# The Kernel Inner Product Space: Dimensionality Reduction as Kernel Alignment

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Dimensionality reduction has splintered into families (spectral methods such as PCA, kernel PCA, Isomap, and Laplacian eigenmaps on one side, neighbor embeddings such as t-SNE and UMAP on the other), each carrying its own objective and its own solver. We show that these methods are often a single operation seen from different angles. Representing both the data and its embedding by centered *kernels* and scoring their agreement with the *RV coefficient*, a cosine in matrix space, casts dimensionality reduction as the projection of a fixed input kernel onto the set of output kernels that a $q$-dimensional configuration can realize. When the output kernel is linear this achievable set is a convex cone, the projection is a closed-form spectral truncation with an exact alignment ceiling, and the classical spectral methods are its optima for different input kernels; a class-label target extends the same projection to a continuous soft-LDA. Beyond the linear output, an invariance argument leaves exactly two canonical readout families, dot-product and distance, and each turns the cone into a smooth manifold whose dimension we compute, set by the motions its readout cannot see; on the distance manifold of the Student-t readout, the RV gradient becomes force-directed and its attraction a precise relative of t-SNE's. We further show that the kernel diagonal is a degree metric that tethers the embedding's spread, and that repulsion is not part of the objective but the gradient of the one mode, the embedding's global volume, that the centering discards, which places t-SNE, but not UMAP, within the framework's reach. A sequence of experiments verifies each of these predictions in turn.

## 1 Introduction

Dimensionality reduction is the workhorse of exploratory data analysis, and it is also a crowded field. A practitioner who wants to see a high-dimensional dataset in two dimensions must choose among principal component analysis (Pearson, 1901), kernel PCA (Schölkopf et al., 1998), Isomap (Tenenbaum et al., 2000), locally linear embedding (Roweis & Saul, 2000), Laplacian eigenmaps (Belkin & Niyogi, 2001), diffusion maps (Coifman & Lafon, 2006), and, for visualization above all, t-SNE (van der Maaten & Hinton, 2008) and UMAP (McInnes et al., 2020), each arriving with its own objective and its own solver. These methods fall into two cultures that rarely speak. The *spectral* family solves a single eigenproblem, returns a global optimum, and is geometrically well understood; the *neighbor-embedding* family minimizes a non-convex loss by gradient descent, produces the cluster-revealing pictures that made the field popular, and remains comparatively opaque about what, exactly, it computes.

That the spectral methods share a common core is an old insight. A kernel reading places them all under one eigenproblem, differing only in the Gram matrix they diagonalize (Ham et al., 2004; Fouss et al., 2005); the graph-embedding framework of Yan et al. (2007), the linear-dimensionality-reduction survey of Cunningham & Ghahramani (2015), and the spectral-clustering literature (von Luxburg, 2007) organize large subfamilies the same way. These unifications are powerful but bounded: they fix the output to a linear (spectral) readout and stop at the neighbor-embedding border. On the far side of that border, t-SNE and UMAP have grown their own theory: an attraction–repulsion reading (Böhm et al., 2022), an optimization and majorization view (Yang et al., 2014; 2015), a dissection of UMAP's actual loss (Damrich & Hamprecht, 2021), and

the earlier elastic embedding (Carreira-Perpiñán, 2010). However, this theory lives apart from the spectral account. The two literatures share data and goals, not a common object.

This paper supplies that object, and it is a quantity the field already knows. Represent the data and its embedding each by a centered kernel; their agreement is then the cosine between the two in matrix space, which is the *RV coefficient* of Robert & Escoufier (1976), the multivariate analogue of a correlation (Holmes, 2008). The same centered cosine underlies kernel–target alignment (Cristianini et al., 2001; Cortes et al., 2012), the Hilbert–Schmidt independence criterion (Gretton et al., 2005), the centered kernel alignment used to compare neural representations (Kornblith et al., 2019), and, through the distance/RKHS equivalence, the distance-covariance statistics (Sejdinovic et al., 2013). Everywhere it appears as a *criterion*: a quantity to test, or to optimize over a family of kernels. What is new here is its role, and we obtain it by taking the word *cosine* literally. Centered kernels form a finite-dimensional inner-product space, which we construct formally; in it, the data and any candidate embedding are two points, and the RV coefficient is the cosine of the angle between them. The alignment problem thereby becomes geometric: fix the input kernel as a target point in this space, and characterize the *achievable set*, the region swept out by the kernels that a $q$-dimensional configuration can produce at all. Dimensionality reduction then collapses to a single operation, the projection of the input kernel onto that set, and a method is nothing but a choice of input kernel and output geometry.

The shape that the achievable set takes inside the kernel space is what makes this geometry productive. When the output kernel is linear, the set is a convex cone, and the projection onto it is the classical truncated eigendecomposition (Eckart–Young), available in closed form and equipped with a computable ceiling on how well any $q$-dimensional embedding can align with its target. The entire spectral zoo reappears as the projection of *different* input kernels onto the *same* cone (a unification we make exact, ceiling included), and, because the target may be any centered kernel, a class-label kernel yields a continuous *soft-LDA* that dials from an unsupervised embedding to full class separation without ever leaving closed form.

Beyond the cone, an invariance argument narrows the admissible output kernels to two canonical families: readouts of the *dot product* and readouts of the *distance*, the latter hosting the heavy-tailed kernels of the neighbor embeddings. Both bend the cone into a smooth manifold whose exact dimension we compute, and the two answers differ tellingly: the dimension counts the motions the readout cannot see, all rigid motions for a distance readout, rotations alone for a dot-product one, whose sheet is therefore thicker by exactly the $q$ translations; the linear kernel, where the families meet, degenerates back to the cone. On these manifolds the alignment gradient is force-directed, and, at the Student-t readout of t-SNE, its attractive part is a precise and unexpectedly close relative of t-SNE's own: the same output kernel, a different objective. Two further findings, both in the distance family where the neighbor embeddings live, complete the picture. First, the kernel *diagonal* proves to be a degree metric whose energy tethers the embedding's spread; dropping it, in a hollow variant of the cosine, releases exactly the spread that a bounded readout needs, making the diagonal the algebraic dial between multidimensional scaling and neighbor embedding. Second, and most consequential, the objective contains *no repulsion at all*: the centering that defines the kernel space annihilates the one mode, the global volume of the embedding, whose gradient is precisely the repulsion of t-SNE. Repulsion is therefore not a primitive to be posited but a term to be *derived*, and the single form it can take places t-SNE, whose push is global, firmly apart from UMAP, whose push is per-pair.

The framework is developed in a weighted setting throughout (individuals carry arbitrary masses, the uniform case being one instance), extending a line of weighted-kernel analysis (Bavaud, 2024; Bavaud & Guex, 2026) to dimensionality reduction. Section 2 sets up the kernel inner-product space and the alignment objective. Section 3 treats the linear cone: its closed-form projection, its alignment ceiling, the spectral unification, and the soft-LDA dial. Section 4 crosses beyond the cone: it narrows the non-linear readouts to the two canonical families, proves each achievable set a smooth manifold whose dimension it computes, and derives the force-directed gradient. Section 5 isolates the two modes that the centering singles out: the diagonal, a degree metric whose energy tethers the embedding, and the discarded volume mode, from which repulsion is derived and its possible seats across methods taxonomized. Section 6 verifies each prediction empirically, one claim at a time, with a GitHub repository[1] hosting the code and reproducing every figure.

[1]Code provided as supplementary material; a public repository will be linked upon acceptance

## 2 The Kernel Inner Product Space

This section lays the groundwork for our unified framework. The central idea is a change of perspective: instead of describing a dataset by its coordinates, we describe it by the *kernel* that encodes all of its pairwise geometry. Such kernels live in a vector space, and once this space is equipped with an inner product, comparing two datasets becomes the measurement of an *angle* between their kernels. Dimensionality reduction is then recast not as a spectral decomposition *per se*, but as an *alignment* problem: among all the kernels that a low-dimensional configuration can possibly produce, which one is best aligned with the kernel of the data? This viewpoint, rooted in the duality at the heart of multivariate analysis (Robert & Escoufier, 1976; Bavaud, 2024), organizes the entire paper. We first construct the space and its inner product carefully, then pose dimensionality reduction as a concrete optimization problem whose two regimes (solved spectrally in Section 3 and iteratively in Section 4) structure the rest of the paper.

### 2.1 Data Representation with Kernels

Let us consider a data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ describing $n$ individuals measured on $p$ variables. In the general framework of multivariate analysis, observations may differ in importance and can typically represent aggregated objects of different weights. Let $\boldsymbol{f} = (f_i)$ be a vector of *relative weights* such that $\sum_{i=1}^{n} f_i = 1,\ f_i > 0$ (strict positivity is required to recover coordinates from kernels below), and let $\boldsymbol{\Pi} := \mathrm{diag}(\boldsymbol{f})$ be the associated diagonal matrix. The *centering matrix* $\boldsymbol{H}$ is defined as $\boldsymbol{H} := \boldsymbol{I}_n - \boldsymbol{1}_n\boldsymbol{f}^\top$, where $\boldsymbol{I}_n$ is the identity matrix of dimension $n$ and $\boldsymbol{1}_n$ is the vector of ones of length $n$. This operator allows us to obtain the *centered data matrix* $\boldsymbol{X}^c := \boldsymbol{H}\boldsymbol{X}$, where the weighted mean of every variable is zero.

While $\boldsymbol{X}^c$ captures the coordinates, the intrinsic geometry of the point cloud is described by the matrix of pairwise squared Euclidean distances $\boldsymbol{D} = (D_{ij})$, where $D_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$ or equivalently with the matrix of *weighted inner products*, or **Kernel matrix** $\boldsymbol{K}$ (Bavaud, 2024). The latter can be derived from the coordinates or the Euclidean distances:

$$\begin{aligned} \boldsymbol{K} &:= \boldsymbol{Q}\boldsymbol{G}\boldsymbol{Q}^\top = -\frac{1}{2}\boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\top \\ \boldsymbol{G} &:= \boldsymbol{X}\boldsymbol{X}^\top \qquad \boldsymbol{Q} := \boldsymbol{\Pi}^{1/2}\boldsymbol{H} \end{aligned} \tag{1}$$

where $\boldsymbol{G} = (G_{ij})$ is the *Gram matrix* of inner products between individuals, i.e. $G_{ij} = \boldsymbol{x}_i^\top \boldsymbol{x}_j$, and $\boldsymbol{Q}$ is an operator which centers and weights these inner products. Crucially, when the weights $\boldsymbol{f}$ are non-uniform, neither $\boldsymbol{H}$ nor $\boldsymbol{Q}$ are symmetric matrices. This asymmetry necessitates careful tracking of the transpose operators in subsequent gradient derivations.

It is important to note that any matrix $\boldsymbol{K}$ constructed in this manner satisfies the *weighted centering condition*:

$$\boldsymbol{K}\sqrt{\boldsymbol{f}} = \boldsymbol{0} \tag{2}$$

The mapping $\boldsymbol{X} \mapsto \boldsymbol{K}$ is bijective up to an isometry. Since $\boldsymbol{K}$ encodes only relative distances and angles, the absolute orientation of the coordinate system is irrelevant. Given any valid kernel matrix $\boldsymbol{K}$, one can recover a coordinate representation $\tilde{\boldsymbol{X}}$ that generates this specific geometry via classical *Multidimensional Scaling (MDS)*. By computing the eigendecomposition $\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, the coordinates in the principal axes are given by:

$$\tilde{\boldsymbol{X}} = \boldsymbol{\Pi}^{-1/2}\boldsymbol{U}\boldsymbol{\Lambda}^{1/2} \tag{3}$$

This duality bridges the gap between the *coordinate space* and the *kernel space*, the structure defined in the next section.

### 2.2 The Kernel Inner Product Space

We define the **Kernel Space** $\mathcal{K}_n$ as the set of all symmetric matrices $\boldsymbol{K} \in \mathbb{R}^{n\times n}$ that satisfy the weighted centering condition $\boldsymbol{K}\sqrt{\boldsymbol{f}} = \boldsymbol{0}$. It is easily verifiable that this set forms a *vector space*, closed under linear

combinations $a\boldsymbol{K}_1 + b\boldsymbol{K}_2$, $\forall a, b \in \mathbb{R}$. Its dimension is $\binom{n}{2}$: a symmetric $n \times n$ matrix has $\binom{n+1}{2}$ free entries, and the centering condition $\boldsymbol{K}\sqrt{\boldsymbol{f}} = \boldsymbol{0}$ removes exactly $n$ of them. We will return to this count in Section 4, where it serves as the ambient dimension against which the size of the reachable configurations is measured.

To measure the similarity between two configurations, we endow $\mathcal{K}_n$ with the *Frobenius inner product*:

$$\langle \boldsymbol{K}_1, \boldsymbol{K}_2 \rangle := \mathrm{Tr}(\boldsymbol{K}_1 \boldsymbol{K}_2) \tag{4}$$

This inner product induces the *Frobenius norm*:

$$\|\boldsymbol{K}\| := \sqrt{\mathrm{Tr}(\boldsymbol{K}^2)} \tag{5}$$

which makes $\mathcal{K}_n$ an **inner product space**.

The relationship between two kernels is captured by the *cosine of the angle* between them in this space. This quantity is the **RV coefficient** (Robert & Escoufier, 1976)[2] as defined below:

$$\mathrm{RV}(\boldsymbol{K}_1, \boldsymbol{K}_2) := \frac{\langle \boldsymbol{K}_1, \boldsymbol{K}_2 \rangle}{\|\boldsymbol{K}_1\| \|\boldsymbol{K}_2\|} = \frac{\mathrm{Tr}(\boldsymbol{K}_1 \boldsymbol{K}_2)}{\sqrt{\mathrm{Tr}(\boldsymbol{K}_1^2)\,\mathrm{Tr}(\boldsymbol{K}_2^2)}} \tag{6}$$

Maximizing this coefficient amounts to maximizing the structural alignment between two point clouds, normalized by their respective norms.

It is worth naming this quantity from the outset, to forestall a natural objection. Because our kernels are centered, the RV coefficient is precisely the *centered kernel alignment* (CKA) of Cortes et al. (2012), used by Kornblith et al. (2019) to compare representations, and for linear kernels it reduces to the classical RV coefficient of Robert & Escoufier (1976). The cosine itself is therefore *not* where our contribution lies. What is new, and what occupies the remainder of the paper, is the *geometry of the set of kernels that a configuration can actually reach*: a cone in the linear case (Section 3), a thin curved manifold in the non-linear case (Section 4), a distinction that turns out to explain why some reduction methods admit a closed-form spectral solution while others must be solved iteratively.

With this definition, two kernels are *orthogonal* if $\langle \boldsymbol{K}_1, \boldsymbol{K}_2 \rangle = 0$. The *orthogonal projection* of $\boldsymbol{K}_1$ on $\boldsymbol{K}_2$ is:

$$\mathrm{proj}_{\boldsymbol{K}_2}(\boldsymbol{K}_1) := \frac{\langle \boldsymbol{K}_1, \boldsymbol{K}_2 \rangle}{\|\boldsymbol{K}_2\|^2} \boldsymbol{K}_2 = \frac{\mathrm{Tr}(\boldsymbol{K}_1 \boldsymbol{K}_2)}{\mathrm{Tr}(\boldsymbol{K}_2^2)} \boldsymbol{K}_2 \tag{7}$$

We can also identify a special element in this space: the **Neutral Kernel** $\boldsymbol{K}_0$, defined as the projection of the identity matrix into $\mathcal{K}_n$:

$$\boldsymbol{K}_0 = \boldsymbol{I}_n - \sqrt{\boldsymbol{f}}\sqrt{\boldsymbol{f}}^\top \tag{8}$$

Geometrically, $\boldsymbol{K}_0$ represents a simplex (regular when the weights are uniform) where all distinct observations are separated by the squared distance $D_{ij} = \frac{1}{f_i} + \frac{1}{f_j}$, embodying a state of "structural neutrality". It acts as the identity element for the product of kernels, obeying $\boldsymbol{K}\boldsymbol{K}_0 = \boldsymbol{K}_0\boldsymbol{K} = \boldsymbol{K}$ for all $\boldsymbol{K} \in \mathcal{K}_n$.

Within this vector space, a critical geometric distinction must be made. While $\mathcal{K}_n$ contains all centered symmetric matrices, only those that are **positive semi-definite** (PSD) correspond to genuine Euclidean configurations. These valid kernels form a *convex cone* $\mathcal{K}_n^+ \subset \mathcal{K}_n$. In practical applications, one may encounter a matrix $\boldsymbol{K} \in \mathcal{K}_n$ that lies outside this cone. The closest valid kernel $\boldsymbol{K}^+ \in \mathcal{K}_n^+$ to $\boldsymbol{K}$, in the sense of the Frobenius norm, is obtained by spectral projection (Higham, 1988). If $\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ is the eigendecomposition of the matrix, the projection is given by clipping the negative eigenvalues:

$$\boldsymbol{K}^+ = \boldsymbol{U}\boldsymbol{\Lambda}^+\boldsymbol{U}^\top \tag{9}$$

where $\boldsymbol{\Lambda}^+$ replaces each eigenvalue $\lambda_i$ of $\boldsymbol{\Lambda}$ by $\lambda_i^+ := \max(\lambda_i, 0)$. We state this classical fact here without proof; Theorem 2 will re-derive it, strengthened with a rank constraint, as the engine of the linear theory.

[2]Strictly, Robert & Escoufier (1976) define it in the dual space of variables rather than of individuals.

Finally, it is instructive to consider the action of a kernel $\boldsymbol{K} = (K_{ij})$ as a linear operator on a single variable vector $\boldsymbol{y} \in \mathbb{R}^n$. If $K_{ij}$ is interpreted as an attraction-repulsion coefficient between $i$ and $j$, then:

$$\tilde{y}_i = (\boldsymbol{K}\boldsymbol{y})_i = \sum_j K_{ij} y_j \tag{10}$$

is the weighted consensus of the neighbors of $i$ as defined by the kernel. It resembles a lag operator (with possibly negative weights) acting on $\boldsymbol{y}$ according to the structure encoded by $\boldsymbol{K}$, a reading that returns with the spatial (lag) kernels of the next subsection.

### 2.3 The RV Coefficient as a Measure

Before making the RV coefficient the engine of dimensionality reduction, it is worth recalling what it usually *is*: a descriptive measure. Given two kernels over the same $n$ objects, it returns a single number, the cosine of their angle in $\mathcal{K}_n$, summarizing how similarly they arrange those objects. In this role it has been rediscovered, under different names, across quite different fields.

In *multivariate statistics* it is the original object: Robert & Escoufier (1976) introduced it as a coefficient of association between two sets of variables, a multivariate analogue of the squared correlation, and it became the cornerstone of multi-table analysis: STATIS, multiple factor analysis, and the broader "French" school of data analysis (Holmes, 2008). In *machine learning* the same cosine, computed on centered kernels, is the *centered kernel alignment* we already met: introduced to select kernels by their alignment with a target (Cristianini et al., 2001), it now serves to compare the internal representations of neural networks (Kornblith et al., 2019), while its unnormalized numerator $\mathrm{Tr}(\boldsymbol{K}_X \boldsymbol{K}_Y)$ is the Hilbert–Schmidt Independence Criterion, a measure of statistical dependence (Gretton et al., 2005; Sejdinovic et al., 2013) that also drives kernel dimension reduction (Fukumizu et al., 2004; Song et al., 2007). And in *spatial statistics* that same numerator, read with $\boldsymbol{K}_X$ a *spatial* (lag) kernel and $\boldsymbol{K}_Y$ a *feature* kernel, is a multivariate index of Moran spatial autocorrelation, which Bavaud (2024) formalizes as the Frobenius product of a feature and a spatial kernel (Loup & Bavaud, 2025), and which underlies non-parametric tests of association between dissimilarities (Bavaud & Guex, 2026). Its spectral side is kin to Moran eigenvector maps (Dray et al., 2006) and to MULTISPATI (Dray et al., 2008), though the latter maximizes a variance–autocorrelation compromise rather than autocorrelation alone.

