A Variational Manifold Embedding Framework for Nonlinear Dimensionality Reduction

John J. Vastola JOHN_VASTOLA@HMS.HARVARD.EDU
Samuel J. Gershman GERSHMAN@FAS.HARVARD.EDU
Kanaka Rajan KANAKA_RAJAN@HMS.HARVARD.EDU
Harvard University, Cambridge, MA, USA

Editors: List of editors' names

Abstract

Dimensionality reduction algorithms like principal component analysis (PCA) are workhorses of machine learning and neuroscience, but each has well-known limitations. Variants of PCA are simple and interpretable, but not flexible enough to capture nonlinear data manifold structure. More flexible approaches have other problems: autoencoders are generally difficult to interpret, and graph-embedding-based methods can produce pathological distortions in manifold geometry. Motivated by these shortcomings, we propose a variational framework that casts dimensionality reduction algorithms as solutions to an optimal manifold embedding problem. By construction, this framework permits nonlinear embeddings, allowing its solutions to be more flexible than PCA. Moreover, the variational nature of the framework has useful consequences for interpretability: each solution satisfies a set of partial differential equations, and can be shown to reflect symmetries of the embedding objective. We discuss these features in detail and show that solutions can be analytically characterized in some cases. Interestingly, one special case exactly recovers PCA.

Keywords: dimensionality reduction, manifold embedding, principal component analysis

1. Introduction

Dimensionality reduction—the problem of identifying useful low-dimensional representations of high-dimensional data—is a central concern across machine learning, statistics, and neuroscience. But existing approaches have issues with accuracy, interpretability, or both. For example, principal component analysis (PCA) (Pearson, 1901; Hotelling, 1933; Bishop, 2006) is computationally cheap and can be interpreted in terms of a simple generative model (Tipping and Bishop, 1999), but cannot capture nonlinear structure in data.

Two complementary perspectives have motivated lines of work that go beyond PCA. The *geometric* perspective on dimensionality reduction casts the problem as identifying the low-dimensional structure (or *data manifold*) along which data tend to lie. Methods like Locally Linear Embedding (Roweis and Saul, 2000), Isomap (Tenenbaum et al., 2000), and diffusion maps (Coifman and Lafon, 2006) exploit this perspective to learn a low-dimensional data manifold embedded in the high-dimensional ambient space. Work on so-called neural manifolds, which seeks low-dimensional representations of neural data, also tends to take this perspective (Cunningham and Yu, 2014; Chung and Abbott, 2021; Perich et al., 2025).

The *probabilistic generative modeling* perspective on dimensionality reduction casts the problem as minimizing a divergence between the data distribution (or a suitably transformed version of it) and a distribution on a low-dimensional 'latent' space. Neural-network-based

© 2025 J.J. Vastola, S.J. Gershman & K. Rajan.

autoencoders (LeCun, 1987; Hinton and Zemel, 1993) and variational autoencoders (Kingma and Welling, 2013; Higgins et al., 2017) exploit this perspective in order to capture arbitrarily complex (smooth) structure in data, but may yield results that are difficult to interpret. Methods which involve embedding samples into a graph, like Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) and t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton, 2008), are nonlinear and somewhat easier to interpret, but have been observed to produce pathologies in downstream statistical and clustering analyses (Wattenberg et al., 2016; Chari and Pachter, 2023).

Motivated by these two perspectives, here we present a novel, nonlinear framework for dimensionality reduction that (i) involves minimizing a divergence between latent space distributions, like variational autoencoders; and (ii) constrains that optimization problem by making geometry-related *nonparametric* assumptions about the mapping between latent and ambient space. We show that this framework has a variety of reasonable features, is amenable to mathematical analysis, and includes PCA as a special case. Critically, the variational structure of the problem permits us to use tools from physics to reason about properties of optimal embeddings, including that symmetries of the data distribution constrain them, and that they satisfy a certain set of partial differential equations (PDEs).

2. A tractable variational framework for dimensionality reduction

When are two data manifolds the same? Let $\boldsymbol{x} \in \mathbb{R}^D$ denote an observation in a D-dimensional space, and $p_{\text{data}}(\boldsymbol{x})$ the associated data likelihood. When is p_{data} more or less identical to some other likelihood q? Or, in more explicitly geometric language, when are the 'data manifolds' associated with these likelihoods equivalent? One reasonable notion of equivalence comes from demanding that q and p_{data} can be obtained from one another via a smooth bijection (i.e., a *diffeomorphism*) $\boldsymbol{g} : \mathbb{R}^D \to \mathbb{R}^D$. In particular, we ask that

$$q(\boldsymbol{x})\ d\boldsymbol{x} = p_{\text{data}}(\boldsymbol{g}(\boldsymbol{x}))\ |\det \boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{x})|\ d\boldsymbol{x} \tag{1}$$

almost everywhere, where $\boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{x})$ is the $D \times D$ Jacobian matrix whose entries are $J_{ij}(\boldsymbol{x}) := \partial g_i(\boldsymbol{x})/\partial x_j$. We will call the distribution $(\boldsymbol{g}^* p_{\text{data}})(\boldsymbol{x}) := p_{\text{data}}(\boldsymbol{g}(\boldsymbol{x}))\ |\det \boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{x})|$ the *pullback* of p_{data} by $\boldsymbol{g}$. Intuitively, the above equation says that one can obtain p_{data} from q by a smooth deformation of the state space $\mathbb{R}^D$.

Dimensionality reduction involves attempting to capture *most* of the structure of p_{data} by relating it to a distribution defined on a generally lower-dimensional space $\mathbb{R}^d$ for some $d \leq D$. Let $\boldsymbol{z} \in \mathbb{R}^d$ denote an element of this 'latent' space. Since it is usually impossible to capture *all* high-dimensional structure of p_{data} via a lower-dimensional approximation, we can only ask that $q(\boldsymbol{z})d\boldsymbol{z}$ is *approximately* the same as the pullback of p_{data} through a smooth embedding map $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^D$ (which we denote $\boldsymbol{\phi}^* p_{\text{data}}$). More precisely, we want

$$q(\boldsymbol{z})\ d\boldsymbol{z} \approx p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z}))\sqrt{\det(\boldsymbol{J}(\boldsymbol{z})^T \boldsymbol{J}(\boldsymbol{z}))}\ d\boldsymbol{z} \tag{2}$$

almost everywhere, where $\boldsymbol{J}(\boldsymbol{z})$ is the $D \times d$ Jacobian with entries $J_{ij}(\boldsymbol{z}) := \partial \phi_i / \partial z_j$. Notice that, since we may have $d < D$, the change of variables correction involves the Riemannian volume form $\sqrt{\det(\boldsymbol{J}(\boldsymbol{z})^T \boldsymbol{J}(\boldsymbol{z}))}$ (we cannot use $|\det \boldsymbol{J}(\boldsymbol{z})|$, since it may equal zero).

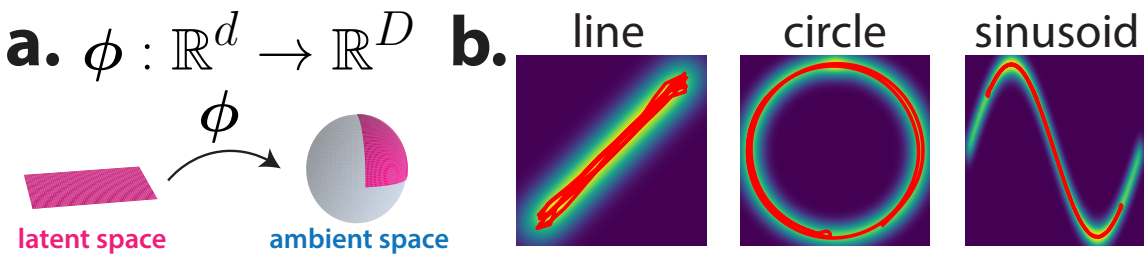

Figure 1: **a.** The function $\boldsymbol{\phi}$ maps the latent space to the (higher-dimensional) ambient space. **b.** Optimal one-dimensional embeddings when p_{data} is a mixture of Gaussians centered at points along a line (left), circle (middle), or sinusoid (right).

A variational objective for dimensionality reduction. Let $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^D$ be a smooth injection from a 'latent' space to an 'ambient' space (with $d \leq D$), and suppose that $\boldsymbol{z} \in \mathbb{R}^d$ is distributed according to a *prior* $q(\boldsymbol{z})$, and $\boldsymbol{x} \in \mathbb{R}^D$ is distributed according to a likelihood p_{data}. Motivated by Eq. 2, we define an *optimal dimensionality reduction algorithm* (or *optimal embedding*) as a map $\boldsymbol{\phi}$ which *maximizes*

$$\boxed{J[\boldsymbol{\phi}] = \int_{\mathbb{R}^d} d\boldsymbol{z}\ q(\boldsymbol{z}) \left[\frac{1}{2} \log \det(\boldsymbol{J}(\boldsymbol{z})^T \boldsymbol{J}(\boldsymbol{z})) + \log p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z})) - \log q(\boldsymbol{z}) \right]}\ , \tag{3}$$

i.e., the negative of the Kullback-Leibler (KL) divergence between q and $\boldsymbol{\phi}^* p_{\text{data}}$, while satisfying the finiteness conditions (see Appendix A)

$$\mathbb{E}_{\boldsymbol{z}\sim q} \left\{ \sqrt{\det(\boldsymbol{J}(\boldsymbol{z})^T \boldsymbol{J}(\boldsymbol{z}))} \right\} < \infty \qquad \text{and} \qquad \mathbb{E}_{\boldsymbol{z}\sim q} \left\{ \|\boldsymbol{\phi}(\boldsymbol{z})\|_2^2 \right\} < \infty\ . \tag{4}$$

Because $J = -D_{\text{KL}}(q \| \boldsymbol{\phi}^* p_{\text{data}})$, we immediately know that J is upper bounded by zero,[1] which corresponds to perfect performance. See Figure 1a for a schematic.

This objective has two terms, each of which perform slightly different roles. The log-likelihood term encourages $\boldsymbol{\phi}$ to map into regions of high likelihood, while the log-determinant term encourages embeddings to have a nontrivial volume. (If only the log-likelihood term were present, an optimal embedding may map all $\boldsymbol{z}$ near a global maximum of p_{data}.) Interestingly, these terms also have clear physical analogues: the log-likelihood term functions like a *potential energy*, and the log-determinant term, which depends on derivatives of the embedding, functions like a *kinetic energy*.

