Thermodynamic inference of data manifolds

Purushottam D. Dixit [*]
Department of Physics, University of Florida, Gainesville, Florida 32611, USA

(Received 18 November 2019; accepted 16 April 2020; published 20 May 2020)

The Gibbs-Boltzmann distribution offers a physically interpretable approach to massively reduce the dimensionality of high dimensional probability distributions where the extensive variables represent "features" of the state space and the intensive variables represent realization-specific "coefficients." However, not all probability distributions can be modeled using the Gibbs-Boltzmann form. Here, we present Thermodynamic Manifold Inference (TMI), a thermodynamic approach to approximate arbitrary distributions using the Gibbs-Boltzmann form. TMI simultaneously learns from data the intensive and the extensive variables and achieves dimensionality reduction using a multiplicative, positive valued, and interpretable decomposition of the data. Importantly, the reduced dimensional space of intensive parameters is not homogeneous. The Gibbs-Boltzmann form defines an analytically tractable Riemannian metric on the space of intensive variables. We show applications of TMI and its comparison with other related dimensionality reduction approaches. Possible extensions to TMI are discussed as well.

DOI: 10.1103/PhysRevResearch.2.023201

I. INTRODUCTION

Over the past few years, our ability to collect high dimensional data has improved substantially. This has been accompanied by a flurry of dimensionality reduction methods. The central goal of these methods is to uncover important lower dimensional features in the high dimensional data. These methods usually belong to one of two broad classes. Methods such as principal component analysis (PCA), singular value decomposition (SVD), and non-negative matrix factorization (NMF) [1,2] are examples of matrix factorization based methods. Here, the high dimensional data (in the form of a matrix) are expressed as a product of two or more *simpler* (for example, sparse or low rank) matrices. In contrast methods such as diffusion maps [3], Laplacian Eigenmaps [4], Isomaps [5], tSNE (t-stochastic neighborhood embedding) [6], and UMAP (uniform manifold approximation and projection) [7] are based on manifold learning. These methods rely on the assumption that the high dimensional data lie on a much lower dimensional embedded manifold. These methods infer the manifold using estimation of local density of data points in the higher dimensions using kernel based approaches. Matrix-based methods do not infer data manifolds, they do allow us to approximate the data using "feature vectors" and "coefficients." In contrast, the latter manifold learning-based class allows inference of a data manifold but cannot represent the data using approximate reconstruction. Notably, no current dimensionality approach achieves both approximate reconstruction of the data and manifold learning.

Orthogonal to these modern approaches, statistical physics offers a physically interpretable solution to dimensionality reduction; albeit for a restricted class of distributions. Consider a system at thermodynamic equilibrium with a surrounding bath that can exchange K types of extensive variables with it. Let the number of states in the system be d. Typically, $d \gg 1$ ($d = 2^{N_s}$ for an Ising model with N_s spins) and $K \sim o(1) \ll d$ ($K = 1, 2$ for the canonical and the grand canonical ensemble respectively). Imagine that there are N different realizations of the bath. Each realization (α) ($\alpha \in [1, N]$) is characterized by K intensive variables $\lambda_k^{(\alpha)}$ ($k \in [1, K]$). At thermodynamic equilibrium, the probability $q_a^{(\alpha)}$ of observing the system in state "a" is given by the Gibbs-Boltzmann distribution:

$$q_a^{(\alpha)} = \frac{1}{Z^{(\alpha)}} \exp\left(-\sum_{k=1}^{K} \lambda_k^{(\alpha)} Y_{ka}\right). \tag{1}$$

In Eq. (1), $\lambda_k^{(\alpha)}$ are realization-specific intensive variables, Y_{ka} are state-dependent extensive variables, and

$$Z^{(\alpha)} = \sum_a \exp\left(-\sum_{k=1}^{K} \lambda_k^{(\alpha)} Y_{ka}\right) \tag{2}$$

is the partition function.

Importantly, recent work has shown that the Gibbs-Boltzmann form has a much broader applicability, even beyond thermal systems at thermodynamic equilibrium. Motivating the Gibbs-Boltzmann distribution using the maximum entropy principle [8,9] has allowed us to employ it to model probabilities in a variety of complex systems such as ensembles of protein sequences [10], parameters of signaling networks [11,12], collective firing of neurons [13], and collective motions of birds [14]. More recently, this approach approach has also been used to approximate dynamics of chemical reaction networks [15,16].