What unites these uses is that the coefficient is *evaluated*: both kernels are given, and RV reports how aligned they are, whether to describe, to test, or to compare. This paper takes the complementary step. We fix one kernel, the geometry of the data, and treat RV not as a quantity to report but as an *objective* to maximize over the second, which a low-dimensional configuration is free to vary. Measuring becomes optimizing, and that optimization, as we now formulate it, is exactly dimensionality reduction.

### 2.4 Dimensionality Reduction as Kernel Alignment

Turning the RV coefficient from a measure into an objective, we can now state the problem the rest of the paper solves. The geometry to be reproduced is carried by a fixed *input* (or target) kernel $\boldsymbol{K}_X \in \mathcal{K}_n$, while a low-dimensional configuration $\boldsymbol{Y} \in \mathbb{R}^{n \times q}$, with $q < p$, induces an *output* kernel $\boldsymbol{K}_Y \in \mathcal{K}_n$. As in (1), both are double-centered raw similarity matrices,

$$\boldsymbol{K}_X = \boldsymbol{Q}\,\boldsymbol{G}_X\,\boldsymbol{Q}^\top, \qquad \boldsymbol{K}_Y = \boldsymbol{Q}\,\boldsymbol{G}_Y\,\boldsymbol{Q}^\top, \tag{11}$$

and dimensionality reduction is the search for the configuration $\boldsymbol{Y}$ whose output kernel is best aligned, in the RV sense, with the target:

$$\max_{\boldsymbol{Y}} \quad \mathrm{RV}(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \frac{\mathrm{Tr}(\boldsymbol{K}_X \boldsymbol{K}_Y)}{\sqrt{\mathrm{Tr}(\boldsymbol{K}_X^2)\ \mathrm{Tr}(\boldsymbol{K}_Y^2)}}. \tag{12}$$

The two raw Gram matrices $\boldsymbol{G}_X$ and $\boldsymbol{G}_Y$ are the modeller's two design choices, and each can be built in the same two ways. The *linear* choice takes the ordinary matrix of inner products, $\boldsymbol{G} = \boldsymbol{X}\boldsymbol{X}^\top$ (or $\boldsymbol{Y}\boldsymbol{Y}^\top$ for the

output), as in (1). The *kernelized* choice replaces it with a similarity function evaluated pairwise,

$$[\boldsymbol{G}]_{ij} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{13}$$

where $\kappa$ is a kernel in the sense used by support-vector methods (a radial basis function, a graph affinity, a diffusion or geodesic similarity), or, equivalently, any matrix of squared distances read through the double-centering $-\frac{1}{2}\boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\top$. The linear kernel is simply the special case $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^\top \boldsymbol{x}_j$.

These two choices play very different roles, and the distinction organizes the entire paper. The **input** kernel $\boldsymbol{G}_X$ sets *what* we approximate: it fixes the target geometry. Whether it is linear or kernelized leaves the *nature* of the optimization untouched, and through this single lever one objective already subsumes a whole family of methods (Section 3.3). The **output** kernel $\boldsymbol{G}_Y$ is more consequential: its choice changes the very *nature* of the problem, and this is what splits the two sections that follow.

- A *linear* output, $\boldsymbol{G}_Y = \boldsymbol{Y}\boldsymbol{Y}^\top$, makes the set of achievable kernels a *cone* with a closed-form spectral solution (Section 3). *Any* input kernel, linear or kernelized, is handled here, which is why PCA, classical MDS, and the kernelized-input methods (Kernel PCA, Isomap, LLE, Laplacian Eigenmaps, Diffusion Maps) all belong to this regime.
- A *non-linear* output (a dot-product or a distance readout, the two canonical families derived in Section 4) makes it a curved *manifold* with no closed-form solution, to be optimized iteratively. The bounded, heavy-tailed distance readouts are the regime of modern neighbor embeddings such as t-SNE.

We state this contrast here without proof; establishing it, in each of the two regimes, is the task of the two sections that follow.

## 3 Linear Reduction as Projection onto a Cone

Section 2.4 split dimensionality reduction into two regimes according to the output kernel. We develop the first here: the *linear* output, $\boldsymbol{G}_Y = \boldsymbol{Y}\boldsymbol{Y}^\top$, so that $\boldsymbol{K}_Y = \boldsymbol{Q}\boldsymbol{Y}\boldsymbol{Y}^\top\boldsymbol{Q}^\top$. The input kernel $\boldsymbol{K}_X$ may be anything in $\mathcal{K}_n$, linear or kernelized; only the output is constrained. This single restriction, we will see, is enough to make the set of achievable kernels a *cone* and the alignment (12) solvable in closed form by one eigendecomposition.

The restriction shifts the difficulty. Written over $\boldsymbol{Y}$, the objective looks like a free optimization; in reality all of its structure lies in one prior question: as $\boldsymbol{Y}$ varies, *which kernels $\boldsymbol{K}_Y$ can a linear output actually produce?* Answering it turns the abstract objective into a concrete, solvable problem, and it is where we begin.

### 3.1 The Achievable Set of Linear Output Kernels

With a linear output kernel, the achievable set admits a remarkably clean description, and this description is what makes the linear problem solvable in closed form. A configuration living in $q$ dimensions has only $q$ axes to work with, so it can only generate a positive semi-definite kernel of rank at most $q$. The content of the following theorem is the converse: *every* centered PSD kernel of rank $\leq q$ is produced by some $\boldsymbol{Y}$. The reachable kernels are thus precisely the low-rank slice of the PSD cone.

**Theorem 1** (Achievable set)**.** *Let the output kernel be linear, i.e. $\boldsymbol{K}_Y = \boldsymbol{Q}\boldsymbol{Y}\boldsymbol{Y}^\top\boldsymbol{Q}^\top$ with $\boldsymbol{Y} \in \mathbb{R}^{n\times q}$. As $\boldsymbol{Y}$ ranges over $\mathbb{R}^{n\times q}$, the set of achievable kernels is exactly*

$$\mathcal{S}_q = \{\boldsymbol{K} \in \mathcal{K}_n \mid \boldsymbol{K} \text{ is PSD},\ \operatorname{rank}(\boldsymbol{K}) \leq q\}. \tag{14}$$

The set $\mathcal{S}_q$ is a *cone*: if $\boldsymbol{K} \in \mathcal{S}_q$ then $t\boldsymbol{K} \in \mathcal{S}_q$ for every $t \geq 0$, because scaling preserves both the rank and the positive semi-definiteness. It is *not* convex (the sum of two rank-$q$ kernels generally has rank $2q$), but it is this conic, scale-invariant geometry that the RV cosine exploits.

### 3.2 Closed-Form Solution and the Alignment Ceiling

Maximizing a cosine over a cone is a problem with a classical answer. Because the RV coefficient is a cosine, the maximization is insensitive to scale: it amounts to finding the *direction* inside $\mathcal{S}_q$ that is closest to the target, that is, the orthogonal Frobenius projection of $\boldsymbol{K}_X$ onto the cone. Approximating a symmetric matrix by a PSD matrix of rank $\leq q$ is the textbook low-rank problem, whose solution is to keep the $q$ largest *positive* eigenpairs and discard the rest.

**Theorem 2** (Closed-form projection)**.** *Let $\boldsymbol{K}_X = \sum_{i=1}^{n} \lambda_i \, \boldsymbol{u}_i \boldsymbol{u}_i^\top$ be the eigendecomposition of the target, ordered $\lambda_1 \geq \cdots \geq \lambda_n$, and suppose $\lambda_1 > 0$ (when $\lambda_1 \leq 0$, no output kernel attains a positive alignment). The maximizer of $\mathrm{RV}(\boldsymbol{K}_X, \boldsymbol{K}_Y)$ over $\boldsymbol{K}_Y \in \mathcal{S}_q$ is*

$$\boldsymbol{K}_Y^\star = \sum_{j=1}^{q} \lambda_j^+ \, \boldsymbol{u}_j \boldsymbol{u}_j^\top, \qquad \boldsymbol{Y}^\star = \boldsymbol{\Pi}^{-1/2} \boldsymbol{U}_q (\boldsymbol{\Lambda}_q^+)^{1/2}, \tag{15}$$

*where $\lambda_j^+ = \max(\lambda_j, 0)$ is the clipping of (9), $\boldsymbol{U}_q := [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q]$ and $\boldsymbol{\Lambda}_q^+ := \mathrm{diag}(\lambda_1^+, \ldots, \lambda_q^+)$. The optimum is global and obtained in closed form by a single eigendecomposition; it coincides with classical (weighted) multidimensional scaling on $\boldsymbol{K}_X$.*

This theorem means that the whole linear theory collapses onto a single eigendecomposition. There are no local optima and no iterations: the optimal $q$-dimensional embedding is read directly off the leading eigenvectors of $\boldsymbol{K}_X$. This is exactly why the classical methods are called *spectral*.

The same computation also quantifies *how well* the data can be represented in $q$ dimensions.

**Proposition 1** (Alignment ceiling)**.** *In the setting of Theorem 2, the optimal alignment achievable in dimension $q$ is*

$$\mathrm{RV}_{\max}(q) := \mathrm{RV}(\boldsymbol{K}_X, \boldsymbol{K}_Y^\star) = \sqrt{\frac{\sum_{j=1}^{q} (\lambda_j^+)^2}{\sum_{i=1}^{n} \lambda_i^2}}. \tag{16}$$

$\mathrm{RV}_{\max}(q)$ is therefore the Frobenius analogue of the cumulative *explained variance* of PCA: its square is the fraction of structural energy (here *squared* eigenvalues, because the RV is a Hilbert–Schmidt cosine) captured by the best $q$-dimensional embedding. It is an interpretable, data-dependent ceiling on what any method with a linear output can reach, and it provides the natural yardstick against which the non-linear embeddings of the following sections will be measured.

### 3.3 A Unified View of Spectral Methods

Theorems 1 and 2 make no assumption on how the target kernel was built: any $\boldsymbol{K}_X \in \mathcal{K}_n$ with a positive leading eigenvalue is handled by the same projection. The achievable set and the closed-form optimum depend only on the choice of output (here, linear), not on the construction of the input. This single observation organizes a large family of classical methods.

**Corollary 1** (Spectral unification)**.** *Let $\boldsymbol{K}_X \in \mathcal{K}_n$ with $\lambda_1 > 0$. Any method that embeds the data through the $q$ leading eigenpairs of $\boldsymbol{K}_X$ (that is, returns $\boldsymbol{Y}^\star = \boldsymbol{\Pi}^{-1/2} \boldsymbol{U}_q (\boldsymbol{\Lambda}_q^+)^{1/2}$) solves the linear-output RV problem with target $\boldsymbol{K}_X$, exactly and globally. Conversely, this problem depends on the modelling choices only through $\boldsymbol{K}_X$: two constructions producing the same target kernel pose the very same optimization problem, and share the same optimum.*

The linear Gram $\boldsymbol{G}_X = \boldsymbol{X}\boldsymbol{X}^\top$ recovers PCA and classical MDS (Theorem 2). Building $\boldsymbol{K}_X$ instead from a non-linear similarity or a graph-based affinity (a radial basis kernel, geodesic distances, locally linear reconstructions, a graph Laplacian, or a diffusion kernel) recovers Kernel PCA, Isomap, LLE, Laplacian Eigenmaps, and Diffusion Maps, respectively: each of these methods embeds through the leading eigenpairs of its own centered kernel, and each is therefore, by the corollary, the exact RV optimum for its target, up to one axis-scaling convention: reference implementations that return orthonormal (balanced-axis) coordinates rather than the $\boldsymbol{\Lambda}^{1/2}$-scaled ones of $\boldsymbol{Y}^\star$, as LLE and Laplacian Eigenmaps do, renounce a fraction of the attainable alignment, a gap measured in Section 6.2. All of them share the *same* achievable cone $\mathcal{S}_q$ and the

*same* closed-form projection; they differ only in how the input kernel is built, following the linear/kernelized menu of Section 2.4. This is the sense in which the kernel space unifies the spectral zoo: not a new method, but a single geometric statement of which all of them are special cases.

### 3.4 A Supervised Extension: From LDA to Soft-LDA

The projection of Theorem 2 does more than re-describe existing methods. Because it accepts *any* target in $\mathcal{K}_n$, we may feed it a kernel that no dataset alone provides, one built from class labels, and obtain a supervised embedding. Given $m$ classes described by an indicator matrix $\boldsymbol{Z} = (z_{ig})$, with $z_{ig} = 1$ if individual $i$ belongs to class $g$ and 0 otherwise, the *inter-class kernel*

$$\boldsymbol{K}_Z := \boldsymbol{P}_Z \boldsymbol{K}_X \boldsymbol{P}_Z, \qquad \boldsymbol{P}_Z := \boldsymbol{\Pi}^{1/2} \boldsymbol{Z} (\boldsymbol{Z}^\top \boldsymbol{\Pi} \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{\Pi}^{1/2}, \tag{17}$$

relocates every point onto the ($\boldsymbol{f}$-weighted) centroid of its class, retaining only the between-class geometry. Here $\boldsymbol{P}_Z$ is the class-averaging operator conjugated into kernel coordinates: a symmetric orthogonal projector satisfying $\boldsymbol{P}_Z \sqrt{\boldsymbol{f}} = \sqrt{\boldsymbol{f}}$, so that $\boldsymbol{K}_Z$ stays in $\mathcal{K}_n$ and coincides with the kernel of the configuration of class centroids.

Classical **Linear Discriminant Analysis** aligns the embedding with this class kernel, but under a constraint: the configuration must remain a linear *isometry* of the data, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{A}$ with $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}_q$. Maximizing $\mathrm{RV}(\boldsymbol{K}_Z, \boldsymbol{K}_Y)$ over such $\boldsymbol{A}$ recovers LDA (Robert & Escoufier, 1976), the constraint turning the problem into a generalized eigenproblem of between- against within-class scatter. It is not cosmetic: $\boldsymbol{K}_Z$ has rank at most $m - 1$, so a *free* maximization would simply collapse each class onto its centroid. Tying $\boldsymbol{Y}$ to a rigid image of $\boldsymbol{X}$ is what keeps the within-class spread alive, and, as a bonus, yields an out-of-sample map $\boldsymbol{A}$. Supervised and semi-supervised embeddings of this kind (supervised PCA (Barshan et al., 2011), local Fisher discriminant analysis (Sugiyama, 2007), and their semi-supervised variants (Zhang et al., 2007)) all sharpen class structure by constraining or reweighting the configuration; our route instead leaves the output free and softens the target.

Our formalism offers a different remedy for the collapse, one that stays in closed form. Rather than constrain the output, we *soften the target*, mixing the class kernel back with the data kernel,

$$\boldsymbol{K}_\beta := (1 - \beta)\, \boldsymbol{K}_X + \beta\, \boldsymbol{K}_Z, \qquad \beta \in [0, 1], \tag{18}$$

which stays in $\mathcal{K}_n$ and, both terms being PSD, retains the full rank of $\boldsymbol{K}_X$ for $\beta < 1$. The plain cone projection of Theorem 2, with no constraint and no generalized eigenproblem, then already returns a sensible *soft-LDA*: at $\beta = 0$ it is unsupervised projection, at $\beta = 1$ it degenerates to the pure class centroids, and in between $\beta$ dials the strength of supervision, pulling same-class points together while preserving their scatter. The isometry constraint and the mixture are two remedies for the same degeneracy; the framework prefers the mixture, because it is a *continuous* dial that leaves the solution spectral. We put this dial to work in the case studies (Section 6), where a non-linear analogue carries the same idea beyond the linear regime: a single $\beta$ then softens *both* kernels together, the input target toward $\boldsymbol{K}_Z$ and the heavy-tailed output toward the linear $\boldsymbol{G}_Y = \boldsymbol{Y}\boldsymbol{Y}^\top$, so that $\beta = 1$ recovers the linear soft-LDA endpoint above and $\beta = 0$ a purely unsupervised neighbor embedding. Why the output too must move, and how far the supervision carries to held-out points, are taken up there.

## 4 Beyond the Cone: The Manifold

The closed form of Section 3 rested on one pillar: the linear output kernel. Kernelizing the *input* (Corollary 1) stretched its reach without touching it; to reproduce the dynamics of modern neighbor embeddings, whose heavy-tailed readout mitigates the crowding problem, the pillar itself must go, the step flagged in Section 2.4. The spectral guarantee does not survive: the task shifts from a single eigenproblem to a continuous, gradient-based pursuit of the alignment. What replaces the cone is the subject of this section. An invariance argument first narrows the admissible readouts to two canonical families, *dot-product* and *distance*; for both, the achievable set turns out to be a smooth manifold whose dimension is dictated by the motions its readout cannot see; and the gradient of the alignment on these manifolds is force-directed, its attractive part a close relative of t-SNE's own.

### 4.1 The Non-Linear Output Kernel: Two Canonical Readouts

Concretely, we enforce a geometric structure on $\boldsymbol{Y}$ through a *model kernel function* $\kappa_y$,

$$[\boldsymbol{G}_Y^{\kappa_y}]_{ij} = \kappa_y(\boldsymbol{y}_i, \boldsymbol{y}_j), \tag{19}$$

so that the problem becomes finding the configuration whose induced geometry, seen through $\kappa_y$, aligns best with the target:

$$\max_{\boldsymbol{Y}} \quad \mathrm{RV}\left(\boldsymbol{K}_X,\ \boldsymbol{Q}\,\boldsymbol{G}_Y^{\kappa_y}\,\boldsymbol{Q}^\top\right). \tag{20}$$

Which functions $\kappa_y$ are admissible? The model kernel is meant to encode the geometry of the configuration, not the coordinates chosen to write it down, so the minimal requirement is invariance under orthogonal transformations of the embedding space, $\kappa_y(\boldsymbol{R}\boldsymbol{y}_i, \boldsymbol{R}\boldsymbol{y}_j) = \kappa_y(\boldsymbol{y}_i, \boldsymbol{y}_j)$. This requirement bites: two pairs of vectors with equal norms and inner product are exchanged by such a transformation, so an invariant $\kappa_y$ sees the pair only through the three scalars $\|\boldsymbol{y}_i\|^2$, $\|\boldsymbol{y}_j\|^2$ and $\boldsymbol{y}_i^\top \boldsymbol{y}_j$. Kernels in practice go one step further and read a *single* scalar through a smooth profile, and the triple offers exactly two canonical ways of producing one:

$$\text{(A)} \quad [\boldsymbol{G}_Y^{\kappa_y}]_{ij} = \kappa_y\big(\boldsymbol{y}_i^\top \boldsymbol{y}_j\big), \qquad\qquad \text{(B)} \quad [\boldsymbol{G}_Y^{\kappa_y}]_{ij} = \kappa_y\big(\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2\big), \tag{21}$$

where $\kappa_y$ now denotes the scalar profile. The *dot-product* readout (A) keeps the origin as a landmark; it hosts the polynomial and sigmoid profiles, the output model of matrix-factorization embeddings. The *distance* readout (B) is the one reduction of the triple that is also translation-invariant (a translation-invariant function of the pair depends on $\boldsymbol{y}_i - \boldsymbol{y}_j$ alone, and orthogonal invariance then retains only its norm); it hosts the Gaussian and Student-t profiles of the neighbor embeddings, and setting it to a heavy-tailed profile makes the output behave like t-SNE within this single alignment objective.

The two families are not strangers: they meet exactly at the cone. After centering, the linear member of (A), $\kappa_y(t) = t$, and the linear member of (B), $\kappa_y(t) = -t/2$, produce the same kernel $\boldsymbol{K}_Y = \boldsymbol{Q}\,\boldsymbol{Y}\boldsymbol{Y}^\top\boldsymbol{Q}^\top$ (the double-centering identity of classical MDS): both families pass through the linear world of Section 3. Beyond that shared member they part ways, and the invariance that separates them (a dot-product readout registers translations, a distance readout does not) will reappear in the next subsection as an exact difference of dimensions. We derive the gradient-based scheme that both families call for below, but only after first characterizing the geometry of the sets the optimization explores.

