Visualizing probabilistic models in Minkowski space with intensive symmetrized Kullback-Leibler embedding

Han Kheng Teoh,[1] Katherine N. Quinn,[2,3] Jaron Kent-Dobias,[1] Colin B. Clement,[1]
Qingyang Xu,[4] and James P. Sethna[1]

[1]*LASSP, Physics Department, Cornell University, Ithaca, New York 14853-2501, USA*
[2]*Center for the Physics of Biological Function, Department of Physics, Princeton University, Princeton, New Jersey 08544, USA*
[3]*Initiative for Theoretical Sciences, The Graduate Center CUNY, New York, New York 10031, USA*
[4]*MIT Operations Research Center, Cambridge, Massachusetts 02139, USA*

(Received 15 January 2020; accepted 13 July 2020; published 7 August 2020)

We show that the predicted probability distributions for any N-parameter statistical model taking the form of an exponential family can be explicitly and analytically embedded isometrically in a $N+N$-dimensional Minkowski space. That is, the model predictions can be visualized as control parameters are varied, preserving the natural distance between probability distributions. All pairwise distances between model instances are given by the symmetrized Kullback-Leibler divergence. We give formulas for these intensive symmetrized Kullback-Leibler (isKL) coordinate embeddings, and illustrate the resulting visualizations with the Bernoulli (coin-toss) problem, the ideal gas, n-sided die, the nonlinear least-squares fit, and the Gaussian fit. We highlight how isKL can be used to determine the minimum number of parameters needed to describe probabilistic data, and conclude by visualizing the prediction space of the two-dimensional Ising model, where we examine the manifold behavior near its critical point.

DOI: 10.1103/PhysRevResearch.2.033221

I. CONTEXT

Many features of multiparameter models are best understood by studying the manifold of model predictions [1]. Within this paradigm, a *model manifold* is constructed, representing the space of possible model predictions. The manifold is embedded in a larger *behavior space*, representing the space of all possible observables and experimental measurements. Surprisingly, model manifolds are usually observed to be well approximated by relatively flat *hyperribbons*, defined as objects whose successive cross sectionals are successively smaller by a roughly constant factor [2,3]. This has now been found in numerous nonlinear least-squares models [4] and helps explain the parameter indeterminacy or "sloppiness" observed in systems biology [5], quantum Monte Carlo [6], and critical phenomena [7]. The hyperribbon geometry of the model manifold has inspired new algorithms for nonlinear least-squares fits [2,3,8,9] and for the control of complex instrumentation such as particle accelerators [10].

Many statistical models are not of least-squares form. For example, the Ising model of magnetism and the lambda cold dark matter (ΛCDM) model of the cosmic microwave background predict the underlying statistics for experimental observation or, more generally, a distribution of possible observations. Local analysis of parameter sensitivity shows that the Ising model [7] and the ΛCDM model [11] are *sloppy*, in the sense that they have a hierarchy of sensitivity eigenvalues spanning many decades. These local sensitivities are quantitatively measured by the natural distance in the space of probability distributions, the Fisher information metric (FIM) [12].

In Ref. [11], it was shown that the model manifold of probability distributions can be visualized using Intensive Principle Component Analysis (InPCA) by embedding in a Minkowski space. For a model whose parameters $\boldsymbol{\theta}$ correspond to a probability distribution $P_{\boldsymbol{\theta}}(x)$ over observable data x, InPCA allows visualization of the model manifold with pairwise distances between models with parameters $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ given by the Bhattacharyya divergence [13],

$$D^2_{\text{Bhat}}(P_{\boldsymbol{\theta}}, P_{\tilde{\boldsymbol{\theta}}}) = -\ln\left(\sum_x \sqrt{P_{\boldsymbol{\theta}}(x) P_{\tilde{\boldsymbol{\theta}}}(x)}\right). \qquad (1)$$

For the Ising and ΛCDM models, x runs over spin configurations and observed spatial cosmic microwave background (CMB) maps, respectively. The manifold visualized with InPCA reveals its hyperribbon structure, thereby capturing most of the model variation with only a few principal components. The key trick in InPCA, where the limit of zero data is considered to extract an *intensive* property, can be applied using a more general class of pairwise distances given by the f divergences [14] and, in return, yields a collection of intensive distance measures, expressed as a linear combinations of the Rényi divergences [15] (details of which are provided in Appendix A). All Rényi divergences locally reproduce the

FIM, so distances in behavior space reflect how sensitive the model predictions are to shifts in the model parameters.

Here we show, for a large class of important multiparameter models, that a different intensive embedding, built on the symmetrized Kullback-Leibler divergence [16],

$$D^2_{sKL}(P_{\boldsymbol{\theta}}, P_{\tilde{\boldsymbol{\theta}}}) = \sum_x [P_{\boldsymbol{\theta}}(x) - P_{\tilde{\boldsymbol{\theta}}}(x)] \ln \left[\frac{P_{\boldsymbol{\theta}}(x)}{P_{\tilde{\boldsymbol{\theta}}}(x)}\right], \tag{2}$$

generates an explicit, analytically tractable embedding in a Minkowski space of dimension equal to twice the number of parameters. We call this the intensive symmetrized Kullback-Leibler (isKL) (pronounced "icicle") embedding, and provide the corresponding isKL coordinates in Sec. III. Our result is obtained for models which form the *exponential families* [17]:

$$P_{\boldsymbol{\theta}}(x) = h(x) \exp\left[\sum_i \eta_i(\boldsymbol{\theta})\Phi_i(x) - A(\boldsymbol{\theta})\right], \tag{3}$$

where $h(x)$ is the base measure, $\eta_i(\boldsymbol{\theta})$ is the ith natural parameter, $\Phi_i(x)$ is the ith sufficient statistic, and $A(\boldsymbol{\theta})$ is the log partition function. Many models in statistical mechanics form exponential families, e.g., the Boltzmann distribution defined on most Hamiltonians. Moreover, while our method can be used to visualize the manifolds of probabilistic models described by exponential families, we explain in Sec. V how we can use this method to determine the minimum number of parameters needed to describe probabilistic data.

II. CURSE OF DIMENSIONALITY

Both large data sets and multiparameter probabilistic models of large systems suffer from the curse of dimensionality [18]: as the dimension of the system increases, it becomes more difficult to establish meaningful relationships between points as the distance measure becomes saturated. This effect obscures meaningful features within the data set and renders contrast in distances between different data points nonexistent [19].

Intensive embeddings such as InPCA and isKL break the curse of dimensionality for probabilistic models, allowing for low-dimensional projections of model manifolds in a suitable Minkowski space [11]. Big data applications have attempted to resolve this dimensionality issue by embedding the manifold in a curved space [20–22] or in an Euclidean space with an alternative distance measure [23–26], which can yield lower-dimensional projections that capture dominant components of the variation in the data set. For example, Ref. [26] makes use of the extensive [27] and nonisometric [28] *potential distance* in generating useful visualizations of large data sets for biological data in Euclidean space. Our methods suggest an alternative approach.

To prove the utility of embedding probability distributions in a Minkowski space, we consider discrete probability distributions, $\sum_x P(x) = 1$, for simplicity. We first introduce the following three type of distances: (1) A geodesic distance d_G between two distributions, defined as the shortest path through the space of all possible probability distributions. Because probability distributions are normalized and nonnegative, they can be interpreted as unit vectors $y(x) = \sqrt{P(x)}$ in a high-dimensional space, thus forming a high-dimensional sphere. Thus, the path between two distributions would be the great-circle distance between them. (2) The shortest path-length distance through the model manifold, which we call the manifold distance d_M. (3) The pairwise straight-line distance d_S in the embedding space, with a metric which here will be a particular divergence measure. These distances are illustrated in Fig. 1(c). First, note that the path-length distance for d_G and d_M is computed by integrating the Fisher information metric (FIM) along said path,

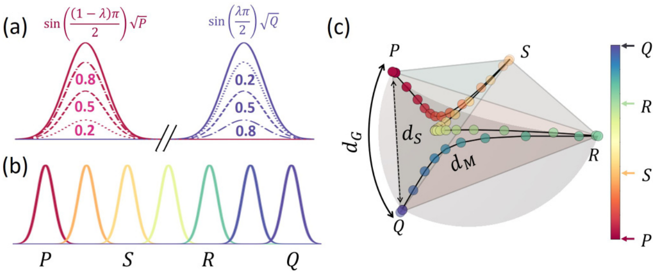
(a)
$\sin\left(\frac{(1-\lambda)\pi}{2}\right)\sqrt{P}$
0.8
0.5
0.2
$\sin\left(\frac{\lambda\pi}{2}\right)\sqrt{Q}$
0.2
0.5
0.8
(b)
P
S
R
Q
(c)
P
S
R
Q
d_G
d_S
d_M
Q
R
S
P

FIG. 1. (a) The geodesic path (with path length d_G) between two probability distributions P and Q is given by an interpolation: $\sqrt{P^*_\lambda(x)} = \frac{\sin[\frac{(1-\lambda)d_G}{2}]}{\sin(\frac{d_G}{2})}\sqrt{P(x)} + \frac{\sin(\frac{\lambda d_G}{2})}{\sin(\frac{d_G}{2})}\sqrt{Q(x)}$. This equals $\sin[\frac{(1-\lambda)\pi}{2}]\sqrt{P(x)} + \sin(\frac{\lambda\pi}{2})\sqrt{Q(x)}$ in the limit when P and Q are orthogonal. As $0 \leqslant \lambda \leqslant 1$, the interpolation remains positive and normalized. The length of this path under the Fisher information metric (FIM) equals the arc length of the great circle, which is $d_G(P,Q) = 2\arccos\sum_x \sqrt{P(x)}\sqrt{Q(x)}$. (b) The shortest path through the model manifold, with path length given by d_M between two Gaussian distributions with fixed σ, is given by sliding the Gaussian μ_P to μ_Q, $d_M = \sigma^{-1}|\mu_P - \mu_Q|$. (c) The global pairwise distance, d_S, between distributions as compared to d_G and d_M. The pairwise distance d_S is determined by the Euclidean distance between points and is represented here as a straight line from P to Q. The octant of the sphere schematically represents the space of all possible probability distributions (due to the normalized, non-negative nature of distributions, discussed further in Sec. II), and so the great-circle path d_G is the arc-length distance from P to Q. The manifold path length d_M is the minimum distance between the two distributions when one considers the path through the model manifold of a complex, nonlinear model. When $d_M \gg d_G$, the path must curl around in multiple dimensions to fit inside the sphere; mutually orthogonal distributions [as in (b)] will form a hypertetrahedron inside the model manifold. Note that (c) represents a 3D projection of a much higher-dimensional space.

$$I_{\alpha,\beta}(\boldsymbol{\theta}) = -\left\langle \frac{\partial^2 \ln P(x)}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle_x, \tag{4}$$

giving

$$d(P,Q) = \int \sqrt{\sum_x \frac{1}{P^*_\lambda(x)}\left[\frac{dP^*_\lambda(x)}{d\lambda}\right]^2}\, d\lambda, \tag{5}$$

where λ parametrizes the path between $P(x)$ and $Q(x)$. Upon letting $P^*_\lambda(x) = y^2_\lambda(x)$, Eq. (5) simplifies to

$$d(P,Q) = \int 2\sqrt{\sum_x \left[\frac{dy_\lambda(x)}{d\lambda}\right]^2}\, d\lambda, \tag{6}$$

representing the familiar path length in Euclidean space. The requirement $\sum_x P_\lambda^*(x) = \sum_x y_\lambda^2(x) = 1$ restricts the path to lie on a sphere, and thus Eq. (5) yields the arc length of a great circle connecting the two distributions [29],

$$d_G(P, Q) = 2 \arccos \sum_x \sqrt{P(x)}\sqrt{Q(x)}. \tag{7}$$

Alternatively, one could perform a variational calculation on Eq. (5) to find the shortest path connecting $P(x)$ and $Q(x)$ and its length. This has been worked out in [30] and the path is given by

$$\sqrt{P_\lambda^*(x)} = \frac{\sin\left[\frac{(1-\lambda)d_G}{2}\right]}{\sin\left(\frac{d_G}{2}\right)}\sqrt{P(x)} + \frac{\sin\left(\frac{\lambda d_G}{2}\right)}{\sin\left(\frac{d_G}{2}\right)}\sqrt{Q(x)}, \tag{8}$$

where $\lambda \in [0, 1]$ and d_G is given by Eq. (7). Thus, the geodesic path between two probability distributions P and Q is a linear interpolation between $\sqrt{P}$ and $\sqrt{Q}$, renormalized to unit length.

When considering a more specific path through a model manifold generated by a specific physical model (e.g., the Ising model), Eq. (5) no longer reduces to such a simple form. Instead, we obtain a more complicated expression, which represents a path through the manifold, d_M. However, since the manifold is confined to the surface of the hypersphere, it is bounded from below by the geodesic distance $d_G \leqslant d_M$. To illustrate the difference between d_G and d_M, consider the example illustrated in Fig. 1: If P and Q are nonoverlapping Gaussians of mean μ_P and μ_Q, the geodesic path d_M along the model manifold of Gaussians of fixed width σ is given by sliding the Gaussian from μ_P to μ_Q, while the shortest path in the space of all probability distributions d_G is given by shrinking P and growing Q in place [see Figs. 1(a) and 1(b)].