While the Gibbs-Boltzmann form offers attractive dimensionality reduction possibilities, unfortunately, however, not

[*] pdixit@ufl.edu

every arbitrary collection $\{\mathbf{x}^{(\alpha)}\}$, $\alpha \in [1, N]$ of probability distributions can be described using it. To that end, we ask the following question: Given data in the form of N arbitrary distributions $\{\mathbf{x}^{(\alpha)}\}$, can we infer approximate extensive variables Ys and intensive variables λs such that the Gibbs-Boltzmann form in Eq. (1) approximates the data?

We introduce TMI: thermodynamic manifold Iinference. In TMI, we simultaneously infer from data the extensive and the intensive variables. The extensive variables represent state space features while the intensive variables embed the data points in a lower dimensional space. TMI achieves several key objectives. By enforcing the number of extensive variables to be much smaller than the dimension of the state space (the data dimension), it achieves dimensionality reduction. Notably, unlike principal component analysis (PCA) or singular value decomposition, but similar to non-negative matrix factorization [1,2], TMI-based approximation of the data leads to interpretable positive-valued factorization [see Eq. (1)]. Importantly, TMI defines a Riemannian manifold with an analytically tractable distance metric on the space of intensive variables. This metric allows us to define geodesic distances between arbitrary points in the space of intensive variables as well as volume elements.

Below, we first describe the theoretical developments of TMI and its numerical implementation. Then, we illustrate its applications using several data sets. Finally, we discuss future extensions and applications.

II. TMI APPROXIMATES ARBITRARY DISTRIBUTIONS

Consider data in the form of discrete distributions $\{\mathbf{x}^{(\alpha)}\}$, $\alpha \in [1, N]$ defined on a d-dimensional state space. We assume that $\mathbf{x}_a^{(\alpha)} > 0\, \forall\, a \in [1, d]$ and $\forall\, \alpha \in [1, N]$. We want to find K d-dimensional extensive variables $\{\bar{Y}_k\} \equiv \{Y_{ka}\}$ and N K-dimensional intensive bath parameters $\{\bar{\lambda}^{(\alpha)}\} \equiv \{\lambda_k^{(\alpha)}\}$ such that the Gibbs-Boltzmann distributions $q^{(\alpha)}$ in Eq. (1) approximate the original distributions $\mathbf{x}^{(\alpha)}$.

In TMI, we enforce $K \ll N$ to obtain an approximate lower dimensional representation of each distribution. For a given K, we minimize the sum of Kullback-Leibler divergences between $\mathbf{x}^{(\alpha)}$ and $q^{(\alpha)}$:

$$C = \sum_{\alpha} \sum_{a} \mathbf{x}_a^{(\alpha)} \ln \frac{\mathbf{x}_a^{(\alpha)}}{q_a^{(\alpha)}}. \tag{3}$$

The first term in the expanded Kullback—Leibler (KL) divergence depends only on the distributions $\mathbf{x}^{(\alpha)}$ and can be dropped. We have

$$C = -\sum_{\alpha} \left(\sum_{a} \mathbf{x}_a^{(\alpha)} \ln q_a^{(\alpha)} \right) \tag{4}$$

$$= \sum_{\alpha} \left[\sum_{a} \mathbf{x}_a^{(\alpha)} \left(\sum_{k=1}^{K} \lambda_k^{(\alpha)} Y_{ka} + \ln Z^{(\alpha)} \right) \right] \tag{5}$$

$$= \sum_{\alpha} \ln Z^{(\alpha)} + \sum_{\alpha, a, k} \mathbf{x}_a^{(\alpha)} \lambda_k^{(\alpha)} Y_{ka}. \tag{6}$$

The cost is convex with respect to λs when Ys are fixed and vice versa. However, similar to non-negative matrix factorization [1] it is not guaranteed to be globally convex (see Appendix A for a discussion on convexity). We can minimize C with respect to the intensive and the extensive variables to find a local minimum. Differentiating with respect to the intensive and the extensive variables and setting the derivative to zero, we find that the intensive variables are fixed points of nonlinear equations

$$\sum_{a} q_a^{(\alpha)} Y_{ka} = \sum_{a} \mathbf{x}_a^{(\alpha)} Y_{ka} \tag{7}$$

and the extensive variables are fixed points of nonlinear equations

$$\sum_{\alpha} \lambda_k^{(\alpha)} q_a^{(\alpha)} = \sum_{\alpha} \mathbf{x}_a^{(\alpha)} \lambda_k^{(\alpha)}. \tag{8}$$