### 4.2 The Achievable Set is a Manifold

When the output kernel is non-linear, the clean conic picture of Section 3 is lost. To see what replaces it, write, for either readout of (21), the map sending a configuration to its output kernel,

$$\Phi : \mathbb{R}^{n\times q} \to \mathcal{K}_n, \qquad \Phi(\boldsymbol{Y}) = \boldsymbol{Q}\,\boldsymbol{G}_Y^{\kappa_y}\,\boldsymbol{Q}^\top, \qquad \mathcal{S}_q^\kappa := \Phi(\mathbb{R}^{n\times q}). \tag{22}$$

The achievable set is its image $\mathcal{S}_q^\kappa$. Scaling $\boldsymbol{Y}$ no longer scales $\boldsymbol{K}_Y$ (a non-linear profile bends or saturates), so this set is no longer a cone. It is instead a thin, curved sheet: each of the $n$ points contributes $q$ coordinates, but the displacements the readout cannot see, all rigid motions for the distance readout, rotations alone for the dot-product one, leave $\boldsymbol{K}_Y$ unchanged. Quotienting out these invisible motions is what fixes the dimension, and the two families will not have the same.

To make this precise, we work with the *differential* of $\Phi$ at a configuration $\boldsymbol{Y}$, written $D\Phi(\boldsymbol{Y})$. This is the Jacobian of $\Phi$: the linear map that sends an infinitesimal displacement $\dot{\boldsymbol{Y}}$ of the points to the first-order change $D\Phi(\boldsymbol{Y})[\dot{\boldsymbol{Y}}]$ it produces in the output kernel. Two features of this map carry the whole result. Its *kernel* $\ker D\Phi(\boldsymbol{Y})$ collects the displacements that leave $\boldsymbol{K}_Y$ unchanged, the directions along which the configuration can move without the geometry noticing. Its *rank* counts the remaining, effective directions, in which $\boldsymbol{K}_Y$ genuinely moves; that number is precisely the local dimension of the achievable set. Part of the kernel is visible by inspection: the displacements that leave the readout's *argument* unchanged. These are the *infinitesimal rigid motions* $\mathcal{I}_{\boldsymbol{Y}}$ for the distance readout, but only the *infinitesimal rotations* $\mathcal{R}_{\boldsymbol{Y}}$ for the dot-product one, which registers translations:

$$\mathcal{I}_{\boldsymbol{Y}} := \{\boldsymbol{1}\boldsymbol{v}^\top + \boldsymbol{Y}\boldsymbol{\Omega}^\top \mid \boldsymbol{v} \in \mathbb{R}^q,\ \boldsymbol{\Omega} \in \mathfrak{so}(q)\}, \qquad \mathcal{R}_{\boldsymbol{Y}} := \{\boldsymbol{Y}\boldsymbol{\Omega}^\top \mid \boldsymbol{\Omega} \in \mathfrak{so}(q)\},$$

where $\mathbf{\Omega}$ ranges over the $q \times q$ skew-symmetric matrices $\mathfrak{so}(q)$, so that $\dim \mathcal{I}_{\boldsymbol{Y}} = \binom{q+1}{2}$ and $\dim \mathcal{R}_{\boldsymbol{Y}} = \binom{q}{2}$. In case (A) we further call a configuration *regular* if it has full column rank and the overdetermined linear system (47) of Appendix A, whose coefficients depend on $\boldsymbol{Y}$ and $\kappa_y'$, admits only the zero solution; regularity is an open condition on $\boldsymbol{Y}$. The proposition states when these visible motions are the whole kernel, and what follows for the achievable set; the distance case comes first, its conclusion being uniform where the dot-product one is generic.

**Proposition 2** (Achievable manifolds)**.** *Let the scalar profile $\kappa_y$ act through one of the two readouts (21), and be smooth with $\kappa_y'(t) \neq 0$ on an open interval containing its arguments ($\mathbb{R}$ in case (A), $[0, \infty)$ in case (B)); let $n \geq q + 1$.*

*(i)* (Distance.) *In case (B), at every affinely spanning configuration, $\mathcal{S}_q^\kappa$ is locally a smooth embedded submanifold of $\mathcal{K}_n$, with*

$$\ker D\Phi(\boldsymbol{Y}) = \mathcal{I}_{\boldsymbol{Y}}, \qquad \dim \mathcal{S}_q^\kappa = nq - \binom{q+1}{2}.$$

*(ii)* (Dot-product.) *In case (A), at every regular configuration, $\mathcal{S}_q^\kappa$ is locally a smooth embedded submanifold of $\mathcal{K}_n$, with*

$$\ker D\Phi(\boldsymbol{Y}) = \mathcal{R}_{\boldsymbol{Y}}, \qquad \dim \mathcal{S}_q^\kappa = nq - \binom{q}{2}.$$

*If $\kappa_y$ is real-analytic, either no configuration is regular, or the non-regular ones form a closed null set.*

*(iii)* (Affine member.) *In case (A) with $\kappa_y(t) = \alpha t + c$, $\alpha \neq 0$, no configuration is regular, and*

$$\mathcal{S}_q^\kappa = \alpha\, \mathcal{S}_q,$$

*the cone of Section 3.*

*(iv)* (Boundedness.) *For either readout, if $\kappa_y$ is bounded, then $\mathcal{S}_q^\kappa$ is bounded and contains nonzero kernels; in particular, it is not a cone.*

The proof, deferred to Appendix A, factors $\Phi = C \circ R \circ S$ (readout argument, entrywise profile, linear centering) through one shared computation, the kernel of the centering. For the distance readout, the differentials of the squared distances are *hollow* matrices while that kernel demands a nonzero diagonal, so the centering loses no dimension, and a displacement invisible to every pairwise distance must be an infinitesimal rigid motion, the classical infinitesimal rigidity of the complete graph. For the dot-product readout the Gram differentials do carry a diagonal, and their interaction with the centering reduces exactly to the linear system (47): its solutions are the displacements that the centering absorbs, none at a regular configuration, the $q$ translations in the affine case. In cases (i) and (ii) the differential keeps a constant rank near the configuration, and the constant-rank theorem (a smooth map whose differential keeps a constant rank $r$ maps a sufficiently small neighborhood onto an embedded submanifold of dimension $r$) supplies the sheet; the statement is local, and sheets issued from far-apart configurations may still globally intersect. Boundedness in (iv) follows from the boundedness of $\kappa_y$ itself; the Student-t readout $\kappa_y(t) = (1 + t)^{-1}$, smooth and bounded on an open interval containing $[0, \infty)$ (say $t > -\frac{1}{2}$), satisfies every hypothesis of the proposition, boundedness included.

The reachable kernels therefore form, at the Student-t readout, a sheet of dimension $2n - 3$ (for $q = 2$) inside a space of dimension $\binom{n}{2} \approx n^2/2$: vanishingly thin. A non-affine dot-product readout traces a sheet thicker by exactly the $q$ translations it registers, and still vanishingly so. For either family there is no scale freedom and no global spectral solution, and the embedding must be found by following the gradient along the curved sheet, the task of the next subsections. The boundedness in (iv) returns in Section 5: it is precisely because the achievable set is bounded that a separate repulsive force is needed to spread clusters apart. Both dimension formulas are confirmed numerically in Section 6: the rank of $D\Phi$ equals $nq - \binom{q+1}{2}$ at the Student-t readout, and $nq - \binom{q}{2}$ at an exponential dot-product readout, across sample sizes; the latter checks supply the regular configurations that item (ii) calls for.

### 4.3 Derivation of the Gradient

To maximize the RV coefficient, we compute $\frac{\partial \mathrm{RV}}{\partial \boldsymbol{Y}}$, the $n \times q$ gradient matrix of the objective function regarding coordinates in the embedding space, via the chain rule:

$$\frac{\partial \mathrm{RV}}{\partial \boldsymbol{Y}} = \underbrace{\frac{\partial \mathrm{RV}}{\partial \boldsymbol{K}_Y} \cdot \frac{\partial \boldsymbol{K}_Y}{\partial \boldsymbol{G}_Y^{\kappa_y}}}_{\text{Structural gradient}} \cdot \underbrace{\frac{\partial \boldsymbol{G}_Y^{\kappa_y}}{\partial \boldsymbol{Y}}}_{\text{Parametric gradient}} \tag{23}$$

In this formula, $\frac{\partial \mathrm{RV}}{\partial \boldsymbol{K}_Y}$ is an $n \times n$ matrix, while $\frac{\partial \boldsymbol{K}_Y}{\partial \boldsymbol{G}_Y^{\kappa_y}}$ and $\frac{\partial \boldsymbol{G}_Y^{\kappa_y}}{\partial \boldsymbol{Y}}$ are tensors with dimensions $n \times n \times n \times n$ and $n \times n \times n \times q$ respectively (the symbol "$\cdot$" denotes here tensor contraction). We structure this derivation in two levels: first, the *structural gradient* defined in the space of kernels, and second, the *parametric gradient* which brings back the gradient to the coordinate level. Assembled, these factors are the differential $D\Phi(\boldsymbol{Y})$ of Section 4.2 written out in coordinates; the two-level computation that follows is exactly its adjoint $D\Phi(\boldsymbol{Y})^*$, applied to the kernel-space gradient, a reading we make precise in Section 4.4 (Proposition 4). In what follows, $d$ denotes an infinitesimal increment (of a scalar or a matrix), so that $d\boldsymbol{K}_Y = D\Phi(\boldsymbol{Y})[d\boldsymbol{Y}]$: the differential $D\Phi$ is the map, $d\boldsymbol{K}_Y$ its action on a displacement.

#### 4.3.1 Structural Gradient

The gradient with respect to the configuration $\boldsymbol{Y}$ depends first on how the RV coefficient changes with the kernel matrix $\boldsymbol{K}_Y$, and then on how that kernel matrix changes with the raw similarities $\boldsymbol{G}_Y^{\kappa_y}$.

Using the quotient rule on the definition of the RV coefficient (Eq. 6), the partial derivative with respect to the kernel $\boldsymbol{K}_Y$ is:

$$\begin{aligned}\frac{\partial \mathrm{RV}}{\partial \boldsymbol{K}_Y} &= \frac{\boldsymbol{K}_X}{\sqrt{\mathrm{Tr}(\boldsymbol{K}_X^2)\,\mathrm{Tr}(\boldsymbol{K}_Y^2)}} - \frac{\mathrm{Tr}(\boldsymbol{K}_X\boldsymbol{K}_Y)\boldsymbol{K}_Y}{\sqrt{\mathrm{Tr}(\boldsymbol{K}_X^2)\,\mathrm{Tr}(\boldsymbol{K}_Y^2)}\,\mathrm{Tr}(\boldsymbol{K}_Y^2)} \\ &= \frac{1}{\|\boldsymbol{K}_X\| \cdot \|\boldsymbol{K}_Y\|}\left(\boldsymbol{K}_X - \frac{\langle \boldsymbol{K}_X, \boldsymbol{K}_Y\rangle}{\|\boldsymbol{K}_Y\|^2}\boldsymbol{K}_Y\right)\end{aligned} \tag{24}$$

**Geometric Interpretation:** The term in parentheses has an interpretable geometric meaning. Let $b = \frac{\langle \boldsymbol{K}_X, \boldsymbol{K}_Y\rangle}{\|\boldsymbol{K}_Y\|^2}$ be the scalar projection coefficient of the target onto the model. The gradient is proportional to:

$$\boldsymbol{R} = \boldsymbol{K}_X - b\,\boldsymbol{K}_Y \tag{25}$$

This matrix $\boldsymbol{R}$ is the *residual of the projection* of $\boldsymbol{K}_X$ onto the line spanned by $\boldsymbol{K}_Y$. Intuitively, the gradient pushes the model kernel to cover the "unexplained" geometric structure of the target. It aligns the model with the component of the target that is currently orthogonal to it.

The derivation of the gradient with respect to the Gram matrix $\boldsymbol{G}_Y^{\kappa_y}$ follows from the properties of the total differential and the duality of the Frobenius inner product. Let RV be a scalar function of the kernel $\boldsymbol{K}_Y$. The total differential $d\mathrm{RV}$ is defined as:

$$d\mathrm{RV} = \left\langle \frac{\partial \mathrm{RV}}{\partial \boldsymbol{K}_Y}, d\boldsymbol{K}_Y \right\rangle = \mathrm{Tr}\left(\left(\frac{\partial \mathrm{RV}}{\partial \boldsymbol{K}_Y}\right)^\top d\boldsymbol{K}_Y\right) \tag{26}$$

Because $\boldsymbol{K}_Y$ and its gradient are symmetric matrices, the trace inner product satisfies $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \mathrm{Tr}(\boldsymbol{AB})$. From the linear mapping $\boldsymbol{K}_Y = \boldsymbol{Q}\boldsymbol{G}_Y^{\kappa_y}\boldsymbol{Q}^\top$, we obtain the differential:

$$d\boldsymbol{K}_Y = \boldsymbol{Q}(d\boldsymbol{G}_Y^{\kappa_y})\boldsymbol{Q}^\top \tag{27}$$

Substituting this into (26) and applying the cyclic property of the trace operator, $\mathrm{Tr}(\boldsymbol{ABC}) = \mathrm{Tr}(\boldsymbol{CAB})$, yields:

$$\begin{aligned}d\mathrm{RV} &= \mathrm{Tr}\left(\frac{\partial \mathrm{RV}}{\partial \boldsymbol{K}_Y}\boldsymbol{Q}\,d\boldsymbol{G}_Y^{\kappa_y}\boldsymbol{Q}^\top\right) \\ &= \mathrm{Tr}\left(\boldsymbol{Q}^\top\frac{\partial \mathrm{RV}}{\partial \boldsymbol{K}_Y}\boldsymbol{Q}\,d\boldsymbol{G}_Y^{\kappa_y}\right)\end{aligned} \tag{28}$$

By the principle of identification in matrix calculus, the gradient $\boldsymbol{\Gamma}$ is the term in the trace inner product that is dual to $d\boldsymbol{G}_Y^{\kappa_y}$:

$$\boldsymbol{\Gamma} := \frac{\partial \mathrm{RV}}{\partial \boldsymbol{G}_Y^{\kappa_y}} = \boldsymbol{Q}^\top \left( \frac{\partial \mathrm{RV}}{\partial \boldsymbol{K}_Y} \right) \boldsymbol{Q} \tag{29}$$

Expanding this with the derivative of the RV coefficient with respect to $\boldsymbol{K}_Y$, we arrive at the gradient:

$$\boldsymbol{\Gamma} = \frac{1}{\|\boldsymbol{K}_X\| \cdot \|\boldsymbol{K}_Y\|} \boldsymbol{Q}^\top \left( \boldsymbol{K}_X - \frac{\langle \boldsymbol{K}_X, \boldsymbol{K}_Y \rangle}{\|\boldsymbol{K}_Y\|^2} \boldsymbol{K}_Y \right) \boldsymbol{Q} \tag{30}$$

We call $\boldsymbol{\Gamma} = (\Gamma_{ij})$ the *structural gradient*. It is the adjoint transformation of the kernel-space residual into the uncentered similarity space, providing the necessary pairwise attraction-repulsion weights for the parametric update.

#### 4.3.2 The Parametric Gradient

The final step turns the structural requirement $\boldsymbol{\Gamma}$ into a rule for moving the coordinates $\boldsymbol{Y}$. This depends only on the readout $\kappa_y$, and for the two standard families it takes a clean, force-directed form.

**Proposition 3** (The RV gradient)**.** *Write $\boldsymbol{\Gamma}$ for the structural gradient (30) and $\circ$ for the Hadamard (entrywise) product. The gradient of the alignment with respect to the configuration is:*

*(A) for the* dot-product *readout of (21),*

$$\frac{\partial \mathrm{RV}}{\partial \boldsymbol{Y}} = 2\big(\boldsymbol{\Gamma} \circ \kappa_y'(\boldsymbol{Y}\boldsymbol{Y}^\top)\big)\boldsymbol{Y}; \tag{31}$$

*(B) for the* distance *readout of (21), at each point $\boldsymbol{y}_i$,*

$$\frac{\partial \mathrm{RV}}{\partial \boldsymbol{y}_i} = 4 \sum_{j=1}^{n} \Gamma_{ij}\, \kappa_y'(\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2)\, (\boldsymbol{y}_i - \boldsymbol{y}_j). \tag{32}$$

Both readouts act as forces on the point cloud. In the dot-product case the linear identity kernel $\kappa_y(x) = x$ has $\kappa_y' \equiv 1$, so the gradient collapses to $2\boldsymbol{\Gamma}\boldsymbol{Y}$, the classical spectral update, whose fixed points are the leading eigenvectors of $\boldsymbol{\Gamma}$: the degenerate member of Proposition 2(iii) recovers, in gradient form, the spectral world of Section 3. In the distance case each pair $(i, j)$ is pushed along the axis $\boldsymbol{y}_i - \boldsymbol{y}_j$ with a signed weight $\Gamma_{ij}$ (attracting or repelling according to the residual error) modulated by a *stiffness* $\kappa_y'(\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2)$ that depends on the pair's current proximity. These two cases are all an iterative solver requires; the resulting scheme, with implementation notes, is collected in Appendix B.

### 4.4 The Attractive Force and its Kinship with t-SNE

The gradient of Section 4.3 was derived by the chain rule, as if $\boldsymbol{Y}$ were a free parameter: assembled, its two levels compute the *pullback* $\partial \mathrm{RV}/\partial \boldsymbol{Y} = D\Phi(\boldsymbol{Y})^* \nabla_{\boldsymbol{K}_Y} \mathrm{RV}$ of the ambient Frobenius gradient (24) into the coordinates $\boldsymbol{Y}$. Here and below, $\nabla_{\boldsymbol{K}_Y} \mathrm{RV}$ denotes the matrix $\partial \mathrm{RV}/\partial \boldsymbol{K}_Y$ of (24) read no longer as an array of partial derivatives but as the element of $\mathcal{K}_n$ representing the differential against the Frobenius inner product, $d\mathrm{RV} = \langle \nabla_{\boldsymbol{K}_Y} \mathrm{RV}, d\boldsymbol{K}_Y \rangle$; the two coincide entrywise because the metric is Frobenius, and it is this metric reading that the tangent projection below requires. But Proposition 2 shows that $\boldsymbol{Y}$ is not free: the output kernels are confined to the manifold $\mathcal{S}_q^\kappa$. One may therefore worry that following $\partial \mathrm{RV}/\partial \boldsymbol{Y}$ chases an artifact of the parametrization rather than a genuine feature of the alignment. It does not.

**Proposition 4** (The parametric gradient as a Riemannian gradient)**.** *Let $\boldsymbol{Y}$ be a configuration at which Proposition 2 yields a local sheet, with $\boldsymbol{K}_Y \neq \boldsymbol{0}$, and let $P_T$ be the orthogonal projection of $\mathcal{K}_n$ onto the tangent space $T = \operatorname{Im} D\Phi(\boldsymbol{Y})$ of that sheet at $\boldsymbol{K}_Y$. Then:*

*(i) the Riemannian gradient of* RV *restricted to $\mathcal{S}_q^\kappa$ is*

$$\operatorname{grad} \mathrm{RV} = P_T\, \nabla_{\boldsymbol{K}_Y} \mathrm{RV}; \tag{33}$$

*(ii) it vanishes exactly where the parametric gradient does:*

$$\partial \mathrm{RV}/\partial \boldsymbol{Y} = 0 \iff P_T \, \nabla_{\boldsymbol{K}_Y} \mathrm{RV} = 0. \tag{34}$$

The ambient gradient is proportional to the residual $\boldsymbol{R} = \boldsymbol{K}_X - b\,\boldsymbol{K}_Y$ of Section 4.3: the part of the target not yet explained. Two things can be done with it. The Riemannian gradient (Absil et al., 2008; Boumal, 2023) *projects* it onto the tangent plane of the manifold; the parametric gradient *pulls it back* into the coordinate chart $\boldsymbol{Y}$. The two are readings of the same object in different metrics: they vanish together (exactly when the residual is orthogonal to the manifold, at a true critical point of the alignment on the sheet, not at an accident of the chart) and differ by the preconditioner $\boldsymbol{B_Y} = D\Phi(\boldsymbol{Y})^* D\Phi(\boldsymbol{Y})$. On the whole chart, $\boldsymbol{B_Y}$ is only positive semi-definite, its kernel being the motions the readout cannot see ($\mathcal{I}_{\boldsymbol{Y}}$ or $\mathcal{R}_{\boldsymbol{Y}}$, by family); but the parametric gradient automatically lives in the *horizontal* space $(\ker D\Phi)^\perp$, on which $\boldsymbol{B_Y}$ is positive definite: the preconditioner rescales but never reverses the ascent, and the invisible directions receive no force at all. The gradient scheme of Appendix B therefore solves the geometric problem, not a surrogate; the proof is given in Appendix A.