A key point is that for any embedding that takes general families of probability distributions isometrically into a Euclidean space, *the straight-line distance d_S is constrained by the diameter of the hypersphere containing the probability distributions* [Eq. (7)]. To further illustrate the differences between the three types of distances, we embedded simple Gaussians with a fixed variance on a hypersphere by using the Hellinger divergence $d_{\text{Hel}}^2 = 1 - \sum_x \sqrt{P(x)}\sqrt{Q(x)}$ as the straight-line distance d_S. Figure 1(c) depicts the three-dimensional projection of the Hellinger embedding. Here, the octant represents the space of all possible probability distributions schematically. In our simple example, if μ_P and μ_Q are many standard deviations apart, the path-length distance d_M between them on the fixed-Gaussian model manifold has length

$$d_M = \int_{\mu_P}^{\mu_Q} \frac{d\mu}{\sigma} = \frac{|\mu_P - \mu_Q|}{\sigma}. \tag{9}$$

When $d_M \gg d_G$, the path must curl around to fit inside the sphere of radius 2. Thus, low-dimensional projection will, at best, show a crumpled tangle that usually rapidly escapes into higher, undisplayed dimensions. In other words, a useful low-dimensional projection should be able to take any set of M probability distributions and project them in a way that maintains their distinguishability.

More generally, consider any type of general-purpose Euclidean embedding, derived from any divergence which locally agrees with the FIM (hence isometric). The straight-line divergence distance d_S between any two distributions is bounded above by the "great-circle" distance $d_G < \pi$ [Fig. 1(c)]. Consider now a low-dimensional visualization of a physical model, where D is the number of dimensions. Since all pairs in this visualization must have distances bounded by π, the low-dimensional visualization must be contained in a sphere in dimension D of diameter π. We certainly would want a good visualization to keep nearly orthogonal probability distributions at least some minimum distance Δ apart, where we could wish that $\Delta \lesssim \pi$. Hence, the minimum acceptable embedding dimension is determined by whether one can pack spheres of diameter Δ into a sphere of diameter π in D dimensions. Whenever $\Delta \sim \pi$, one needs $D \sim M - 1$ projection directions—the curse of dimensionality.

Note that in a Euclidean space, the global pairwise distance d_S is always smaller than the geodesic path through the hypersphere (great-circle length) d_G [bounded by π; see Eq. (7)]. The geodesic distance d_G sets a lower bound on the manifold path length, d_M, since the manifold is confined to the surface of the hypersphere. We shall illustrate many times in the rest of this manuscript that this bound no longer holds when one considers embeddings in Minkowski space. These Minkowski space embeddings can be constructed by defining a pairwise distance between probability distributions d_S that violates the triangle inequality, which in turn breaks the curse of dimensionality, as noted in [11]. For example, Fig. 4 in Ref. [11] shows the InPCA model manifold for the coin-flip problem (different from the isKL embedding in Sec. V A). The straight-line distance between the two endpoints (all heads and all tails) in Minkowski space goes to infinity, but the model manifold hugs a light cone, and the embedding distances from either endpoint to a fair coin are finite. We have shown here how the curse of dimensionality manifests in the Euclidean space of probability distributions. To circumvent this problem, we instead consider embeddings in a Minkowski space, and develop our isKL method in the following section.

III. isKL COORDINATES

In this section, we derive the isKL coordinates for a general exponential family, giving an explicit isometric embedding for probability distributions in a Minkowski space. Our embedding space is similar to Minkowski space but not identical to it, in that it has N spacelike coordinates (positive metric elements) with N corresponding timelike coordinates (negative metric elements), forming an $N + N$-dimensional space. We shall generate two coordinates $\mathcal{S}_i(\boldsymbol{\theta})$ and $\mathcal{T}_i(\boldsymbol{\theta})$ for each natural parameter $\eta(\boldsymbol{\theta})$, one spacelike (with positive squared distance) and one timelike (with negative squared distance), such that

$$D_{sKL}^2(P_{\boldsymbol{\theta}}, P_{\tilde{\boldsymbol{\theta}}}) = \sum_i [\mathcal{S}_i(\boldsymbol{\theta}) - \mathcal{S}_i(\tilde{\boldsymbol{\theta}})]^2 - \sum_i [\mathcal{T}_i(\boldsymbol{\theta}) - \mathcal{T}_i(\tilde{\boldsymbol{\theta}})]^2, \tag{10}$$

where $P_{\boldsymbol{\theta}}$ and $P_{\tilde{\boldsymbol{\theta}}}$ are two probability distributions produced by the model for parameters evaluated at $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$. The squared term with a positive sign is thus a spacelike coordinate, and the term with a negative sign is the corresponding timelike coordinate. Since the symmetrized Kullback-Leibler distance is non-negative, no pair of points can be timelike separated. The length of the model manifold projection along the timelike coordinates will typically be smaller than the length of its projection along the spacelike coordinates. However, the timelike coordinates are both physical and important, as we shall illustrate in particular using the two-dimensional (2D) Ising model.

The symmetrized Kullback-Leibler (K-L) divergence (D^2_{sKL}) from Eq. (2), evaluated for the exponential families considered in this manuscript [shown in Eq. (3)], reduces to

$$D^2_{sKL}(P_{\boldsymbol{\theta}}, P_{\tilde{\boldsymbol{\theta}}}) = \sum_i [\eta_i(\boldsymbol{\theta}) - \eta_i(\tilde{\boldsymbol{\theta}})](\langle\Phi_i\rangle_{\boldsymbol{\theta}} - \langle\Phi_i\rangle_{\tilde{\boldsymbol{\theta}}}). \quad (11)$$

Now, notice that we can rearrange the terms in the above equation, and we obtain

$$\begin{aligned}
&[\eta_i(\boldsymbol{\theta}) - \eta_i(\tilde{\boldsymbol{\theta}})](\langle\Phi_i\rangle_{\boldsymbol{\theta}} - \langle\Phi_i\rangle_{\tilde{\boldsymbol{\theta}}})\\
&\quad = (1/4)\{[\eta_i(\boldsymbol{\theta}) + \langle\Phi_i\rangle_{\boldsymbol{\theta}}] - [\eta_i(\tilde{\boldsymbol{\theta}}) + \langle\Phi_i\rangle_{\tilde{\boldsymbol{\theta}}}]\}^2\\
&\qquad - (1/4)\{[\eta_i(\boldsymbol{\theta}) - \langle\Phi_i\rangle_{\boldsymbol{\theta}}] - [\eta_i(\tilde{\boldsymbol{\theta}}) - \langle\Phi_i\rangle_{\tilde{\boldsymbol{\theta}}}]\}^2\\
&\quad = [\mathcal{S}_i(\boldsymbol{\theta}) - \mathcal{S}_i(\tilde{\boldsymbol{\theta}})]^2 - [\mathcal{T}_i(\boldsymbol{\theta}) - \mathcal{T}_i(\tilde{\boldsymbol{\theta}})]^2, \qquad (12)
\end{aligned}$$

with the two Minkowski coordinates for the ith statistic, determined from model parameters from Eq. (3):

$$\begin{aligned}
\mathcal{S}_i(\boldsymbol{\theta}) &= (1/2)[\eta_i(\boldsymbol{\theta}) + \langle\Phi_i\rangle_{\boldsymbol{\theta}}],\\
\mathcal{T}_i(\boldsymbol{\theta}) &= (1/2)[\eta_i(\boldsymbol{\theta}) - \langle\Phi_i\rangle_{\boldsymbol{\theta}}], \qquad (13)
\end{aligned}$$

now summing to $D^2_{sKL}(P_{\boldsymbol{\theta}}, P_{\tilde{\boldsymbol{\theta}}})$. The terms quadratic in the parameters and quadratic in the expectation values all cancel, and the cross terms give the contribution of statistic [defined from model parameters in Eq. (3)]. From Eq. (13), the spacelike coordinate is indeed greater than the timelike coordinate for each parameter, $[\mathcal{S}_i(\boldsymbol{\theta}) - \mathcal{S}_i(\tilde{\boldsymbol{\theta}})]^2 \geqslant [\mathcal{T}_i(\boldsymbol{\theta}) - \mathcal{T}_i(\tilde{\boldsymbol{\theta}})]^2$. This is our main result.

IV. FAMILIES OF EMBEDDINGS: ISOMETRIES OF MINKOWSKI SPACE

The coordinates produced from isKL represent projections. Just as any rotation or translation of an object isometrically embedded in Euclidean space forms another isometric embedding, so also there is a family of isKL embeddings formed by the isometries of Minkowski space. Translating the coordinates can be used to center the sampled points of the model manifold; certain Lorentz boosts can be valuable in minimizing the total squared length of the coordinates (and hence reducing the importance of the timelike coordinates). The rotational isometries within the spacelike and timelike subspaces can then be used to focus attention on the directions of the model manifold that show the largest variations.

As a first step in considering the effects of these isometries, let us consider other embeddings, similar to Eq. (13), that also preserve pairwise distances. Clearly, one can add a constant $C_i^{\pm}$ to each coordinate (translations in Minkowski space). One also notes that the two terms $\eta_i(\boldsymbol{\theta})$ and $\langle\Phi_i\rangle_{\boldsymbol{\theta}}$ being subtracted may have different units. This can be fixed by rescaling these two terms up and down by a scale factor λ_i with units $\sqrt{[\langle\Phi_i\rangle_{\boldsymbol{\theta}}]/[\eta_i(\boldsymbol{\theta})]}$:

$$\begin{aligned}
\mathcal{S}_i(\boldsymbol{\theta}) &= (1/2)[\lambda_i\eta_i(\boldsymbol{\theta}) + (1/\lambda_i)\langle\Phi_i\rangle_{\boldsymbol{\theta}} + C_i^+],\\
\mathcal{T}_i(\boldsymbol{\theta}) &= (1/2)[\lambda_i\eta_i(\boldsymbol{\theta}) - (1/\lambda_i)\langle\Phi_i\rangle_{\boldsymbol{\theta}} + C_i^-], \qquad (14)
\end{aligned}$$

with different rescaling parameter λ_i and shifts $C_i^{\pm}$ for each pair of coordinates.

We can view Eq. (14) as a composition of two transformations—a translation and a rescaling. The translation is, of course, one of our isometries. The average of Φ_i over parameters $\boldsymbol{\theta}$ is written as $\langle\Phi_i\rangle_{\boldsymbol{\theta}} = \langle\Phi_i\rangle$ in the subsequent discussion for brevity. Rescaling by λ_i corresponds to a Lorentz boost $\mathcal{T}' = \gamma(\mathcal{T} - v\mathcal{S})$, $\mathcal{S}' = \gamma(\mathcal{S} - v\mathcal{T})$ of our timelike and spacelike coordinates, where $\gamma = 1/\sqrt{1 - v^2}$:

$$\begin{aligned}
\mathcal{T}' &= (1/2)\gamma\{[\eta_i(\boldsymbol{\theta}) - \langle\Phi_i\rangle] - v[\eta_i(\boldsymbol{\theta}) + \langle\Phi_i\rangle]\}\\
&= (1/2)[\gamma(1 - v)\eta_i(\boldsymbol{\theta}) - \gamma(1 + v)\langle\Phi_i\rangle]\\
&= (1/2)[\lambda_i\eta_i(\boldsymbol{\theta}) - (1/\lambda_i)\langle\Phi_i\rangle],\\
\mathcal{S}' &= (1/2)\gamma\{[\eta_i(\boldsymbol{\theta}) + \langle\Phi_i\rangle] - v[\eta_i(\boldsymbol{\theta}) - \langle\Phi_i\rangle]\}\\
&= (1/2)\big[\gamma(1 - v)\eta_i(\boldsymbol{\theta}) + \gamma(1 + v)\langle\Phi_i\rangle\big]\\
&= (1/2)[\lambda_i\eta_i(\boldsymbol{\theta}) + (1/\lambda_i)\langle\Phi_i\rangle]. \qquad (15)
\end{aligned}$$

A natural criterion for a good projection of the model manifold would be one which minimizes the sum of squares of the coordinates. In Euclidean space, this just translates the manifold so that its center of mass sits at the origin. Indeed, using C_i^+ and C_i^- to shift our two coordinates to their centers of mass corresponds nicely to shifting the sampled parameters $\eta_i(\boldsymbol{\theta}) \to \eta_i(\boldsymbol{\theta}) - \overline{\eta_i(\boldsymbol{\theta})}$ and resulting means $\langle\Phi_i\rangle - \overline{\langle\Phi_i\rangle}$ to their respective centers of mass. Now, presuming for simplicity that the data is centered, let us examine the sum of the squares of our two coordinates $\mathcal{S}_i$ and $\mathcal{T}_i$,

$$[\mathcal{S}_i(\boldsymbol{\theta})]^2 + [\mathcal{T}_i(\boldsymbol{\theta})]^2 = \frac{1}{2}\left[\lambda_i^2\eta_i^2(\boldsymbol{\theta}) + \frac{1}{\lambda_i^2}\langle\Phi_i\rangle^2\right]. \qquad (16)$$

To get a good point of view in Minkowski space, we seek to minimize the sum of squares of the coordinates by optimizing λ_i. This yields $\lambda_i^4 = \langle\Phi_i\rangle^2/\eta_i^2(\boldsymbol{\theta})$. As the parameters are shifted with respect to their centers of mass, we can recast $\lambda_i = [\mathrm{Var}(\langle\Phi_i\rangle)/\mathrm{Var}(\eta_i)]^{1/4}$, where the variance is averaged over the ensemble of parameters and the mean $\langle\Phi_i\rangle$ is taken at a fixed parameter $\boldsymbol{\theta}$. It turns out our isKL embedding has a close connection to principal component analysis (PCA) and multidimensional scaling (MDS) techniques. We refer interested readers to Appendix B for an in-depth discussion.

