—	64	0.961	0.65	0.61	0.06	0.000	1.00
random metric	—	—	64	—	—	0.24	0.23	0.112	—
spread, $\alpha = 0.01$	$\tau = 0.5$	0.010	64	0.893	0.84	0.19	0.23	0.058	0.63
	0.45	0.014	64	0.937	0.75	0.25	0.09	0.027	0.93
	0.4	0.023	64	0.983	0.58	0.56	0.07	0.010	0.98
	0.35	0.037	64	0.992	0.52	0.58	0.07	0.008	0.99
	$\tau \in [0.025, 0.3]$	0.047	64	0.994	0.50	0.59	0.07	0.009	0.99
sparsity	$\alpha = 0$	0.000	64	0.509	0.99	0.11	0.16	0.081	0.58
	0.003	0.000	64	0.662	0.98	0.14	0.19	0.079	0.65
	0.01	0.001	64	0.755	0.95	0.07	0.16	0.068	0.77
	0.03	0.004	64	0.835	0.89	0.16	0.12	0.038	0.94
	0.1	0.021	64	0.968	0.67	0.50	0.07	0.010	0.97
	0.2	0.039	64	0.987	0.56	0.58	0.06	0.009	0.98
	0.3	0.057	64	0.988	0.55	0.57	0.06	0.008	0.98
	1	0.100	62	0.993	0.51	0.59	0.07	0.010	0.98

and 0.23); H_2 is the longest bar of the rank-transformed d_Y in degree two. Width 64, $\alpha = 0.01$, τ from 0.5 to 0.025, the sparsity penalty alone along α, three seeds.

On the sphere. The codes trained with the spread penalty tile the sphere from $\tau = 0.4$ down: $H_2 = 0.55$–0.59, $H_1 \leq 0.07$, $d_{\mathrm{GW}} \leq 0.011$, $\rho_s \geq 0.98$, 64 live latents. Unlike the circle, the floor penalty alone does not tile it: at $\tau = 0.5$ the hinge is inactive and the code stays dense ($L_0 = 32$–40, spread 0.83–0.87, feature side at the random level); and lowering τ trades reconstruction for spread: from $\tau = 0.5$ to 0.3 the spread falls from 0.85 to 0.49, L_0 from 36 to 3 and FVU rises from 0.010 to 0.047, and below 0.3 every τ gives the same code. Sparsity alone is at the random level up to $\alpha = 0.03$, in transition at 0.1 and a sphere from $\alpha = 0.2$ on, at FVU 0.039–0.10. Pooled, every code with relative spread ≤ 0.59 is a sphere and no code with spread ≥ 0.74 is; the true features sit at 0.65. A dense code does not recover the sphere on the sample side either ($\rho_s = 0.55$–0.67 at $\alpha \leq 0.003$).