We can now make the pulling force explicit. From here on the distance family takes the lead, as the home of the neighbor embeddings whose dynamics we set out to explain; the dot-product family, whose non-affine members are standard in matrix-factorization embeddings, is left as an open direction. Specializing the distance gradient (case (B) above) to the Student-t readout reveals a precise, and slightly surprising, relation to t-SNE.

**Corollary 2** (Attractive force)**.** *With the Student-t readout, write* $\tilde{q}_{ij} = (1 + \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2)^{-1}$ *and* $\tilde{\boldsymbol{K}}_X = \boldsymbol{Q}^\top \boldsymbol{K}_X \boldsymbol{Q}$*. The attractive (numerator) part of the gradient on a point* $\boldsymbol{y}_k$ *is*

$$\frac{\partial \langle \boldsymbol{K}_X, \boldsymbol{K}_Y \rangle}{\partial \boldsymbol{y}_k} = 4 \sum_j (\tilde{\boldsymbol{K}}_X)_{kj} \, \tilde{q}_{kj}^2 \, (\boldsymbol{y}_j - \boldsymbol{y}_k), \tag{35}$$

*that is, each pair is pulled together with the* signed *weight (centered input affinity) × (output affinity)*$^2$*, pairs with a negative centered affinity being pushed apart.*

This is a close relative of the attractive force of t-SNE, $4 \sum_j p_{kj} \, \tilde{q}_{kj} \, (\boldsymbol{y}_j - \boldsymbol{y}_k)$, but *not* the same one: t-SNE carries one power of $\tilde{q}$ fewer. The two share the very same Student-t output; the discrepancy comes entirely from the *objective*. The logarithm in the Kullback–Leibler divergence of t-SNE linearizes the output dependence ($\tilde{q}^1$), whereas the bilinear RV alignment preserves the readout derivative $\kappa'_y = -\tilde{q}^2$. Our attraction is therefore strictly more local (distant pairs are crushed by the square), and identity (35) is one that the framework *predicts* and that we confirm against the implementation by automatic differentiation. The matching repulsive force, and the precise sense in which our objective is a *cousin* of t-SNE rather than t-SNE itself, are taken up in Section 5.

## 5 The Diagonal and the Volume: Tether and Repulsion

The gradient of Section 4 navigates the manifold; two structural questions about the *objective* remain before the picture of neighbor embedding is complete, and both turn on modes that the centering singles out. We take them up here, specializing throughout to the bounded distance readout that hosts the neighbor embeddings (Section 4.4), with the linear cone kept as the point of contrast.

The first mode is the *diagonal*. Every kernel in $\mathcal{K}_n$ is centered, and centering ties its diagonal to its off-diagonal entries: the diagonal is never free. Yet the working currency of neighbor embeddings is a *hollow* matrix, pairwise affinities with an empty diagonal, as in the input affinities of t-SNE (van der Maaten & Hinton, 2008). What, then, does the full Frobenius cosine weigh that a cosine on the off-diagonal entries alone would not? The answer is exact and compact: *one metric*, an energy of *degrees*, which is signal on the cone but a spread-suppressing *tether* on the bounded manifold.

The second mode is the *volume*. Once that tether is lifted, something must still push clusters apart; yet the RV objective contains no repulsion at all, and this is a theorem, not an oversight. The centering discards

exactly one mode of the raw similarities, the global *volume* of the embedding, and the repulsion of t-SNE is nothing but the gradient of that discarded mode. Where prior accounts treat repulsion as a primitive (an empirical dial on an attraction–repulsion spectrum (Böhm et al., 2022), or a term unified at the solver level (Yang et al., 2014; 2015)), here it is *derived*.

The two modes sit on opposite sides of the kernel space: the diagonal is redundant *inside* $\mathcal{K}_n$, tied to the off-diagonal by a degree metric; the volume is discarded *outside* it. Together they are the algebraic distance between kernel alignment and neighbor embedding.

### 5.1 Hollow Coordinates and the Degree Metric

All the objects of this section stem from one decomposition. Split any $\boldsymbol{K} \in \mathcal{K}_n$ into its diagonal and its hollow (off-diagonal) part, and name the *weighted degrees* of the latter:

$$\boldsymbol{K} = \operatorname{diag}(\boldsymbol{d}) + \mathring{\boldsymbol{K}}, \qquad \mathring{K}_{ii} := 0, \qquad r_i := \sum_{j \neq i} \mathring{K}_{ij} \sqrt{f_j}. \tag{36}$$

Read row by row, the centering condition $\boldsymbol{K}\sqrt{\boldsymbol{f}} = \boldsymbol{0}$ says that the diagonal carries the degrees, $d_i = -r_i/\sqrt{f_i}$, and nothing else. Writing $\mathcal{D}$ for this reconstruction (the linear map $(\mathcal{D}\mathring{\boldsymbol{K}})_i := -r_i/\sqrt{f_i}$ on $\operatorname{Hollow}(n)$, the space of symmetric matrices with zero diagonal), a symmetric matrix therefore belongs to $\mathcal{K}_n$ if and only if $\boldsymbol{d} = \mathcal{D}\mathring{\boldsymbol{K}}$, and $\mathring{\boldsymbol{K}} \mapsto \operatorname{diag}(\mathcal{D}\mathring{\boldsymbol{K}}) + \mathring{\boldsymbol{K}}$ is a linear bijection from $\operatorname{Hollow}(n)$ onto $\mathcal{K}_n$: the hollow part is a free and complete coordinate system, recovering the dimension $\binom{n}{2}$ of Section 2. Note that $\mathring{\boldsymbol{K}}$ itself is no longer a kernel, neither centered nor positive semi-definite: the centering of $\boldsymbol{K}$ lives entirely in its diagonal.

This bookkeeping is not new. With uniform weights it is the graph Laplacian: $\boldsymbol{K}\sqrt{\boldsymbol{f}} = \boldsymbol{0}$ says that $\boldsymbol{K}$ has zero row sums, i.e. $\boldsymbol{K} = \operatorname{diag}(\boldsymbol{W}\boldsymbol{1}_n) - \boldsymbol{W}$ with (possibly signed) weights $W_{ij} = -\mathring{K}_{ij}$ and diagonal $K_{ii} = -\rho_i$, minus the plain degree $\rho_i = \sum_{j \neq i} \mathring{K}_{ij}$ (Chung, 1997; von Luxburg, 2007). What has not been written, to our knowledge, is the exact role this degree-diagonal plays inside the RV cosine, and the fact that its legitimacy there depends on the output readout.

To make that role explicit, define the *hollow RV coefficient* as the plain cosine of the hollow parts, $\mathrm{RV}_{\mathrm{hol}}(\boldsymbol{K}_X, \boldsymbol{K}_Y) := \langle \mathring{\boldsymbol{K}}_X, \mathring{\boldsymbol{K}}_Y \rangle / \|\mathring{\boldsymbol{K}}_X\| \|\mathring{\boldsymbol{K}}_Y\|$ (the objective of a method that compares affinity patterns and ignores diagonals), and rewrite the inner product of $\mathcal{K}_n$ on the hollow coordinates.

**Proposition 5** (The diagonal as a degree metric)**.** *On hollow coordinates, the Frobenius inner product of $\mathcal{K}_n$ becomes*

$$\langle \boldsymbol{K}_X, \boldsymbol{K}_Y \rangle = \langle \mathring{\boldsymbol{K}}_X, \mathring{\boldsymbol{K}}_Y \rangle + \sum_{i=1}^{n} \frac{r_i(X)\, r_i(Y)}{f_i} = \langle \mathring{\boldsymbol{K}}_X, \boldsymbol{M}\, \mathring{\boldsymbol{K}}_Y \rangle, \quad \boldsymbol{M} := \boldsymbol{I} + \mathcal{D}^*\mathcal{D}, \tag{37}$$

*where $r_i(X), r_i(Y)$ are the degrees (36) of the two kernels and $\boldsymbol{M}$ is self-adjoint with $\boldsymbol{M} \succeq \boldsymbol{I} \succ \boldsymbol{0}$, i.e. with eigenvalues at least one, depending only on the weights $\boldsymbol{f}$. In particular $\|\boldsymbol{K}\|^2 = \|\mathring{\boldsymbol{K}}\|^2 + \sum_i r_i^2/f_i$ (the diagonal contributes exactly the* degree energy $\sum_i r_i^2/f_i = \sum_i K_{ii}^2$*), and the full and hollow RV coefficients are the same cosine on the same coordinates, evaluated in two metrics: writing $\langle \cdot, \cdot \rangle_{\boldsymbol{M}} := \langle \cdot, \boldsymbol{M} \cdot \rangle$, with induced norm $\| \cdot \|_{\boldsymbol{M}}$ and cosine $\cos_{\boldsymbol{M}}$,*

$$\mathrm{RV}(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \cos_{\boldsymbol{M}}\left(\mathring{\boldsymbol{K}}_X, \mathring{\boldsymbol{K}}_Y\right), \qquad \mathrm{RV}_{\mathrm{hol}}(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \cos_{\boldsymbol{I}}\left(\mathring{\boldsymbol{K}}_X, \mathring{\boldsymbol{K}}_Y\right). \tag{38}$$

A direct computation of the adjoint makes the metric concrete: entrywise,

$$(\boldsymbol{M}\mathring{\boldsymbol{K}})_{ij} = \mathring{K}_{ij} + \frac{1}{2}\left(\frac{\sqrt{f_j}}{f_i}\, r_i + \frac{\sqrt{f_i}}{f_j}\, r_j\right) \quad (i \neq j), \tag{39}$$

each pair is surcharged with the degrees of its two endpoints, the factors $1/f_i$ amplifying low-mass points; with uniform weights, $(\boldsymbol{M}\mathring{\boldsymbol{K}})_{ij} = \mathring{K}_{ij} + \frac{1}{2}(\rho_i + \rho_j)$ and the degree energy is $\sum_i \rho_i^2$. The identity (37) also says that the coordinate map is a linear *isometry* between $(\operatorname{Hollow}(n), \langle \cdot, \cdot \rangle_{\boldsymbol{M}})$ and $(\mathcal{K}_n, \langle \cdot, \cdot \rangle)$: aligning full centered kernels and aligning bare affinity patterns are the *same* problem up to this one metric, and everything that separates the full RV from its hollow variant must pass through the degree term.

### 5.2 Signal on the Cone, Tether on the Manifold

The metric $\boldsymbol{M}$ is derived, not chosen: it is forced by the centering. The genuine question is whether its surcharge weighs information that the output can actually *control*, and this is where the linear cone and the distance manifold differ. For the linear readout, the raw diagonal $[\boldsymbol{G}_Y]_{ii} = \|\boldsymbol{y}_i\|^2$ holds the squared radii of the configuration, which the embedding varies freely; after centering, $K_{ii} = f_i \|\boldsymbol{y}_i^c\|^2$, so the degree energy is precisely the energy of the (weighted) centered radii. For a distance readout, the raw diagonal $[\boldsymbol{G}_Y^{\kappa_y}]_{ii} = \kappa_y(0)$ is the same constant for every point and every configuration: all the information sits off the diagonal, and the diagonal of $\boldsymbol{K}_Y$ is a pure artifact of the centering. The following definition and lemma make the dichotomy formal.

**Definition 1** (Data-justification)**.** The metric $\boldsymbol{M}$ is *data-justified* for an output family when maximizing the full RV coefficient over the family's achievable set coincides with the metric projection of the target onto that set (in the $\boldsymbol{M}$-norm on hollow coordinates or, equivalently by Proposition 5, in the Frobenius norm on $\mathcal{K}_n$), and this projection is solved globally, in closed form, by spectral truncation. In words: maximizing in the metric $\boldsymbol{M}$ *is* projecting onto the data, exactly.

**Lemma 1** (Justification of the metric)**.** *Between the linear output of Section 3 and the distance readouts of Section 4, $\boldsymbol{M}$ is data-justified exactly for the former; within this pair, the criterion is whether the diagonal of the raw output Gram varies with the configuration.*

The operative content of the lemma is its criterion: a constant raw diagonal versus a variable one. A non-affine dot-product readout sits between the two regimes: its raw diagonal $\kappa_y(\|\boldsymbol{y}_i\|^2)$ does vary, so the surcharge weighs genuine radial information, yet its curved sheet (Proposition 2(ii)) offers no spectral closed form; we do not pursue it.

On the manifold, the surcharge does more than lose its justification: it actively distorts the optimization. For a bounded readout such as the Student-t, spreading the configuration makes every off-diagonal affinity decay together, so the informative energy $\|\mathring{\boldsymbol{K}}_Y\|^2$ collapses; the degree energy does not follow, because the constant diagonal $\kappa_y(0)$ cannot be centered away: in the limit of an infinite spread, $\boldsymbol{K}_Y \to \kappa_y(0)\,\boldsymbol{Q}\boldsymbol{Q}^\top$, a fixed matrix whose energy sits mostly on the diagonal ($\frac{\kappa_y(0)}{n}\,\boldsymbol{K}_0$ for uniform weights: the *structural neutrality* of Section 2). In the full cosine (38), whose denominator carries $\|\mathring{\boldsymbol{K}}_Y\|_{\boldsymbol{M}}^2 = \|\mathring{\boldsymbol{K}}_Y\|^2 + \sum_i r_i^2/f_i$, the degree term thus behaves as a spread-insensitive *normalization floor*: once $\|\mathring{\boldsymbol{K}}_Y\|$ has fallen to its order, further spreading costs alignment no matter how much it improves the affinity pattern: a global *tether* on the embedding, which the hollow cosine simply removes. Section 6 measures this directly: under the full RV, the diagonal share of $\|\boldsymbol{K}_Y\|^2$ locks and the spread caps; under the hollow RV, both are released, converging toward the regime of t-SNE.

The story of the diagonal closes on a division of labor rather than a replacement. The full RV remains the canonical objective: the entire linear theory of Section 3 is the theory of *that* cosine, in the exact regime where Lemma 1 certifies the diagonal as signal. The hollow RV is its faithful specialization to the neighbor-embedding regime, measuring the off-diagonal information the readout actually encodes, and nothing else. The diagonal, in short, is the algebraic dial between multidimensional scaling and neighbor embedding: kept, it anchors the embedding to its radii; dropped, it releases the spread that a bounded readout needs. The force that actively *separates* clusters once the tether is lifted cannot be found in $\mathcal{K}_n$ at all: it is the subject of the next two subsections.

### 5.3 The Volume Mode

The tether is lifted; what must still *push*? Corollary 2 made the attractive force of the alignment explicit and left its repulsive counterpart pending, and Proposition 2(iv) warned that the Student-t sheet is bounded: separated clusters do not come for free. Work in the neighbor-embedding regime of Section 4, with the Student-t readout and its affinities $\tilde{q}_{ij} = (1 + \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2)^{-1}$, and define the *volume* of a configuration as the off-diagonal mass of its raw similarities,

$$Z(\boldsymbol{Y}) := \sum_{i \neq j} [\boldsymbol{G}_Y^{\kappa_y}]_{ij} = \sum_{i \neq j} \tilde{q}_{ij}, \tag{40}$$

the (constant) self-affinities being excluded, as in the normalization of t-SNE. The volume measures compactness: it is largest when the points coincide and decays to zero as the configuration spreads. The following proposition shows that the kernel space is structurally blind to it, and that this blindness is where the missing force was hiding.

**Proposition 6** (Volume blindness and the origin of repulsion)**.** *(a)* (The centering discards the volume.) *Since $\boldsymbol{Q}\mathbf{1}_n = \mathbf{0}$, the output kernel $\boldsymbol{K}_Y = \boldsymbol{Q}\,\boldsymbol{G}_Y\,\boldsymbol{Q}^\top$ is invariant under the volume shifts $\boldsymbol{G}_Y \mapsto \boldsymbol{G}_Y + \alpha\,\mathbf{1}\mathbf{1}^\top$, $\alpha \in \mathbb{R}$, while the off-diagonal mass moves, $Z \mapsto Z + \alpha\, n(n-1)$: every objective built on $\boldsymbol{K}_Y$ alone (the RV coefficient in particular) carries no information about $Z$.*

*(b)* (The discarded gradient is the repulsion of t-SNE.) *For $\tau > 0$, the gradient of the size penalty $-\tau \log Z$ is*

$$\frac{\partial}{\partial \boldsymbol{y}_i}\,(-\tau \log Z) = \frac{4\tau}{Z} \sum_j \tilde{q}_{ij}^{\,2}\,(\boldsymbol{y}_i - \boldsymbol{y}_j), \tag{41}$$

*pushing every pair apart; at $\tau = 1$, it is exactly the repulsive force of t-SNE.*

*(c)* (No repulsion from within $\mathcal{K}_n$.) *For every differentiable $F : \mathcal{K}_n \to \mathbb{R}$, $\frac{d}{d\alpha} F\big(\boldsymbol{Q}(\boldsymbol{G}_Y + \alpha \mathbf{1}\mathbf{1}^\top)\boldsymbol{Q}^\top\big) = 0$: no objective of the centered kernel can reward spreading; a repulsive term must be a function of the* raw *similarities, one that the volume mode moves.*

This is best read through the lens of shape statistics (Kendall, 1984; Dryden & Mardia, 2016). A normalized cosine between centered kernels captures the *form* of the configuration (invariant to rigid motions and to scale) while $Z$ measures its *size*. Neighbor embedding then decomposes exactly as *form alignment plus size control*,

$$\max_{\boldsymbol{Y}} \quad \mathrm{RV}(\boldsymbol{K}_X, \boldsymbol{K}_Y) \;-\; \tau \log Z(\boldsymbol{Y}), \tag{42}$$

the repulsion reinjecting the one dimension that the cosine discards by construction. Note the symmetry: the attraction of the alignment (Corollary 2) and the push (41) are both in $\tilde{q}^{\,2}$, two faces of the same chain rule through the bounded readout $\kappa_y' = -\tilde{q}^{\,2}$, whereas t-SNE is asymmetric, $\tilde{q}^{\,1}$ in attraction and $\tilde{q}^{\,2}$ in repulsion. The composite objective (42) is therefore not t-SNE but its RV-cosine *cousin*: identical repulsion, an analogous input weighting (signed here, where t-SNE's $p_{ij}$ are nonnegative), strictly more local attraction; the divergence is located, and explained, in Section 4.4.

The proposition also completes the accounting of the centering: three components, three outcomes. The hollow differentials of the distances pass through it in full (case (B) in the proof of Proposition 2); the diagonal is redundant *inside* $\mathcal{K}_n$, tied to the off-diagonal by the degree metric (Proposition 5); the volume mode is discarded *outside* $\mathcal{K}_n$. The dials compose accordingly: under the full cosine, the tether of Section 5.2 caps the very spread that the size control tries to create; under the hollow cosine, the tether is lifted and the volume push separates cleanly. And the deeper reason a counter-force is needed at all is Proposition 2(iv): unlike the linear cone, where scale is free, the Student-t sheet is bounded: its affinities saturate, and only the size mode can spread the clusters apart.

### 5.4 Where Repulsion Can Live

Proposition 6(c) organizes the neighbor-embedding families by the seat of their repulsion (Table 1). The decisive axis is the *domain*: on the centered kernel, the only carrier of global mass, the component $\mathbf{1}\mathbf{1}^\top$, is annihilated by $\boldsymbol{Q}$, so no push can exist; every push is a functional of the raw similarities. Among raw-domain methods, the *aggregation* then separates two subfamilies: a *global* normalization, one partition function $Z$ coupling all pairs, versus a *per-pair* repulsion, a sum of independent edge terms.

The global-versus-per-pair distinction is precisely the *place of the logarithm.* t-SNE takes $\log\big(\sum_{k\neq l} \tilde{q}_{kl}\big)$ (the logarithm outside the sum), so a single $Z$ couples every pair, and the force self-adjusts through the factor $1/Z$; UMAP takes $\sum_{i\neq j} \log(1 - \nu_{ij})$ (the logarithm inside), so its edges repel independently, and no global mass exists. A family of volume penalties $g(Z)$ therefore parametrizes only the *global* subfamily: a linear $g$ gives a constant-strength push in the spirit of elastic embedding, $g = \log$ gives t-SNE.