V. EXAMPLES

To demonstrate how isKL embeddings optimize the total squared distance of coordinates to produce a good visualization, we consider several probabilistic models that form exponential families: the Bernoulli (coin-toss) problem (Sec. V A), the ideal gas model (Sec. V B), the n-sided die (Sec. V C) the nonlinear least-squares problem (Sec. V D), Gaussian fits to the data (Sec. V E), and the two-dimensional Ising model (Sec. V F). We will be using $T_i^{\pm}(\boldsymbol{\theta})$ to denote spacelike $\mathcal{S}_i(\boldsymbol{\theta})$ and timelike $\mathcal{T}_i(\boldsymbol{\theta})$ coordinates, respectively, for subsequent discussion for brevity.

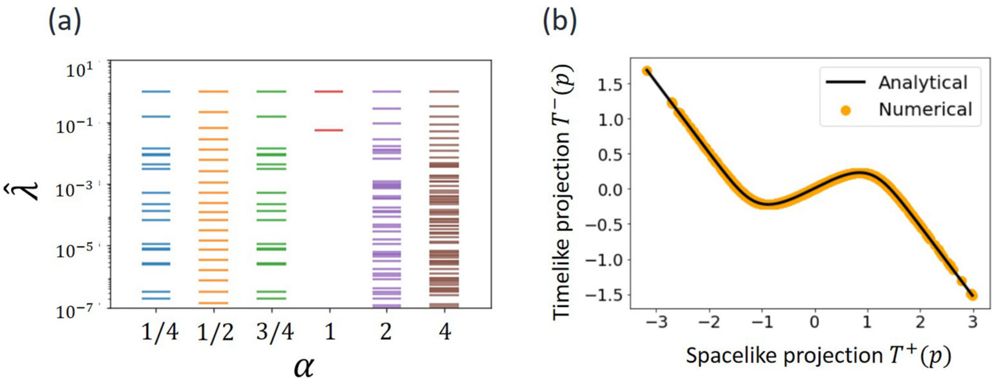

FIG. 2. (a) Squared principal length of intensive embedding with different symmetrized Rényi choices for the coin-toss manifold. $\alpha = 1/2$ corresponds to Bhattacharyya divergence and $\alpha \to 1$ leads to the symmetrized Kullback-Leibler (sKL) divergence. Throughout the exponential family models considered in subsequent sections, sKL provides the lowest embedding dimension, while other Rényi choices gives an embedding into which the manifold projection widths decrease geometrically over several decades. This implies that the sloppiness of the embedding is influenced by the choice of divergence used. (b) Model manifold for the Bernoulli (coin-toss) problem is visualized with our isKL embedding. The analytical calculation matches well with the numerical result returned from multidimensional scaling (MDS).

Before diving into the examples, it is worth highlighting that the finite embedding dimension for exponential families appears to be a unique feature of D^2_{sKL}. As D^2_{sKL} is part of a family of intensive distance measures known as the symmetrized Rényi divergence,

$$D^2_\alpha(P, Q) = \frac{1}{\alpha - 1}\left[\sum_x \ln P(x)^\alpha Q(x)^{1-\alpha} + \sum_x \ln Q(x)^\alpha P(x)^{1-\alpha}\right], \tag{17}$$

with $\alpha \to 1$, we embed the coin-toss manifold with other symmetrized Rényi divergences by varying α to illustrate this uniqueness. As shown in Fig. 2(a), the embedding is sloppy for all α (geometrically decreasing manifold widths that span several decades), but only for $\alpha = 1$ does it truncate after two dimensions. This exact truncation is true for all the probabilistic models considered in this paper. This also serves to illustrate that while the symmetrized Rényi divergences locally reproduce the FIM that describes the local structure of a model manifold, they have a varying degree of performance in utilizing the number of dimensions to embed a model manifold isometrically. Therefore, we can reduce the embedding dimension significantly by choosing an optimal divergence. In principle, we could perform experiments or simulations without knowing the number of parameters that the exponential family distribution needs to describe the behavior. If the isKL embedding gives a cutoff after $N + N$ dimensions, it suggests that a hidden N-parameter exponential family describes the experiment.

A. Bernoulli problem

The Bernoulli problem or coin-toss experiment is one of the simplest probabilistic models. As a function of the fairness parameter p, the result $x \in \{0, 1\}$ of a coin toss is distributed by $P(x|p) = p^x(1-p)^{1-x}$. This probability distribution can be written in the form of an exponential family with $\eta(p) = \ln[p/(1-p)]$, $\Phi(x) = x$, $h(x) = 1$, and $A(p) = -\ln(1-p)$. The FIM for this model is given by

$$(ds)^2 = \frac{(dp)^2}{p(1-p)}. \tag{18}$$

Known embeddings. By defining $p = \sin^2\theta$, we have $ds = 2d\theta$. This produces a one-dimensional embedding onto a Hellinger quarter-circle of radius 2 with $\theta \in [0, \pi/2]$. Upon taking the limit of zero data, the Hellinger distance transforms into the Bhattacharyya divergence. It is known that with the Bhattacharyya divergence, the coin-toss manifold is embedded into a Minkowski space [11]. This embedding is illustrated in Fig. 2(a) with $\alpha = 1/2$. We worked out the analytical expression for each projection coordinate in Appendix C. Our analytical calculation suggests that the embedding is at least high dimensional. We would presume the inPCA embedding does not truncate and continues to have a smaller and smaller amount of variation out to an infinite number of dimensions.

With isKL embedding, the coin-toss manifold can be isometrically embedded into (1 + 1) dimensions. As $\langle\Phi\rangle = p$, its pairwise distance is given by

$$D^2_{sKL}(p, q) = (p - q)\ln\frac{p(1-q)}{q(1-p)}. \tag{19}$$

Here, we will illustrate the utility of Eq. (13) in obtaining the analytical expression for each embedding coordinate. With the Jeffrey's prior as the sampling measure, the centers of mass are $\overline{\eta} = 0$ and $\overline{\langle\Phi\rangle} = 1/2$, respectively. Furthermore, $\mathrm{Var}(\eta) = \pi^2$ and $\mathrm{Var}(\langle\Phi\rangle) = 1/8$, and we have $\lambda = [\mathrm{Var}(\langle\Phi\rangle)/\mathrm{Var}(\eta)]^{1/4} = (2^{3/4}\sqrt{\pi})^{-1}$. With these, the spacelike and timelike $T^\pm(p)$ projection coordinates are determined to be

$$\begin{aligned} T^\pm(p) &= \frac{1}{2}\left[\lambda(\eta - \overline{\eta}) \pm \frac{1}{\lambda}(\Phi - \overline{\langle\Phi\rangle})\right] \\ &= \frac{1}{2^{7/4}\sqrt{\pi}}\ln\left(\frac{p}{1-p}\right) \pm \frac{\sqrt{\pi}}{2^{1/4}}\left(p - \frac{1}{2}\right). \end{aligned} \tag{20}$$

Figure 2(b) shows the coin-toss manifold.

B. Ideal gas

The ideal gas is a model of noninteracting particles. At pressure P and temperature β^{-1}, the probability that N particles will be found in a configuration with momenta $\mathbb{P}$, positions $\mathbb{Q}$, and container volume V is

$$p(\mathbb{P}, \mathbb{Q}, V|P, \beta) = Z^{-1}(P, \beta) \exp\left(-\beta\mathbb{P}^2/2m - \beta PV\right), \tag{21}$$

where the partition function $Z(P, \beta) = (2\pi m/\beta)^{3N/2}(\beta P)^{-(N+1)}$ normalizes the distribution. This probability distribution is in the form of an exponential family with $[\eta_1(\boldsymbol{\theta}), \eta_2(\boldsymbol{\theta})] = (-\beta, -\beta P)$, $[\Phi_1(x), \Phi_2(x)] = (\mathbb{P}^2/2m, V)$, $h(x) = 1$, and $A(\boldsymbol{\theta}) = \ln[Z(P, \beta)]$. Using the coordinates (p, β), where $p = \beta P$, its FIM is $(ds)^2 = (N+1)(dp/p)^2 + (3N/2)(d\beta/\beta)^2$. The scalar curvature of the resulting manifold is zero everywhere, implying that it is a developable surface.

Known embedding. By defining a new pair of coordinates $(x, y) = [\sqrt{1+N}\log(p), \sqrt{3N/2}\log(\beta)]$, we have a two-dimensional Euclidean embedding. However, the pairwise distance in this embedding is not given by D^2_{sKL} and, in fact, it is not obtainable from any symmetrized Rényi divergence [31].

The isKL isometrically embeds the ideal gas into $(2+2)$ dimensions. To determine the axis of projection analytically, note that the ideal-gas law $PV = N/\beta$ yields the sufficient statistics $\langle\mathbb{P}^2/2m\rangle = N/\beta$ and $\langle V\rangle = N/p$. Hence, the pairwise KL divergence between two distributions is

$$D^2_{sKL}(p_1, p_2, \beta_1, \beta_2) = N(p_2 - p_1)\left(\frac{1}{p_1} - \frac{1}{p_2}\right) + N(\beta_2 - \beta_1)\left(\frac{1}{\beta_1} - \frac{1}{\beta_2}\right). \tag{22}$$

Letting the centers of mass be $\overline{\langle\eta\rangle} = \langle\eta\rangle$ and $\overline{\langle\Phi\rangle} = \langle\Phi\rangle$, the projection coordinates are given by

$$\begin{aligned} T^{\pm}_p(p) &= \tfrac{1}{2}\left[\lambda_p(-p + \langle p\rangle) \pm N\lambda_p^{-1}(p^{-1} - \langle p^{-1}\rangle)\right], \\ T^{\pm}_\beta(\beta) &= \tfrac{1}{2}\left[\lambda_\beta(-\beta + \langle\beta\rangle) \pm N\lambda_\beta^{-1}(\beta^{-1} - \langle\beta^{-1}\rangle)\right]. \end{aligned} \tag{23}$$

From Eq. (23), the coordinate pairs yield $(T^+_k - C^+_k)^2 - (T^-_k - C^-_k)^2 = -r^2$, where $k = \{p, \beta\}$, $r^2 = N$ and $C^{\pm}_k = (1/2)(\lambda_k\langle k\rangle \mp N\lambda_k^{-1}\langle k^{-1}\rangle)$ are constants that depend on the sampling range. Therefore, the ideal-gas manifold is a four-dimensional Minkowskian torus (topologically a hyperboloid) with radii $r_1 = r_2 = i\sqrt{N}$. Its projections are illustrated in Fig. 3(b). Just as the 4D Euclidean torus has zero curvature [32], so it does in Minkowski space.

We can map our isKL embedding onto the known embedding into $\mathbb{R}^2$ above. Roughly speaking, this works because our torus is the Cartesian product of two circles with zero Gaussian curvature. We are thus able to provide a mapping to the Euclidean embedding discussed by shifting the coordinates, $T^{\pm}_k \to T^{\pm}_k - C^{\pm}_k$, and parametrizing the coordinate pairs as $(T^+_k, T^-_k) = [\sqrt{N}\sinh(\phi_k), \sqrt{N}\cosh(\phi_k)]$, where $\phi_k = \ln(k\lambda_k/\sqrt{N})$ and $k \in \{p, \beta\}$. This gives

$$(x, y) = \left(\sqrt{1+N}\left[\ln\left(\frac{\sqrt{N}}{\lambda_p}\right) + \phi_p\right], \sqrt{\frac{3N}{2}}\left[\ln\left(\frac{\sqrt{N}}{\lambda_\beta}\right) + \phi_\beta\right]\right), \tag{24}$$

where the "circles" have been unwound to straight lines through the hyperbolic angle ϕ_k.

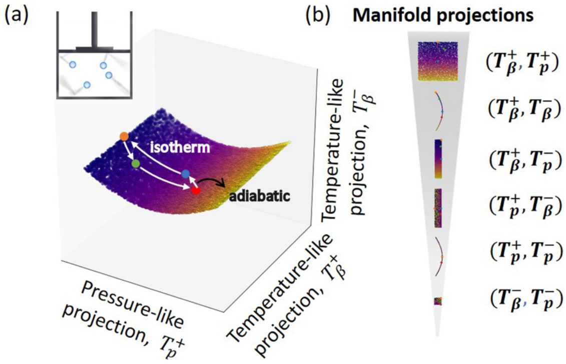

FIG. 3. Model manifold for the ideal gas. The flat ideal-gas manifold is embedded into a $(2+2)$-dimensional Minkowski space. The manifold is "rolled" twice in the four-dimensional space, giving it a torus appearance in Minkowski space. (a) The three-dimensional projection of the ideal-gas manifold is colored based on the inverse temperature β and the Carnot cycle is illustrated on the manifold. (b) The manifold projections are depicted in a descending order based on the manifold widths along the spacelike/timelike components. The spacelike directions are denoted by T^+ and the timelike directions are denoted by T^-. The analytical expression for each projection is given by Eq. (23). The torus appearance in Minkowski space can be deduced from the curves in (T^+_β, T^-_β) and (T^+_p, T^-_p) coordinate pairs, both of which have the form of $(T^+_k - C^+_k)^2 - (T^-_k - C^-_k)^2 = -r^2$ for some constants $C^{\pm}_k$ and r.

Figure 3(a) shows the three-dimensional projection of the ideal-gas manifold, which is colored based on the inverse temperature β. Discussion of the ideal gas is often accompanied by that of the thermodynamic cycles with which it can be used to extract work from a heat bath. The Carnot cycle, which is often considered to cost no entropy, was recently shown [33] to have a subextensive entropy cost proportional to the arc length of the cycle's path on the model manifold. This challenges Szilard's argument that information entropy and thermodynamic entropy can be freely exchanged. The path of a Carnot cycle is shown on the model manifold in Fig. 3(a).