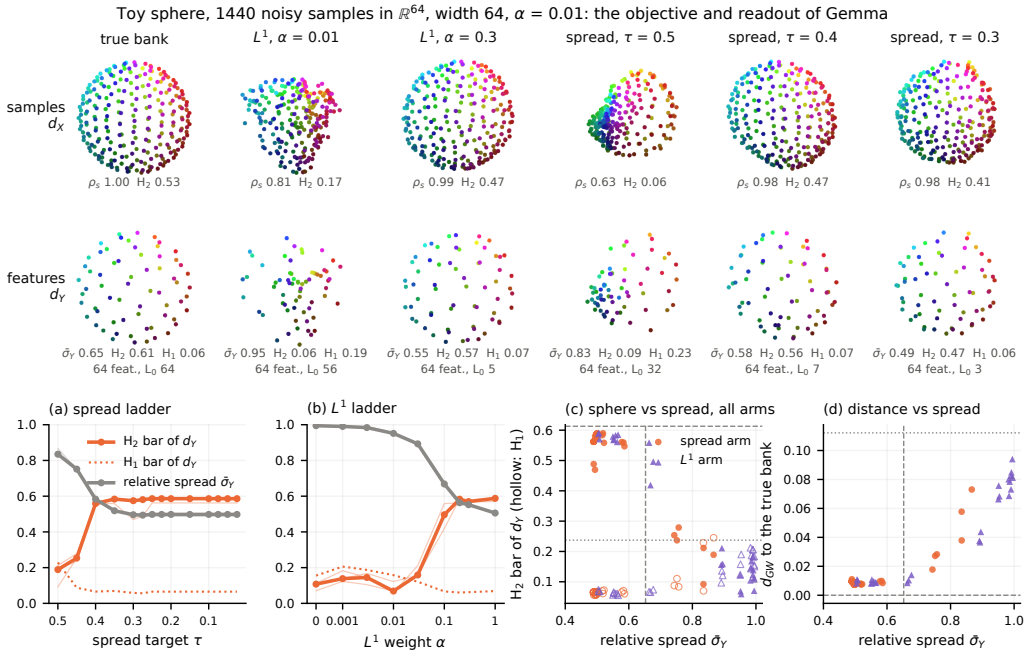

Figure 6: Toy sphere. Top: classical MDS in $\mathbb{R}^3$ of (X, d_X) and of (Y, d_Y), aligned to the true positions and seen from one side (hue: longitude; light to dark: north to south). Under each panel ρ_s and the H_2 bar of d_X on 100 landmarks, resp. relative spread, the H_2 and H_1 bars of d_Y, live latents and L_0. Bottom: (a) the spread across values of τ and (b) the sparsity penalty alone against α, medians of three seeds (thin: seeds); (c, d) every code of both families against its relative spread

E Details for Section 4.2

Prompts and activations. Gemma-2-2b (`google/gemma-2-2b`, bfloat16, no fine-tuning), the hidden state after block ℓ at the last token of the prompt. Activations are centred and rescaled to the mean squared norm of the toy data. d_{ref} is the Euclidean distance between the centred activations scaled to unit length, $\sqrt{2-2\cos}$ of their angle, scaled to unit mean over pairs. Readout on every sixth prompt (every second for year, all for temperature), 100 solver iterations. The raw activations are read on their 64 highest-variance coordinates after centring.

The position. The residual at the period of "The time is now t." at block 18 has $\rho_s = 0.68$ with the day and governs the token predicted there (replacing it by another time's moves the decoded time by 0.98 of the shift). At the token *Good* of "The time is now t. Good" the residual decides the greeting: patching it at blocks 12, 18, 24 makes the greeting the target's in 0.08, 0.49, 0.98 of the trials, and its raw ρ_s with the day is 0.62, 0.76, 0.78. The four greeting words carry 0.62 of the mass at that token. Their natural frequencies over the day are 0.34, 0.07, 0.09, 0.09. Section 4.2 uses this position at block 24.

Along τ. Table 4: at *Good*, block 24, rings on both sides at $\tau = 0.4, 0.35, 0.3$ in all three seeds ($\rho_s = 0.95$–0.97. $H_1 = 0.21$–0.47 on the samples and 0.34–0.57 on the features), the ring breaking at 0.275 ($\rho_s = 0.88$–0.91, $H_1 = 0.27$–0.41) and cut at 0.25 in two seeds ($\rho_s = 0.71, 0.78$; $H_1 \leq 0.06$), the third keeping it (0.89; 0.30, 0.39). The same objective at the period, block 18 (Figure 7): discs ordered by time at $\tau = 0.4$ and 0.35, rings on both sides at 0.3 ($\rho_s = 0.89$, $H_1 = 0.35$ and 0.46, 210 live latents, $L_0 = 17$, FVU 0.005), cut from 0.25 on ($\rho_s \leq 0.66$, $H_1 \leq 0.05$), the transition at the same τ in three seeds. From $\tau = 0.4$ to 0.25, λ rises from 0.986 to 0.999.