Table 1: The seat of the repulsion across embedding objectives: this framework, elastic embedding (Carreira-Perpiñán, 2010), t-SNE (van der Maaten & Hinton, 2008) and UMAP (McInnes et al., 2020). All sums run over $i \neq j$; $w_{ij}^-$ are data-dependent repulsive weights, and $\mu_{ij}, \nu_{ij}$ are the input and output fuzzy memberships of UMAP, whose per-pair term is optimized by negative sampling and whose effective loss is analyzed by Damrich & Hamprecht (2021).

| method | domain | repulsive term | coupling |
|---|---|---|---|
| RV alignment (this paper) | centered $\boldsymbol{K}_Y$ | none (volume annihilated) | — |
| elastic embedding | raw $\boldsymbol{G}_Y$ | $\tau \sum w_{ij}^- \, e^{-\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2}$ | per-pair, fixed |
| t-SNE | raw $\boldsymbol{G}_Y$ | $\log Z$, force $\propto \tilde{q}_{ij}^2 / Z$ | global (one $Z$) |
| UMAP | raw $\boldsymbol{G}_Y$ | $-\sum (1-\mu_{ij}) \log(1-\nu_{ij})$ | per-pair |

This positions the framework precisely. The only repulsion that $\mathcal{K}_n$ allows to be reinjected is a function of its own discarded mode, $g(Z)$, which lands, by construction, in the global subfamily. The completed objective (42) is therefore a close relative of t-SNE, and *cannot* be one of UMAP: the per-pair push is of a different functional nature, not a different tuning. What separates t-SNE from UMAP is thus not the output kernel (both are heavy-tailed) but the structure of the repulsion; just as what separates the RV attraction from t-SNE's is not the kernel but the objective (Corollary 2). The framework does not blur these methods into one: it says where they differ, and why.

Two reservations and a programme close the section. The derivation concerns the push, not the pull, and the composite objective (42) is still optimized locally; moreover, the size reward $-\log Z$ grows without bound as the embedding spreads, so $\tau$ is a genuine tuning parameter. As a programme, the volume mode connects to the majorization–minimization view of neighbor embeddings (Yang et al., 2014; 2015): for a concave penalty $g(Z)$, the instantaneous strength $g'(Z)$ is at once the dual variable of a size constraint and the slope of a majorization step, a connection we leave open, noting only that the gradient identities of this section and of Corollary 2 appear in neither line of work.

## 6 Case Studies

The preceding sections build a single geometric object and read off its consequences: the achievable cone and its closed-form projection (Section 3), the manifold that replaces it once the output kernel bends (Section 4), the gradient that navigates it (Section 4.3), the diagonal that tethers it (Section 5.1), and the discarded volume mode that alone can push it apart (Section 5.3). Each of these is a *claim*, and each claim makes a measurable prediction. This section verifies them one at a time, in the order of the theory. The aim is validation, not benchmarking: we do not ask whether the framework beats t-SNE, but whether the embedding it computes is the one the geometry says it must be.[3]

### 6.1 Experimental Conditions

The conditions below are shared by every study; each subsection states only its own departures from them.

**Datasets.** Three datasets span the regimes of interest, at the sample sizes noted: **MNIST** (784 raw pixels, 200 images per digit, $n = 2000$; LeCun et al., 1998), the **PBMC3k single-cell** transcriptome (50 leading principal components of the log-normalized counts, $n = 2638$; Wolf et al., 2018), and the **Swiss roll** (3 ambient coordinates, $n = 2000$). Individuals carry the uniform weight $\boldsymbol{f} = \mathbf{1}/n$, and every embedding is to $q = 2$.

**Kernels.** Input targets are built from the menu of Section 2.4, and Table 2 collects their explicit forms. The six spectral methods differ only in the similarity $\boldsymbol{G}$ they center into $\mathcal{K}_n$; the neighbor-embedding studies use the perplexity-calibrated Gaussian affinity of t-SNE, symmetrized and *softened* by the entrywise power

[3] All experiments are reproducible from the companion repository under fixed seeds; the material cut for space (the spectral scatter galleries, the $t$-SNE/UMAP side-by-sides, and the full hyperparameter and $\gamma$ sweeps) is hosted there as an online gallery.

Table 2: The kernels used across the case studies. Each input similarity $\boldsymbol{G}$ enters the RV objective double-centered into $\mathcal{K}_n$ (as $\boldsymbol{QGQ}^\top$); the table lists $\boldsymbol{G}$ itself. Notation: $\boldsymbol{D}_{\text{geo}}$ is the matrix of $k$-NN graph shortest-path distances and $\boldsymbol{D}_{\text{geo}}^{\circ 2}$ its entrywise square; $\boldsymbol{W}_{\text{lle}}$ the local reconstruction weights; $\boldsymbol{L}_{\text{sym}} = \boldsymbol{I} - \boldsymbol{D}_{\text{deg}}^{-1/2}\boldsymbol{S}\boldsymbol{D}_{\text{deg}}^{-1/2}$ the normalized Laplacian of a $k$-NN affinity $\boldsymbol{S}$ ($\boldsymbol{S} = \boldsymbol{A}$, a binary connectivity, for Laplacian Eigenmaps; $\boldsymbol{S} = \boldsymbol{W}$, heat weights $\mathrm{e}^{-\|\cdot\|^2/\sigma^2}$, for Diffusion Maps), with $\boldsymbol{D}_{\text{deg}}$ the degree matrix of $\boldsymbol{S}$; $(\cdot)^\dagger$ the Moore–Penrose pseudo-inverse; and $p_{j|i}$ the perplexity-calibrated conditional affinities of t-SNE. Every method is then read out through the shared linear or Student-t output.

| Method | Kernel $\boldsymbol{G}$ | Parameters |
|---|---|---|
| *Input: spectral methods (linear output)* | | |
| PCA | $\boldsymbol{X}\boldsymbol{X}^\top$ | — |
| Kernel PCA (RBF) | $\exp\big(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/\sigma^2\big)$ | $\sigma^2$: median heuristic |
| Isomap | $-\frac{1}{2}\boldsymbol{D}_{\text{geo}}^{\circ 2}$ | $k = 15$ |
| LLE | $\boldsymbol{\Phi}^\dagger, \quad \boldsymbol{\Phi} = (\boldsymbol{I} - \boldsymbol{W}_{\text{lle}})^\top(\boldsymbol{I} - \boldsymbol{W}_{\text{lle}})$ | $k = 15$ |
| Laplacian Eigenmaps | $\boldsymbol{L}_{\text{sym}}^\dagger \quad (\boldsymbol{S} = \boldsymbol{A})$ | $k = 15$ |
| Diffusion Maps | $\exp(-t\,\boldsymbol{L}_{\text{sym}}) \quad (\boldsymbol{S} = \boldsymbol{W})$ | $k = 15,\ t = 10$ |
| *Input: neighbor embedding and supervision* | | |
| t-SNE affinity $\boldsymbol{K}_{\text{ag}}$ | $\big(\boldsymbol{P}/\max \boldsymbol{P}\big)^\gamma, \quad P_{ij} = \frac{p_{j\|i} + p_{i\|j}}{2n}$ | perplexity 30, $\gamma = 0.5$ |
| Class kernel $\boldsymbol{K}_Z$ | $\boldsymbol{P}_Z\boldsymbol{K}_X\boldsymbol{P}_Z$ (Section 3.4) | labels $\boldsymbol{Z}$ |
| *Output kernels* | | |
| Linear (cone) | $\boldsymbol{Y}\boldsymbol{Y}^\top$ | — |
| Student-t (manifold) | $\big(1 + \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2\big)^{-1}$ | — |

$(\,\cdot\,)^\gamma$, $\gamma = 0.5$ (a concave reshaping that lifts weak affinities), written $\boldsymbol{K}_{\text{ag}}$, together with the class kernel $\boldsymbol{K}_Z$ of Section 3.4. The output kernel is the linear Gram $\boldsymbol{G}_Y = \boldsymbol{Y}\boldsymbol{Y}^\top$ on the cone and the Student-t readout $\kappa_y(t) = (1 + t)^{-1}$ on the manifold.

**Optimization.** The spectral studies use the closed-form projection (one eigendecomposition, Theorem 2). The manifold studies maximize the RV coefficient by autograd gradient ascent with Adam (500 iterations, learning rate 0.1; Kingma & Ba, 2015) from a shared PCA initialization; the neighbor-embedding studies use the hollow cosine of Section 5.

**Indices.** We report, as needed: the *Procrustes* disparity to a reference embedding (residual after optimal similarity alignment, 0 = identical); the *k-NN overlap* and *trustworthiness* (Venna & Kaski, 2001) at $k = 15$ (neighborhood preservation against the input); the *spread* (root-mean-square radius $\sqrt{\sum_i f_i \|\boldsymbol{y}_i\|^2}$); the *diagonal fraction* $\sum_i K_{ii}^2/\|\boldsymbol{K}_Y\|^2$ (the degree energy of Proposition 5, relative to the total); and, on labeled data, the *adjusted Rand index* (ARI) (Hubert & Arabie, 1985) of a $k$-means clustering against the true classes.

### 6.2 The Cone: Exact Spectral Recovery and the Ceiling

The unification corollary (Corollary 1) claims that six classical methods are *exactly* the closed-form RV projection of their respective input kernels, and the alignment ceiling (Proposition 1) claims that a linear-readout embedding can align with its target no better than the truncated spectrum allows, a bound the method attains. Table 3 tests both. For each method we compute the closed-form projection of its input kernel and compare it to the reference library implementation.

Two readings confirm the theory. First, the Procrustes column is essentially zero for every method (the closed form reproduces the library embedding to numerical noise), the lone exception being Laplacian Eigenmaps (0.097 on single-cell, 0.021 on MNIST). That exception is a normalization convention, not an error: the reference recovers its coordinates as $\boldsymbol{D}_{\text{deg}}^{-1/2}\boldsymbol{u}_i$ from the eigenvectors $\boldsymbol{u}_i$ of $\boldsymbol{L}_{\text{sym}}$, a degree factor our symmetric readout omits; the discrepancy vanishes when degrees are uniform (Swiss roll, 0.003) and grows

Table 3: Closed-form spectral recovery. For each method (a single eigendecomposition) against its reference library: Procrustes disparity, $k$-NN overlap ($k = 15$), and the alignment ratio $\mathrm{RV}/\mathrm{RV}_{\max}$ of Proposition 1. The four linear-readout methods sit *exactly* at the ceiling; LLE and Laplacian Eigenmaps use an orthonormal (balanced-axis) readout and therefore fall below it: the reference's convention, not the RV optimum.

| Dataset | Method | Procrustes | $k$-NN | $\mathrm{RV}/\mathrm{RV}_{\max}$ |
|---|---|---|---|---|
| MNIST | PCA | 0.000 | 1.00 | **1.000** |
| | Kernel PCA (RBF) | 0.000 | 1.00 | **1.000** |
| | Isomap | 0.004 | 0.64 | **1.000** |
| | LLE | 0.000 | 1.00 | 0.949 |
| | Diffusion Maps | 0.001 | 0.85 | **1.000** |
| | Laplacian Eigenmaps | 0.021 | 0.34 | 0.991 |
| Single-cell | PCA | 0.000 | 1.00 | **1.000** |
| | Kernel PCA (RBF) | 0.000 | 1.00 | **1.000** |
| | Isomap | 0.006 | 0.48 | **1.000** |
| | LLE | 0.000 | 1.00 | 0.719 |
| | Diffusion Maps | 0.000 | 0.74 | **1.000** |
| | Laplacian Eigenmaps | 0.097 | 0.09 | 0.978 |
| Swiss roll | PCA | 0.000 | 1.00 | **1.000** |
| | Kernel PCA (RBF) | 0.000 | 1.00 | **1.000** |
| | Isomap | 0.000 | 0.97 | **1.000** |
| | LLE | 0.000 | 1.00 | 0.712 |
| | Diffusion Maps | 0.000 | 0.98 | **1.000** |
| | Laplacian Eigenmaps | 0.003 | 0.44 | 0.852 |

with their heterogeneity (single-cell, whose degrees are the most uneven). Second, and this is the column that ties the table to Proposition 1, the ratio $\mathrm{RV}/\mathrm{RV}_{\max}$ is exactly 1.000 for PCA, Kernel PCA, Isomap, and Diffusion Maps across all three datasets: each is the global RV optimum on the cone $\mathcal{S}_q$, not merely a good embedding. LLE and Laplacian Eigenmaps fall below the ceiling (0.71–0.95 and 0.85–0.99) precisely because their balanced-axis (orthonormal) readout trades alignment for equalized variance; the gap measures that trade, not a failure of recovery. The low $k$-NN overlap of Isomap and Laplacian Eigenmaps likewise follows from their nature: both reorganize the neighborhood graph globally.

A separate check closes the section. Feeding the softened Gaussian affinity $\boldsymbol{K}_{\mathrm{ag}}$ (perplexity 30, $\gamma = 0.5$) to the *gradient* optimizer with a linear output, the attained alignment reaches the closed-form ceiling $\mathrm{RV}_{\max}(2)$ to seven digits: 0.3027 (MNIST), 0.3770 (single-cell), 0.2143 (Swiss roll), each within $3 \times 10^{-7}$ of the eigendecomposition value. The manifold optimizer, run on a target whose optimum lies on the cone, lands on that optimum.

### 6.3 The Manifold: Dimension of the Sheet

Once the output kernel is non-linear, the achievable set is no longer a cone but a smooth manifold whose local dimension Proposition 2(i) pins down: $\operatorname{rank} D\Phi = nq - \binom{q+1}{2}$, the ambient $nq$ coordinates minus the $q$ translations and $\binom{q}{2}$ rotations that leave every pairwise distance, hence the kernel, unchanged. Because the predicted rank must hold at every $n$, and the Jacobian $D\Phi$ is a $\binom{n}{2} \times nq$ object infeasible to form at $n = 2000$, we verify it on small configurations ($n \in \{30, 50, 80, 120\}$, $q \in \{1, 2, 3\}$), computing the numerical rank of $D\Phi$ by singular-value decomposition of the autograd Jacobian at five random configurations each. The same protocol run at a non-affine dot-product readout (an exponential profile) returns $nq - \binom{q}{2}$ in all twelve cases: the sheet thicker by exactly the $q$ translations predicted by Proposition 2(ii), each configuration being a regular certificate that settles the generic alternative for that profile.

Table 4 reports a perfect match in all twelve cells: the measured rank equals $nq - \binom{q+1}{2}$ without exception, and a spectral cliff of $10^{13}$–$10^{14}$ separates the retained singular values from the invariance null space, so the

Table 4: Local dimension of the Student-t sheet: measured rank $D\Phi$ (SVD of the autograd Jacobian, five random configurations) against the prediction $nq - \binom{q+1}{2}$ of Proposition 2. The two agree exactly in all 12 cases; a spectral gap of $10^{13}$–$10^{14}$ between the last retained and first discarded singular value makes the rank unambiguous.

| | $q = 1$ | $q = 2$ | $q = 3$ |
|---|---|---|---|
| $n = 30$ | 29 | 57 | 84 |
| $n = 50$ | 49 | 97 | 144 |
| $n = 80$ | 79 | 157 | 234 |
| $n = 120$ | 119 | 237 | 354 |

Table 5: Cross-Procrustes disparities among four embeddings: the framework with a t-SNE and with a UMAP output, and the two reference libraries (perplexity $= k = 30$, $\gamma = 0.5$). The two framework variants are closer to *each other* than either is to its own library reference, and switching in the UMAP kernel moves the embedding no closer to UMAP: the objective, not the kernel, sets the geometry.

| Embedding pair | MNIST | Single-cell |
|---|---|---|
| framework-tSNE – framework-UMAP | **0.122** | **0.190** |
| reference-tSNE – reference-UMAP | 0.152 | 0.218 |
| framework-tSNE – reference-tSNE | 0.254 | 0.362 |
| framework-UMAP – reference-UMAP | 0.322 | 0.469 |
| framework-tSNE – reference-UMAP | 0.293 | 0.469 |
| framework-UMAP – reference-tSNE | 0.279 | 0.367 |

count is not a thresholding artifact. The headline case $q = 2$ gives dim $= 2n - 3$ $(57, 97, 157, 237$ for the four sizes), and the addition of $n = 120$ confirms the formula is no small-sample coincidence. The Student-t sheet is a genuine manifold of exactly the predicted dimension, and its null directions are the invariances the theory identifies: the rigid motions of the plane.

### 6.4 The Gradient: Two Forces, One Prediction

The gradient of Section 4.3 carries two quantitative claims. The attraction claim (Corollary 2) is that, with the Student-t readout, each pair is pulled with weight (centered input affinity) $\times\, \tilde{q}_{ij}^2$, an output affinity raised to the *second* power, where t-SNE's own attraction carries only the first. The repulsion claim (Proposition 6) is that the volume push $-\log Z$ likewise scales as $\tilde{q}_{ij}^2$. We check both against the autograd gradient of the actual code at $n = 2000$, $q = 2$. All three identities hold to machine precision: the RV attraction matches $\tilde{q}^2$ to $4 \times 10^{-19}$, and is cleanly distinguished from the $\tilde{q}^1$ alternative, which it misses by $10^{-4}$, five orders of magnitude above the match; the t-SNE attraction matches its own $(p - q)\tilde{q}$ form to $9 \times 10^{-9}$; and the volume repulsion matches $\tilde{q}^2$ to $4 \times 10^{-18}$. The attraction of the RV objective and the push of the volume mode share the same $\tilde{q}^2$ law, while t-SNE's attraction stands apart at $\tilde{q}^1$: the framework differs from t-SNE not in its output kernel but in its objective, exactly as Corollary 2 states.

This algebraic kinship has a geometric consequence, which the taxonomy of Section 5.4 predicts: because the reinjectable repulsion is a global $g(Z)$, the completed objective is a close relative of t-SNE and *cannot* be UMAP. Table 5 puts this to the test with a cross-Procrustes design. On each dataset we build four embeddings (the framework with a t-SNE output and with a UMAP output, and the two reference libraries) and measure all pairwise disparities, at matched hyperparameters (perplexity and $k$ equal to 30, $\gamma = 0.5$).