As in physics, the fact that the objective depends not just on the embedding coordinates $\phi_i(\boldsymbol{z})$, but also on the derivatives $J_{ij} := \partial \phi_i / \partial z_j$, allows us to directly apply results from the calculus of variations (Goldstein, 1980; Elsgolc, 2007) that cannot be usefully leveraged in more generic (e.g., autoencoder) settings. For more discussion of this objective, including a Bayesian interpretation of it that casts optimizing J as identifying the maximum a posteriori estimate of $\boldsymbol{\phi}$, see Appendix A.

1. Note that this is only true if $\boldsymbol{\phi}$ is an injection. See Appendix A.

Examples of optimal one-dimensional embeddings. What do optimal embeddings look like? The naive way to estimate them is to parameterize $\boldsymbol{\phi}$ and numerically optimize J via a gradient-descent-like algorithm. If we parameterize $\boldsymbol{\phi}$ as a multilayer perceptron and choose p_{data} to be different Gaussian mixtures, optimizing J in the standard way tells us that optimal *one-dimensional* embeddings are generally nonlinear, and correspond to what one might expect: if there is an obvious path through the points, the optimal embedding follows it (Figure 1b; code: https://github.com/john-vastola/varembed-neurreps25).

3. Interpreting optimal one-dimensional embeddings

We can say more about optimal one-dimensional embeddings, and optimal embeddings more generally, by exploiting results from the calculus of variations. The most basic such result is that optimal embeddings satisfy the *Euler-Lagrange* (EL) *equations*, which we define next.

3.1. Optimal one-dimensional embeddings follow score vectors

If we approximate p_{data} with a one-dimensional manifold ($d = 1$), the objective becomes

$$J[\boldsymbol{\phi}] = \int_{-\infty}^{\infty} dz \; q(z) \left[\frac{1}{2} \log \|\boldsymbol{\phi}'(z)\|_2^2 + \log p_{\text{data}}(\boldsymbol{\phi}(z)) - \log q(z) \right] \tag{5}$$

where $\boldsymbol{\phi}'(z) := (\phi_1'(z), ..., \phi_D'(z))^T$ is the derivative of the embedding. This version of the objective looks extremely similar to a classical mechanics problem, with the main qualitative difference being that the kinetic term is $\frac{1}{2} \log \|\boldsymbol{\phi}'(z)\|_2^2$ instead of $\frac{1}{2} \|\boldsymbol{\phi}'(z)\|_2^2$.

We can view $\boldsymbol{\phi}$ as tracing out a one-dimensional trajectory through $\mathbb{R}^D$. According to a standard result in the calculus of variations, these trajectories satisfy the EL equations

$$\frac{d}{dz} \left(\frac{\partial L}{\partial \phi_i'} \right) = \frac{\partial L}{\partial \phi_i} \quad \text{, i.e.,} \quad \frac{d}{dz} \left[q(z) \frac{\phi_i'(z)}{\|\boldsymbol{\phi}'(z)\|_2^2} \right] = q(z) \frac{\partial}{\partial \phi_i} \log p_{\text{data}}(\boldsymbol{\phi}(z)) \tag{6}$$

for each $i = 1, ..., D$ (Goldstein, 1980; Elsgolc, 2007), where the *Lagrangian* L is defined as

$$L(\boldsymbol{\phi}, \boldsymbol{\phi}', z) := q(z) \left[\frac{1}{2} \log \|\boldsymbol{\phi}'(z)\|_2^2 + \log p_{\text{data}}(\boldsymbol{\phi}(z)) - \log q(z) \right] . \tag{7}$$

These equations place nontrivial constraints on the form of optimal embeddings; satisfying them is necessary, but not sufficient, for $\boldsymbol{\phi}$ to be a global maximizer of J. If we define momentum-like variables $p_i(z) := \phi_i'(z) / \|\boldsymbol{\phi}'(z)\|_2^2$, we find that the 'dynamics' defined by the EL equations looks like that of a massive particle moving through the landscape determined by the potential $-\log p_{\text{data}}$:

$$\phi_i'(z) = \frac{p_i(z)}{\|\boldsymbol{p}(z)\|_2^2} \qquad p_i'(z) = s_i(\boldsymbol{\phi}(z)) - \left[\frac{d}{dz} \log q(z) \right] p_i(z) \tag{8}$$

where $s_i(\boldsymbol{\phi}(z)) := \frac{\partial}{\partial \phi_i} \log p_{\text{data}}(\boldsymbol{\phi}(z))$ is the *score vector* associated with p_{data} at the point $\boldsymbol{\phi}(z) \in \mathbb{R}^D$. These vectors, which play a foundational role in how diffusion models sample (Song et al., 2021) and generalize (Wang and Vastola, 2024; Vastola, 2025a), point in the direction in which the data likelihood increases most. Eq. 8 says that the trajectory $\boldsymbol{\phi}(z)$

moves in the direction of the score $\boldsymbol{s}(\boldsymbol{\phi}(\boldsymbol{z}))$, and that the influence of the score vector is controlled by a q-dependent 'friction' factor and the norm of $\boldsymbol{p}(z)$. The norm $\|\boldsymbol{p}(z)\|_2^2$ plays the role of a 'mass'. When $\boldsymbol{p}(z)$ is large (which can happen when the score vectors become large), the trajectory changes very slowly; when it is small, the trajectory changes quickly.

3.2. Connection to diffusion models

We can make the connection between optimal one-dimensional embeddings and diffusion models more concrete by observing that one significant difference between Eq. 8 and a probability flow ODE (Song et al., 2021) is that the former involves *inertia*: $\boldsymbol{\phi}(z)$ does not *only* move in the direction of the score vector. Since the amount of inertia that appears in Eq. 8 depends on the prior q, we can modulate it by choosing a special prior. If we pick $q(z) = \gamma e^{\gamma z}$ for $z \in (-\infty, 0]$, in the large γ limit we obtain dynamics (see Appendix C)

$$\phi_i'(z) = \gamma \frac{s_i(\boldsymbol{\phi}(z))}{\|\boldsymbol{s}(\boldsymbol{\phi}(z))\|_2^2} \tag{9}$$

which follows score vectors, but moves faster or slower depending on their magnitude. If we reparameterize 'time', we recover an ODE that follows score vectors without the norm-dependent distortion, and hence looks like a probability flow ODE at a fixed noise scale.

4. Useful properties of the variational objective

What other properties of optimal embeddings can we reason about without actually optimizing the objective? In this section, we discuss two kinds: properties related to *flexibility*, and properties related to *interpretability*. Unlike in the previous section, here our comments apply for any choice of latent dimensionality (i.e., any $d \leq D$).

4.1. Flexibility-related properties

Objective is geometry-sensitive but not coordinate-sensitive. Our framework ought to reflect the geometry of the data manifold and prior, but not our incidental choice of coordinates;[2] in particular, if we relabeled each $\boldsymbol{z} \in \mathbb{R}^d$ according to a diffeomorphism $g : \mathbb{R}^d \to \mathbb{R}^d$, up to this relabeling, we would like J to have the same value and solutions. This reparameterization-invariance property trivially holds, since J equals a (negative) KL divergence, and the property holds for KL divergences. It is also straightforward to show that this property holds through direct computation (see Appendix D).

Objective has expected trivial solutions. Because J is bounded from above by zero, we know that any choice of $\boldsymbol{\phi}$ that makes it achieve this bound is a global optimum. This property makes it clear that J has two types of reasonable trivial solutions. First, when $d = D$ and the prior q equals the data likelihood p_{data}—i.e., when there is no dimensionality reduction—the identity map $\boldsymbol{\phi}(\boldsymbol{z}) = \boldsymbol{z}$ is optimal since it makes J equal zero. Second, if p_{data} is genuinely low-dimensional (say, d-dimensional), and there exists a smooth injection $\boldsymbol{\psi}$ for which the pullback $\boldsymbol{\psi}^* p_{\text{data}}$ is normalized on $\mathbb{R}^d$, then $\boldsymbol{\phi} = \boldsymbol{\psi}$ is a global optimum. Importantly, this is true even if $\boldsymbol{\psi}$ is highly nonlinear.

2. In this respect, embedding is similar to other geometry-sensitive problems, like close-packing (Viazovska, 2017; Vastola, 2024).

4.2. Interpretability-related properties

Optimal embeddings satisfy EL equations. The Lagrangian density corresponding to our objective (i.e., its integrand) is

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{J}, \boldsymbol{z}) := q(\boldsymbol{z}) \left[\frac{1}{2} \log \det(\boldsymbol{J}^T \boldsymbol{J}) + \log p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z})) - \log q(\boldsymbol{z}) \right] . \tag{10}$$

A useful consequence of our variational formulation is that it allows us to exploit mathematical results from the calculus of variations. In particular, optimal mappings $\boldsymbol{\phi}$ satisfy the Euler-Lagrange (EL) equations (see Appendix B)

$$\sum_{j=1}^{d} \partial_j \left(\frac{\partial \mathcal{L}}{\partial(\partial_j \phi_i)} \right) = \frac{\partial \mathcal{L}}{\partial \phi_i} \text{ , i.e., } \sum_j \left[\partial_j (J^+_{ji}) + J^+_{ji} \partial_j [\log q(\boldsymbol{z})] \right] = \frac{\partial \log p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z}))}{\partial \phi_i} , \tag{11}$$

where $\boldsymbol{J}^+ := (\boldsymbol{J}^T \boldsymbol{J})^{-1} \boldsymbol{J}^T \in \mathbb{R}^{d \times D}$ is the Moore-Penrose pseudoinverse of $\boldsymbol{J}$ (which satisfies $\boldsymbol{J}^+ \boldsymbol{J} = \boldsymbol{I}_d$), and where we have relied on a well-known result on the derivative of a log-determinant (Petersen and Pedersen, 2008).

The EL equations define a system of D coupled nonlinear PDEs: there is one equation for each $\phi_i(\boldsymbol{z})$, and only up to second derivatives of the ϕ_k appear. (Unfortunately, due to the presence of $\boldsymbol{J}^+$, one may encounter derivatives of the ϕ_k in denominators.) Satisfying the EL equations is necessary but not sufficient for $\boldsymbol{\phi}$ to maximize J, which allows one to use these PDEs to strongly constrain the solution space.