Along α, sparsity alone. At width 1024, $\alpha \in \{0, 0.003, 0.01, 0.03, 0.1, 0.3\}$: spread $0.78, 0.71, 0.61, 0.53, 0.44, 0.44$, live latents $330 \to 59$, $L_0 = 82 \to 6$, FVU $0.002 \to 0.023$,

Table 4: Time of day at the token *Good*, block 24, width 1024, $\alpha = 0.01$, medians over three seeds. FVU: reconstruction error; live: latents entering the coupling; $\bar{\sigma}_Y$: relative feature spread; ρ_s: Spearman correlation of d_X with the day; H_1: longest bar of d_X and of d_Y.

τ	FVU	live	L_0	λ	$\bar{\sigma}_Y$	ρ_s	$H_1(d_X)$	$H_1(d_Y)$
0.4	0.004	226	17	0.990	0.49	0.96	0.42	0.52
0.35	0.004	231	18	0.990	0.49	0.96	0.39	0.44
0.3	0.004	235	18	0.989	0.49	0.96	0.41	0.49
0.275	0.005	121	11	0.990	0.48	0.90	0.30	0.30
0.25	0.013	118	9	0.996	0.36	0.78	0.06	0.04

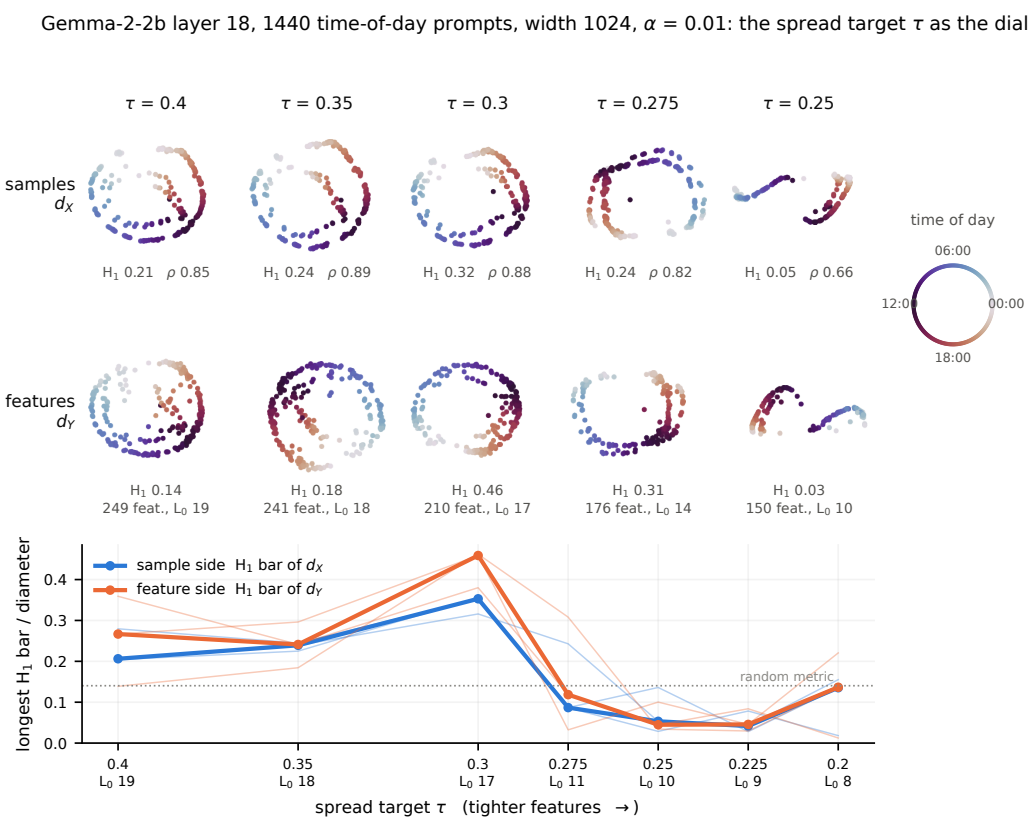

Figure 7: Time of day at the period, block 18, width 1024, seed 0. Top: MDS of (X, d_X) over the 240 probe prompts and of (Y, d_Y) over the live latents along τ, coloured by the time of day or by the circular mean of a latent's profile (colour only). Under each panel H_1 bar, ρ_s, live latents, L_0. Bottom: the H_1 bars of both sides against τ, median of three seeds (thin: seeds). Dotted: a random metric.

$\rho_s = 0.83$–0.94 with $H_1 \leq 0.35$ on the samples (a ring at $\alpha = 0.03$), and on the features $H_1 = 0.10, 0.14, 0.18, 0.20, 0.19, 0.12$.

The second ring. Sampling residual stream activations at the period yields two concentric copies of the day in the WSV geometry of spread regularized representations. The inner one corresponds to the prompts at a full or half hour. In the raw residual a `:00` or `:30` prompt sits 5.2 times farther from its same-hour neighbours than they sit from each other (3.4 for `:15` and `:45`). At *Good* the ratio is 3.4 and the trained code has one ring.