The prediction holds cleanly. The two framework variants sit at disparity 0.122 (MNIST) and 0.190 (single-cell) from one another, closer than either is to *any* reference, and closer even than the two references are to each other. Swapping the framework's output from a t-SNE to a UMAP kernel barely moves the embedding, because the objective, not the kernel, sets the geometry. The sharpest reading is the pair of framework-to-reference rows: the framework with a UMAP output is *farther* from reference-UMAP (0.322, 0.469) than

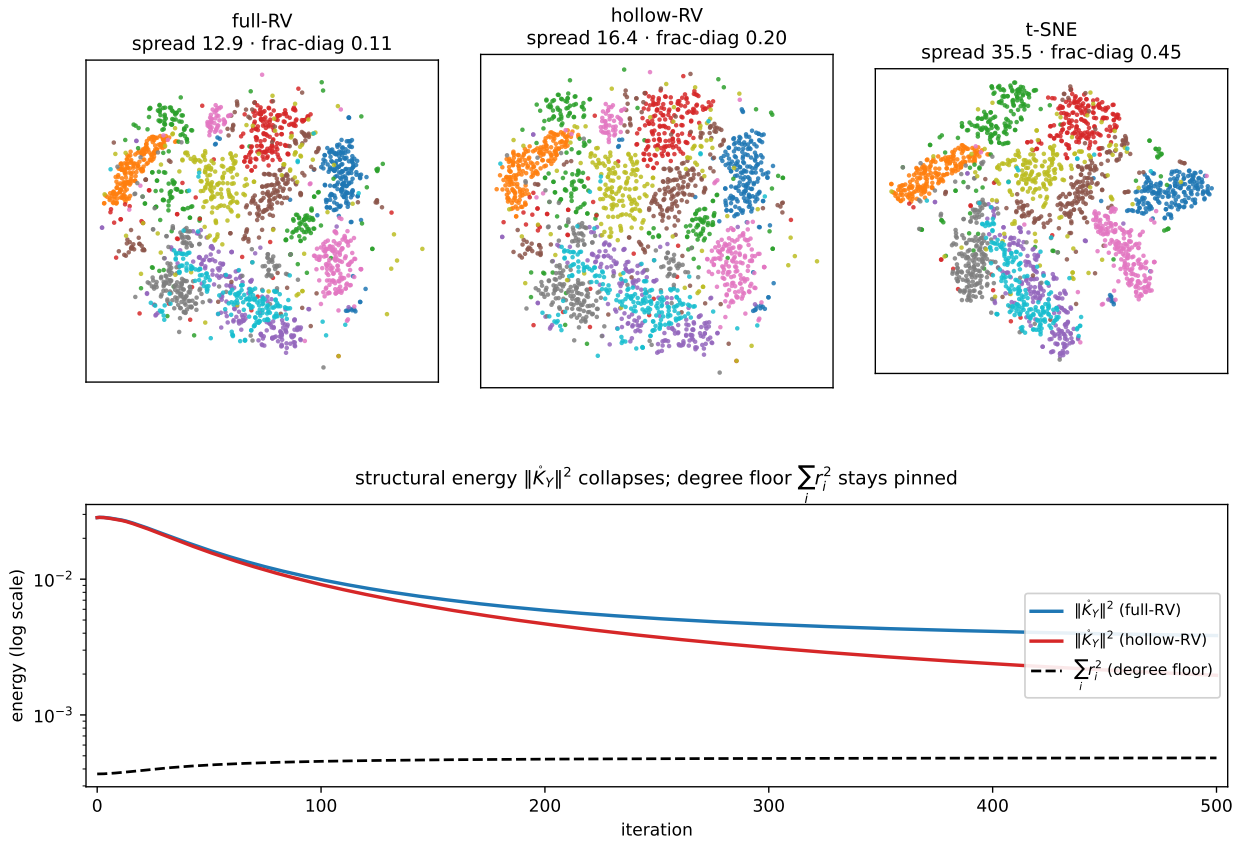


Figure 1: The diagonal tether (MNIST, $n = 2000$, perplexity 30, $\gamma = 0.5$, seed 0). *Top:* the three embeddings, full-RV $\rightarrow$ hollow-RV $\rightarrow$ t-SNE, each with its spread and diagonal fraction. *Bottom:* along the optimization, the structural energy $\|\mathring{\boldsymbol{K}}_Y\|^2$ collapses (by $\sim 7\times$ for the full RV) while the degree floor $\sum_i K_{ii}^2$ stays essentially pinned: the mechanism of Proposition 5 and the justifying Lemma 1.

the framework with a t-SNE output is from reference-tSNE (0.254, 0.362). Plugging in UMAP's kernel does not carry the framework toward UMAP; the per-pair repulsion that defines UMAP is a different functional object, out of reach of an objective whose only reinjectable push is global: the content of Table 1.

### 6.5 The Diagonal: Tether and Release

Section 5 argues that the kernel diagonal is a degree metric (Proposition 5) whose energy acts as a spread-suppressing tether on the full cosine, and that dropping it, in the hollow objective, releases the spread a bounded readout needs. The signature is an energy split: $\|\boldsymbol{K}_Y\|^2 = \|\mathring{\boldsymbol{K}}_Y\|^2 + \sum_i K_{ii}^2$, the structural (off-diagonal) energy plus the degree energy of Proposition 5, acting here as a floor. Figure 1 traces it along the optimization for three embeddings of the same MNIST input ($n = 2000$, perplexity 30, $\gamma = 0.5$, seed 0): the full RV, the hollow RV, and a t-SNE reference.

The split behaves exactly as predicted. The degree floor $\sum_i K_{ii}^2$ is essentially invariant across all three objectives ($4.8, 4.9, 5.0 \times 10^{-4}$ for full-RV, hollow-RV, t-SNE) and barely moves along the optimization: it is a near-constant pedestal in $\|\boldsymbol{K}_Y\|^2$, not a quantity the embedding can spend. The structural energy $\|\mathring{\boldsymbol{K}}_Y\|^2$, by contrast, collapses, from $2.84 \times 10^{-2}$ to $3.84 \times 10^{-3}$ under the full RV, and further to $1.96 \times 10^{-3}$ under the hollow RV. Under the full cosine, maximizing alignment therefore spends the structural energy against a pinned floor, and the embedding stays tethered: spread 12.9, diagonal fraction 0.11. The hollow cosine, blind to that floor, lets the configuration expand toward the t-SNE regime (spread 16.4 then 35.5; diagonal fraction 0.20 then 0.45). The diagonal behaves exactly as the dial that Section 5 predicted.

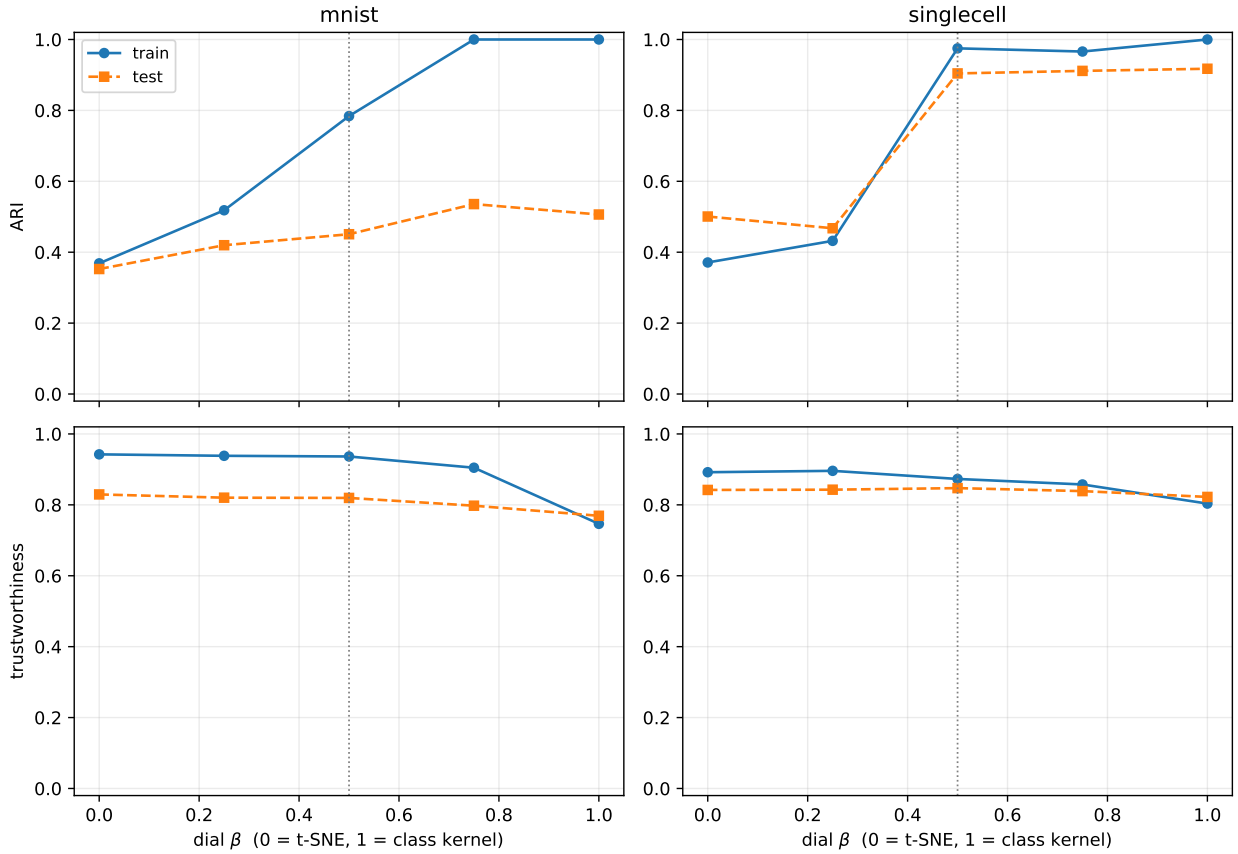


Figure 2: The supervised dial (MNIST, $n = 2000$; single-cell, $n = 2638$; 70/30 split). ARI (top) and trustworthiness (bottom) on the training (solid) and held-out test (dashed) sets, against the dial $\beta$ (0 = t-SNE, 1 = class kernel). Test ARI rises on points whose labels were never used (supervision generalizes) while test trustworthiness holds near its t-SNE level through the intermediate regime; $\beta = 1$ overfits (train ARI $\to 1$).

### 6.6 The Supervised Dial: From t-SNE to Classes

The last study puts to work the supervised dial promised in Section 3.4, in its non-linear form. A single $\beta$ moves both kernels together: the input target $\boldsymbol{K}_{\text{in}}(\beta) = (1-\beta)\boldsymbol{K}_{\text{ag}} + \beta\boldsymbol{K}_Z$ from the neighbor affinity toward the class kernel, and the output $\boldsymbol{K}_{\text{out}}(\beta) = (1-\beta)\,\text{Student-t} + \beta\,\text{linear}$ from the heavy-tailed readout toward the linear Gram (each component unit-Frobenius before mixing), under the hollow objective. At $\beta = 0$ this is an unsupervised t-SNE; at $\beta = 1$ it is the linear soft-LDA endpoint of Section 3.4, a classical scaling of the class centroids. To measure whether supervision *generalizes* rather than merely memorizes, we adopt a strict out-of-sample protocol on the two labeled datasets: a stratified 70/30 train/test split, with $\boldsymbol{K}_Z$ built from the training labels only, the embedding optimized on the training set, and each test point placed out-of-sample from its features alone (a Gaussian-weighted mean of its nearest training neighbors' positions, using no test label). Figure 2 reports ARI and trustworthiness, on both splits, against $\beta$.

Supervision generalizes. On held-out points, whose labels enter neither the kernel nor the placement, the test ARI climbs with $\beta$: from 0.50 to 0.92 on single-cell and from 0.35 to a peak of 0.54 at $\beta = 0.75$ on MNIST, while test trustworthiness holds near its unsupervised t-SNE level ($\approx$ 0.80–0.85) through the intermediate regime. The training ARI reaches 1.000 at $\beta = 1$ on both datasets, the pure class kernel collapsing each training class to a point; the held-out metric, by contrast, exhibits the gap that signals generalization rather than memorization, and does so differently on the two datasets. On MNIST the test ARI peaks at $\beta = 0.75$ and then relents at $\beta = 1$, the textbook overfitting turn; on single-cell it keeps climbing to $\beta = 1$, where

the class geometry is simple enough that the test metric never turns, and the overfitting shows only in the train–test gap (train 1.000 against test 0.92). Either way, the held-out points never collapse onto the class centroids. That the output must move with the input is not incidental: only a linear output can realize the rank-$(m-1)$ centroid geometry of $\boldsymbol{K}_Z$, so that $\beta = 1$ reproduces the closed-form soft-LDA of Section 3.4; holding the output at a bounded Student-t sheet (Proposition 2(iv)) leaves the class structure unrealizable and the supervision fails to transfer. The dial is a continuous supervised interpolation with exact spectral endpoints, a construction the framework supplies by blending kernels, and one with no counterpart in the standard neighbor-embedding libraries.

## 7 Conclusion

We have argued that dimensionality reduction, across its linear and non-linear branches, is a single operation: the projection of an input kernel onto the set of output kernels a low-dimensional configuration can achieve, measured by the RV cosine of the kernel inner-product space. From this one object the structure of the field follows. In the linear regime the achievable set is a convex cone, the projection is a closed-form spectral truncation with an exact alignment ceiling, and PCA, kernel PCA, Isomap, LLE, Laplacian eigenmaps, and diffusion maps are its optima for different input kernels (up to the axis-scaling convention of their reference implementations), with a continuous soft-LDA extending the same projection to supervision. Letting the output kernel bend turns the cone into a smooth manifold, of dimension $nq - \binom{q+1}{2}$ for the distance readout of the neighbor embeddings, on which the RV gradient becomes force-directed and its attraction a close cousin of t-SNE's. Two mechanisms that neighbor-embedding practice treats as implementation choices are, here, consequences of the geometry: the kernel diagonal is a degree metric whose energy tethers the spread, so that dropping it recovers the neighbor-embedding regime; and repulsion is not part of the alignment objective but the gradient of the very volume mode the centering discards, which is why the framework is a relative of t-SNE and cannot be UMAP. Section 6 confirms each of these as a measured prediction, from the exact spectral ceiling to the generalizing supervised dial.

Several limitations mark the boundary of the account, and each is also an opening. Off the cone the optimization is local: the manifold results characterize the achievable set and the gradient, not a global optimum, and the embeddings inherit the non-convexity common to all neighbor methods. The derivation isolates the *push*, while the pull is optimized rather than solved, and the size reward $-\log Z$ grows without bound as the embedding spreads, leaving its strength a genuine tuning parameter. The out-of-sample extension used for the supervised dial is a feature-space heuristic, not a learned map. On the constructive side, the majorization–minimization connection sketched in Section 5.4 invites a fuller development. The framework likewise invites other targets (supervised, semi-supervised, or spatially weighted kernels (Bavaud, 2024; Loup & Bavaud, 2025)) carried, unchanged, by the same projection. What the kernel inner-product space offers is not another method but a map of the ones we have: a single geometry in which their objectives, their solvers, and even their tricks become locations rather than inventions.

## References

P.-A. Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008. doi: 10.1515/9781400830244.

Leonard Asimow and Ben Roth. The rigidity of graphs. *Transactions of the American Mathematical Society*, 245:279–289, 1978. doi: 10.1090/S0002-9947-1978-0511410-9.

Elnaz Barshan, Ali Ghodsi, Zohreh Azimifar, and Mansoor Zolghadri Jahromi. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, 44(7):1357–1371, July 2011. ISSN 00313203. doi: 10.1016/j.patcog.2010.12.015.

François Bavaud. Measuring and Testing Multivariate Spatial Autocorrelation in a Weighted Setting: A Kernel Approach. *Geographical Analysis*, 56(3):573–599, July 2024. ISSN 0016-7363, 1538-4632. doi: 10.1111/gean.12390.

François Bavaud and Guillaume Guex. A Non-Parametric Test of Association and Partial Association Between Dissimilarities. *Submitted*, 2026.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In T. Dietterich, S. Becker, and Z. Ghahramani (eds.), *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

Jan Niklas Böhm, Philipp Berens, and Dmitry Kobak. Attraction-repulsion spectrum in neighbor embeddings. *Journal of Machine Learning Research*, 23(95):1–32, 2022.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023. doi: 10.1017/9781009166164.

Miguel Á. Carreira-Perpiñán. The elastic embedding algorithm for dimensionality reduction. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010.

Fan R. K. Chung. *Spectral Graph Theory*. Number 92 in CBMS Regional Conference Series in Mathematics. American Mathematical Society, Providence, RI, 1997.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, July 2006. ISSN 10635203. doi: 10.1016/j.acha.2006.04.006.

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

Nello Cristianini, John Shawe-Taylor, André Elisseeff, and Jaz Kandola. On Kernel-Target Alignment. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

John P. Cunningham and Zoubin Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *J. Mach. Learn. Res.*, 16(1):2859–2900, January 2015. ISSN 1532-4435.

Sebastian Damrich and Fred A. Hamprecht. On UMAP's true loss function. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Stéphane Dray, Pierre Legendre, and Pedro R. Peres-Neto. Spatial modelling: A comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling*, 196(3–4):483–493, 2006.

Stéphane Dray, Sonia Saïd, and François Débias. Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. *Journal of Vegetation Science*, 19(1):45–56, 2008.

Ian L. Dryden and Kanti V. Mardia. *Statistical Shape Analysis, with Applications in R*. Wiley, Chichester, 2nd edition, 2016.

F. Fouss, A. Pirotte, and M. Saerens. A Novel Way of Computing Similarities between Nodes of a Graph, with Application to Collaborative Recommendation. In *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pp. 550–556, Compiegne, France, 2005. IEEE. ISBN 978-0-7695-2415-3. doi: 10.1109/WI.2005.9.

Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita (eds.), *Algorithmic Learning Theory*, pp. 63–77, Berlin, Heidelberg, 2005. Springer. ISBN 978-3-540-31696-1. doi: 10.1007/11564089_7.

Jihun Ham, Daniel D. Lee, Sebastian Mika, and Bernhard Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Twenty-First International Conference on Machine Learning - ICML '04*, pp. 47, Banff, Alberta, Canada, 2004. ACM Press. doi: 10.1145/1015330.1015417.

Nicholas J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, 103:103–118, May 1988. ISSN 00243795. doi: 10.1016/0024-3795(88)90223-6.

Susan Holmes. Multivariate data analysis: The French way. In *Institute of Mathematical Statistics Collections*, pp. 219–233. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2008. ISBN 978-0-940600-74-4. doi: 10.1214/193940307000000455.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 2nd edition, 2013.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. doi: 10.1007/BF01908075.

David G. Kendall. Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, 16(2):81–121, 1984.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3519–3529. PMLR, May 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Romain Loup and François Bavaud. Spatial autocorrelation of political opinions: A kernel approach. *Journal of Geographical Systems*, April 2025. ISSN 1435-5949. doi: 10.1007/s10109-025-00463-6.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020.

Karl Pearson. Principal components analysis. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 6(2):559, 1901.

P. Robert and Y. Escoufier. A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient. *Applied Statistics*, 25(3):257, 1976. ISSN 00359254. doi: 10.2307/2347233.

Sam T. Roweis and Lawrence K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, December 2000. doi: 10.1126/science.290.5500.2323.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319, July 1998. ISSN 0899-7667, 1530-888X. doi: 10.1162/089976698300017467.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5), October 2013. ISSN 0090-5364. doi: 10.1214/13-AOS1140.

Le Song, Arthur Gretton, Karsten Borgwardt, and Alex Smola. Colored Maximum Variance Unfolding. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Masashi Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of machine learning research*, 8(5), 2007.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, December 2000. doi: 10.1126/science.290.5500.2319.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Jarkko Venna and Samuel Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In *Artificial Neural Networks (ICANN 2001)*, volume 2130 of *Lecture Notes in Computer Science*, pp. 485–491. Springer, 2001. doi: 10.1007/3-540-44668-0_68.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. doi: 10.1186/s13059-017-1382-0.

Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-jiang Zhang, Qiang Yang, and Stephen Lin. Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, January 2007. ISSN 0162-8828, 2160-9292. doi: 10.1109/TPAMI.2007.250598.

Zhirong Yang, Jaakko Peltonen, and Samuel Kaski. Optimization equivalence of divergences improves neighbor embedding. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014.

Zhirong Yang, Jaakko Peltonen, and Samuel Kaski. Majorization-minimization for manifold embedding. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

Daoqiang Zhang, Zhi-Hua Zhou, and Songcan Chen. Semi-Supervised Dimensionality Reduction. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pp. 629–634. Society for Industrial and Applied Mathematics, April 2007. ISBN 978-0-89871-630-6 978-1-61197-277-1. doi: 10.1137/1.9781611972771.73.

## A Proofs

This appendix collects the proofs of the results stated in the main text, in their order of appearance.

*Proof of Theorem 1.* The idea is a two-way inclusion: every linear output is a centered, PSD, rank-$\leq q$ kernel, and conversely every such kernel is realized by reading a configuration off its eigenvectors.

($\subseteq$) For any $\boldsymbol{Y}$, the kernel $\boldsymbol{K}_Y = (\boldsymbol{QY})(\boldsymbol{QY})^\top$ is symmetric and positive semi-definite with $\operatorname{rank}(\boldsymbol{K}_Y) \leq \operatorname{rank}(\boldsymbol{Y}) \leq q$. It is centered, since $\boldsymbol{Q}^\top\sqrt{\boldsymbol{f}} = \boldsymbol{H}^\top\boldsymbol{\Pi}^{1/2}\sqrt{\boldsymbol{f}} = \boldsymbol{H}^\top\boldsymbol{f} = \boldsymbol{f} - \boldsymbol{f}\,(\boldsymbol{1}^\top\boldsymbol{f}) = \boldsymbol{0}$, so that $\boldsymbol{K}_Y\sqrt{\boldsymbol{f}} = \boldsymbol{QYY}^\top(\boldsymbol{Q}^\top\sqrt{\boldsymbol{f}}) = \boldsymbol{0}$. Hence $\boldsymbol{K}_Y \in \mathcal{S}_q$.