Continuous symmetries of the objective constrain solutions. We expect that symmetries of the data likelihood constrain how one ought to perform dimensionality reduction. For example, the dimensionality-reduced version of a radially symmetric likelihood is also probably radially symmetric. *Noether's theorem* (Noether, 1918; Neuenschwander, 2017; Vastola, 2025b), a powerful result frequently used in field theory settings, makes this intuition precise. It states that if a small perturbation $\boldsymbol{\phi}(\boldsymbol{z}) \to \boldsymbol{\phi}(\boldsymbol{z}) + \delta\boldsymbol{\phi}(\boldsymbol{z})$ of the embedding causes the Lagrangian density to change by a total derivative, i.e., $\mathcal{L} \to \mathcal{L} + \nabla \cdot \boldsymbol{K}$ for some vector $\boldsymbol{K}$, then the quantity (the *Noether current*)

$$j_a := \sum_{b=1}^{D} \frac{\partial \mathcal{L}}{\partial(\partial_a \phi_b)} (\delta \phi_b) - K_a \tag{12}$$

is conserved (i.e., $\nabla \cdot \boldsymbol{j} = 0$) when the EL equations are satisfied (see Appendix E). Noether's theorem is useful because it allows us to make concrete statements about how continuous symmetries constrain optimal embeddings, even if and especially when the EL equations are too difficult to solve analytically.

5. Conserved quantities and consequences of continuous symmetries

In this section, we study explicit examples of how continuous symmetries constrain optimal embeddings. We consider three kinds of continuous symmetry: *reparameterization invariance*, which yields an analogue of energy conservation; *embedding translation invariance*, which yields an analogue of momentum conservation; and *embedding rotation invariance*, which yields an analogue of angular momentum conservation.

5.1. Embedding energy is conserved for optimal embeddings

If q is uniform, the Lagrangian density is locally reparameterization invariant, since an infinitesimal change of coordinates changes $\mathcal{L}$ by a total derivative. By Noether's theorem, this implies that the *stress-energy tensor*

$$T_{ij} = \sum_k \frac{\partial \mathcal{L}}{\partial(\partial_i \phi_k)}(\partial_j \phi_k(\boldsymbol{z})) - \delta_{ij}\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{J}, \boldsymbol{z}) = \sum_k J^+_{ik} J_{kj} q - \delta_{ij}\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{J}, \boldsymbol{z}) = \delta_{ij}(q - \mathcal{L}) \quad (13)$$

is conserved ($\sum_i \partial_i T_{ij} = 0$). But since T_{ij} is only nonzero when $i = j$, conservation just means that $\mathcal{L} = \text{const}$. Hence, $\mathcal{L}$ being conserved is analogous to the conservation of energy.

Usefully, this result can be combined with the reparameterization invariance of the objective to establish a more general result that holds even for non-uniform q (one can also derive this result through a direct computation; see Appendix E): the **embedding energy**

$$\boxed{\mathcal{E} := -\frac{1}{2}\log\det(\boldsymbol{J}^T\boldsymbol{J}) - \log p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z})) + \log q(\boldsymbol{z})} \quad (14)$$

is constant (i.e., does not depend on $\boldsymbol{z}$) for all solutions to the EL equations (and in particular, for the global maximizers of J). Interestingly, since

$$0 \geq J[\boldsymbol{\phi}] = -\int_{\mathbb{R}^d} d\boldsymbol{z}\ q(\boldsymbol{z})\ \mathcal{E} = -\mathcal{E}\ , \quad (15)$$

embedding energy is always nonnegative, and global maximizers of J *minimize* it.

5.2. Translation invariance yields momentum conservation analogue

If p_{data} is uniform along some ambient space direction, so that it does not depend on ϕ_i, Noether's theorem implies that the *canonical momentum* along that direction is conserved:

$$\sum_j \partial_j \left(\frac{\partial \mathcal{L}}{\partial(\partial_j \phi_i)} q(\boldsymbol{z}) \right) = \sum_j \partial_j \left[J^+_{ji} q(\boldsymbol{z}) \right] = 0\ . \quad (16)$$

This conservation law almost fully characterizes the optimal embedding when p_{data} is uniform (on a ball of radius one-half centered at $\boldsymbol{x} = (1/2, ..., 1/2)^T$, say) and $d = 1$. Together with embedding energy conservation, it implies $\phi_i'(z) = \text{const.}/q(\boldsymbol{z})$, so

$$\phi_i(z) = \int_{-\infty}^{z} q(y)\ dy \quad (17)$$

for all $i = 1, ..., D$. Interestingly, this mapping exactly matches the one used by efficient codes to convert a generally non-uniform stimulus distribution to a uniform distribution.

5.3. Rotation invariance yields angular momentum conservation analogue

If p_{data} is invariant to rotations in the ϕ_i-ϕ_j plane, Noether's theorem implies that

$$\sum_a \partial_a L^a_{ij} = 0 \qquad \text{where} \qquad L^a_{ij} := q(\boldsymbol{z}) \left[J^+_{ai}\phi_j - J^+_{aj}\phi_i \right]\ . \quad (18)$$

When $d = 1$, $\boldsymbol{J} \to \boldsymbol{\phi}'$ and $\boldsymbol{J}^+ \to \boldsymbol{\phi}'/\|\boldsymbol{\phi}'\|_2^2$, so angular momentum conservation becomes

$$q(z)\frac{[\phi_1'(z)\phi_2(z) - \phi_1(z)\phi_2'(z)]}{\|\boldsymbol{\phi}'(z)\|_2^2} = \text{const.} \tag{19}$$

When combined with embedding energy conservation, this conservation law can be used to show that optimal embeddings lie along circles when $D = 2$. Hence, the optimal embedding for a ring attractor structure, as expected, follows the ring.

6. A solvable case: PCA is optimal for Gaussian prior and likelihood

Optimal embeddings are difficult to determine analytically, in part because the EL equations are generally a system of highly nonlinear coupled PDEs. Do there exist any interesting solvable cases? Perhaps surprisingly, the answer is yes: the optimal embedding exactly corresponds to PCA when (i) the prior is an isotropic normal distribution $q(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \sigma_0^2\boldsymbol{I})$, with $\sigma_0 > 0$; and (ii) the likelihood is a multivariate Gaussian $p_{\text{data}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ a positive definite $D \times D$ covariance matrix.

The PCA solution corresponds to the linear mapping $\boldsymbol{\phi}(\boldsymbol{z}) = \sum_{k=1}^d \frac{\sqrt{\lambda_k}}{\sigma_0}(\boldsymbol{q}_k^T\boldsymbol{z})\boldsymbol{v}_k$, where the $\boldsymbol{v}_k$ are the top d (normalized) eigenvectors of $\boldsymbol{\Sigma}$, and $\{\boldsymbol{q}_k\}_{k=1,\dots,d}$ is any set of orthonormal vectors (i.e., $\boldsymbol{q}_k^T\boldsymbol{q}_\ell = \delta_{k\ell}$). Although it is not mathematically obvious that the EL equations admit such a simple solution, the intuition is fairly clear: one can obtain an anisotropic Gaussian from an isotropic one by uniformly stretching it in different directions.

The one-dimensional case. Let us begin with the relatively simple $d = 1$ case, with

$$J[\boldsymbol{\phi}] := \int_{-\infty}^{\infty} dz\, \frac{e^{-\frac{1}{2\sigma_0^2}z^2}}{\sqrt{2\pi\sigma_0^2}} \left\{\frac{1}{2}\log\|\boldsymbol{\phi}'(z)\|_2^2 - \frac{1}{2}\boldsymbol{\phi}(z)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(z)\right\} + \text{const.} \tag{20}$$

If the optimal $\boldsymbol{\phi}$ *does* correspond to PCA, it should pick out the first principal component, i.e., an eigenvector of $\boldsymbol{\Sigma}$ whose eigenvalue is largest.[3] The EL equations read

$$\frac{d}{dz}\left[q(z)\frac{\boldsymbol{\phi}'(z)}{\|\boldsymbol{\phi}'(z)\|_2^2}\right] = -q(z)\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(z)\ . \tag{21}$$

Consider a linear ansatz $\boldsymbol{\phi}(z) = \boldsymbol{v}z$ for some vector $\boldsymbol{v} \in \mathbb{R}^D$. Since $\boldsymbol{\phi}'(z) = \boldsymbol{v}$, we require

$$\frac{d}{dz}\left[q(z)\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2^2}\right] = -q(z)\boldsymbol{\Sigma}^{-1}\boldsymbol{v}z \qquad \Longrightarrow \qquad \boldsymbol{\Sigma}\boldsymbol{v} = (\sigma_0^2\|\boldsymbol{v}\|_2^2)\ \boldsymbol{v} \tag{22}$$

i.e., that $\boldsymbol{v}$ must be an eigenvector of $\boldsymbol{\Sigma}$ with eigenvalue $\lambda = (\sigma_0^2\|\boldsymbol{v}\|_2^2)$. Hence, the EL equations force $\boldsymbol{v}$ to be *one* of the eigenvectors of $\boldsymbol{\Sigma}$, but not any specific eigenvector. The symmetry is broken by asking that $\boldsymbol{\phi}$ maximize J (Eq. 20). Its $\boldsymbol{\phi}$-dependent terms equal

$$\mathbb{E}_{z\sim q}\left\{-\frac{z^2}{2}\boldsymbol{v}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{v} + \frac{1}{2}\log(\boldsymbol{v}^T\boldsymbol{v})\right\} = -\frac{\sigma_0^2}{2}\boldsymbol{v}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{v} + \frac{1}{2}\log(\boldsymbol{v}^T\boldsymbol{v}) = -\frac{1}{2} + \frac{1}{2}\log\left(\frac{\lambda}{\sigma_0^2}\right) \tag{23}$$

given our ansatz, which is maximized when λ is as large as possible. This happens precisely when $\boldsymbol{v}$ is the top eigenvector of $\boldsymbol{\Sigma}$, which (almost) finishes our argument. Do the EL equations admit nonlinear solutions? We can show the answer is no by exploiting embedding energy conservation (see Appendix F), which means $\boldsymbol{\phi}(z) = z\boldsymbol{v}$ is indeed optimal.