C. The *n*-sided die

The n-sided die is a model for a process with n outcomes. It has a discrete probability distribution of n states, with p_i as the probability of the ith state. This distribution can be written as $P(x|\boldsymbol{p}) = \prod_{i=1}^n p_i^{[x=i]}$, where $[x=i]$ is the Iverson bracket which evaluates to 1 if $x = i$, 0 otherwise, and $\sum_{i=1}^n p_i = 1$. The probability distribution can be written in the form of an exponential family with $\eta_i(\boldsymbol{p}) = \ln(p_i/p_n)$, $\Phi_i = [x]$, $h(x) = 1$, and $A(\boldsymbol{p}) = -\ln(1 - \sum_{i=1}^{n-1} p_i)$. Its FIM is $(ds)^2 = \sum_{i=1}^n (dp_i)^2/p_i$.

Known embedding. Taking $\sqrt{p_i}$ as parameters instead of p_i gives an embedding onto a Hellinger n sphere. This implies that in the Hellinger embedding, the n-sided die manifold has both permutation and spherical symmetry. Moreover, since this mapping is a universal cover of n sphere, its scalar curvature must be positive [34]. For example, the scalar curvature of a three-sided die and a four-sided die are 1/2 and 2, respectively.

The isKL produces an embedding in $(n-1)+(n-1)$ dimensions. As $\langle\Phi_i\rangle = p_i$, the pairwise KL divergence between $P_{\boldsymbol{p}}$ and $P_{\boldsymbol{a}}$ is

$$D^2_{sKL}(\boldsymbol{p}, \boldsymbol{a}) = \sum_{i=1}^{n} (p_i - a_i) \ln\left(\frac{p_i}{a_i}\right). \tag{25}$$

By letting $\overline{\langle\eta_i\rangle} = \langle\eta_i\rangle$ and $\overline{\langle\Phi_i\rangle} = \langle\Phi_i\rangle$, the projection coordinates are

$$T^{\pm}_k(p_1, \ldots, p_n) = \frac{1}{2}\left\{\lambda_k\left[\ln\left(\frac{p_k}{p_n}\right) - \left\langle \ln\left(\frac{p_k}{p_n}\right)\right\rangle\right] \pm \frac{1}{\lambda_k}(p_k - \langle p\rangle)\right\}, \tag{26}$$

where $k = 1, \ldots, n-1$ and $p_n = 1 - \sum_{i=1}^{n-1} p_i$. As examples, we consider three- and four-sided dice. The isKL gives $(2+2)$- and $(3+3)$-dimensional embeddings in Minkowski space. There are only two eigenvalues returned in both cases, signaling the existence of symmetries in our embeddings. With uniform sampling of the parameter space, for $n = 3$,

$$T^{(k)}_{\pm}(p_1, p_2) = \frac{1}{2}\left(\frac{1}{6^{1/4}\sqrt{\pi}} \ln\left(\frac{p_k}{p_3}\right) \pm 6^{1/4}\sqrt{\pi}\left(p_k - \frac{1}{3}\right)\right), \tag{27}$$

where $k = 1, 2$ and $p_3 = 1 - p_1 - p_2$. For $n = 4$,

$$T^{(k)}_{\pm}(p_1, p_2, p_3) = \frac{1}{2}\left(\frac{1}{5^{1/4}}\sqrt{\frac{3}{4\pi}} \ln\left(\frac{p_k}{p_4}\right) \pm 5^{1/4}\sqrt{\frac{4\pi}{3}}\left(p_k - \frac{1}{4}\right)\right), \tag{28}$$

where $k = 1, 2, 3$ and $p_4 = 1 - p_1 - p_2 - p_3$. Finally, the projection coordinates for $n = 2$ (a coin toss) are

$$T^{(k)}_{\pm}(p_1, p_2) = \frac{1}{2}\left(\frac{1}{\sqrt{2\pi}} \ln\left(\frac{p}{1-p}\right) \pm \sqrt{2\pi}\left(p - \frac{1}{2}\right)\right). \tag{29}$$

As expected, comparing Eq. (29) with Eq. (20), the form does not depend on the sampling choice, while the constant λ_p does. Figure 4(b) shows the numerically calculated manifold projections. The manifold is colored based on the fairness parameter p_1. Unlike the Hellinger embedding, the lack of spherical symmetry is manifest. We do, however, see a permutation symmetry among p_i's and a reflection symmetry along $T^{\pm}_{p_1} = T^{\pm}_{p_2}$ in the $(T^{\pm}_{p_1}, T^{\pm}_{p_2})$ coordinate pairs. One can extract the submanifold of a coin-toss problem by restricting $p_2 = 0$. This submanifold is shown by the red line in Fig. 4(a). In general, any discrete probability distribution is a subset of the n-sided die distribution, implying that other discrete exponential family distributions may have hidden low-dimensional representation within the n-sided die model manifold.

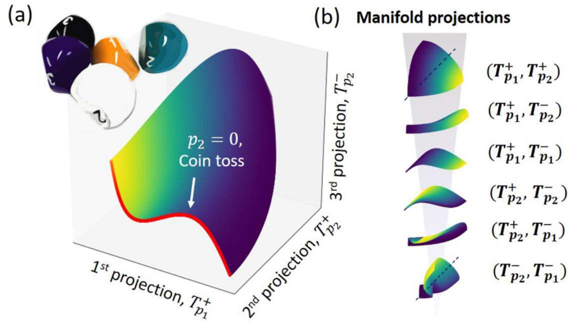

FIG. 4. Model manifold for the three-sided die is embedded into $(2+2)$ dimensions with isKL embedding. (a) The three-dimensional projection of the three-sided die manifold is colored according to the fairness parameter p_1. Depicted also is the coin-toss submanifold in red. (b) The manifold projections are arranged based on the manifold widths. The spacelike directions are denoted by T^+ and the timelike directions are denoted by T^-. The analytical expression for each projection is given by Eq. (27). We have permutation symmetry in $(T^+_{p_i}, T^-_{p_j})$ coordinate pairs and reflection symmetry along the $p_1 = p_2$ line (dotted line) in $(T^{\pm}_{p_i}, T^{\pm}_{p_j})$ coordinate pairs.

D. Nonlinear least-squares models

Nonlinear least-squares models are ubiquitous in fitting deterministic models to data with noise. These models take the form of a nonlinear vector-valued function $f_i(\boldsymbol{\theta})$ predicting the value of experimental data points x_i with uncertainties σ_i. Their associated probability distribution is

$$P(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\left\{\frac{[f_i(\boldsymbol{\theta}) - x_i]^2}{2\sigma_i^2}\right\}\right). \tag{30}$$

This probability distribution takes the form of an exponential family with $\eta_i(\boldsymbol{\theta}) = f_i(\boldsymbol{\theta})/\sigma_i$, $\Phi(x_i) = x_i/\sigma_i$, $h(\boldsymbol{x}) = -\sum_i x_i^2/\sigma_i^2$, and $A(\boldsymbol{\theta}) = \sum_i f_i^2(\boldsymbol{\theta})/2\sigma_i^2 + \ln(2\pi\sigma_i^2)/2$. Unlike the other models discussed, which have the same number of natural parameters $\eta_i(\boldsymbol{\theta})$ and model parameters θ_i, here the number of natural parameters is given by the number of data points being fit. The FIM is given by $J^{\top}_{\beta i} J_{i\alpha}$, where $J_{i\alpha} = \partial f_i(\boldsymbol{\theta})/\partial\theta_\alpha$ is the Jacobian.

Known embedding. Least-squares models with N data points have a natural "prediction embedding" into N-dimensional Euclidean space, with one coordinate per data point x_i given by the error-normalized model prediction $f_i(\boldsymbol{\theta})/\sigma_i$. While the number of data points can be much larger than the number of parameters, this embedding remains valuable because the model predictions are surprisingly often well approximated by low-dimensional, flat model manifolds that we call *hyperribbons* [2–4]. Hyperribbons have a hierarchy of manifold widths—like a ribbon, their dimensions (length, width, thickness, etc.) become geometrically smaller—yielding predictions that depend mostly on the first few principal components. Our least-squares model has N natural parameters, so isKL will produce an embedding into an $N+N$-dimensional Minkowski space. Can we find one

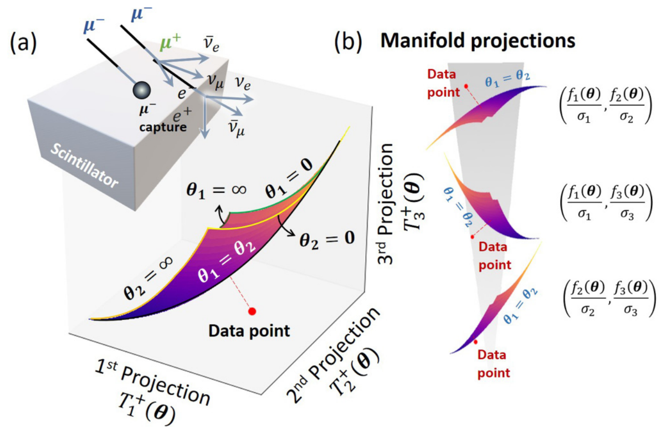
(a)
μ^-
μ^-
μ^+
$\bar{\nu}_e$
ν_μ
ν_e
e^-
e^+
μ^- capture
$\bar{\nu}_\mu$
Scintillator
$\theta_1 = \infty$
$\theta_1 = 0$
$\theta_2 = 0$
$\theta_1 = \theta_2$
$\theta_2 = \infty$
Data point
1st Projection $T_1^+(\boldsymbol{\theta})$
2nd Projection $T_2^+(\boldsymbol{\theta})$
3rd Projection $T_3^+(\boldsymbol{\theta})$
(b) Manifold projections
Data point
$\theta_1 = \theta_2$
$\left(\frac{f_1(\boldsymbol{\theta})}{\sigma_1}, \frac{f_2(\boldsymbol{\theta})}{\sigma_2}\right)$
$\theta_1 = \theta_2$
Data point
$\left(\frac{f_1(\boldsymbol{\theta})}{\sigma_1}, \frac{f_3(\boldsymbol{\theta})}{\sigma_3}\right)$
$\theta_1 = \theta_2$
Data point
$\left(\frac{f_2(\boldsymbol{\theta})}{\sigma_2}, \frac{f_3(\boldsymbol{\theta})}{\sigma_3}\right)$

FIG. 5. Model manifold for the muon lifetime, our two-parameter least-squares model, evaluated at three time points. The isKL embedding is confined to three Euclidean dimensions, with the three timelike coordinates identically zero. (a) The manifold is colored with the muon lifetime θ. The model manifold is bounded with four edges at $\theta_k = 0$ and $\theta_k = \infty$ and a tight fold along $\theta_1 = \theta_2$. Also depicted is the experimental data point in red, which is in close proximity to the $\theta_1 = \theta_2$ boundary; see Fig. 1 in Ref. [2]. (b) The manifold projections of the muon lifetime model manifold are arranged based on the manifold widths. The analytical expression for each axis is given by Eq. (32).

that makes the timelike distances equal to zero, reproducing the N-dimensional prediction embedding?

The symmetrized Kullback-Leibler divergence between two models is indeed given by the Euclidean distance between the two model predictions,

$$D_{sKL}^2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sum_{i=1}^{N} \frac{[f_i(\boldsymbol{\theta}_1) - f_i(\boldsymbol{\theta}_2)]^2}{\sigma_i^2}. \tag{31}$$

This appears promising: the isKL distance is the same as that of the prediction embedding above. Interestingly, any Rényi divergence (such as the Bhattacharyya distance used by InPCA [11]) gives the same pairwise distance measure. Since $\langle \Phi(x_i) \rangle = f_i(\boldsymbol{\theta})/\sigma$, the projection coordinates are

$$T_i^{\pm}(\boldsymbol{\theta}) = \frac{1}{2\sigma_i}\left(\lambda \pm \frac{1}{\lambda}\right)[f_i(\boldsymbol{\theta}) - \langle f_i(\boldsymbol{\theta}) \rangle]. \tag{32}$$

By taking $\lambda = 1$, the timelike coordinates vanish and we reproduce the N-dimensional prediction embedding.

Figure 5 shows this prediction embedding for the classical nonlinear least-squares model of two exponential decays, here in the context of a cosmic muon lifetime experiment. Approximately half of the muons generated by cosmic ray collisions are negative muons, which can be captured by a proton of the host nuclei. The effective negative muon lifetime $1/\theta_2$ (including capture) is therefore expected to be shorter than the decay-only lifetime of positive muons $1/\theta_1$. The model prediction for the number of muons surviving after some time $N(t)$ is thus the sum of two exponentials. Mathematically, we have

$$\hat{N}(\theta_1, \theta_2, r, t) = \frac{1}{1+r}(re^{-\theta_1 t} + e^{-\theta_2 t}), \tag{33}$$

where $\hat{N}(t)$ is the normalized number of muons and $r = N_{\mu^+}/N_{\mu^-} = 1.18 \pm 0.12$ is the ratio of incident positive muons to negative muons formed by the cosmic rays [35]. Figure 5 shows the muon lifetime model manifold via the isKL embedding (identical to the prediction embedding), with three sampled time points. The manifold is colored based on the muon lifetime θ_1. The projection coordinates are $\hat{N}(t_i)/\sigma_i$. Since $r \approx 1$, there is a tight fold in the model manifold along $\theta_1 = \theta_2$. The experimental data point is close to the manifold fold, implying that the negative muon capture event only leads to a slight change in negative muon lifetime.