Released dictionaries. Gemma Scope (Lieberum et al., 2024) releases JumpReLU dictionaries for the residual stream of every layer. We applied it to the raw residual of the 1440 prompts as trained,

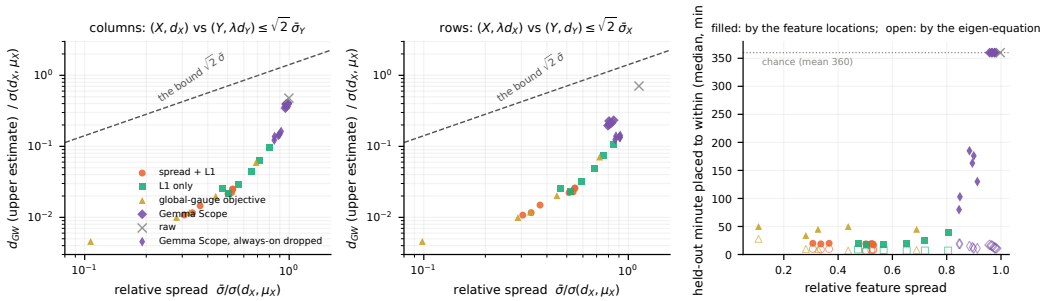

Figure 8: The bound and the locations on the activations, every time-of-day code of the period position, block 18. (a, b) The Gromov–Wasserstein distance between the two sides in the column and in the row data of Theorem 3, an upper estimate, against the relative spread, in units of $\sigma(d_X, \mu_X)$. Dashed, the bounds $\sqrt{2}\,\bar{\sigma}_Y$ (a) and $\sqrt{2}\,\bar{\sigma}_X$ (b). (c) A held-out minute placed by the locations of the latents it fires (filled) and by the eigen-equation (open), median cyclic error, against the spread.

$a = \mathrm{pre}\,[\mathrm{pre} > \theta]$, $\mathrm{pre} = xW_e + b_e$, and read exactly as the trained codes. At *Good*, block 24, the five dictionaries at width 16k ($L_0 = 20, 38, 73, 158, 457$ on text) and one at 65k ($L_0 = 19$) have 28–908 live latents on the prompts, 6–203 of them always on, relative spread 0.95–0.97, $\lambda = 0.63$–0.68, $\rho_s = 0.58$–0.84, $H_1 \le 0.26$ on the samples and ≤ 0.19 on the features. The fraction of the between-prompt variance a dictionary explains is negative (FVU 2.1–5.5), and dropping the latents active on every prompt lowers the spread to 0.75–0.89 and changes nothing else.