3. If this is not unique, we expect multiple solutions.

The general case. For arbitrary $d \leq D$, the objective is

$$J[\boldsymbol{\phi}] := \int_{\mathbb{R}^d} d\boldsymbol{z} \; \frac{e^{-\frac{1}{2\sigma_0^2}\|\boldsymbol{z}\|_2^2}}{(2\pi\sigma_0^2)^{d/2}} \left\{ \frac{1}{2} \log\det[\boldsymbol{J}(\boldsymbol{z})^T \boldsymbol{J}(\boldsymbol{z})] - \frac{1}{2}\boldsymbol{\phi}(\boldsymbol{z})^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\phi}(\boldsymbol{z}) \right\} + \text{const.} \tag{24}$$

Although the EL equations are more complicated than in the one-dimensional case, if we assume a linear solution $\boldsymbol{\phi}(\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{z}$ for some $D \times d$ matrix $\boldsymbol{A}$, we recover an analogous constraint:

$$\boldsymbol{\Sigma}\boldsymbol{A} = \sigma_0^2 \boldsymbol{A}\boldsymbol{A}^T\boldsymbol{A} \; . \tag{25}$$

If we exploit a singular value decomposition $\boldsymbol{A} = \sum_{k=1}^d s_k \boldsymbol{u}_k \boldsymbol{q}_k^T$ of $\boldsymbol{A}$, where $\boldsymbol{u}_k^T \boldsymbol{u}_\ell = \boldsymbol{q}_k^T \boldsymbol{q}_\ell = \delta_{k\ell}$ and the s_k are the singular values of $\boldsymbol{A}$, then this condition becomes

$$\boldsymbol{\Sigma}\boldsymbol{u}_k = (\sigma_0^2 s_k^2)\boldsymbol{u}_k \; . \tag{26}$$

for all $k = 1, ..., d$. As before, this implies each $\boldsymbol{u}_k$ is an eigenvector of $\boldsymbol{\Sigma}$—although they must be *different* eigenvectors, since they are pairwise orthogonal. Also as before, the symmetry is broken by asking that $\boldsymbol{\phi}$ maximize J. The $\boldsymbol{\phi}$-dependent part of J equals

$$\mathbb{E}_{\boldsymbol{z}\sim q} \left\{ -\frac{1}{2}\boldsymbol{z}^T \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A}\boldsymbol{z} + \frac{1}{2}\log\det(\boldsymbol{A}^T\boldsymbol{A}) \right\} = -\frac{d}{2} + \frac{1}{2}\sum_{k=1}^d \log\left(\frac{\lambda_k}{\sigma_0^2}\right) \tag{27}$$

given our ansatz, which is maximized when each λ_k is chosen to be as large as possible. This happens precisely when the $\{\boldsymbol{u}_k\}$ are chosen to be the top d eigenvectors of $\boldsymbol{\Sigma}$. Finally, we can argue that the solutions of the EL equations *must* be linear (see Appendix F).

7. Discussion

We introduced a variational framework for dimensionality reduction that is more flexible than PCA—and which includes it as a special case—and which has various features that aid interpretability, including that solutions satisfy PDEs, and that Noether's theorem links data manifold symmetries to conservation laws. Since each dimensionality reduction algorithm associated with the framework optimizes the objective for specific choices of d, q and p_{data}, we can use these quantities to reason about the inductive biases of solutions. Our work is in the spirit of recent studies that leverage insights from physics to clarify the structure of machine learning algorithms (Tanaka and Kunin, 2021; Vastola et al., 2025).

Limitations. At least in the framework's current form, it assumes that both q and p_{data} are distributions on continuous rather than discrete spaces, making it impossible to directly apply it to (for example) Poisson-like spiking data. On the other hand, since distributions of spikes are plausibly related by Bayes' rule to the distributions of continuous latent variables (e.g., an animal's allocentric position in space), an intuition made precise in encoding models like generalized linear models (Weber and Pillow, 2017) and probabilistic population codes (Ma et al., 2006; Vastola et al., 2023), there is still a sense in which the framework can be applied indirectly. Also, while the framework could in principle be used to approximate data manifolds of any dimensionality, it is unclear how scalable it is, and how useful certain analytic insights (e.g., related to the EL equations) are for high-dimensional data sets of interest. Symmetries, for example, strongly constrain low-dimensional embeddings, but not necessarily high-dimensional ones.

Acknowledgments

SJG was funded by the Kempner Institute for the Study of Natural and Artificial Intelligence, and a Polymath Award from Schmidt Sciences. KR was funded by the NIH (RF1DA056403, U01NS136507), James S. McDonnell Foundation (220020466), Simons Foundation (Pilot Extension-00003332-02), McKnight Endowment Fund, CIFAR Azrieli Global Scholar Program, NSF (2046583), a Harvard Medical School Neurobiology Lefler Small Grant Award, and a Harvard Medical School Dean's Innovation Award. This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University.

References

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

Tara Chari and Lior Pachter. The specious art of single-cell genomics. *PLOS Computational Biology*, 19(8):e1011288, August 2023. doi: 10.1371/journal.pcbi.1011288. URL https://doi.org/10.1371/journal.pcbi.1011288. Publisher: Public Library of Science.

SueYeon Chung and L. F. Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology*, 70: 137–144, 2021. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2021.10.010. URL https://www.sciencedirect.com/science/article/pii/S0959438821001227.

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

John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, November 2014. ISSN 1546-1726. doi: 10.1038/nn.3776. URL https://doi.org/10.1038/nn.3776.

Lev D. Elsgolc. *Calculus of variations*. Dover, 2007.

Herbert Goldstein. *Classical mechanics*. Addison-Wesley, 1980.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Geoffrey E Hinton and Richard Zemel. Autoencoders, Minimum Description Length and Helmholtz Free Energy. In J. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann, 1993. URL https://proceedings.neurips.cc/paper_files/paper/1993/file/9e3cfc48eccf81a0d57663e129aef3cb-Paper.pdf.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv e-prints*, art. arXiv:1312.6114, December 2013. doi: 10.48550/arXiv.1312.6114.

Yann LeCun. PhD thesis: Modeles connexionnistes de l'apprentissage (connectionist learning models). 1987.

Wei Ji Ma, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438, November 2006. ISSN 1546-1726. doi: 10.1038/nn1790. URL https://doi.org/10.1038/nn1790.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv e-prints*, art. arXiv:1802.03426, February 2018. doi: 10.48550/arXiv.1802.03426.

Dwight E Neuenschwander. *Emmy Noether's wonderful theorem*. JHU Press, 2017.

E. Noether. Invariante variationsprobleme. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 1918:235–257, 1918. URL http://eudml.org/doc/59024.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11): 559–572, 1901. doi: 10.1080/14786440109462720. URL https://doi.org/10.1080/14786440109462720.

Matthew G. Perich, Devika Narain, and Juan A. Gallego. A neural manifold view of the brain. *Nature Neuroscience*, 28(8):1582–1597, August 2025. ISSN 1546-1726. doi: 10.1038/s41593-025-02031-z. URL https://doi.org/10.1038/s41593-025-02031-z.

K. B. Petersen and M. S. Pedersen. The matrix cookbook, October 2008. URL http://www2.imm.dtu.dk/pubdb/p.php?3274. Version 20081110.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. doi: 10.1126/science.290.5500.2323.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Hidenori Tanaka and Daniel Kunin. Noether's learning dynamics: Role of symmetry breaking in neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25646–25660. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/d76d8deea9c19cc9aaf2237d2bf2f785-Paper.pdf.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319.

Michael E. Tipping and Christopher M. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3): 611–622, September 1999. ISSN 1369-7412. doi: 10.1111/1467-9868.00196. URL https://doi.org/10.1111/1467-9868.00196.

John Vastola. Optimal packing of attractor states in neural representations. In Sophia Sanborn, Christian Shewmake, Simone Azeglio, and Nina Miolane, editors, *Proceedings of the 2nd NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, volume 228 of *Proceedings of Machine Learning Research*, pages 425–442. PMLR, 16 Dec 2024. URL https://proceedings.mlr.press/v228/vastola24a.html.

John Vastola. Generalization through variance: how noise shapes inductive biases in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=7lUdo8Vuqa.

John Vastola, Samuel J. Gershman, and Kanaka Rajan. Gradient descent as loss landscape navigation: a normative framework for deriving learning rules. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=oMi4uyNOlL.

John J. Vastola. Dynamical symmetries in the fluctuation-driven regime: an application of Noether's theorem to noisy dynamical systems. In *Proceedings of the 3rd NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, 2025b. URL https://openreview.net/forum?id=lLiIJc7oCJ.

John J. Vastola, Zach Cohen, and Jan Drugowitsch. Is the information geometry of probabilistic population codes learnable? In Sophia Sanborn, Christian Shewmake, Simone Azeglio, Arianna Di Bernardo, and Nina Miolane, editors, *Proceedings of the 1st NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, volume 197 of *Proceedings of Machine Learning Research*, pages 258–277. PMLR, 03 Dec 2023. URL https://proceedings.mlr.press/v197/vastola23a.html.

Maryna S. Viazovska. The sphere packing problem in dimension 8. *Annals of Mathematics*, 185(3):991–1015, 2017. ISSN 0003486X. URL http://www.jstor.org/stable/26395747.

Binxu Wang and John Vastola. The unreasonable effectiveness of gaussian score approximation for diffusion models and its applications. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=I0uknSHM2j.

Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 2016. doi: 10.23915/distill.00002. URL http://distill.pub/2016/misread-tsne.

Alison I. Weber and Jonathan W. Pillow. Capturing the Dynamical Repertoire of Single Neurons with Generalized Linear Models. *Neural Computation*, 29(12):3260–3289, December 2017. ISSN 0899-7667. doi: 10.1162/neco_a_01021. URL https://doi.org/10.1162/neco_a_01021. _eprint: https://direct.mit.edu/neco/article-pdf/29/12/3260/1026921/neco_a_01021.pdf.

Appendix A. Motivating the objective function

In this appendix, we present an alternative Bayesian argument to motivate our objective function

$$J[\boldsymbol{\phi}] = \int_{\mathbb{R}^d} d\boldsymbol{z}\ q(\boldsymbol{z}) \left[\frac{1}{2} \log \det(\boldsymbol{J}(\boldsymbol{z})^T \boldsymbol{J}(\boldsymbol{z})) + \log p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z})) - \log q(\boldsymbol{z}) \right] \ , \tag{28}$$

and also briefly comment on technical subtleties related to this objective. As described in Sec. 2, D is the ambient space dimensionality, $d \leq D$ is the latent space dimensionality, p_{data} is a likelihood on $\mathbb{R}^D$, q is a distribution (the 'prior') on $\mathbb{R}^d$, and $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^D$ is a smooth injection chosen to maximize J, subject to certain finiteness conditions.

A.1. A Bayesian argument for the objective function

Assumed ground truth distribution. Let $\boldsymbol{x} \in \mathbb{R}^D$ denote an observation. Assume that observations genuinely have low-dimensional structure, in the sense that they can be viewed as generally lying on or near some d-dimensional manifold embedded in $\mathbb{R}^D$ (with $d \leq D$).