E. Gaussian fits to data

The Gaussian distribution is an exceptionally good approximation for many physical problems and thus serves as a good model to explore in the context of manifold visualization. For example, the distribution of women's heights with mean height μ and variance in height σ^2 in a country is fitted to a normal (Gaussian) distribution. The Gaussian distribution $P(x|\mu, \sigma) = (2\pi\sigma^2)^{-1/2} \exp[-(x-\mu)^2/2\sigma^2]$ has two parameters: the mean μ and the variance σ^2. It can be written in the form of an exponential family with $[\eta_1(\boldsymbol{\theta}), \eta_2(\boldsymbol{\theta})] = (\mu/\sigma^2, -1/2\sigma^2)$, $[\Phi_1(x), \Phi_2(x)] = (x, x^2)$, $h(x) = (2\pi)^{-1/2}$, and $A(\mu, \sigma) = \mu^2/2\sigma^2 + \ln\sigma$. Its FIM is given by $(ds)^2 = \sigma^{-2}[(d\mu)^2 + 2(d\sigma)^2]$.

Known embeddings. The Gaussian distribution FIM has a close resemblance to the Poincaré half-plane metric $(ds)^2 = y^{-2}[(dx)^2 + (dy)^2]$, both of which have a constant negative scalar curvature: -1 and -2, respectively. In differential geometry, it is known [36] that the Poincaré half plane has an isometric canonical embedding into $(2+1)$-dimensional Minkowski space and takes the form of an imaginary sphere with radius squared equal to minus one. By rescaling, the corresponding embedding for the Gaussian fit manifold is therefore an imaginary sphere of radius squared equal to -2. Its spacelike components are given by $X_1^+(\mu, \sigma) = (\mu^2 + 2\sigma^2 + 2)/2\sqrt{2}\sigma^2$, $X_2^+(\mu, \sigma) = \mu/\sigma$ and its timelike component is given by $X_3^-(\mu, \sigma) = (\mu^2 + 2\sigma^2 - 2)/(2\sqrt{2}\sigma^2)$. The pairwise distance which generates such an embedding is therefore

$$D^2(\mu_1, \sigma_1, \mu_2, \sigma_2) = \frac{(\mu_1 - \mu_2)^2 + 2(\sigma_1 - \sigma_2)^2}{2\sigma_1\sigma_2}. \tag{34}$$

However, there is no obvious way of writing Eq. (34) in terms of $P(x|\mu, \sigma)$.

With the isKL embedding, the Gaussian distribution can be isometrically embedded into $(2+2)$ dimensions. As $\langle \Phi_1(x) \rangle = \mu$ and $\langle \Phi_2(x) \rangle = \mu^2 + \sigma^2$, the pairwise distance is given by

$$\begin{aligned} D_{sKL}^2&(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \\ &= \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right)(\mu_1 - \mu_2) \\ &\quad - \frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)(\mu_1^2 + \sigma_1^2 - \mu_2^2 - \sigma_2^2). \end{aligned} \tag{35}$$

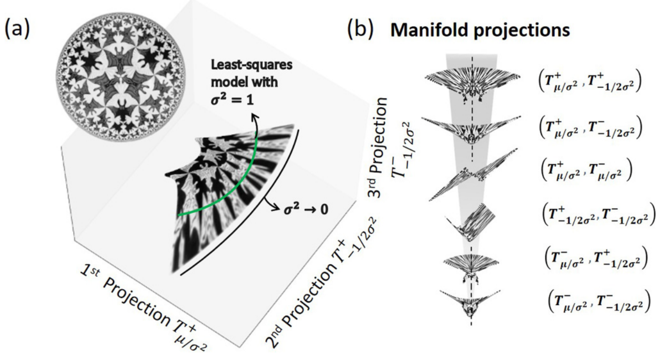
(a)
Least-squares model with $\sigma^2 = 1$
$\sigma^2 \to 0$
1st Projection T^+_{μ/σ^2}
2nd Projection $T^+_{-1/2\sigma^2}$
3rd Projection $T^-_{-1/2\sigma^2}$
(b) Manifold projections
$(T^+_{\mu/\sigma^2}, T^+_{-1/2\sigma^2})$
$(T^+_{\mu/\sigma^2}, T^-_{-1/2\sigma^2})$
$(T^+_{\mu/\sigma^2}, T^-_{\mu/\sigma^2})$
$(T^+_{-1/2\sigma^2}, T^-_{-1/2\sigma^2})$
$(T^-_{\mu/\sigma^2}, T^+_{-1/2\sigma^2})$
$(T^-_{\mu/\sigma^2}, T^-_{-1/2\sigma^2})$

FIG. 6. Viewing "Heaven and Hell" in Minkowski space. The Gaussian fit manifold is embedded into $(2+2)$ dimensions with isKL embedding. (a) The three-dimensional projection of the Gaussian fit manifold is decorated with Escher's art, Circle Limit IV, which is also known as "Heaven and Hell." The submanifold of a least-squares model with a single Gaussian distribution of fixed $\sigma^2 = 1$ is depicted in green. (b) The manifold projections are depicted in a descending order based on the manifold widths along the spacelike/timelike components. The spacelike directions are denoted by T^+ and the timelike directions are denoted by T^-. The analytical expression for each axis is given by Eq. (36). Reflection symmetry is illustrated with a dashed line along projections with a μ/σ^2 component.

Letting $\overline{\langle\eta\rangle} = \langle\eta\rangle$ and $\overline{\langle\Phi\rangle} = \langle\Phi\rangle$, the coordinates are given by

$$T^{\pm}_{\mu/\sigma^2}(\mu, \sigma^2) = \frac{1}{2}\left[\lambda_{\mu/\sigma^2}\left(\frac{\mu}{\sigma^2} - \left\langle\frac{\mu}{\sigma^2}\right\rangle\right) \pm \frac{1}{\lambda_{\mu/\sigma^2}}(\mu - \langle\mu\rangle)\right],$$

$$T^{\pm}_{-1/2\sigma^2}(\mu, \sigma^2) = \frac{1}{2}\left[\lambda_{-1/2\sigma^2}\left(\frac{1}{\sigma^2} - \left\langle\frac{1}{\sigma^2}\right\rangle\right) \pm \frac{1}{\lambda_{-1/2\sigma^2}}\left(\mu^2 + \sigma^2 - \langle\mu^2 + \sigma^2\rangle\right)\right]. \quad (36)$$

Upon closer inspection, the coordinate pairs can be written as

$$\left(T^+_{\mu/\sigma^2} - C^+_{\mu/\sigma^2}\right)^2 - \left(T^-_{\mu/\sigma^2} - C^-_{\mu/\sigma^2}\right)^2 - \left(T^+_{-1/2\sigma^2} - C^+_{-1/2\sigma^2}\right)^2 + \left(T^-_{-1/2\sigma^2} - C^-_{-1/2\sigma^2}\right)^2 = 1, \quad (37)$$

where $C^{\pm}$ are constants. This suggests the isKL embedding is a four-dimensional hyperboloid in Minkowski space. To get a good pictorial sense of how the probability distributions are arranged, we embedded "Heaven and Hell" (Escher's Circle Limit IV 1960, depicting a Poincaré disk) in Minkowski space via our isKL embedding. Figure 6(a) shows the three-dimensional projection of the manifold and, in Fig. 6(b), the manifold projections along the spacelike (black) and timelike (red) axes are to scale and accurately capture the manifold widths. The probabilistic manifold projection along the $(T^+_{\mu/\sigma^2}, T^+_{-1/2\sigma^2})$, $(T^+_{\mu/\sigma^2}, T^-_{-1/2\sigma^2})$, $(T^+_{-1/2\sigma^2}, T^-_{\mu/\sigma^2})$, and $(T^-_{-1/2\sigma^2}, T^-_{\mu/2\sigma^2})$ components exhibit a reflection symmetry about $\mu = 0$, manifesting the even-parity coordinates. Moreover, the bats become stretched as $\sigma^2 \to 0$, along the projected edge of the Poincaré disk. The submanifold of a least-squares model with a single Gaussian distribution of fixed $\sigma^2 = 1$ from Sec. II in shown in green.

F. 2D Ising model

Most statistical mechanics models form exponential families, and of particular interest is the behavior of their model manifolds near phase transitions. Here we show how the two-dimensional Ising model manifold is embedded using our method. The Ising model is a model of magnetism comprised of a lattice of n spins that can take the values ±1, "pointing up" or "pointing down." At temperature β^{-1} and in an external magnetic field H, the probability of observing a particular configuration $\boldsymbol{s} = (s_1, \ldots, s_n)$ of the spins is given by the Boltzmann distribution,

$$P(\boldsymbol{s}|\beta, h) = \frac{\exp\left(\beta \sum_{\langle ij\rangle} s_i s_j + h \sum_i s_i\right)}{Z(\beta, h)}, \quad (38)$$

where $h = \beta H$, $\langle ij\rangle$ denotes a sum over neighboring sites, and the partition function $Z(\beta, h)$ normalizes the distribution. The Ising model is an exponential family with $[\eta_1(\boldsymbol{\theta}), \eta_2(\boldsymbol{\theta})] = (\beta, h)$, $[\Phi_1(\boldsymbol{s}), \Phi_2(\boldsymbol{s})] = (\sum_{\langle ij\rangle} s_i s_j, \sum_i s_i)$, $h(\boldsymbol{s}) = 1$, and $A(\boldsymbol{\theta}) = \ln Z$. The Fisher information metric is given by the mixed partial derivatives $g_{ij} = \partial_i \partial_j \ln Z$, with $i, j \in \{\beta, h\}$.

Known embeddings. The Hellinger embedding of the Ising model manifold is 2^n dimensional. The curse of dimensionality manifests through an increase of "wrapping" around the unit hypersphere as the number of spins increases, rendering low-dimensional projections increasingly useless for visualization [4]. The wrapping phenomenon can be ameliorated by using the InPCA embedding. Though InPCA still embeds the Ising model manifold in a high-dimensional Minkowski space, the length scales of adjacent principal components are well separated.

The isKL embeds the Ising model manifold into $(2+2)$ dimensions. Not only is the curse of dimensionality broken, the Ising model manifold is embedded into *finite*-dimensional Minkowski space. The expectation values of the sufficient statistics can be related directly to the Ising average energy E and magnetization M by $(\langle\Phi_1\rangle, \langle\Phi_2\rangle) = (HM - E, M)$. The pairwise distance is then

$$D^2_{sKL}(\beta_1, \beta_2, h_1, h_2) = (\beta_2 - \beta_1)(M_2 h_2/\beta_2 - E_2 - M_1 h_1/\beta_1 + E_1) + (h_2 - h_1)(M_2 - M_1). \quad (39)$$

The Ising model manifold is centered at the critical point $(\beta, h) = (\beta_c, 0)$, with the projection coordinates being

$$T^{\pm}_{\beta} = \frac{1}{2}\left[\lambda_\beta(\beta - \beta_c) \pm \frac{1}{\lambda_\beta}(Mh/\beta - E + E_c)\right],$$

$$T^{\pm}_{h} = \frac{1}{2}\left(\lambda_h h \pm \frac{1}{\lambda_h} M\right), \quad (40)$$

where E_c is the average energy at the critical point. Figure 7 shows the isKL embedding of the 2D Ising manifold with E and M estimated from Monte Carlo simulations at $n = 128 \times 128$ spins using a rejection-free variant of the Wolff algorithm in an external field [37]. The exact solution for the zero field is included in the embedding as well and is illustrated with a black line [38,39]. For completeness, we also show all the manifold projections. The first and third principal components are fieldlike directions, and the second and the fourth components are temperaturelike directions. Reflection

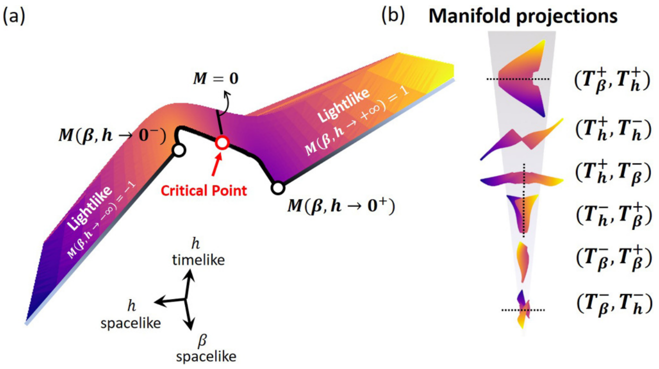

FIG. 7. Two-dimensional Ising model isKL embedding is used to illustrate the geometric structure of statistical models with a phase transition. The Ising model manifold is embedded into $(2+2)$ dimensions. (a) The three-dimensional projection of the Ising model manifold is colored based on the external magnetic field h. For $\beta > \beta_c$, there is an opening on the manifold due to the spontaneous magnetization. The two illustrated arms correspond to magnetization $M(\beta, h) = \pm 1$ with $\beta > \beta_c$ and are lightlike. The values of Ising average energy E and magnetization M used were estimated from simulations with $n = 128 \times 128$ spins. The exact solution at zero field is depicted by the black line. (b) The Ising model manifold projections are shown in a descending order based on the manifold widths along the spacelike/timelike directions. The spacelike directions are denoted by T^+ and the timelike directions are denoted by T^-. The analytical expression for each axis of projection is given by Eq. (40). Reflection symmetry is illustrated with a dotted line along projections with a external magnetic field h component.

symmetry along $H = 0$ is depicted with a dotted line. This observation is further highlighted by having the Ising model manifold colored based on the external magnetic field h.