At the period, block 18**.** Six Gemma Scope dictionaries at spread 0.96–0.98, $\rho_s \le 0.66$, $H_1 \le 0.20$ and 0.18; the sparsity penalty alone, along α, at spread $0.81 \to 0.47$, $\rho_s \le 0.78$, $H_1 \le 0.15$ on the samples and ≤ 0.33 on the features (a partial loop once, at $\alpha = 0.1$). Over all 40 codes of the period position the feature-side ring occurs only in a band of relative spread 0.50–0.53, and not for every code in it: below the band the target has been met by straightening ($\rho_s \le 0.66$) or, without the local normalization and the floor, by collapsing the day onto a few latents ($\lambda \to 1$). Across these codes, low spread is necessary for the ring and not sufficient. What it certifies is that the feature geometry mirrors the sample geometry, whatever that geometry is.

The bound. For each code of the period position, λd_Y is computed directly as $\mathcal{W}_2^{d_X}(\psi_y, \psi_{y'})$ over the latent pairs, so no normalization enters, and d_{GW} between the two sides as the objective of Definition 9 at the coupling returned by a square-loss solver (Flamary et al., 2021) from two starts, a feasible coupling and so an upper estimate. On all 31 codes of the period position, trained, released and raw, both Gromov–Wasserstein inequalities of Theorem 3 hold, with a factor 2–40 of room, and d_{GW} lies on one curve against the spread, from 0.005 at spread 0.11 to 0.48 at 1.00 in units of $\sigma(d_X, \mu_X)$ (Figure 8).

Locations as a decoder. Every latent y has a Fréchet center $\iota(y)$ of Section 3, the probe prompt that best stands in for its profile. To decode a prompt outside the probe, take the latents it fires and place the prompt at the Fréchet mean of their centers, $\arg\min_x \sum_y \varphi(y)\, d_X(x, \iota(y))^2$ with φ the prompt's code, and read off the time of that probe prompt. No label enters the placement. As a control, the eigen-equation extends d_X to the new prompt without any locations, $\lambda\, d_X(x_{\mathrm{new}}, x) = \mathcal{W}_2^{d_Y}(\varphi_{x_{\mathrm{new}}}, \varphi_x)$, and the nearest probe prompt is taken. Errors are cyclic minutes, median over the 1200 prompts outside the probe, where chance is 360 min. Every trained code with spread ≤ 0.65 places a minute to 17–20 min by the locations and to 5–10 min by the eigen-equation. The released dictionaries place at chance by the locations and to 5–17 min by the eigen-equation. Year, with the odd years held out: 3 years by the locations and 1 by the eigen-equation. So the time is in every code, released ones included, and what the spread buys is that each latent stands for one place on the clock. The released latents have no such place, and the decoder that uses only places finds nothing.

Steering at *Good*, block 24**.** Each live latent gets an angle from the two-dimensional classical MDS of the rank-transformed d_Y (circular correlation 0.99 with its preferred time, the labels fixing

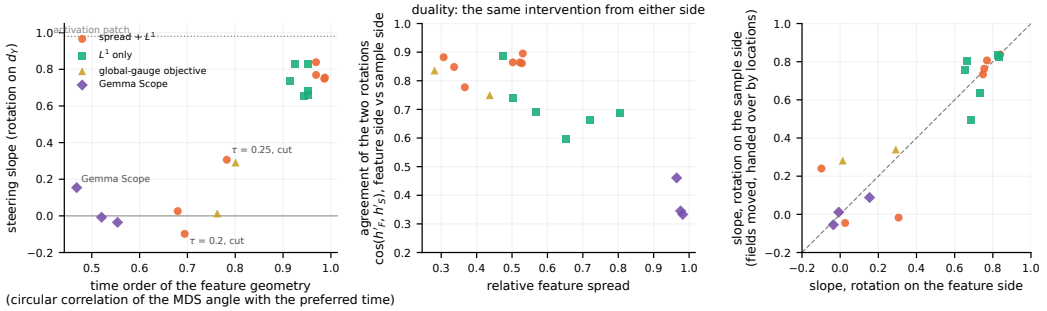

Figure 9: Steering across codes at the period, block 18, and the dual rotation. (a) Steering slope of the rotation on each code's own d_Y against the cyclic order of its feature geometry (circular correlation of the MDS angle with the preferred time). (b) Agreement of the feature-side and the sample-side rotation, the cosine of the two rotated codes, against the relative spread. (c) The slope of the sample-side rotation against that of the feature-side one.