We can make the idea of a low-dimensional manifold more concrete by imagining that there exists some smooth injection $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^D$ which embeds any particular location $\boldsymbol{z} \in \mathbb{R}^d$ on this manifold (given in coordinates that may differ from those of the ambient space) into $\mathbb{R}^D$. We can make our demand that observations generally lie 'near' the manifold more concrete by assuming a noise model $p_{\text{noise}}(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\phi}) := \mathcal{N}(\boldsymbol{x}; \boldsymbol{\phi}(\boldsymbol{z}), \epsilon \boldsymbol{I}_D)$, where $\epsilon > 0$ is small. Note that this noise model allows observations to be off-manifold, but that this only occurs with a low probability.

If we assume that latent states are sampled according to a distribution $q(\boldsymbol{z})$, then the joint distribution of observations and latent states is

$$p(\boldsymbol{x}, \boldsymbol{z}) := p_{\text{noise}}(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\phi}) q(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\phi}(\boldsymbol{z}), \epsilon \boldsymbol{I}_D) q(\boldsymbol{z}) \ . \tag{29}$$

Assume that q and $\epsilon > 0$ are known, but the mapping $\boldsymbol{\phi}$ is not. How can we infer it?

Assumed model. One sensible way to proceed is to exploit Bayes' rule. If we model observations as distributed according to $p_{\text{data}}(\boldsymbol{x})$, our knowledge of the noise model $p_{\text{noise}}(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\phi})$ allows us to infer latent states from observations via

$$p(\boldsymbol{z}|\boldsymbol{\phi}) := \int p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\phi}) p_{\text{data}}(\boldsymbol{x})\ d\boldsymbol{x} = \int \frac{p_{\text{noise}}(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\phi}) q(\boldsymbol{z})}{\int p_{\text{noise}}(\boldsymbol{x}|\boldsymbol{z}', \boldsymbol{\phi}) q(\boldsymbol{z}') d\boldsymbol{z}'} p_{\text{data}}(\boldsymbol{x})\ d\boldsymbol{x} \ . \tag{30}$$

Let $p(\boldsymbol{\phi})$ denote a prior over mappings that encodes our preferences about it (e.g., that it is continuous). By Bayes' rule, the posterior distribution of $\boldsymbol{\phi}$ given a latent state $\boldsymbol{z}$ is

$$p(\boldsymbol{\phi}|\boldsymbol{z}) = \frac{p(\boldsymbol{z}|\boldsymbol{\phi}) p(\boldsymbol{\phi})}{\int p(\boldsymbol{z}|\boldsymbol{\phi}') p(\boldsymbol{\phi}')\ d\boldsymbol{\phi}'} \ , \tag{31}$$

where we note that, since the distributions over $\boldsymbol{\phi}$ are defined over *function spaces*, there are technical subtleties related to normalization. Fortunately, these difficulties are irrelevant from the point of view of identifying the most likely mapping $\boldsymbol{\phi}$.

Using Bayes' rule to estimate the embedding map. Let $\{\boldsymbol{z}_i\}_{i=1}^N$ be $N \geq 1$ independent and identically distributed samples from q, the ground truth distribution of latent states. We can estimate $\boldsymbol{\phi}$ by maximizing the scaled log-posterior probability

$$\begin{aligned} \frac{1}{N}\log p(\boldsymbol{\phi}|\{\boldsymbol{z}_i\}) &= \frac{1}{N}\log p(\{\boldsymbol{z}_i\}|\boldsymbol{\phi}) + \frac{1}{N}\log p(\boldsymbol{\phi}) + \text{const.} \\ &= \frac{1}{N}\sum_{i=1}^{N}\log p(\boldsymbol{z}_i|\boldsymbol{\phi}) + \frac{1}{N}\log p(\boldsymbol{\phi}) + \text{const.} \end{aligned} \tag{32}$$

where 'const.' denotes a $\boldsymbol{\phi}$-independent term that comes from the normalization of the posterior. In the large N limit, the central limit theorem implies that J' approaches its average value, which is

$$J'[\boldsymbol{\phi}] = \int_{\mathbb{R}^d} d\boldsymbol{z}\ q(\boldsymbol{z})\ \log p(\boldsymbol{z}|\boldsymbol{\phi}) + \frac{1}{N}\log p(\boldsymbol{\phi})\ , \tag{33}$$

where we have dropped the constant term since it does not affect our optimization. Our argument so far indicates that maximizing J' with respect to $\boldsymbol{\phi}$ yields a Bayes-optimal estimate of it. The final remaining step, after which we will recover an objective equivalent to the one we study in the main text, is to simplify J'.

Taking the small-noise limit. We can simplify J' by using Laplace's method to approximate $p(\boldsymbol{z}|\boldsymbol{\phi})$ in the small noise ($\epsilon \to 0$) limit. The probability $p_{\text{noise}}(\boldsymbol{x}|\boldsymbol{z})$ is obviously maximized when $\boldsymbol{x} = \boldsymbol{\phi}(\boldsymbol{z})$; similarly, if $\boldsymbol{x} \in \boldsymbol{\phi}(\boldsymbol{R}^d)$, then $p(\boldsymbol{z}|\boldsymbol{x})$ is maximized at the unique $\boldsymbol{z}$ for which $\boldsymbol{\phi}(\boldsymbol{z}) = \boldsymbol{x}$ (this is unique since $\boldsymbol{\phi}$ is an injection).

Denote this unique point by $\boldsymbol{z}_*(\boldsymbol{x})$. At a small displacement $\Delta\boldsymbol{z}$ away from it,

$$\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\phi}(\boldsymbol{z}_*(\boldsymbol{x}) + \Delta\boldsymbol{z})\|_2^2 \approx \frac{1}{2}(\Delta\boldsymbol{z})^T\boldsymbol{J}(\boldsymbol{z}_*(\boldsymbol{x}))^T\boldsymbol{J}(\boldsymbol{z}_*(\boldsymbol{x}))(\Delta\boldsymbol{z})\ , \tag{34}$$

i.e., the Hessian near $\boldsymbol{z}_*(\boldsymbol{x})$ is $\boldsymbol{J}(\boldsymbol{z}_*(\boldsymbol{x}))^T\boldsymbol{J}(\boldsymbol{z}_*(\boldsymbol{x}))$. By Laplace's method, this implies

$$\int \frac{1}{(2\pi\epsilon)^{D/2}} e^{-\frac{1}{2\epsilon}\|\boldsymbol{x}-\boldsymbol{\phi}(\boldsymbol{z}')\|_2^2}\ q(\boldsymbol{z}')\ d\boldsymbol{z}' \xrightarrow{\epsilon\to 0} \frac{1}{(2\pi\epsilon)^{(D-d)/2}} \frac{1}{\sqrt{\det[\boldsymbol{J}(\boldsymbol{z}_*(\boldsymbol{x}))^T\boldsymbol{J}(\boldsymbol{z}_*(\boldsymbol{x}))]}}\ q(\boldsymbol{z}_*(\boldsymbol{x}))\ . \tag{35}$$

Furthermore, applying Laplace's method again tells us that

$$\begin{aligned} p(\boldsymbol{z}|\boldsymbol{\phi}) &\approx \frac{1}{(2\pi\epsilon)^{d/2}} \int e^{-\frac{1}{2\epsilon}\|\boldsymbol{x}-\boldsymbol{\phi}(\boldsymbol{z})\|_2^2}\ q(\boldsymbol{z}) \frac{\sqrt{\det[\boldsymbol{J}(\boldsymbol{z}_*(\boldsymbol{x}))^T\boldsymbol{J}(\boldsymbol{z}_*(\boldsymbol{x}))]}}{q(\boldsymbol{z}_*(\boldsymbol{x}))}\ p_{\text{data}}(\boldsymbol{x})\ d\boldsymbol{x} \\ &\approx \sqrt{\det[\boldsymbol{J}(\boldsymbol{z})^T\boldsymbol{J}(\boldsymbol{z})]}\ p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z})) \end{aligned} \tag{36}$$

in the $\epsilon \to 0$ limit. This gives us the result we want: maximizing the average log-posterior probability of $\boldsymbol{\phi}$ yields our objective, possibly plus an extra $\log p(\boldsymbol{\phi})$ term if the prior over $\boldsymbol{\phi}$ is non-uniform. More explicitly,

$$J'[\boldsymbol{\phi}] \xrightarrow{\epsilon\to 0} \int_{\mathbb{R}^d} d\boldsymbol{z}\ q(\boldsymbol{z})\ \left\{\ \frac{1}{2}\log\det[\boldsymbol{J}(\boldsymbol{z})^T\boldsymbol{J}(\boldsymbol{z})] + \log p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z}))\ \right\} + \frac{1}{N}\log p(\boldsymbol{\phi})\ , \tag{37}$$

which only differs from Eq. 28 by a constant offset when $p(\boldsymbol{\phi})$ is uniform (or when we eliminate the prior term by taking $N \to \infty$).

A.2. Additional comments about the objective

The embedding map needs to be an injection. The latent states $\boldsymbol{z}$ index locations on the d-dimensional manifold embedded in the ambient space, and hence function as coordinates for it. This intuition suggests that $\boldsymbol{\phi}$ ought to be an injection; any given point on the manifold should only correspond to a single value of $\boldsymbol{z}$.