At the critical point, there is an opening that corresponds to the growing spontaneous magnetization. This resolves a serious-seeming problem with any embedding based on the Fisher information metric. The FIM can be written in terms of the free energy, and the free energies for the two zero-field branches $\pm M(T)$ agree: the two magnetizations are zero distance apart, even though they manifestly are far apart in probability space. Any Euclidean embedding will place them at the same point. The embedding in Minkowski space resolves this: the zero distributional distance manifests itself in a large, physically sensible opening in the embedding, along a line of lightlike separation. This highlights the crucial role of timelike coordinates in qualitatively differentiating unlike systems that have the same free energy. This is not the whole story of lightlike separations, however: the two arms highlighted at large β in Fig. 7 are also lightlike. These have a more conventional interpretation: for sufficiently high field, the configuration with all spins in the direction of the field becomes the most probable and the resulting distributions are difficult to distinguish. The isKL spreads these points out as well.

The connection between phase transitions and differential geometry has been widely investigated [40–43]. Researchers have argued that the scalar curvature R can be viewed as a measurement of interactions and that the divergence of the scalar curvature signals a phase transition. The leading singularity in the scalar curvature of the 2D Ising model manifold as the critical point is approached can be computed from the metric above and the asymptotic scaling form $-\log Z \simeq t^2 \mathcal{F}(ht^{-15/8}) + t^2 \log t^2$ for $t = \beta_c - \beta$ to be $R \sim -t^{-2}/\log(t^2)$. For small $\beta - \beta_c$, R diverges. Near the critical point, one might expect to see a cusp as a result. Instead, there is an opening near the critical point in our embedding and the surrounding manifold looks smooth. The identification of each point along the opening with an opposing point suggests that we may have disguised the cusp in our embedding by "cutting" the manifold with lightlike displacements, the way one might remove the point of a cone by cutting up the side. The connection between the geometry of our manifold and the singularity of its scalar curvature will be further explored in future work.

VI. NONEXPONENTIAL FAMILIES: CAUCHY DISTRIBUTION

The success of the isKL embedding in obtaining an analytical expression for each coordinate is special to exponential family distributions. As an example of a nonexponential family, we consider the long-tailed Cauchy distribution,

$$P(x|x_0, \gamma) = \frac{\gamma}{\pi[\gamma^2 + (x - x_0)^2]}. \tag{41}$$

Interestingly, its FIM, $(ds)^2 = (2\gamma^2)^{-1}[(dx_0)^2 + (d\gamma)^2]$, has a constant negative scalar curvature just as the Gaussian fit in Sec. V E. In fact, there is a deeper connection between the Gaussian and Cauchy distributions: they both belong to the location scale family distributions, $f(x) = c^{-1} f((x - \delta)/c)$, where δ is the location parameter and c is the scale parameter. It is known that any location scale distribution has a constant negative curvature [44]. That the Gaussian and Cauchy distributions share this property but are distinct indicates that being locally isometric is not enough to distinguish them. This demands the use of a global distance as an additional measure to characterize the model manifold. We embed the Cauchy distribution manifold using the isKL embedding with the distance measure [45], which gives

$$D^2_{sKL}(x_1, \gamma_1, x_2, \gamma_2) = 2 \ln \left[\frac{(\gamma_1 + \gamma_2)^2 + (x_1 - x_2)^2}{4\gamma_1\gamma_2} \right]. \tag{42}$$

Interestingly, the isKL embedding returns a Euclidean embedding for the Cauchy manifold (Fig. 8) to the number of components we have explored. To compare it with the Gaussian fits manifold, we have colored the Cauchy manifold with Escher's art, Circle Limit IV, as well. Here, we observe well-preserved bat shapes as compared to Fig. 6. Strikingly, not only is this also true for any symmetrized Rényi choices, as shown in Fig. 8(b), but the projections obtained from different symmetrized Rényi choices appear to be virtually the same. Thus, D^2_{sKL} is not obviously better than other intensive Rényi divergences for models not in exponential families.

VII. SUMMARY

In this paper, we demonstrate that any N-parameter probabilistic model that takes the form of an exponential family can be embedded isometrically into a low-dimensional $(N + N)$

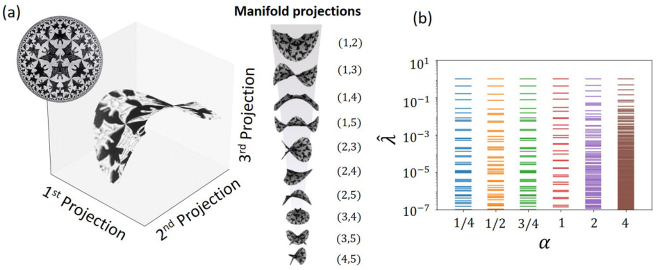

FIG. 8. The Cauchy distribution is considered to exemplify the rough equivalence of the isKL embedding with various other Minkowski embeddings for visualizing non-exponential-family distributions. (a) The three-dimensional projection of the Cauchy distribution manifold is shown on the left. To compare it with the Gaussian fits manifold, we have colored the Cauchy manifold with Escher's art, Circle Limit IV. Here, the bat shapes are well preserved as compared to Fig. 6. The first five manifold projections are shown on the right in a descending order based on the manifold widths along the (m, n) principal components. (b) Squared principal length of intensive embedding with different symmetrized Rényi divergences for the Cauchy manifold. Here, we observe geometrically decreasing manifold widths that spans many decades for all α's.

Minkowski space via the isKL embedding technique. This is done by using the symmetrized Kullback-Leibler divergence (sKL) as the pairwise distance between model predictions. This could potentially be used to determine the number of parameters needed to describe an experiment or a simulation should the underlying distribution belong to the exponential family. To illustrate how the isKL embedding technique can be used to visualize the exponential family probabilistic manifold in a simple and tractable way, we consider the Bernoulli (coin-toss) problem, the ideal gas, the n-sided die, the nonlinear least-squares models, Gaussian fits to data, and the two-dimensional Ising model. Additionally, we use the nonexponential Cauchy distribution to illustrate the importance of preserving both global and local structures in embeddings.

ACKNOWLEDGMENTS

We thank Pankaj Mehta and Anirvan Sengupta for suggesting the possible importance of MDS. H.K.T. was supported by the Army Research Office through ARO Grant No. W911NF-18-1-0032. H.K.T., J.K.-D., C.C.B., and J.P.S. was supported by the National Science Foundation through Grant No. NSF DMR-1719490. K.N.Q. was supported by a fellowship from the Natural Sciences and Engineering Research Council of Canada.

APPENDIX A: REPLICA ZERO LIMIT OF f DIVERGENCE

To visualize the underlying geometry of probabilistic model data, a distance measure in probability space is needed. In this Appendix, we will generalize the limit of zero data procedure in obtaining an intensive distance measure to a family of divergences, specifically from f divergence to Rényi divergence. f divergence measures the difference between two probability distributions P and Q with a convex function f such that $f(1) = 0$ and takes the form

$$D_f(P, Q) = \int f\left(\frac{p(x)}{q(x)}\right) q(x) d\mu(x). \tag{A1}$$

By assuming f is analytic [46], we can Taylor expand it about $x = 1$, $f(x) = \sum_{m=0}^{\infty} \frac{1}{m!} f^{(m)}(1)(x-1)^m$. Thus, f divergence takes the form

$$\begin{aligned} D_f(P, Q) &= \int f\left(\frac{p(x)}{q(x)}\right) q(x) dx \\ &= \sum_{m=0}^{\infty} \int \frac{1}{m!} f^{(m)}(1) \left[\frac{p(x)}{q(x)} - 1\right]^m q(x) dx \\ &= \sum_{m=0}^{\infty} \frac{1}{m!} f^{(m)}(1) \chi^m_{1,q}(P, Q), \end{aligned} \tag{A2}$$

where

$$\chi^m_{1,q}(P, Q) = \int \frac{[p(x) - q(x)]^m}{q^{m-1}(x)} dx \tag{A3}$$

is the χ^k divergence with parameter 1. Expanding the polynomial and simplifying,

$$\begin{aligned} \chi^m_{1,q}(P, Q) &= \int \sum_{k=0}^{m} \binom{m}{k} (-1)^{m-k} q^{1-k}(x) p^k(x) dx \\ &= \sum_{k=0}^{m} \binom{m}{k} (-1)^{m-k} \int q^{1-k}(x) p^k(x) dx. \end{aligned} \tag{A4}$$

Suppose we increase the number of data samples by N, which amounts to having an N-replicated system,

$$\begin{aligned} \chi^m_{1,q}(P_N, Q_N) &= \sum_{k=0}^{m} \binom{m}{k} (-1)^{m-k} \left[\int \dots \int q^{1-k}(x_1, \dots, x_N) p^k(x_1, \dots, x_N) dx_1 \dots dx_N\right] \\ &\qquad \Big| \text{ Since } p(x_1, \dots, x_N) = \prod_{i=1}^{N} p(x_i) \text{ and } q(x_1, \dots, x_N) = \prod_{i=1}^{N} q(x_i), \\ &= \sum_{k=0}^{m} \binom{m}{k} (-1)^{m-k} \left[\int q^{1-k}(x) p^k(x) dx\right]^N \\ &= \sum_{k=0}^{m} \binom{m}{k} (-1)^{m-k} \left\{\left[\int q^{1-k}(x) p^k(x) dx\right]^N - 1\right\} + \sum_{k=0}^{m} \binom{m}{k} (-1)^{m-k}. \end{aligned}$$

$$\Bigg|\ \text{Note that } (1-x)^n = \sum_{n=0}^{\infty}\binom{n}{k}(-x)^n,\ \text{ so} \sum_{k=0}^{m}\binom{m}{k}(-1)^{m-k} = 0.$$

$$= \sum_{k=0}^{m}\binom{m}{k}(-1)^{m-k}\left\{\left[\int q^{1-k}(x)p^{k}(x)dx\right]^{N} - 1\right\}. \tag{A5}$$

Upon closer inspection, each χ^m term contains partition-function-like terms $(\int q^{1-k}p^k dx)^N$ that are known as Hellinger divergence of the order of k that increase geometrically with N. Upon sending N continuously to zero, we have

$$\lim_{N\to 0}\frac{\chi^{m}_{1,q}(P_N, Q_N)}{N} = \sum_{k=0}^{m}\binom{m}{k}(-1)^{m-k}\ln\left[\int q^{1-k}(x)p^{k}(x)dx\right]. \tag{A6}$$

As $D_\alpha(P,Q) = \frac{1}{\alpha-1}\ln(\int p^\alpha q^{1-\alpha}dx)$ is the Rényi divergence,

$$\lim_{N\to 0}\frac{\chi^{m}_{1,q}(P_N, Q_N)}{N} = \sum_{k=0}^{m}\binom{m}{k}(-1)^{m-k}(k-1)D_k(P,Q). \tag{A7}$$

Thus, for any f divergences,

$$\lim_{N\to 0}\frac{D_f(P_N, Q_N)}{N} = \sum_{m=1}^{\infty}\sum_{k=0}^{m}\frac{f^{(m)}(1)}{m!}\binom{m}{k}(-1)^{m-k}(k-1)D_k(P,Q). \tag{A8}$$

APPENDIX B: CONNECTIONS TO PRINCIPAL COMPONENT ANALYSIS (PCA) AND MULTIDIMENSIONAL SCALING (MDS)

The interested reader will note a connection to both principal component analysis (PCA) [47] and multidimensional scaling (MDS) [48]. Principal component analysis uses the isometries of Euclidean space to optimally display data in a space of many dimensions. PCA translates the data to center it, then uses singular-value decomposition to rotate and diagonalize the "moment of inertia" tensor of the data set. The data remain many dimensional, but PCA allows one to examine the directions for which the data varies the most. The principal components are the orthogonal directions which best describe the data set—minimizing the sum of squared distances of the remaining data from an approximation restricted to the subspace that they span.

Multidimensional scaling generalizes these ideas to situations where the data vectors are not known, but some measure of the pairwise distance is available. MDS generates an isometric embedding maintaining the pairwise distances, usually in a vector space of dimension equal to the number of data points. Again, this manifold can rotate or translate for a given system, depending on the sampling used. Indeed, the eigensystem solved in MDS often has negative eigenvalues [49–51] corresponding to timelike coordinates, and changing the sampling can also induce Lorentz boosts. MDS, using the symmetrized Kullback-Leibler divergence D^2_{sKL} as the pairwise distance, in fact produces an isKL embedding [52]. Our main result [Eq. (13)] implies that MDS applied with D^2_{sKL} to high-dimensional data produced by an N-parameter exponential family will embed its predictions in a much smaller space, with only N spacelike and N timelike nonzero coordinates. Furthermore, the resulting manifold will be given by the explicit isKL embedding of Eq. (13) up to isometries.