only the orientation). A rotation by $\Delta\theta$ moves each latent's activation to the latents near its rotated point through a Gaussian kernel of width 0.15 of the median radius, normalized so that the total activation is preserved, and the residual becomes $x' = x + (h' - h)\, W_d / s_x$. The time is decoded from the next-token distribution at *Good*: the natural distributions of the other 1439 times are compared with it by Kullback–Leibler divergence and the nearest wins (leave-one-out median error 12 min, 99% within an hour). The greeting readout is the argmax over the four greeting words, scored against the model's own greeting at the target time. Over 120 source times and rotations of $\pm 1, 2, 3, 6, 9, 12$ h: slope 0.97 ($R^2 = 0.92$), patch 1.00, the same operator on the cosine geometry of the decoder rows 0.96 (where the MDS angle has circular correlation 0.69 with the preferred time), and a random permutation of the live latents at $\Delta\theta = 0$ scatters the decoded time over an interquartile range of -7.5 to $+6$ h. When the target's greeting differs from the source's, the argmax becomes the target's in 0.30–0.52 of the trials at 1–6 h and 0.67–0.71 at 9–12 h (patch 0.95–0.98), and under the rotation it is only ever *morning* or *evening*. The mean probability of the target's word by class is 0.49, 0.12, 0.23, 0.15 (patch 0.53, 0.29, 0.30, 0.24) against the natural 0.34, 0.07, 0.09, 0.09, so the default wins between two classes. Dividing each word by its natural frequency before the argmax gives 0.82 (patch 0.86) on 60 sources at $\pm 3, 6, 9$ h.

Steering at the period, block 18. The same operator on the $\tau = 0.3$ code moves the decoded time by 0.69 of the rotation ($R^2 = 0.56$, patch 0.98), though the rotated code reads as 0.93 of the shift by its own locations and kernel widths 0.08–0.3 give slopes 0.69–0.77. On every code of this position with its own readout geometry (40 source times, ± 3 and ± 6 h, Figure 9a) the codes trained with the spread penalty at τ from 0.4 to 0.275 steer at 0.75–0.84 and those trained with the sparsity penalty alone at 0.65–0.83 down to $\alpha = 0$. The spread codes at $\tau \leq 0.25$, whose ring is cut open, the two codes trained without the local normalization and the floor, which collapse the day onto a few latents, and the released dictionaries steer at -0.10 to 0.31. Steering follows the cyclic order of the feature geometry (Spearman 0.83), not the spread (-0.02): a code whose sample geometry is cut open or collapsed mirrors that geometry faithfully, and rotating it moves nothing. The dual rotation never reads d_Y. It rotates each latent's field on the sample ring, $\tilde{\psi}_y = K_X \psi_y$ with K_X the same kernel on the MDS of d_X, and hands the activation of y' to the latents y in proportion to $\tilde{\psi}_{y'}(\iota(y))$, normalized. The cosine of the two rotated codes is 0.86–0.90 for the spread codes at τ from 0.4 to 0.275 and 0.89 for the sparsity code at $\alpha = 0.3$, 0.60–0.74 for the sparsity codes at $\alpha \leq 0.1$ and 0.33–0.46 for the released dictionaries (Spearman -0.70 with the spread, Figure 9b), and the two rotations steer alike (Spearman 0.88 between their slopes, Figure 9c).