($\supseteq$) Conversely, take any $\boldsymbol{K} \in \mathcal{S}_q$ and let $r = \operatorname{rank}(\boldsymbol{K}) \leq q$. Since $\boldsymbol{K}$ is PSD, its $r$ nonzero eigenvalues are strictly positive, and we may write $\boldsymbol{K} = \boldsymbol{U}_r\boldsymbol{\Lambda}_r\boldsymbol{U}_r^\top$, where $\boldsymbol{\Lambda}_r$ collects these positive eigenvalues and the columns of $\boldsymbol{U}_r$ the corresponding orthonormal eigenvectors. Because $\boldsymbol{K}\sqrt{\boldsymbol{f}} = \boldsymbol{0}$, the vector $\sqrt{\boldsymbol{f}}$ is an eigenvector of $\boldsymbol{K}$ with eigenvalue 0; eigenvectors of a symmetric matrix associated with distinct eigenvalues being orthogonal, every column of $\boldsymbol{U}_r$ is orthogonal to it, $\sqrt{\boldsymbol{f}}^\top\boldsymbol{U}_r = \boldsymbol{0}$. (Restricting the decomposition to the *positive* eigenvalues is what secures this orthogonality: a rank-$q$ decomposition with $r < q$ would include null-space eigenvectors, among which the unit vector $\sqrt{\boldsymbol{f}}$ itself could figure.) Set

$$\boldsymbol{Y} = \left[\, \boldsymbol{\Pi}^{-1/2}\boldsymbol{U}_r\boldsymbol{\Lambda}_r^{1/2} \;\middle|\; \boldsymbol{0}_{n\times(q-r)} \,\right] \in \mathbb{R}^{n\times q}, \tag{43}$$

padding with zero columns when $r < q$. The centering acts trivially on the non-zero block: using $\boldsymbol{f}^\top\boldsymbol{\Pi}^{-1/2} = \sqrt{\boldsymbol{f}}^\top$,

$$\boldsymbol{H}\,\boldsymbol{\Pi}^{-1/2}\boldsymbol{U}_r = \boldsymbol{\Pi}^{-1/2}\boldsymbol{U}_r - \boldsymbol{1}_n\,(\sqrt{\boldsymbol{f}}^\top\boldsymbol{U}_r) = \boldsymbol{\Pi}^{-1/2}\boldsymbol{U}_r, \tag{44}$$

so that $\boldsymbol{QY} = \boldsymbol{\Pi}^{1/2}\boldsymbol{H}\,\boldsymbol{Y} = \left[\, \boldsymbol{U}_r\boldsymbol{\Lambda}_r^{1/2} \,\middle|\, \boldsymbol{0} \,\right]$. Hence $\boldsymbol{K}_Y = \boldsymbol{QYY}^\top\boldsymbol{Q}^\top = \boldsymbol{U}_r\boldsymbol{\Lambda}_r\boldsymbol{U}_r^\top = \boldsymbol{K}$, and $\boldsymbol{Y}$ is automatically centered, $\boldsymbol{f}^\top\boldsymbol{Y} = \left[\, \sqrt{\boldsymbol{f}}^\top\boldsymbol{U}_r\boldsymbol{\Lambda}_r^{1/2} \,\middle|\, \boldsymbol{0} \,\right] = \boldsymbol{0}$. □

*Proof of Theorem 2.* The idea is to remove the scale, which turns the cosine maximization into a nearest-point problem on the cone; an eigenvalue comparison then shows that the nearest point is the clipped truncation.

We first justify the hypothesis: for any PSD $\boldsymbol{K}$, $\langle \boldsymbol{K}_X, \boldsymbol{K}\rangle = \sum_i \lambda_i \boldsymbol{u}_i^\top \boldsymbol{K}\boldsymbol{u}_i$ with every $\boldsymbol{u}_i^\top \boldsymbol{K}\boldsymbol{u}_i \geq 0$, so if $\lambda_1 \leq 0$ the alignment is non-positive on all of $\mathcal{S}_q$ and no positive-alignment maximizer exists.

*From cosine to projection.* The objective is scale-invariant in $\boldsymbol{K}_Y$, so we may impose $\|\boldsymbol{K}_Y\| = 1$ and maximize the linear functional $\langle \boldsymbol{K}_X, \boldsymbol{K}_Y\rangle$. For a unit-norm $\boldsymbol{K}$ and $t \geq 0$, the quadratic $\|\boldsymbol{K}_X - t\boldsymbol{K}\|^2 = \|\boldsymbol{K}_X\|^2 - 2t\langle \boldsymbol{K}_X, \boldsymbol{K}\rangle + t^2$ is minimized at $t = \max(\langle \boldsymbol{K}_X, \boldsymbol{K}\rangle, 0)$, so the squared distance from $\boldsymbol{K}_X$ to the ray $\{t\boldsymbol{K} \mid t \geq 0\}$ equals $\|\boldsymbol{K}_X\|^2 - \max(\langle \boldsymbol{K}_X, \boldsymbol{K}\rangle, 0)^2$. The cone being the union of its rays, maximizing the inner product over unit-norm elements and minimizing the Frobenius distance to $\mathcal{S}_q$ select the *same* direction, provided the optimal inner product is positive. It is: $\boldsymbol{u}_1\boldsymbol{u}_1^\top$ is a unit-norm element of $\mathcal{S}_q$ (its eigenvalue $\lambda_1 > 0$ is nonzero, forcing $\boldsymbol{u}_1 \perp \sqrt{\boldsymbol{f}}$, hence centering) and achieves $\langle \boldsymbol{K}_X, \boldsymbol{u}_1\boldsymbol{u}_1^\top\rangle = \lambda_1 > 0$. It remains to project $\boldsymbol{K}_X$ onto the cone.

*The projection is the clipped truncation.* For symmetric matrices with eigenvalues sorted in decreasing order, the Hoffman–Wielandt inequality (see, e.g., Horn & Johnson, 2013) bounds the Frobenius distance from below by the distance between spectra: for any $\boldsymbol{K} \in \mathcal{S}_q$ with eigenvalues $\mu_1 \geq \cdots \geq \mu_n$,

$$\|\boldsymbol{K}_X - \boldsymbol{K}\|^2 \;\geq\; \sum_{i=1}^{n} (\lambda_i - \mu_i)^2. \tag{45}$$

Membership in $\mathcal{S}_q$ constrains this spectrum alone: $\mu_i \geq 0$ for all $i$, and $\mu_i = 0$ for $i > q$. Term by term, the right-hand side is then minimized at $\mu_i = \lambda_i^+$ for $i \leq q$ (an admissible, decreasing choice), so $\|\boldsymbol{K}_X - \boldsymbol{K}\|^2 \geq \sum_{j\leq q}(\lambda_j - \lambda_j^+)^2 + \sum_{j>q}\lambda_j^2$ over the whole cone. The bound is attained by $\boldsymbol{K}_Y^\star = \sum_{j\leq q}\lambda_j^+\, \boldsymbol{u}_j\boldsymbol{u}_j^\top$, which does belong to $\mathcal{S}_q$: it is PSD of rank $\leq q$ by construction, and centered, since every $\boldsymbol{u}_j$ retained with $\lambda_j^+ > 0$ carries a nonzero eigenvalue and is therefore orthogonal to $\sqrt{\boldsymbol{f}}$. Because the bound holds over the entire cone, the optimum is global, and unique whenever the cut is unambiguous ($\lambda_q > \lambda_{q+1}$ or $\lambda_q \leq 0$). The realizing configuration is read off the construction of Theorem 1, $\boldsymbol{Y}^\star = \boldsymbol{\Pi}^{-1/2}\boldsymbol{U}_q(\boldsymbol{\Lambda}_q^+)^{1/2}$, the clipped eigenvalues contributing the zero-padded columns. Retaining the leading eigenpairs of $\boldsymbol{K}_X$ is classical weighted multidimensional scaling; and taking $q = n$ removes the rank constraint, so the same comparison also recovers the PSD projection (9) of Section 2. □

*Proof of Proposition 1.* The idea is a direct substitution: on the orthonormal eigenbasis, the inner product and both norms collapse to sums of squared eigenvalues.

Since the $\boldsymbol{u}_i$ are orthonormal, $\langle \boldsymbol{u}_i\boldsymbol{u}_i^\top, \boldsymbol{u}_j\boldsymbol{u}_j^\top\rangle = (\boldsymbol{u}_i^\top\boldsymbol{u}_j)^2 = \delta_{ij}$, and substituting $\boldsymbol{K}_Y^\star$ into (6) reduces everything to eigenvalues. The inner product becomes $\langle \boldsymbol{K}_X, \boldsymbol{K}_Y^\star\rangle = \sum_{j\leq q}\lambda_j\lambda_j^+ = \sum_{j\leq q}(\lambda_j^+)^2$, the last equality holding term by term ($\lambda_j\lambda_j^+ = \lambda_j^2$ when $\lambda_j > 0$; both sides vanish otherwise). Likewise $\|\boldsymbol{K}_X\|^2 = \sum_i \lambda_i^2$ and $\|\boldsymbol{K}_Y^\star\|^2 = \sum_{j\leq q}(\lambda_j^+)^2$, the latter positive since $\lambda_1^+ = \lambda_1 > 0$, so the cosine is well defined. Hence

$$\mathrm{RV}(\boldsymbol{K}_X, \boldsymbol{K}_Y^\star) = \frac{\sum_{j\leq q}(\lambda_j^+)^2}{\sqrt{\sum_i \lambda_i^2}\,\sqrt{\sum_{j\leq q}(\lambda_j^+)^2}} = \sqrt{\frac{\sum_{j\leq q}(\lambda_j^+)^2}{\sum_i \lambda_i^2}}, \tag{46}$$

and this value is the *optimum* over the achievable set because $\boldsymbol{K}_Y^\star$ is the maximizer of Theorem 2. □

*Proof of Corollary 1.* Immediate from Theorems 1 and 2: the achievable set $\mathcal{S}_q$ does not depend on $\boldsymbol{K}_X$, and the global maximizer over it is the clipped truncation, realized by $\boldsymbol{Y}^\star$. □

Two self-contained facts are used repeatedly; we isolate them.

**Lemma 2** (Kernel of the centering)**.** *On symmetric matrices, the centering $C(\boldsymbol{G}) = \boldsymbol{Q}\boldsymbol{G}\boldsymbol{Q}^\top$ has kernel*

$$\ker C = \{\mathbf{1}\boldsymbol{a}^\top + \boldsymbol{a}\mathbf{1}^\top \mid \boldsymbol{a} \in \mathbb{R}^n\},$$

*whose elements have diagonal entries $2a_i$.*

*Proof.* Since $\mathbf{\Pi}^{1/2}$ is invertible, $C(\boldsymbol{G}) = \mathbf{0}$ reduces to $\boldsymbol{H}\boldsymbol{G}\boldsymbol{H}^\top = \mathbf{0}$. Expanding with $\boldsymbol{H} = \boldsymbol{I}_n - \mathbf{1}\boldsymbol{f}^\top$ gives $\boldsymbol{G} = \mathbf{1}\boldsymbol{b}^\top + \boldsymbol{b}\mathbf{1}^\top - c\,\mathbf{1}\mathbf{1}^\top$ with $\boldsymbol{b} = \boldsymbol{G}\boldsymbol{f}$ and $c = \boldsymbol{f}^\top\boldsymbol{G}\boldsymbol{f}$, i.e. $\boldsymbol{G} = \mathbf{1}\boldsymbol{a}^\top + \boldsymbol{a}\mathbf{1}^\top$ with $\boldsymbol{a} = \boldsymbol{b} - \frac{c}{2}\mathbf{1}$; conversely, $\boldsymbol{H}\mathbf{1} = \mathbf{0}$ kills every matrix of this form. □

**Lemma 3** (Infinitesimal rigidity of the complete graph)**.** *Let $S(\boldsymbol{Y}) = \boldsymbol{D}(\boldsymbol{Y})$ be the squared-distance map, $[\boldsymbol{D}(\boldsymbol{Y})]_{ij} = \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2$. At every affinely spanning configuration,*

$$\ker DS(\boldsymbol{Y}) = \mathcal{I}_{\boldsymbol{Y}}, \qquad \dim \mathcal{I}_{\boldsymbol{Y}} = \binom{q+1}{2}.$$

*Proof.* This is the classical infinitesimal rigidity of the complete framework (Asimow & Roth, 1978); we include the short argument. The inclusion $\supseteq$: for $\dot{\boldsymbol{Y}} = \mathbf{1}\boldsymbol{v}^\top + \boldsymbol{Y}\mathbf{\Omega}^\top$ one has $\dot{\boldsymbol{y}}_i - \dot{\boldsymbol{y}}_j = \mathbf{\Omega}(\boldsymbol{y}_i - \boldsymbol{y}_j)$, and the skew-symmetry of $\mathbf{\Omega}$ kills $DS[\dot{\boldsymbol{Y}}]_{ij} = 2(\boldsymbol{y}_i - \boldsymbol{y}_j)^\top(\dot{\boldsymbol{y}}_i - \dot{\boldsymbol{y}}_j)$. For $\subseteq$, center points and displacements at their weighted means ($\boldsymbol{y}_i^c = \boldsymbol{y}_i - \sum_j f_j\,\boldsymbol{y}_j$, likewise $\dot{\boldsymbol{y}}_i^c$; differences are unchanged), so that $DS[\dot{\boldsymbol{Y}}] = \mathbf{0}$ reads $(\boldsymbol{y}_i^c)^\top\dot{\boldsymbol{y}}_i^c + (\boldsymbol{y}_j^c)^\top\dot{\boldsymbol{y}}_j^c = (\boldsymbol{y}_i^c)^\top\dot{\boldsymbol{y}}_j^c + (\boldsymbol{y}_j^c)^\top\dot{\boldsymbol{y}}_i^c$; averaging with the weights $f_j$ over $j$, then over $i$, forces every $(\boldsymbol{y}_i^c)^\top\dot{\boldsymbol{y}}_i^c = 0$, hence, collecting the centered vectors as columns of $\boldsymbol{Y}_c, \dot{\boldsymbol{Y}}_c \in \mathbb{R}^{q\times n}$, the matrix $\boldsymbol{Y}_c^\top\dot{\boldsymbol{Y}}_c$ is antisymmetric. As $\boldsymbol{Y}_c$ has rank $q$ (the points affinely span), antisymmetry first gives $\ker \boldsymbol{Y}_c \subseteq \ker \dot{\boldsymbol{Y}}_c$, so $\dot{\boldsymbol{Y}}_c = \mathbf{\Omega}\boldsymbol{Y}_c$ for some $\mathbf{\Omega}$, and then forces $\mathbf{\Omega} + \mathbf{\Omega}^\top = \mathbf{0}$: every $\dot{\boldsymbol{y}}_i = \mathbf{\Omega}\boldsymbol{y}_i + \boldsymbol{v}$, i.e. $\dot{\boldsymbol{Y}} \in \mathcal{I}_{\boldsymbol{Y}}$. The parametrization by $(\boldsymbol{v}, \mathbf{\Omega})$ is injective at a spanning configuration, whence $\dim \mathcal{I}_{\boldsymbol{Y}} = q + \binom{q}{2}$. □

*Proof of Proposition 2.* Factor the kernel map through its three layers,

$$\Phi = C \circ R \circ S, \qquad S(\boldsymbol{Y}) = \boldsymbol{Y}\boldsymbol{Y}^\top \text{ (A) or } \boldsymbol{D}(\boldsymbol{Y}) \text{ (B)},$$

where $R$ applies $\kappa_y$ entrywise and $C(\boldsymbol{G}) = \boldsymbol{Q}\boldsymbol{G}\boldsymbol{Q}^\top$ is the linear centering. $DR$, the entrywise (Hadamard, $\circ$) multiplication by $\kappa_y'(S(\boldsymbol{Y}))$, is an isomorphism since $\kappa_y' \neq 0$, and $C$ is linear, so by Lemma 2

$$\dot{\boldsymbol{Y}} \in \ker D\Phi \iff \kappa_y'(S(\boldsymbol{Y})) \circ DS[\dot{\boldsymbol{Y}}] = \mathbf{1}\boldsymbol{a}^\top + \boldsymbol{a}\mathbf{1}^\top \text{ for some } \boldsymbol{a} \in \mathbb{R}^n.$$

*Case (B).* The differentials $DS[\dot{\boldsymbol{Y}}]_{ij} = 2(\boldsymbol{y}_i - \boldsymbol{y}_j)^\top(\dot{\boldsymbol{y}}_i - \dot{\boldsymbol{y}}_j)$ are *hollow* (zero diagonal), and the Hadamard multiplication preserves hollowness; the right-hand side has diagonal $2a_i$, so the match forces $\boldsymbol{a} = \mathbf{0}$, then $DS[\dot{\boldsymbol{Y}}] = \mathbf{0}$: the centering loses no dimension, and Lemma 3 gives

$$\ker D\Phi(\boldsymbol{Y}) = \ker DS(\boldsymbol{Y}) = \mathcal{I}_{\boldsymbol{Y}}$$

at every affinely spanning configuration. Rank–nullity yields $\operatorname{rank} D\Phi = nq - \binom{q+1}{2}$, constant on this open set, and the constant-rank theorem provides the embedded submanifold. This proves (i).

*Case (A), kernel of the Gram map:* $\ker DS = \mathcal{R}_{\boldsymbol{Y}}$. Here $\boldsymbol{Y}$ has full column rank and $DS[\dot{\boldsymbol{Y}}] = \dot{\boldsymbol{Y}}\boldsymbol{Y}^\top + \boldsymbol{Y}\dot{\boldsymbol{Y}}^\top$. For $\mathbf{\Omega} \in \mathfrak{so}(q)$, $DS[\boldsymbol{Y}\mathbf{\Omega}^\top] = \boldsymbol{Y}(\mathbf{\Omega}^\top + \mathbf{\Omega})\boldsymbol{Y}^\top = \mathbf{0}$; conversely, right-multiplying $DS[\dot{\boldsymbol{Y}}] = \mathbf{0}$ by $\boldsymbol{Y}(\boldsymbol{Y}^\top\boldsymbol{Y})^{-1}$ gives $\dot{\boldsymbol{Y}} = \boldsymbol{Y}\boldsymbol{N}$, and substituting back forces $\boldsymbol{N} \in \mathfrak{so}(q)$.

*Case (A), reduction to a linear system.* Define $\boldsymbol{E} := \left[1/\kappa_y'\big(\boldsymbol{y}_i^\top\boldsymbol{y}_j\big)\right]_{ij}$, write $\mathbf{\Delta}_{\boldsymbol{a}} := \operatorname{diag}(\boldsymbol{a})$, and let $\boldsymbol{P}_{\boldsymbol{Y}}^\perp$ be the orthogonal projector onto $\operatorname{col}(\boldsymbol{Y})^\perp$. The full-column-rank configuration $\boldsymbol{Y}$ is *regular*, in the sense of Section 4.2, when the homogeneous linear system

$$\boldsymbol{P}_{\boldsymbol{Y}}^\perp\,(\mathbf{\Delta}_{\boldsymbol{a}}\,\boldsymbol{E} + \boldsymbol{E}\,\mathbf{\Delta}_{\boldsymbol{a}})\,\boldsymbol{P}_{\boldsymbol{Y}}^\perp = \mathbf{0}, \tag{47}$$

overdetermined ($\binom{n-q+1}{2}$ equations, the symmetric block that $\boldsymbol{P}_{\boldsymbol{Y}}^\perp$ retains, in the $n$ unknowns $\boldsymbol{a}$), admits $\boldsymbol{a} = \mathbf{0}$ as its only solution. Since $\boldsymbol{E} \circ (\mathbf{1}\boldsymbol{a}^\top + \boldsymbol{a}\mathbf{1}^\top) = \mathbf{\Delta}_{\boldsymbol{a}}\boldsymbol{E} + \boldsymbol{E}\mathbf{\Delta}_{\boldsymbol{a}}$ entrywise, the kernel condition becomes

$$\dot{\boldsymbol{Y}} \in \ker D\Phi \iff DS[\dot{\boldsymbol{Y}}] = \mathbf{\Delta}_{\boldsymbol{a}}\boldsymbol{E} + \boldsymbol{E}\mathbf{\Delta}_{\boldsymbol{a}} \text{ for some } \boldsymbol{a} \in \mathbb{R}^n,$$

and such an $\boldsymbol{a}$ exists exactly when the right-hand side lies in

$$\operatorname{Im} DS = \{\boldsymbol{S} \text{ symmetric} \mid \boldsymbol{P}_{\boldsymbol{Y}}^\perp\,\boldsymbol{S}\,\boldsymbol{P}_{\boldsymbol{Y}}^\perp = \mathbf{0}\}$$

($\subseteq$ is immediate from $\boldsymbol{P}_{\boldsymbol{Y}}^{\perp}\boldsymbol{Y} = \boldsymbol{0}$; equality holds because both sides have dimension $nq - \binom{q}{2}$, by rank–nullity on the left and block-splitting along $\mathrm{col}(\boldsymbol{Y}) \oplus \mathrm{col}(\boldsymbol{Y})^{\perp}$ on the right). The admissible $\boldsymbol{a}$ therefore form the solution space of (47), henceforth written $\mathcal{A}_{\boldsymbol{Y}} \subseteq \mathbb{R}^n$; each solution contributes one coset of $\ker DS = \mathcal{R}_{\boldsymbol{Y}}$, distinct for distinct $\boldsymbol{a}$ (the diagonal of $\boldsymbol{\Delta_a}\boldsymbol{E} + \boldsymbol{E}\boldsymbol{\Delta_a}$ is $2a_iE_{ii}$, with $E_{ii} \neq 0$), whence the count

$$\dim \ker D\Phi(\boldsymbol{Y}) = \binom{q}{2} + \dim \mathcal{A}_{\boldsymbol{Y}}, \tag{48}$$

so that $\ker D\Phi = \mathcal{R}_{\boldsymbol{Y}}$ exactly at the regular configurations, where $\mathcal{A}_{\boldsymbol{Y}} = \{\boldsymbol{0}\}$.