There is a more important technical reason for us to demand that $\boldsymbol{\phi}$ is an injection. If $\boldsymbol{\phi}$ is an injection, the pullback of p_{data} through $\boldsymbol{\phi}$ is not generally normalized, i.e.,

$$\int_{\mathbb{R}^d} p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z}))\sqrt{\boldsymbol{J}(\boldsymbol{z})^T\boldsymbol{J}(\boldsymbol{z})}\; d\boldsymbol{z} \leq 1 \; . \tag{38}$$

But because it is less than or equal to one, we can still invoke Gibbs' inequality to argue

$$J[\boldsymbol{\phi}] = \int_{\mathbb{R}^d} q(\boldsymbol{z}) \log\left[\frac{(\boldsymbol{\phi}^* p_{\text{data}})(\boldsymbol{z})}{q(\boldsymbol{z})}\right] \; d\boldsymbol{z} \leq 0 \; . \tag{39}$$

(The argument is the same as the standard one used to prove that the KL divergence is nonnegative, so we omit it.) If $\boldsymbol{\phi}$ were *not* an injection, J does not necessarily have an upper bound, and hence our optimization problem is not well-defined. For example, suppose

$$\begin{aligned} q(z) &:= 1 \;\; \text{ for } z \in [0,1] \;\; \text{ and} \\ p_{\text{data}}(x,y) &:= \frac{1}{2\pi}\delta(x^2+y^2-1) \;\; \text{ for } (x,y) \in \mathbb{R}^2, \end{aligned} \tag{40}$$

i.e., q is uniform on $[0,1]$ and p_{data} is uniform on the unit circle. Each embedding

$$\phi_1(z;k) := \cos(2\pi k z) \qquad\qquad \phi_2(z;k) := \sin(2\pi k z) \tag{41}$$

for $k \in \mathbb{N}$ yields the same $\log p_{\text{data}}(\boldsymbol{\phi}(z))$ term, but have different Jacobian terms. Note,

$$\frac{1}{2}\log \|\boldsymbol{\phi}'(z)\|_2^2 = \frac{1}{2}\log\left[(2\pi k)^2\right] = \log(2\pi k) \; , \tag{42}$$

which implies that we can make this term arbitrarily large by choosing k to be larger and larger. This type of construction is only possible if we allow each (x,y) on the unit circle to correspond to multiple values of z; such constructions are not possible if we insist that $\boldsymbol{\phi}$ is an injection, as Gibbs' inequality indicates. As an aside, we speculate that issues with enforcing the injectivity constraint are responsible for the 'wiggles' depicted in Figure 1b.

Boundary conditions. We are looking for smooth injections $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^D$ that maximize Eq. 28 while also satisfying the finiteness conditions

$$\mathbb{E}_{\boldsymbol{z}\sim q}\left\{\sqrt{\det(\boldsymbol{J}(\boldsymbol{z})^T\boldsymbol{J}(\boldsymbol{z}))}\right\} < \infty \qquad \text{and} \qquad \mathbb{E}_{\boldsymbol{z}\sim q}\left\{\|\boldsymbol{\phi}(\boldsymbol{z})\|_2^2\right\} < \infty \; . \tag{43}$$

Since a small region of $\mathbb{R}^d$ becomes a small region of $\mathbb{R}^D$ with surface area

$$dV = \sqrt{\det(\boldsymbol{J}(\boldsymbol{z})^T\boldsymbol{J}(\boldsymbol{z}))}d\boldsymbol{z} \tag{44}$$

when projected through $\boldsymbol{\phi}$, the first condition is equivalent to asking that the average surface area (or the total q-weighted 'mass' of the embedding) is finite. This condition prevents 'needle'-like embeddings which project small regions of $\mathbb{R}^d$ to extremely large regions of $\mathbb{R}^D$. It also plays a crucial role in our argument that the PCA objective does not admit nonlinear solutions (see Appendix F). The second condition prevents pathological embeddings in the tails of q (i.e., large $\|\boldsymbol{z}\|_2$) by controlling their norm.

Appendix B. Derivation of Euler-Lagrange equations

In this appendix, we derive the Euler-Lagrange (EL) equations for our objective

$$J[\boldsymbol{\phi}] = \int_{\mathbb{R}^d} d\boldsymbol{z}\ q(\boldsymbol{z}) \left[\frac{1}{2} \log\det(\boldsymbol{J}(\boldsymbol{z})^T \boldsymbol{J}(\boldsymbol{z})) + \log p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z})) - \log q(\boldsymbol{z}) \right] , \tag{45}$$

where $J_{ij} := \partial\phi_i/\partial z_j$ are the elements of the $D \times d$ Jacobian matrix $\boldsymbol{J}$. We will assume that $\boldsymbol{J}$ has full column rank, so that the determinant that appears is nonzero. The Lagrangian density corresponding to this objective is

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{J}, \boldsymbol{z}) = q(\boldsymbol{z}) \left[\frac{1}{2} \log\det(\boldsymbol{J}^T \boldsymbol{J}) + \log p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z})) - \log q(\boldsymbol{z}) \right] . \tag{46}$$

B.1. Euler-Lagrange equations for arbitrary-dimensional embeddings

The EL equations are

$$\sum_{j=1}^{d} \partial_j \left(\frac{\partial \mathcal{L}}{\partial(\partial_j \phi_i)} \right) = \frac{\partial \mathcal{L}}{\partial \phi_i} , \tag{47}$$

so we only need to evaluate these derivatives explicitly and simplify the result. First, note

$$\frac{\partial \mathcal{L}}{\partial(\partial_j \phi_i)} = \frac{\partial \mathcal{L}}{\partial J_{ij}} = q(\boldsymbol{z}) \frac{\partial}{\partial J_{ij}} \left[\frac{1}{2} \log\det(\boldsymbol{J}^T \boldsymbol{J}) \right] = q(\boldsymbol{z}) \left[\boldsymbol{J}(\boldsymbol{J}^T \boldsymbol{J})^{-1} \right]_{ij} , \tag{48}$$

where we have used a matrix cookbook result on the derivative of a log-determinant (Petersen and Pedersen, 2008). Note that $\boldsymbol{J}^+ := (\boldsymbol{J}^T \boldsymbol{J})^{-1} \boldsymbol{J}^T$ is the Moore-Penrose pseudoinverse of $\boldsymbol{J}$, i.e., the $d \times D$ matrix which satisfies $\boldsymbol{J}^+ \boldsymbol{J} = \boldsymbol{I}_d$. Using it allows us to write the above partial derivatives in vector form as

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{J}} = (\boldsymbol{J}^+)^T q(\boldsymbol{z}) . \tag{49}$$

More straightforwardly, the partial derivative of $\mathcal{L}$ with respect to ϕ_i is the score of p_{data}:

$$\frac{\partial \mathcal{L}}{\partial \phi_i} = q(\boldsymbol{z}) \frac{\partial \log p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z}))}{\partial \phi_i} . \tag{50}$$

These results immediately imply that the EL equation for ϕ_i reads

$$\sum_j \partial_j \left(\frac{\partial \mathcal{L}}{\partial(\partial_j \phi_i)} \right) = \frac{\partial \mathcal{L}}{\partial \phi_i} \implies \sum_j \partial_j \left[J^+_{ji} q(\boldsymbol{z}) \right] = q(\boldsymbol{z}) \frac{\partial \log p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z}))}{\partial \phi_i} . \tag{51}$$

Simplifying, we have

$$\sum_j \partial_j (J^+_{ji}) + \sum_j J^+_{ji} \frac{(\partial_j q)}{q} = \frac{\partial \log p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z}))}{\partial \phi_i} , \tag{52}$$

or in vector form,

$$[\boldsymbol{\partial} + \nabla \log q(\boldsymbol{z})]^T \boldsymbol{J}^+ = \boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{z})^T , \tag{53}$$

where we have defined the score vector $s_i(\boldsymbol{z}) := \partial \log p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z}))/\partial \phi_i$.

B.2. Euler-Lagrange equations for one-dimensional embeddings

The one-dimensional ($d = 1$) form of the objective reads

$$J[\boldsymbol{\phi}] = \int_{-\infty}^{\infty} dz \; q(z) \left[\frac{1}{2} \log \|\boldsymbol{\phi}'(z)\|_2^2 + \log p_{\text{data}}(\boldsymbol{\phi}(z)) - \log q(z) \right] \tag{54}$$

and the corresponding Lagrangian is

$$L(\boldsymbol{\phi}, \boldsymbol{\phi}', z) := q(z) \left[\frac{1}{2} \log \|\boldsymbol{\phi}'(z)\|_2^2 + \log p_{\text{data}}(\boldsymbol{\phi}(z)) - \log q(z) \right] \; . \tag{55}$$

In this setting, the EL equations have the somewhat simpler form

$$\frac{d}{dz} \left(\frac{\partial L}{\partial \phi_i'} \right) = \frac{\partial L}{\partial \phi_i} \quad \text{, i.e.,} \quad \frac{d}{dz} \left[q(z) \frac{\phi_i'(z)}{\|\boldsymbol{\phi}'(z)\|_2^2} \right] = q(z) \frac{\partial}{\partial \phi_i} \log p_{\text{data}}(\boldsymbol{\phi}(z)) \; . \tag{56}$$

Simplifying as in the previous subsection, we end up with

$$\frac{d}{dz} \left[\frac{\phi_i'(z)}{\|\boldsymbol{\phi}'(z)\|_2^2} \right] + \frac{q'(z)}{q(z)} \frac{\phi_i'(z)}{\|\boldsymbol{\phi}'(z)\|_2^2} = \frac{\partial}{\partial \phi_i} \log p_{\text{data}}(\boldsymbol{\phi}(z)) \; , \tag{57}$$

or if we prefer to be even more explicit,

$$\frac{\phi_i''(z)}{\|\boldsymbol{\phi}'(z)\|_2^2} - \frac{2\phi_i'(z) \sum_k \phi_k'(z)}{\|\boldsymbol{\phi}'(z)\|_2^4} + \frac{q'(z)}{q(z)} \frac{\phi_i'(z)}{\|\boldsymbol{\phi}'(z)\|_2^2} = \frac{\partial}{\partial \phi_i} \log p_{\text{data}}(\boldsymbol{\phi}(z)) \; . \tag{58}$$

Appendix C. Optimal 1D embeddings and diffusion models

In Sec. 3, we showed that optimal 1D embeddings $\boldsymbol{\phi} : \mathbb{R} \to \mathbb{R}^D$ satisfy a system

$$\phi_i'(z) = \frac{p_i(z)}{\|\boldsymbol{p}(z)\|_2^2} \qquad p_i'(z) = s_i(\boldsymbol{\phi}(z)) - \left[\frac{d}{dz}\log q(z)\right] p_i(z) \tag{59}$$

of first-order ODEs, where the $p_i(z)$ are auxiliary momentum-like variables and $s_i(\boldsymbol{\phi}(z)) := \frac{\partial}{\partial \phi_i}\log p_{\text{data}}(\boldsymbol{\phi}(z))$ are the components of the *score vector* associated with p_{data} at the point $\boldsymbol{\phi}(z) \in \mathbb{R}^D$. In this appendix, we argue that optimal embeddings exactly follow score vectors in a certain limit.