We can now establish a connection with the multidimensional scaling (MDS) technique. Given n sampled points from the parameter space, MDS generates an embedding whose ith projection is given by $\sqrt{\Lambda_i}v_i$, where Λ_i and v_i are the eigenvalue and eigenvector of the double mean centered pairwise distance matrix, $D^2_c = -(1/2)PD^2P$, where $P_{i,j} = 1/n - \delta_{i,j}$ and D^2 is the pairwise distance matrix. Writing out the matrix explicitly, we have $(D^2_c)_{i,j} = -\frac{1}{2}[D^2_{i,j} + \frac{1}{n^2}\sum_{k,k'}D^2_{k,k'} - \frac{1}{n}\sum_k(D^2_{i,k} + D^2_{k,j})]$. We will solve for the eigensolutions in a more general setting by taking a continuous sampling limit. This yields an integral eigenvalue problem,

$$\int D^2_c(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}})v(\boldsymbol{\theta})d\mu(\boldsymbol{\theta}) = \Lambda v(\tilde{\boldsymbol{\theta}}), \tag{B1}$$

with

$$D^2_c(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) = -\frac{1}{2}\Bigg[D^2(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) - \int D^2(\boldsymbol{\theta},\boldsymbol{\xi})d\mu(\boldsymbol{\xi}) - \int D^2(\boldsymbol{\xi},\tilde{\boldsymbol{\theta}})d\mu(\boldsymbol{\xi}) + \iint D^2(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}})d\mu(\boldsymbol{\theta})d\mu(\tilde{\boldsymbol{\theta}})\Bigg], \tag{B2}$$

where $d\mu(\boldsymbol{\theta})$ is the sampling measure, v is the eigenfunction, and Λ is the eigenvalue. One can recover MDS by having a discrete measure $\mu(\boldsymbol{\theta}) = \sum_x c_x\delta_x(\boldsymbol{\theta})d\boldsymbol{\theta}$, where $\delta_x(\boldsymbol{\theta})$ is the Dirac measure. For D^2_{sKL}, the double mean centered distance measure takes the form

$$D^2_c(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) = \frac{1}{2}\sum_i[(\eta_i(\tilde{\boldsymbol{\theta}}) - \overline{\eta}_i)(\langle\Phi_i\rangle_{\boldsymbol{\theta}} - \overline{\langle\Phi_i\rangle}) + (\eta_i(\boldsymbol{\theta}) - \overline{\eta}_i)(\langle\Phi_i\rangle_{\tilde{\boldsymbol{\theta}}} - \overline{\langle\Phi_i\rangle})], \tag{B3}$$

where $\int\eta_i(\boldsymbol{\theta})d\mu(\boldsymbol{\theta}) = \overline{\eta}_i$ and $\int\langle\Phi_i\rangle_{\boldsymbol{\theta}}d\mu(\boldsymbol{\theta}) = \overline{\langle\Phi_i\rangle}$. It turns out that the coordinates $\mathcal{S}_i$ and $\mathcal{T}_i$ discussed in Sec. IV,

$$\begin{aligned}\mathcal{S}_i(\boldsymbol{\theta}) &= \frac{1}{2}\left\{\lambda_i[\eta_i(\boldsymbol{\theta}) - \overline{\eta_i}] + \frac{1}{\lambda_i}\big(\langle\Phi_i\rangle_{\boldsymbol{\theta}} - \overline{\langle\Phi_i\rangle}\big)\right\},\\ \mathcal{T}_i(\boldsymbol{\theta}) &= \frac{1}{2}\left\{\lambda_i[\eta_i(\boldsymbol{\theta}) - \overline{\eta_i}] - \frac{1}{\lambda_i}\big(\langle\Phi_i\rangle_{\boldsymbol{\theta}} - \overline{\langle\Phi_i\rangle}\big)\right\},\end{aligned} \tag{B4}$$

where $\lambda_i^2 = \sqrt{\mathrm{Var}(\langle\Phi_i\rangle)/\mathrm{Var}(\eta_i)}$, are indeed the solutions to Eq. (B1), with the ith eigenvalue pairs being $\Lambda_i^{\pm} = \frac{1}{2}[\mathrm{Cov}(\eta_i,\langle\Phi_i\rangle) \pm \sqrt{\mathrm{Var}(\eta_i)\mathrm{Var}(\langle\Phi_i\rangle)}]$. Here we will prove it

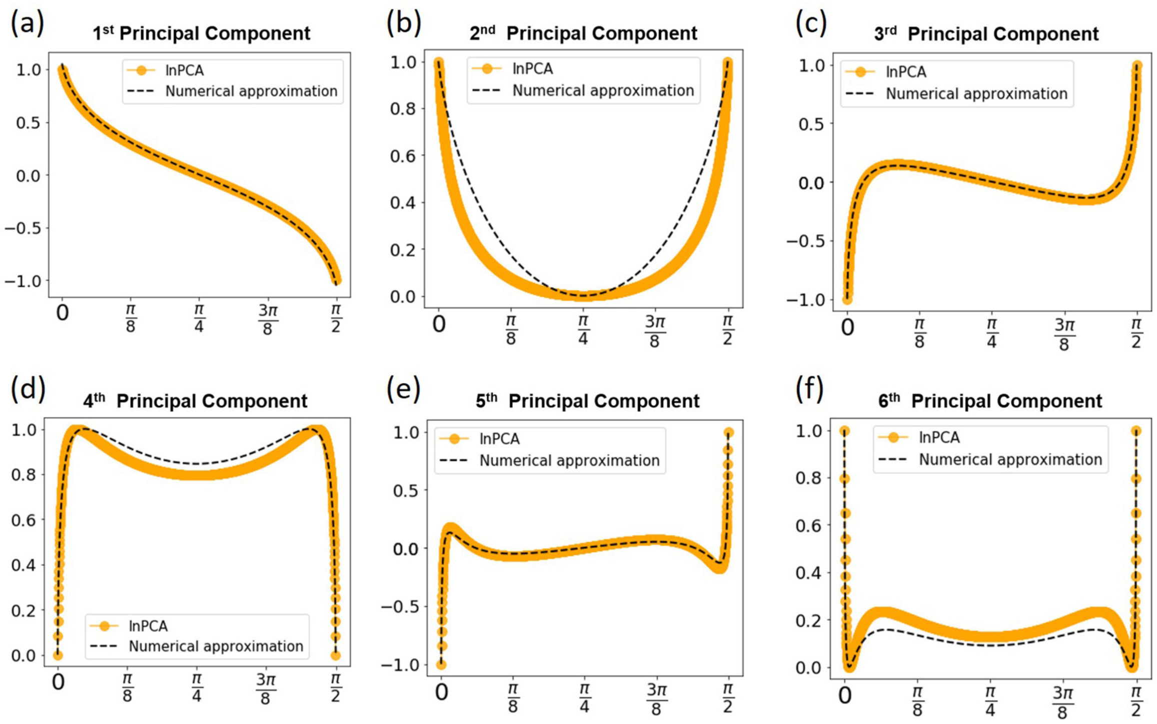

FIG. 9. (a)-(f) Normalized projection of the coin-toss manifold onto the first six principal axes. The dashed line is the numerical approximation of the analytical expressions given in Eq. (C6) and Eq. (C7) with $N = 2000$.

as follows:

$$\int D_c(\boldsymbol{\theta}.\tilde{\boldsymbol{\theta}})\frac{1}{2}\left\{\lambda_i[\eta_i(\boldsymbol{\theta}) - \overline{\eta}_i] \pm \frac{1}{\lambda_i}\big(\langle\Phi_i\rangle_{\boldsymbol{\theta}} - \overline{\langle\Phi_i\rangle}\big)\right\}d\mu(\boldsymbol{\theta}).$$

$$\Bigg|\ \text{Letting} \int \eta_i(\boldsymbol{\theta})\langle\Phi_i\rangle_{\boldsymbol{\theta}} d\mu(\boldsymbol{\theta}) = \overline{\langle\Phi_i\rangle\eta_i}, \quad \int \eta_i^2(\boldsymbol{\theta})d\mu(\boldsymbol{\theta}) = \overline{\eta_i^2}, \text{ and} \int \langle\Phi_i\rangle_{\boldsymbol{\theta}}^2 d\mu(\boldsymbol{\theta}) = \overline{\langle\Phi_i^2\rangle}:$$

$$= \frac{1}{4}\left\{\lambda_i[\eta_i(\tilde{\boldsymbol{\theta}}) - \overline{\eta_i}](\overline{\langle\Phi_i\rangle\eta_i} - \overline{\langle\Phi_i\rangle}\cdot\overline{\eta_i}) + \lambda_i(\langle\Phi_i\rangle_{\tilde{\boldsymbol{\theta}}} - \overline{\langle\Phi_i\rangle})\big(\overline{\eta_i^2} - \overline{\eta_i}^2\big)\right\}$$

$$\pm \frac{1}{4}\left\{\frac{1}{\lambda_i}[\eta_i(\tilde{\boldsymbol{\theta}}) - \overline{\eta_i}](\overline{\langle\Phi_i^2\rangle} - \overline{\langle\Phi_i\rangle}^2) + \frac{1}{\lambda_i}(\langle\Phi_i\rangle_{\tilde{\boldsymbol{\theta}}} - \overline{\langle\Phi_i\rangle})(\overline{\langle\Phi_i\rangle\eta_i} - \overline{\langle\Phi_i\rangle}\cdot\overline{\eta_i})\right\}.$$

$$\Bigg|\ \text{Rewriting } \overline{\langle\Phi_i\rangle\eta_i} - \overline{\langle\Phi_i\rangle}\cdot\overline{\eta_i} = \text{Cov}(\eta_i, \langle\Phi_i\rangle), \quad \overline{\eta_i^2} - \overline{\eta_i}^2 = \text{Var}(\eta_i), \text{ and } \overline{\langle\Phi_i^2\rangle} - \overline{\langle\Phi_i\rangle}^2 = \text{Var}(\langle\Phi_i\rangle):$$

$$= \frac{1}{4}\left[\text{Cov}(\eta_i, \langle\Phi_i\rangle) \pm \frac{1}{\lambda_i^2}\text{Var}(\langle\Phi_i\rangle)\right]\lambda_i\big(\eta_i(\tilde{\boldsymbol{\theta}}) - \overline{\eta}\big) \pm \frac{1}{4}[\text{Cov}(\eta_i, \langle\Phi_i\rangle) \pm \lambda_i^2\text{Var}(\eta_i)]\frac{1}{\lambda_i}(\langle\Phi_i\rangle_{\tilde{\boldsymbol{\theta}}} - \overline{\langle\Phi_i\rangle}).$$

$$\Bigg|\ \text{Since } \lambda_i^2 = \sqrt{\text{Var}(\langle\Phi_i\rangle)/\text{Var}(\eta_i),}$$

$$= \frac{1}{2}[\text{Cov}(\eta_i, \langle\Phi_i\rangle) \pm \sqrt{\text{Var}(\eta_i)\text{Var}(\langle\Phi_i\rangle)}]\frac{1}{2}\left\{\lambda_i[\eta_i(\tilde{\boldsymbol{\theta}}) - \overline{\eta_i}] \pm \frac{1}{\lambda_i}(\langle\Phi_i\rangle_{\tilde{\boldsymbol{\theta}}} - \overline{\langle\Phi_i\rangle})\right\}. \tag{B5}$$

As promised, $\mathcal{S}_i(\boldsymbol{\theta})$ and $\mathcal{T}_i(\boldsymbol{\theta})$ are indeed the solutions to Eq. (B1), with eigenvalues

$$\Lambda_i^{\pm} = \frac{1}{2}\Big[\text{Cov}(\eta_i, \langle\Phi_i\rangle) \pm \sqrt{\text{Var}(\eta_i)\text{Var}(\langle\Phi_i\rangle)}\Big]. \tag{B6}$$

In general, when the eigenvalues are degenerate, the axes of the projections are free to rotate within the degenerate spacelike and timelike subspaces, depending on $d\mu$. Hence, the solution will be a linear combination of the degenerate coordinates described in Eq. (B1), i.e., $\mathcal{S}'(\boldsymbol{\theta}) = \sum_k \alpha_k\mathcal{S}_k(\boldsymbol{\theta})$ and $\mathcal{T}'(\boldsymbol{\theta}) = \sum_k \beta_k\mathcal{T}_k(\boldsymbol{\theta})$, where $\sum_k \alpha_k^2 = 1$ and $\sum_k \beta_k^2 = 1$ and the index k runs over coordinates that share the same eigenvalue. In all of our examples except the generalized die, the symmetry keeps the rotations from mixing directions and the projection coordinates can be calculated from Eq. (B1) regardless of degeneracy.

APPENDIX C: COIN TOSS AND INPCA: THE BERNOULLI PROBLEM MODEL MANIFOLD EMBEDDED WITH THE BHATTACHARYYA DISTANCE

In the Bernoulli problem, the inPCA embedding is given by the following pairwise distance:

$$d^2(\theta_1,\theta_2)=\ln[\cos(\theta_1-\theta_2)]. \tag{C1}$$

To find the embedding, we need to solve the eigenvalue problem discussed in Sec. B. As the double mean centering matrix P gives rotation and boost transformation to the coordinates, for simplicity we proceed our calculation for each projection with just our distance function as an infinite matrix, acting on continuous variables ϕ and θ: $\ln\cos(\phi-\theta)$. This implies the evaluation of the following eigenvalue problem:

$$\int_0^{\pi/2}\ln\cos(\phi-\theta)v_\alpha(\theta)d\theta=\lambda_\alpha v_\alpha(\phi), \tag{C2}$$

where $v_\alpha(\phi)$ are the eigenfunctions with the corresponding eigenvalues λ_α. We solve this numerically by expanding the pairwise distance function in terms of Chebyshev polynomials, $d^2(\theta,\phi)=-\ln(2)+\sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k}\cos[2k(\theta-\phi)]$, and assuming that the eigenfunction $v_\alpha(\theta)$ is odd with respect to $\theta=\pi/4$ and can be expanded as a Fourier series: $\sum_{k=1}^{\infty}b_k\sin[k(\theta-\frac{\pi}{4})]$. Thus we have

$$\sum_{k,m=1}^{\infty}(-1)^{k+1}\frac{b_m}{k}F(\phi)=\lambda_\alpha\sum_{k=1}^{\infty}b_k\sin\left[k\left(\theta-\frac{\pi}{4}\right)\right], \tag{C3}$$

with $F(\phi)=\int_0^{\pi/2}d\theta\cos[2k(\theta-\phi)]\sin[m(\theta-\frac{\pi}{4})]$, where, as $F(\phi)$ only produces terms containing $\sin[2k(\phi-\frac{\pi}{4})]$ and $\cos[2k(\phi-\frac{\pi}{4})]$ for all values of $m\in\mathbb{Z}^+$, it is thus natural to conjecture that the Fourier-series expansion must have its coefficient $b_{2k+1}=0$. Hence,

$$v_\alpha(\theta)=\sum_{k=1}^{\infty}b_{2k}\sin\left[2k\left(\theta-\frac{\pi}{4}\right)\right]. \tag{C4}$$