*Case (A), openness and genericity.* Regularity says that the matrix of the linear map $\boldsymbol{a} \mapsto \boldsymbol{P}_{\boldsymbol{Y}}^{\perp}(\boldsymbol{\Delta_a}\boldsymbol{E} + \boldsymbol{E}\boldsymbol{\Delta_a})\boldsymbol{P}_{\boldsymbol{Y}}^{\perp}$ has full column rank $n$; its entries depend continuously on $\boldsymbol{Y}$, so the condition is open, the rank of $D\Phi$ is constant where it holds, and the constant-rank theorem applies as in case (B). If moreover $\kappa_y$ is real-analytic, these entries are real-analytic on the connected open set of full-column-rank configurations ($q < n$), hence so are the $n \times n$ minors of that matrix: either all vanish identically, or their common zero set, where regularity fails, is a closed null set. This proves (ii).

*Case (A), the affine member.* If $\kappa_y(t) = \alpha t + c$, then $\boldsymbol{E} = \alpha^{-1}\boldsymbol{1}\boldsymbol{1}^{\top}$ and the left-hand side of (47) becomes $\boldsymbol{u}\boldsymbol{w}^{\top} + \boldsymbol{w}\boldsymbol{u}^{\top}$ with $\boldsymbol{u} = \boldsymbol{P}_{\boldsymbol{Y}}^{\perp}\boldsymbol{1}$ and $\boldsymbol{w} = \alpha^{-1}\boldsymbol{P}_{\boldsymbol{Y}}^{\perp}\boldsymbol{a}$. If $\boldsymbol{1} \in \mathrm{col}(\boldsymbol{Y})$, then $\boldsymbol{u} = \boldsymbol{0}$ and every $\boldsymbol{a}$ solves the system; otherwise $\boldsymbol{u} \neq \boldsymbol{0}$ and $\boldsymbol{u}\boldsymbol{w}^{\top} + \boldsymbol{w}\boldsymbol{u}^{\top} = \boldsymbol{0}$ forces $\boldsymbol{w} = \boldsymbol{0}$ (applied to $\boldsymbol{u}$ it gives $\boldsymbol{w} = \lambda\boldsymbol{u}$, then $2\lambda\,\boldsymbol{u}\boldsymbol{u}^{\top} = \boldsymbol{0}$), so $\mathcal{A}_{\boldsymbol{Y}} = \{\boldsymbol{a} \mid \boldsymbol{P}_{\boldsymbol{Y}}^{\perp}\boldsymbol{a} = \boldsymbol{0}\} = \mathrm{col}(\boldsymbol{Y})$, of dimension $q$. In either case no configuration is regular, and by (48) the rank of $D\Phi$ drops to $nq - \binom{q}{2} - q = nq - \binom{q+1}{2}$. These solutions are the translations: $DS[\boldsymbol{1}\boldsymbol{v}^{\top}] = \boldsymbol{1}(\boldsymbol{Y}\boldsymbol{v})^{\top} + (\boldsymbol{Y}\boldsymbol{v})\boldsymbol{1}^{\top} = \boldsymbol{\Delta_a}\boldsymbol{E} + \boldsymbol{E}\boldsymbol{\Delta_a}$ with $\boldsymbol{a} = \alpha\,\boldsymbol{Y}\boldsymbol{v}$. Finally, $\Phi(\boldsymbol{Y}) = \boldsymbol{Q}(\alpha\,\boldsymbol{Y}\boldsymbol{Y}^{\top} + c\,\boldsymbol{1}\boldsymbol{1}^{\top})\boldsymbol{Q}^{\top} = \alpha\,\boldsymbol{Q}\boldsymbol{Y}\boldsymbol{Y}^{\top}\boldsymbol{Q}^{\top}$ since $\boldsymbol{Q}\boldsymbol{1} = \boldsymbol{0}$: the achievable set is $\alpha\,\mathcal{S}_q$. This proves (iii).

For (iv), boundedness of $\kappa_y$ gives $\|\Phi(\boldsymbol{Y})\| \leq \|\boldsymbol{Q}\|_{\mathrm{op}}^2\, n \sup_t |\kappa_y(t)|$ in Frobenius norm: the achievable set is bounded, for either readout. It is not reduced to $\{\boldsymbol{0}\}$. In case (B), for a configuration with two distinct points, $\Phi(\boldsymbol{Y}) = \boldsymbol{0}$ would put $\kappa_y(\boldsymbol{D})$ in the additive form of Lemma 2, forcing $\kappa_y(D_{12}) = \frac{1}{2}\left(\kappa_y(D_{11}) + \kappa_y(D_{22})\right) = \kappa_y(0)$: impossible, since $\kappa_y' \neq 0$ makes $\kappa_y$ strictly monotone and $D_{12} > 0$. In case (A), take $\boldsymbol{y}_1 = \boldsymbol{w} \neq \boldsymbol{0}$ and $\boldsymbol{y}_i = \boldsymbol{0}$ for $i \geq 2$: the additive form forces every $a_i = \kappa_y(0)/2$, and the $(1,1)$ entry then reads $\kappa_y(\|\boldsymbol{w}\|^2) = \kappa_y(0)$, the same contradiction. A bounded set containing a nonzero element cannot be a cone, a nonzero element generating an unbounded ray. □

*Proof of Proposition 3.* Both cases are the chain rule applied through $\boldsymbol{\Gamma} = \partial\mathrm{RV}/\partial\boldsymbol{G}_Y^{\kappa_y}$:

$$d\mathrm{RV} = \sum_{ij} \Gamma_{ij}\, \kappa_y'\big([\boldsymbol{G}]_{ij}\big)\, d[\boldsymbol{G}]_{ij},$$

where $\boldsymbol{G}$ stands for the readout's argument and both $\boldsymbol{\Gamma}$ and the entrywise weights are symmetric. In case (A), $\partial(\boldsymbol{y}_i^{\top}\boldsymbol{y}_j)/\partial\boldsymbol{y}_k = \delta_{ki}\,\boldsymbol{y}_j + \delta_{kj}\,\boldsymbol{y}_i$, so the point $\boldsymbol{y}_k$ receives two contributions, from $i = k$ and from $j = k$, which the symmetry makes equal: $\partial\mathrm{RV}/\partial\boldsymbol{y}_k = 2\sum_j \Gamma_{kj}\,\kappa_y'(\boldsymbol{y}_k^{\top}\boldsymbol{y}_j)\,\boldsymbol{y}_j$, the row form of (31). In case (B), $\partial\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2/\partial\boldsymbol{y}_k = 2(\delta_{ki} - \delta_{kj})(\boldsymbol{y}_i - \boldsymbol{y}_j)$ carries its own factor 2, and the same symmetric pairing doubles it: $2 \times 2 = 4$, the force-directed form (32). □

*Proof of Proposition 4.* The hypothesis that Proposition 2 yields a local sheet at $\boldsymbol{Y}$ means: an affinely spanning configuration for the distance readout (item (i)), a regular one for the dot-product readout (item (ii)). In the distance case $\boldsymbol{K}_Y \neq \boldsymbol{0}$ is automatic: an affinely spanning configuration has two distinct points, and the argument closing the proof of Proposition 2(iv) applies.

That the parametric gradient is the pullback is the chain rule,

$$d\mathrm{RV} = \langle \nabla_{\boldsymbol{K}_Y}\mathrm{RV},\, D\Phi[\dot{\boldsymbol{Y}}]\rangle = \langle D\Phi^{*}\nabla_{\boldsymbol{K}_Y}\mathrm{RV},\, \dot{\boldsymbol{Y}}\rangle,$$

which is precisely the two-level computation of Section 4.3. (i) is the definition of the Riemannian gradient of a function restricted to an embedded submanifold, here the local sheet, whose tangent space at $\boldsymbol{K}_Y$ is $T$; the ambient RV is smooth around $\boldsymbol{K}_Y$, which is nonzero by hypothesis. For (ii), note from (24) that $\nabla_{\boldsymbol{K}_Y}\mathrm{RV}$ is a *positive* multiple of the residual $\boldsymbol{R}$, so

$$D\Phi^{*}\boldsymbol{R} = 0 \iff \boldsymbol{R} \perp \mathrm{Im}\, D\Phi = T \iff P_T\,\boldsymbol{R} = 0.$$

Finally, $\operatorname{Im} D\Phi^* = (\ker D\Phi)^\perp$ ($\mathcal{I}_{\dot{\boldsymbol{Y}}}^\perp$ or $\mathcal{R}_{\dot{\boldsymbol{Y}}}^\perp$, by family) is the horizontal space of the text, and on it $\langle \dot{\boldsymbol{Y}}, \boldsymbol{B}_{\boldsymbol{Y}} \dot{\boldsymbol{Y}} \rangle = \|D\Phi[\dot{\boldsymbol{Y}}]\|^2 > 0$: the preconditioner $\boldsymbol{B}_{\boldsymbol{Y}}$ is positive definite there, as claimed. □

*Proof of Corollary 2.* Case (B) of Proposition 3 is pure chain rule: it holds for any scalar function of $\boldsymbol{G}_Y^{\kappa_y}$, with $\boldsymbol{\Gamma}$ replaced by the corresponding structural weight. For the numerator $F = \langle \boldsymbol{K}_X, \boldsymbol{K}_Y \rangle$, one has $\partial F / \partial \boldsymbol{K}_Y = \boldsymbol{K}_X$, so the identification $\partial F / \partial \boldsymbol{G}_Y^{\kappa_y} = \boldsymbol{Q}^\top (\partial F / \partial \boldsymbol{K}_Y) \boldsymbol{Q}$ gives the weight $\tilde{\boldsymbol{K}}_X = \boldsymbol{Q}^\top \boldsymbol{K}_X \boldsymbol{Q}$. The Student-t derivative is $\kappa_y'(t) = -(1+t)^{-2}$, equal to $-\tilde{q}_{kj}^2$ at $t = \|\boldsymbol{y}_k - \boldsymbol{y}_j\|^2$; the minus sign reverses $(\boldsymbol{y}_k - \boldsymbol{y}_j)$ into $(\boldsymbol{y}_j - \boldsymbol{y}_k)$, giving (35). □

*Proof of Proposition 5.* Diagonal and hollow matrices are Frobenius-orthogonal, so the product splits as

$$\langle \boldsymbol{K}_X, \boldsymbol{K}_Y \rangle = \langle \mathring{\boldsymbol{K}}_X, \mathring{\boldsymbol{K}}_Y \rangle + \langle \mathcal{D}\mathring{\boldsymbol{K}}_X, \mathcal{D}\mathring{\boldsymbol{K}}_Y \rangle;$$

the middle expression of (37) spells out $(\mathcal{D}\mathring{\boldsymbol{K}})_i = -r_i / \sqrt{f_i}$, and the right one is the definition of the adjoint. $\boldsymbol{M}$ is self-adjoint by construction; $\langle \mathring{\boldsymbol{K}}, \boldsymbol{M}\, \mathring{\boldsymbol{K}} \rangle = \|\mathring{\boldsymbol{K}}\|^2 + \|\mathcal{D}\mathring{\boldsymbol{K}}\|^2$ gives both $\boldsymbol{M} \succeq \boldsymbol{I}$ and the norm identity, and (38) follows by substituting into (6). □

*Proof of Lemma 1. (Linear.)* The achievable set is the cone $\mathcal{S}_q$ (Theorem 1); maximizing the scale-invariant cosine over it selects the direction of the Frobenius projection of $\boldsymbol{K}_X$, solved globally and in closed form by the clipped truncation (Theorem 2); and by Proposition 5, this Frobenius projection *is* the $\boldsymbol{M}$-projection of the affinity pattern. The surcharge is moreover bona fide signal: the degree energy $\sum_i K_{ii}^2$ is the energy of the radii, which the truncation matches along with everything else.

*(Distance-based.)* The raw diagonal is frozen at $\kappa_y(0)$. The achievable set is now the bounded manifold of Proposition 2, not a cone: the ray argument of Theorem 2 requires a set closed under scaling, which Proposition 2(iv) rules out, so the maximization is no longer a metric projection; and no spectral truncation solves it: the maximizer is characterized only by stationarity (Proposition 4) and found iteratively, possibly among local optima. Both requirements of Definition 1 fail. □

*Proof of Proposition 6.* (a) One has $\boldsymbol{Q}\mathbf{1} = \boldsymbol{\Pi}^{1/2}\boldsymbol{H}\mathbf{1} = \boldsymbol{\Pi}^{1/2}(\mathbf{1} - \mathbf{1}\,\boldsymbol{f}^\top \mathbf{1}) = \mathbf{0}$, so the shift adds $\alpha\,(\boldsymbol{Q}\mathbf{1})(\boldsymbol{Q}\mathbf{1})^\top = \mathbf{0}$ to $\boldsymbol{K}_Y$, and $\alpha\, n(n-1)$ to the off-diagonal mass. The loss is not accidental: on the kernel of the centering, the symmetric matrices $\mathbf{1}\boldsymbol{a}^\top + \boldsymbol{a}\mathbf{1}^\top$ (Lemma 2), the mass $Z = 2(n-1)\,\mathbf{1}^\top \boldsymbol{a}$ is non-constant: the volume lives exactly in what the centering throws away.

(b) From $\partial \tilde{q}_{kl} / \partial \boldsymbol{y}_i = -2\,\tilde{q}_{kl}^2\,(\delta_{ik} - \delta_{il})(\boldsymbol{y}_k - \boldsymbol{y}_l)$, the two ordered occurrences of $i$ in the sum (40) give $\partial Z / \partial \boldsymbol{y}_i = -4 \sum_j \tilde{q}_{ij}^2 (\boldsymbol{y}_i - \boldsymbol{y}_j)$, whence (41). The identification with t-SNE is a decomposition of its objective: writing its output probabilities $q_{ij} = \tilde{q}_{ij}/Z$, the Kullback–Leibler divergence of van der Maaten & Hinton (2008) splits as

$$\text{const} \;-\; \sum_{i \neq j} p_{ij} \log \tilde{q}_{ij} \;+\; \log Z,$$

a cross-entropy attraction plus the volume penalty at $\tau = 1$, so the repulsive part of its gradient is (41).

(c) The map $\alpha \mapsto \boldsymbol{Q}(\boldsymbol{G}_Y + \alpha \mathbf{1}\mathbf{1}^\top)\boldsymbol{Q}^\top$ is constant by (a). □

## B The Optimization Algorithm

The structural and parametric gradients of Section 4.3 yield the following iterative optimization scheme (Algorithm 1).

The computational complexity of the proposed algorithm involves element-wise operations and matrix products on $n \times n$ kernels, scaling with $\mathcal{O}(n^2)$ per iteration. While scaling this exact $\mathcal{O}(n^2)$ dense formulation to massive datasets would require approximation techniques (e.g., Barnes-Hut or Nyström methods), our current focus is theoretical unification. To ensure scalability and efficiency within the current scope, we implemented the framework in the *PyTorch* library, using GPU acceleration to parallelize the tensor operations. A significant advantage of this implementation is the use of *automatic differentiation* (Autograd).

**Algorithm 1** RV dimensionality reduction

**Inputs:** The target $\boldsymbol{K}_X$, the weights $\boldsymbol{f}$, dimension $q$, the output kernel function $\kappa_y(\boldsymbol{y}_i, \boldsymbol{y}_j)$, the learning rate $\eta$, the convergence threshold $\epsilon$.

**Outputs:** The $n \times q$ matrix of centered coordinates $\boldsymbol{Y}$, the final RV coefficient.

Compute $\boldsymbol{Q} \leftarrow \boldsymbol{\Pi}^{1/2}(\boldsymbol{I}_n - \boldsymbol{1}_n \boldsymbol{f}^\top)$, with $\boldsymbol{\Pi} = \mathrm{diag}(\boldsymbol{f})$.
Compute $\|\boldsymbol{K}_X\| \leftarrow \sqrt{\mathrm{Tr}(\boldsymbol{K}_X^2)}$
Set the initial $\boldsymbol{Y}^{(0)}$ (randomly or with PCA).
Set RV $\leftarrow 0$.
**repeat**
  Set $\mathrm{RV}^{\mathrm{old}} \leftarrow \mathrm{RV}$.
  Compute the output kernel $\boldsymbol{K}_Y \leftarrow \boldsymbol{Q}\, \boldsymbol{G}_Y^{\kappa_y} \boldsymbol{Q}^\top$, with $[\boldsymbol{G}_Y^{\kappa_y}]_{ij} = \kappa_y(\boldsymbol{y}_i, \boldsymbol{y}_j)$.
  Compute $\|\boldsymbol{K}_Y\| \leftarrow \sqrt{\mathrm{Tr}(\boldsymbol{K}_Y^2)}$.
  Compute $\mathrm{RV} \leftarrow \frac{\mathrm{Tr}(\boldsymbol{K}_X \boldsymbol{K}_Y)}{\|\boldsymbol{K}_X\| \|\boldsymbol{K}_Y\|}$.
  Compute $\boldsymbol{\Gamma} \leftarrow \frac{1}{\|\boldsymbol{K}_X\| \cdot \|\boldsymbol{K}_Y\|} \boldsymbol{Q}^\top \left( \boldsymbol{K}_X - \mathrm{RV} \frac{\|\boldsymbol{K}_X\|}{\|\boldsymbol{K}_Y\|} \boldsymbol{K}_Y \right) \boldsymbol{Q}$.
  **if** the output kernel is based on dot-product $\kappa_y(\boldsymbol{y}_i^\top \boldsymbol{y}_j)$ **then**
    $\boldsymbol{\Delta} \leftarrow 2 \left( \boldsymbol{\Gamma} \circ \kappa_y'(\boldsymbol{Y}\boldsymbol{Y}^\top) \right) \boldsymbol{Y}$
  **else if** the output kernel is based on distances $\kappa_y(\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2)$ **then**
    $\boldsymbol{\Delta}_i \leftarrow 4 \sum_j \Gamma_{ij} \kappa_y'(\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2)(\boldsymbol{y}_i - \boldsymbol{y}_j)$
  **end if**
  $\boldsymbol{Y}^{(t+1)} \leftarrow \boldsymbol{Y}^{(t)} + \eta \boldsymbol{\Delta}$
**until** $|\mathrm{RV} - \mathrm{RV}^{\mathrm{old}}| < \epsilon$
**return** $\boldsymbol{Y}$, RV

By defining the RV coefficient purely as a forward-pass objective function, we delegate the backward pass to the computational graph engine. This allows us to avoid the explicit manual construction of the gradient matrices derived in Section 4.3, ensuring both numerical precision and implementation simplicity.