Consider a prior $q(z) = \gamma e^{\gamma z}$ which is nonzero only for $z \in (-\infty, 0]$. For this q,

$$\phi_i'(z) = \frac{p_i(z)}{\|\boldsymbol{p}(z)\|_2^2} \qquad p_i'(z) = s_i(\boldsymbol{\phi}(z)) - \gamma p_i(z) \ . \tag{60}$$

Note that $\gamma > 0$ controls the 'time scale' on which $p_i(z)$ approaches its steady state value, at which $p_i(z) = s_i(\boldsymbol{\phi}(z))/\gamma$. More precisely, Eq. 60 implies that $p_i(z)$ obeys the self-consistent equation

$$p_i(z) = \int_{-\infty}^{z} s_i(\boldsymbol{\phi}(z'))\ e^{-\gamma(z-z')}\ dz' \ . \tag{61}$$

Since the exponential decay factor becomes a delta function in the large γ limit, i.e.,

$$\gamma e^{-\gamma(z-z')} \xrightarrow{\gamma\to\infty} \delta(z-z') \ , \tag{62}$$

in this limit we have

$$p_i(z) \approx \frac{1}{\gamma}\int_{-\infty}^{z} s_i(\boldsymbol{\phi}(z'))\ \delta(z-z')\ dz' = \frac{1}{\gamma} s_i(\boldsymbol{\phi}(z)) \ , \tag{63}$$

which means that the momentum vectors match the score vectors. Eq. 60 then implies

$$\phi_i'(z) \approx \gamma \frac{s_i(\boldsymbol{\phi}(z))}{\|\boldsymbol{s}(z)\|_2^2} \ . \tag{64}$$

As we noted in Sec. 3, this implies that trajectories follow score vectors, albeit with a speed that changes depending on their norm. If we perform a reparameterization of 'time' with

$$\tau(z) := \int_{-\infty}^{z} \frac{1}{\|\boldsymbol{s}(z')\|_2^2}\ dz' \qquad \tau'(z) = \frac{1}{\|\boldsymbol{s}(z)\|_2^2} \ , \tag{65}$$

then we have the equivalent dynamics

$$\phi_i'(\tau) = s_i(\boldsymbol{\phi}(\tau)) \tag{66}$$

which exactly follows the score field.

Appendix D. The objective is reparameterization invariant

Recall that our objective is

$$J[\boldsymbol{\phi}] = \int_{\mathbb{R}^d} d\boldsymbol{z}\ q(\boldsymbol{z}) \left[\frac{1}{2} \log\det[\boldsymbol{J}(\boldsymbol{z})^T \boldsymbol{J}(\boldsymbol{z})] + \log p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z})) - \log q(\boldsymbol{z}) \right] \ . \tag{67}$$

One of its key properties is that it is *invariant to diffeomorphisms*; said differently, it is not sensitive to our choice of latent space coordinates. This property is desirable, since one's choice of coordinates is arbitrary, and different coordinate systems can reflect the same underlying geometry. In this appendix, we show that this property holds via a direct computation.

Let $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^d$ be a diffeomorphism, and consider a change of variables $\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{z}')$ from $\boldsymbol{z}$ to $\boldsymbol{z}'$. Note that

$$\boldsymbol{\phi}(\boldsymbol{z}) = \boldsymbol{\phi}(\boldsymbol{g}(\boldsymbol{z}')) \ , \tag{68}$$

so in the new coordinate system $\boldsymbol{\phi}$ becomes $\tilde{\boldsymbol{\phi}}(\boldsymbol{z}') := \boldsymbol{\phi}(\boldsymbol{g}(\boldsymbol{z}'))$. The Jacobian of $\tilde{\boldsymbol{\phi}}$ in this new coordinate system is

$$\tilde{J}_{ij} := \frac{\partial \tilde{\phi}_i}{\partial z'_j} = \sum_{k=1}^{d} \frac{\partial \tilde{\phi}_i}{\partial z_k} \frac{\partial z_k}{\partial z'_j} = \sum_{k=1}^{d} \frac{\partial \phi_i}{\partial z_k} \frac{\partial z_k}{\partial z'_j} \ , \tag{69}$$

which means that $\tilde{\boldsymbol{J}} = \boldsymbol{J}\boldsymbol{J}_{\boldsymbol{g}}$, i.e., that the new Jacobian equals the old one times the Jacobian of $\boldsymbol{g}$. Finally, the measure $q(\boldsymbol{z})d\boldsymbol{z}$ becomes

$$q(g(\boldsymbol{z}'))|\det \boldsymbol{J}_{\boldsymbol{g}}| d\boldsymbol{z}' \ , \tag{70}$$

so in the new coordinate system we effectively go from $q(\boldsymbol{z})$ to $\tilde{q}(\boldsymbol{z}') := q(\boldsymbol{g}(\boldsymbol{z}'))|\det \boldsymbol{J}_{\boldsymbol{g}}|$. Putting these results together, we find that $J[\boldsymbol{\phi}]$ equals

$$\begin{aligned} J[\boldsymbol{\phi}] &= \int_{\mathbb{R}^d} d\boldsymbol{z}'\ q(\boldsymbol{g}(\boldsymbol{z}'))|\det \boldsymbol{J}_{\boldsymbol{g}}| \left[\frac{1}{2} \log\det[\boldsymbol{J}(\boldsymbol{g}(\boldsymbol{z}'))^T \boldsymbol{J}(\boldsymbol{g}(\boldsymbol{z}'))] + \log p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{g}(\boldsymbol{z}'))) - \log q(\boldsymbol{g}(\boldsymbol{z}')) \right] \\ &= \int_{\mathbb{R}^d} d\boldsymbol{z}'\ \tilde{q}(\boldsymbol{z}') \left[\frac{1}{2} \log\det[\boldsymbol{J}(\boldsymbol{g}(\boldsymbol{z}'))^T \boldsymbol{J}(\boldsymbol{g}(\boldsymbol{z}'))] + \log p_{\text{data}}(\tilde{\boldsymbol{\phi}}(\boldsymbol{z}')) - \log q(\boldsymbol{g}(\boldsymbol{z}')) \right] \\ &= \int_{\mathbb{R}^d} d\boldsymbol{z}'\ \tilde{q}(\boldsymbol{z}') \left[\frac{1}{2} \log\det[\tilde{\boldsymbol{J}}(\boldsymbol{z}')^T \tilde{\boldsymbol{J}}(\boldsymbol{z}')] + \log p_{\text{data}}(\tilde{\boldsymbol{\phi}}(\boldsymbol{z}')) - \log \tilde{q}(\boldsymbol{z}') \right] \ , \end{aligned} \tag{71}$$

where two $\log\det|\boldsymbol{J}_{\boldsymbol{g}}|$ terms cancelled out in the last step. Since the final line is just the objective in the new coordinate system determined by the diffeomorphism $\boldsymbol{g}$, J is indeed reparameterization invariant.

Note that this property also follows straightforwardly from the fact that J equals a (negative) KL divergence.

Appendix E. Review and consequences of Noether's theorem

Noether's theorem, which we first discuss in Sec. 4 and then explore in more detail in Sec. 5, links continuous quasi-symmetries of the objective function to conservation laws (Noether, 1918; Neuenschwander, 2017; Vastola, 2025b). In this appendix, we review its statement and provide a derivation of embedding energy conservation.

E.1. Statement of Noether's theorem

Let $\boldsymbol{\phi} := (\phi_1, ..., \phi_D)^T$ be a vector of D fields defined on $\mathbb{R}^d$, i.e., each field is a smooth function from $\mathbb{R}^d$ to $\mathbb{R}$. Let

$$J[\boldsymbol{\phi}] := \int_{\mathbb{R}^d} d\boldsymbol{z}\ \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{J}, \boldsymbol{z}) \tag{72}$$

be an objective function that depends on such a vector of fields through a Lagrangian density $\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{J}, \boldsymbol{z})$, where the $D \times d$ Jacobian matrix $\boldsymbol{J}$ has components $J_{ij} = \partial\phi_i/\partial z_j$.

We will call a field *extremal* if it makes J stationary, which happens if and only if it satisfies the EL equations

$$\sum_{j=1}^{d} \partial_j \left(\frac{\partial \mathcal{L}}{\partial(\partial_j \phi_i)} \right) = \frac{\partial \mathcal{L}}{\partial \phi_i} \ . \tag{73}$$

Next, define the *stress-energy tensor*

$$T_{ij}(\boldsymbol{z}) := \sum_{k=1}^{D} \frac{\partial \mathcal{L}}{\partial(\partial_i \phi_k)} (\partial_j \phi_k(\boldsymbol{z})) - \delta_{ij} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{J}, \boldsymbol{z}) \ . \tag{74}$$

Following Neuenschwander (2017), Noether's theorem is the following result.

Theorem 1 (Noether's theorem) *Consider a transformation from coordinates $\boldsymbol{z}$ and fields $\boldsymbol{\phi}$ to*

$$\begin{aligned} \boldsymbol{z}' &= \boldsymbol{z} + \epsilon(\delta \boldsymbol{z})(\boldsymbol{z}, \boldsymbol{\phi}) \\ \boldsymbol{\phi}' &= \boldsymbol{\phi} + \epsilon(\delta \boldsymbol{\phi})(\boldsymbol{z}, \boldsymbol{\phi}) \end{aligned} \tag{75}$$

where $\epsilon > 0$ is infinitesimally small. Note that the perturbations $\delta\boldsymbol{z}$ and $\delta\boldsymbol{\phi}$ are allowed to depend on $\boldsymbol{z}$ and $\boldsymbol{\phi}$. If this change of coordinates and fields changes the Lagrangian density by a total derivative, i.e.,

$$\mathcal{L}(\boldsymbol{\phi}', \boldsymbol{J}', \boldsymbol{z}') = \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{J}, \boldsymbol{z}) + \epsilon \nabla \cdot \boldsymbol{K}(\boldsymbol{\phi}, \boldsymbol{J}, \boldsymbol{z}) + \mathcal{O}(\epsilon^2) \tag{76}$$

for some function $\boldsymbol{K}$, then we call such a transformation a (continuous) quasi-symmetry of the objective, and conclude that the ***Noether current***

$$j_a := \sum_{b=1}^{D} \frac{\partial \mathcal{L}}{\partial(\partial_a \phi_b)} (\delta \phi_b) - \sum_{c=1}^{d} T_{ac}(\delta z_c) - K_a \tag{77}$$

is conserved (i.e., $\nabla \cdot \boldsymbol{j} = 0$) whenever $\boldsymbol{\phi}$ satisfies the EL equations.