With this assumption, the eigenvalue equation simplifies into matching the coefficient of each Fourier mode $\sin[2k(\phi-\pi/4)]$,

$$\sum_{m=1}^{\infty}\xi(k,m)b_{2m}=\lambda_\alpha b_{2k}, \tag{C5}$$

or, more succinctly, $\xi\vec{b}=\lambda_\alpha\vec{b}$, where $\vec{b}=(b_2,b_4,\ldots,b_{2N},\ldots)$. The matrix $\xi(k,m)$ is computed via $F(\phi)$ to be

$$\xi(k,m)=\begin{cases}\frac{(-1)^{k+1}}{k}\frac{\pi}{4} & (m=k)\\ \frac{(-1)^{k+1}}{k}\frac{1}{m^2-k^2}\left[k\cos\left(\frac{k\pi}{2}\right)\sin\left(\frac{m\pi}{2}\right)-m\cos\left(\frac{m\pi}{2}\right)\sin\left(\frac{k\pi}{2}\right)\right] & (m\neq k).\end{cases} \tag{C6}$$

For even eigenfunctions $v_\alpha(\theta)=\sum_{k=0}^{\infty}c_k\cos[k(\theta-\pi/4)]$, the argument is almost identical, except we now have an extra contribution from the constant c_0 term which needs to be handled separately. Going through the same derivation, we again have the matrix eigenvalue equation, i.e., $\eta\vec{c}=\lambda_\alpha\vec{c}$, where $\vec{c}=(c_0,c_2,\ldots,c_{2N})$, and we have

$$\eta(k,n)=\begin{cases}-\frac{\pi}{2}\ln(2) & (n=k=0)\\ -\ln(2)\sin\left(\frac{n\pi}{2}\right) & (k=0,n\geqslant 1)\\ \frac{(-1)^{k+1}}{k^2}\sin\left(\frac{k\pi}{2}\right) & (k\geqslant 1,n=0)\\ \frac{(-1)^{k+1}}{k}\frac{\pi}{4} & (k=n\geqslant 1)\\ \frac{(-1)^{k+1}}{k}\frac{1}{n^2-k^2}\left[n\cos\left(\frac{k\pi}{2}\right)\sin\left(\frac{n\pi}{2}\right)-k\cos\left(\frac{n\pi}{2}\right)\sin\left(\frac{k\pi}{2}\right)\right] & (n\geqslant 1,k\geqslant 1,n\neq k).\end{cases} \tag{C7}$$

One could get a numerical approximation for the analytical calculation above by taking η and ξ to be finite-dimensional matrix $N\times N$, where $N\gg 1$, as shown in Fig. 9.

[1] M. K. Transtrum, B. B. Machta, K. S. Brown, B. C. Daniels, C. R. Myers, and J. P. Sethna, Perspective: Sloppiness and emergent theories in physics, biology, and beyond, J. Chem. Phys. **143**, 010901 (2015).

[2] M. K. Transtrum, B. B. Machta, and J. P. Sethna, Why Are Nonlinear Fits to Data So Challenging? Phys. Rev. Lett. **104**, 060201 (2010).

[3] M. K. Transtrum, B. B. Machta, and J. P. Sethna, Geometry of nonlinear least squares with applications to sloppy models and optimization, Phys. Rev. E **83**, 036701 (2011).

[4] K. N. Quinn, H. Wilber, A. Townsend, and J. P. Sethna, Chebyshev Approximation and the Global Geometry of Model Predictions, Phys. Rev. Lett. **122**, 158302 (2019).

[5] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna, Universally sloppy parameter sensitivities in systems biology models, PLoS Comput. Biol. **3**, e189 (2007).

[6] J. J. Waterfall, F. P. Casey, R. N. Gutenkunst, K. S. Brown, C. R. Myers, P. W. Brouwer, V. Elser, and J. P. Sethna, Sloppy-Model Universality Class and the Vandermonde Matrix, Phys. Rev. Lett. **97**, 150601 (2006).

[7] B. B. Machta, R. Chachra, M. Transtrum, and J. P. Sethna, Parameter space compression underlies emergent theories and predictive models, Science **342**, 604 (2013).

[8] M. K. Transtrum and J. P. Sethna, Geodesic acceleration and the small-curvature approximation for nonlinear least squares, arXiv:1207.4999 (2012).

[9] M. K. Transtrum and J. P. Sethna, Improvements to the Levenberg-Marquardt algorithm for nonlinear least-squares minimization, arXiv:1201.5885 (2012).

[10] W. F. Bergan, I. V. Bazarov, C. J. R. Duncan, D. B. Liarte, D. L. Rubin, and J. P. Sethna, Online storage ring optimization using dimension-reduction and genetic algorithms, Phys. Rev. Accel. Beams **22**, 054601 (2019).

[11] K. N. Quinn, C. B. Clement, F. De Bernardis, M. D. Niemack, and J. P. Sethna, Visualizing probabilistic models and data with intensive principal component analysis, Proc. Natl. Acad. Sci. **116**, 13762 (2019).

[12] S.-i. Amari and H. Nagaoka, *Methods of Information Geometry* (American Mathematical Society, Providence, 2007), Vol. 191.

[13] A. Bhattacharyya, On a measure of divergence between two multinomial populations, Sankhyā: Indian J. Stat. (1933–1960) **7**, 401 (1946).

[14] I. Csiszár and P. C. Shields, Information theory and statistics: A tutorial, Found. Trends Commun. Inf. Theory **1**, 417 (2004).

[15] A. Rényi *et al.*, On measures of entropy and information, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1: Contributions to the Theory of Statistics, edited by J. Neyman (University of California Press, Berkeley, CA, 1961), pp. 541–561.

[16] S. Kullback and R. A. Leibler, On information and sufficiency, Ann. Math. Stat **22**, 79 (1951).

[17] F. Nielsen and V. Garcia, Statistical exponential families: A digest with flash cards, arXiv:0911.4863.

[18] H.-P. Kriegel, P. Kröger, and A. Zimek, Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering, TKDD **3**, 1 (2009).

[19] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, When is "nearest neighbor" meaningful? in *International Conference on Database Theory*, edited by C. Beeri and P. Buneman (Springer, Berlin, Heidelberg, 1999), pp. 217–235.

[20] R. C. Wilson, E. R. Hancock, E. Pękalska, and R. P. Duin, Spherical embeddings for non-Euclidean dissimilarities, in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2010), pp. 1903–1910.

[21] M. Boguná, F. Papadopoulos, and D. Krioukov, Sustaining the internet with hyperbolic mapping, Nat. Commun. **1**, 62 (2010).

[22] M. Nickel and D. Kiela, Poincaré embeddings for learning hierarchical representations, in *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017), pp. 6338–6347.

[23] J. B. Tenenbaum, V. De Silva, and J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science **290**, 2319 (2000).

[24] M. Belkin and P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. **15**, 1373 (2003).

[25] L. van der Maaten and G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. **9**, 2579 (2008).

[26] K. R. Moon, D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. van den Elzen, M. J. Hirn, R. R. Coifman *et al.*, Visualizing structure and transitions in high-dimensional biological data, Nat. Biotechnol **37**, 1482 (2019).

[27] The potential distance between N-replicated systems is

$$\begin{aligned} D^2_{\text{pot}}(p^N, q^N) &= \sum_i \log^2\left(p_i^N/q_i^N\right) \\ &= N^2 \sum_i \log^2(p_i/q_i) \\ &= N^2 D^2_{\text{pot}}(p, q). \end{aligned}$$

Thus the potential distance per replica scales linearly with N.

[28] The potential distance between two distributions p and q, $D^2_{\text{pot}} = \sum_i \log^2(p_i/q_i)$, can be shown to disagree with the FIM locally as follows:

$$\begin{aligned} D^2_{\text{pot}}(\theta + \delta\theta, \theta) &= \sum_i \log^2(p_i(\theta + \delta\theta)/p_i(\theta)) \\ &\approx \sum_i \log^2\left(1 + \frac{\partial_\alpha p_i(\theta)}{p_i(\theta)}\delta\theta^\alpha\right) \\ &\approx \sum_i \left(\frac{\partial_\alpha p_i(\theta)}{p_i(\theta)}\right)\left(\frac{\partial_\beta p_i(\theta)}{p_i(\theta)}\right)\delta\theta^\alpha\delta\theta^\beta \\ &= \sum_i \partial_\alpha \log p_i(\theta)\partial_\beta \log p_i(\theta)\delta\theta^\alpha\delta\theta^\beta. \end{aligned}$$

This is different from the FIM that takes the form $I_{\alpha\beta}(\theta) = \sum_i p_i(\theta)\partial_\alpha \log p_i(\theta)\partial_\beta \log p_i(\theta)$.

[29] W. K. Wootters, Statistical distance and Hilbert space, Phys. Rev. D **23**, 357 (1981).

[30] S. Ito and A. Dechant, Stochastic Time-Evolution, Information Geometry and the Cramer-Rao Bound, Phys. Rev. X **10**, 021056 (2020).

[31] The pairwise distance in the 2D Euclidean embedding is $(1+N)\log^2(p_1/p_2) + (3N/2)\log^2(\beta_1/\beta_2)$, whereas the symmetrized Rényi divergence of ideal gas is $D^2_\alpha = (1-\alpha)^{-1}\{F(\theta_1) + F(\theta_2) - F[\alpha\theta_1 + (1-\alpha)\theta_2] - F[\alpha\theta_2 + (1-\alpha)\theta_2]\}$, where $\theta = (p, \beta)$ and $F(p, \beta) = (N+1)\log p + (3N/2)\log\beta$.

[32] The Gauss Bonnet theorem tells us that the integral of the torus curvature is zero; the 4D torus $\mathbb{S}^1 \times \mathbb{S}^1$ has a zero curvature.

[33] B. B. Machta, Dissipation Bound for Thermodynamic Control, Phys. Rev. Lett. **115**, 260603 (2015).

[34] J. A. Wolf, Homogeneous manifolds of constant curvature, Comment. Math. Helvetici **36**, 112 (1961/62).

[35] H. A. Morewitz and M. H. Shamos, The variation of the ratio of positive to negative cosmic-ray μ mesons with momentum and altitude, Phys. Rev. **92**, 134 (1953).

[36] B. Guan, Isometric embedding of negatively curved disks in the Minkowski space, Pure Appl. Math. Quarter. **3**, 827 (2007).

[37] J. Kent-Dobias and J. P. Sethna, Cluster representations and the Wolff algorithm in arbitrary external fields, Phys. Rev. E **98**, 063306 (2018).

[38] L. Onsager, Crystal statistics. I. A two-dimensional model with an order-disorder transition, Phys. Rev. **65**, 117 (1944).

[39] C. N. Yang, The spontaneous magnetization of a two-dimensional Ising model, Phys. Rev. **85**, 808 (1952).

[40] D. Brody and N. Rivier, Geometrical aspects of statistical mechanics, Phys. Rev. E **51**, 1006 (1995).

[41] H. Janyszek, Riemannian geometry and stability of thermodynamical equilibrium systems, J. Phys. A: Math. Gen **23**, 477 (1990).

[42] G. Ruppeiner, Thermodynamics: A Riemannian geometric model, Phys. Rev. A **20**, 1608 (1979).

[43] G. Ruppeiner, Riemannian geometry in thermodynamic fluctuation theory, Rev. Mod. Phys. **67**, 605 (1995).

[44] R. E. Kass and P. W. Vos, *Geometrical Foundations of Asymptotic Inference*, Vol. 908 (Wiley, New York, 2011).

[45] F. Chyzak and F. Nielsen, A closed-form formula for the Kullback-Leibler divergence between Cauchy distributions, arXiv:1905.10965.

[46] F. Nielsen and R. Nock, On the chi square and higher-order chi distances for approximating f-divergences, IEEE Signal Process. Lett. **21**, 10 (2013).

[47] H. Hotelling, Analysis of a complex of statistical variables into principal components., J. Educ. Psychol **24**, 417 (1933).

[48] W. S. Torgerson, Multidimensional scaling: I. Theory and method, Psychometrika **17**, 401 (1952).

[49] A. Harol, E. Pękalska, S. Verzakov, and R. P. W. Duin, Augmented embedding of dissimilarity data into (pseudo-) Euclidean spaces, in *Structural, Syntactic, and Statistical Pattern Recognition*, edited by D.-Y. Yeung, J. T. Kwok, A. Fred, F. Roli, and D. de Ridde (Springer, Berlin, Heidelberg, 2006), pp. 613–621.

[50] E. Pękalska, A. Harol, R. P. W. Duin, B. Spillmann, and H. Bunke, Non-euclidean or non-metric measures can be informative, in *Structural, Syntactic, and Statistical Pattern Recognition*, edited by D.-Y. Yeung, J. T. Kwok, A. Fred, F. Roli, and D. de Ridder (Springer, Berlin, Heidelberg, 2006), pp. 871–880.

[51] E. Pękalska, R. P. W. Duin, S. Günter, and H. Bunke, On not making dissimilarities Euclidean, in *Structural, Syntactic, and Statistical Pattern Recognition*, edited by A. Fred, T. M. Caelli, R. P. W. Duin, A. C. Campilho, and D. de Ridder (Springer, Berlin, Heidelberg, 2004), pp. 1145–1154.

[52] Also, the inPCA embedding [11] is precisely MDS applied to the Bhattacharyya distance d_{sBhat}.