For our *particular* Lagrangian density

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{J}, \boldsymbol{z}) = q(\boldsymbol{z}) \left[\frac{1}{2} \log \det(\boldsymbol{J}^T \boldsymbol{J}) + \log p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z})) - \log q(\boldsymbol{z}) \right] \ , \tag{78}$$

since (see Appendix B)

$$\frac{\partial \mathcal{L}}{\partial J_{ij}} = q(\boldsymbol{z}) J^+_{ji} \ , \tag{79}$$

we have

$$T_{ij}(\boldsymbol{z}) := \sum_k q(\boldsymbol{z}) J^+_{ik} J_{kj} - \delta_{ij} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{J}, \boldsymbol{z}) = \delta_{ij} \left[q(\boldsymbol{z}) - \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{J}, \boldsymbol{z}) \right] \tag{80}$$

and

$$\begin{aligned} j_a &= \sum_{b=1}^{D} q(\boldsymbol{z}) J^+_{ab} (\delta \phi_b) - \sum_{c=1}^{d} \delta_{ac} (q - \mathcal{L})(\delta z_b) - K_a \\ &= \sum_{b=1}^{D} q(\boldsymbol{z}) J^+_{ab} (\delta \phi_b) - (q(\boldsymbol{z}) - \mathcal{L})(\delta z_a) - K_a \ . \end{aligned} \tag{81}$$

E.2. Direct derivation of embedding energy conservation

We can derive embedding energy conservation through a direct computation, without going through Noether's theorem. We want to show that the quantity

$$\boxed{\mathcal{E} := -\frac{1}{2} \log \det(\boldsymbol{J}^T \boldsymbol{J}) - \log p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z})) + \log q(\boldsymbol{z})} \ , \tag{82}$$

which we call the **embedding energy**, is constant (i.e., does not depend on $\boldsymbol{z}$) for all solutions of the EL equations. To see this, multiply Eq. 53 on the right by $\boldsymbol{J}$:

$$\left[\boldsymbol{\partial} + \nabla \log q(\boldsymbol{z}) \right]^T \boldsymbol{J}^+ \boldsymbol{J} = \boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{z})^T \boldsymbol{J} \ . \tag{83}$$

Since $\boldsymbol{J}^+ \boldsymbol{J} = \boldsymbol{I}$, one term is easy to simplify:

$$\left[\nabla \log q(\boldsymbol{z}) \right]^T \boldsymbol{J}^+ \boldsymbol{J} = \left[\nabla \log q(\boldsymbol{z}) \right]^T \ . \tag{84}$$

The derivative term is somewhat more delicate. In terms of components, we have

$$\sum_{j,i} \partial_j (J^+_{ji}) J_{ik} = \sum_{j,i} \partial_j (J^+_{ji} J_{ik}) - J^+_{ji} \partial_j (J_{ik}) = - \sum_{j,i} J^+_{ji} \partial_j (J_{ik}) \ . \tag{85}$$

But note that, since $(\boldsymbol{J}^+)^T$ is the derivative of the kinetic term with respect to $\boldsymbol{J}$, we have

$$\partial_k \left[\frac{1}{2} \log \det(\boldsymbol{J}^T \boldsymbol{J}) \right] = \sum_{i,j} \frac{\partial}{\partial J_{ij}} \left[\frac{1}{2} \log \det(\boldsymbol{J}^T \boldsymbol{J}) \right] \partial_k (J_{ij}) = \sum_{i,j} J^+_{ji} \partial_k (J_{ij}) = \sum_{i,j} J^+_{ji} \partial_j (J_{ik}) \ ,$$

where in the last step we used the fact that $\partial_k (J_{ij}) = \partial^2_{kj} \phi_i = \partial^2_{jk} \phi_i = \partial_j (J_{ik})$. Hence, we conclude that

$$\nabla \left[\frac{1}{2} \log \det(\boldsymbol{J}^T \boldsymbol{J}) \right] + \nabla \log q(\boldsymbol{z}) = \boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{z})^T \boldsymbol{J} = \nabla \log p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z})) \ , \tag{86}$$

or equivalently that

$$\nabla \left\{ \frac{1}{2} \log \det(\boldsymbol{J}^T \boldsymbol{J}) + \log p_{\text{data}}(\boldsymbol{\phi}(\boldsymbol{z})) - \log q(\boldsymbol{z}) \right\} = 0 \ . \tag{87}$$

This implies that the bracketed quantity is constant, i.e., that $\mathcal{E}$ is conserved for solutions of the EL equations.

As an aside, note that

$$0 \geq J[\boldsymbol{\phi}] = - \int_{\mathbb{R}^d} d\boldsymbol{z} \ q(\boldsymbol{z}) \ \mathcal{E} = -\mathcal{E} \ . \tag{88}$$

This implies that we ought to minimize $\mathcal{E}$, and that if J is perfectly maximized (such that the corresponding KL divergence is minimized), $\mathcal{E} = 0$.

Appendix F. There are no nonlinear solutions to the PCA objective

In the main text (Sec. 6), we argue that linear maps satisfy the EL equations associated with the PCA objective, and find that the linear maps that maximize J pick out the top eigenvectors of the data's covariance matrix. Here, we argue that it is impossible for these EL equations to have nonlinear solutions, and hence for J to have any nonlinear global maximizers. This point is important, since otherwise we have not established that linear maps are the global maximizers of the PCA objective.

For technical reasons,[4] we specialize to the prior with $\sigma_0 = 1$ below, but this assumption can easily be relaxed.

F.1. The one-dimensional case

The objective is

$$J[\boldsymbol{\phi}] := \int_{-\infty}^{\infty} dz \, \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}} \, \left\{ \frac{1}{2} \log \|\boldsymbol{\phi}'(z)\|_2^2 - \frac{1}{2} \boldsymbol{\phi}(z)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\phi}(z) \right\} + \text{const.} \tag{89}$$

where we are optimizing over smooth injections $\boldsymbol{\phi} : \mathbb{R} \to \mathbb{R}^D$. Recall from Sec. 2 that we also insist on two finiteness conditions, which here read

$$\mathbb{E}_{z\sim q} \left\{ \|\boldsymbol{\phi}'(z)\|_2 \right\} < \infty \qquad \text{and} \qquad \mathbb{E}_{z\sim q} \left\{ \|\boldsymbol{\phi}\|_2^2 \right\} < \infty \ . \tag{90}$$

Note that the second condition implies that

$$\mathbb{E}_{z\sim q} \left\{ \phi_i(z)^2 \right\} < \infty \tag{91}$$

for each $i = 1, ..., D$. This, in turn, implies that each $\phi_i(z)$ can be written in terms of a (probabilist's) Hermite polynomial expansion

$$\phi_i(z) = \sum_{n=0}^{\infty} c_{ni} H_n(z) = c_{0i} + c_{1i} z + c_{2i}(z^2 - 1) + \cdots \ . \tag{92}$$

By the conservation of embedding energy (see Sec. 5), we have

$$-\frac{1}{2} \|\boldsymbol{\phi}'(z)\|_2^2 + \frac{1}{2} \boldsymbol{\phi}(z)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\phi}(z) - \frac{z^2}{2} = \text{const.} \tag{93}$$

which implies that

$$\begin{aligned} \|\boldsymbol{\phi}'(z)\|_2 &= C \exp \left\{ \frac{1}{2} \boldsymbol{\phi}(z)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\phi}(z) - \frac{z^2}{2} \right\} \\ &= C \exp \left\{ \frac{(\boldsymbol{c}_0^T \boldsymbol{\Sigma}^{-1} \boldsymbol{c}_0)}{2} + (\boldsymbol{c}_0^T \boldsymbol{\Sigma}^{-1} \boldsymbol{c}_1) z + \frac{(\boldsymbol{c}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{c}_1)}{2} z^2 + \mathcal{O}(z^3) - \frac{z^2}{2} \right\} \end{aligned} \tag{94}$$

for some constant $C > 0$, where $\boldsymbol{c}_n$ denotes the vector whose i-th component is c_{ni}. Unless the higher-order coefficients are zero, this norm will grow at least as quickly as $\exp(z^n)$ for some $n > 2$, and hence the first finiteness condition will be violated.

4. For arbitrary σ_0, we must consider Hermite polynomials with scaled arguments, which adds a small amount of notational clutter. Aside from this issue, the argument would be exactly the same.

F.2. The general case

In the general case,

$$J[\boldsymbol{\phi}] := \int_{\mathbb{R}^d} d\boldsymbol{z}\ \frac{e^{-\frac{1}{2}\|\boldsymbol{z}\|_2^2}}{(2\pi)^{d/2}}\ \left\{\frac{1}{2}\log\det[\boldsymbol{J}(\boldsymbol{z})^T\boldsymbol{J}(\boldsymbol{z})] - \frac{1}{2}\boldsymbol{\phi}(\boldsymbol{z})^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(\boldsymbol{z})\right\} + \text{const.} \tag{95}$$

where we are optimizing over smooth injections $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^D$. The general finiteness conditions are

$$\mathbb{E}_{\boldsymbol{z}\sim q}\left\{\sqrt{\det(\boldsymbol{J}(\boldsymbol{z})^T\boldsymbol{J}(\boldsymbol{z}))}\right\} < \infty \qquad \text{and} \qquad \mathbb{E}_{\boldsymbol{z}\sim q}\left\{\|\boldsymbol{\phi}(\boldsymbol{z})\|_2^2\right\} < \infty\ . \tag{96}$$

As above, the second finiteness condition implies that each $\phi_i(z)$ can be written in terms of a Hermite polynomial expansion. By the conservation of embedding energy,

$$-\log\sqrt{\det(\boldsymbol{J}(\boldsymbol{z})^T\boldsymbol{J}(\boldsymbol{z}))} + \frac{1}{2}\boldsymbol{\phi}(\boldsymbol{z})^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(\boldsymbol{z}) - \frac{\|\boldsymbol{z}\|_2^2}{2} = \text{const.} \tag{97}$$

which implies

$$\sqrt{\det(\boldsymbol{J}(\boldsymbol{z})^T\boldsymbol{J}(\boldsymbol{z}))} = C\exp\left\{\frac{1}{2}\boldsymbol{\phi}(\boldsymbol{z})^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(\boldsymbol{z}) - \frac{\|\boldsymbol{z}\|_2^2}{2}\right\} \tag{98}$$

for some constant $C > 0$. By the same argument as before, the argument of this exponential will grow at least as quickly as a polynomial in $\boldsymbol{z}$, and hence not satisfy the first finiteness condition, unless the coefficients of the higher-order Hermite expansion terms are zero. Since only the zeroth and first order terms of the expansion are permitted, we conclude that the global maximizers of J must be linear.