There are several indeterminacies in the cost function in Eq. (6) and the corresponding fixed points in Eqs. (7) and (8). First, for a fixed k, the cost is invariant to to an additive shift $Y_{ka} = Y_{ka} + c\, \forall\, a \in [1, d]$. This corresponds to the translational invariance in energies in a physical system. Second, the cost is invariant with respect to a scaling $\lambda_k^{(\alpha)} \to B \times \lambda_k^{(\alpha)}$ for all distributions $\alpha \in [1, N]$ and a corresponding transformation that scales $Y_{ka} = Y_{ka}/B$ for all $a \in [1, d]$. Physically, this corresponds to the fact that extensive variables (for example, energies) are always multiplied by the corresponding intensive variables (for example, inverse temperatures) when computing probabilities. More generally, if we multiple the $d \times K$ matrix of extensive variables by a $K \times K$ matrix $\mathbf{B}$ and simultaneously multiple the $N \times K$ matrix of intensive variables with $(\mathbf{B}^{-1})^{\mathrm{T}}$, the Gibbs-Boltzmann probabilities don't change. Finally, the cost is invariant to permutations in k, the label of the extensive variables.

It is possible to incorporate information about constraints on the state space in the inference as well (see [17] for a related formalism for non-negative matrix factorization). One such structural constraints is smoothness. Consider the example of gray-scale images. Here, the distributions represent normalized pixel intensities of digitized images. In the images, any state "a" is identified by planar two-dimensional coordinates $a \equiv (i, j)$ which define adjacency in the state space. Let us consider two adjacent states $a \equiv (i, j)$ and $b \equiv a + \hat{e}$ [$\hat{e} \in \{(1, 0), (0, 1), (-1, 0), (0, -1)\}$]. We can ensure that the extensive variables Y_{ka} and Y_{kb} corresponding to neighboring states a and b are similar to each other by introducing regularizing constraints:

$$\sum_{a,b} n_{ab}(Y_{ka} - Y_{kb})^2 < C_k\, \forall\, k, \tag{9}$$

where $n_{ab} = 1$ when a and b are adjacent and zero otherwise. Such constraints will limit the *ruggedness* of the extensive variables. Other constraints on the extensive variables, such as orthogonality, can be imposed as well.

Finally, we note that though the above discussion was restricted to data in the form of normalized distributions, TMI can also be implemented to un-normalized positive valued data. Notably, the equations to determine Ys and λs are identical to those presented above [Eqs. (7) and (8)] (see Appendix B for discussion on un-normalized data].

III. NUMERICAL INFERENCE OF INTENSIVE AND EXTENSIVE VARIABLES

The gradient descent approach is a straightforward way to numerically learn the intensive and extensive variables. The gradient of the cost function with respect to $\lambda_k^{(\alpha)}$ is given by [see Eq. (7)]

$$\frac{\partial C}{\partial \lambda_k^{(\alpha)}} = \sum_a \mathbf{x}_a^{(\alpha)} Y_{ka} - \sum_a q_a^{(\alpha)} Y_{ka}. \tag{10}$$

Similary, the gradient with respect to Y_{ka} is given by [see Eq. (8)]

$$\frac{\partial C}{\partial Y_{ka}} = \sum_\alpha \mathbf{x}_a^{(\alpha)} \lambda_k^{(\alpha)} - \sum_\alpha q_a^{(\alpha)} \lambda_k^{(\alpha)}. \tag{11}$$

To numerically obtain a local minima, we start with a random initialization for both and $\lambda_k^{(\alpha)}$ and Y_{ka}. Next, we iterate till the gradients have reached a small value:

$$\lambda_k^{(\alpha)} \leftarrow \lambda_k^{(\alpha)} - \eta \frac{\partial C}{\partial \lambda_k^{(\alpha)}}, \tag{12}$$

$$Y_{ka} \leftarrow Y_{ka} - \eta \frac{\partial C}{\partial \lambda_k^{(\alpha)}}. \tag{13}$$

In Eq. (13), $\eta > 0$, the learning rate, is a small positive number.

IV. TMI INTRODUCES A RIEMANNIAN DISTANCE METRIC

Using ideas from nonequilibrium statistical physics, TMI defines a Riemannian geometry and a distance metric on the space of intensive variables. We note that the distance metric is defined on the space of intensive variables and not on the data itself.

Consider two different distributions approximated by intensive variables $\bar{\lambda}^{(1)}$ and $\bar{\lambda}^{(2)}$. Consider a smooth and differentiable path $\gamma(t)$ between the two distributions such that $\gamma(t=0) = \bar{\lambda}^{(1)}$ and $\gamma(t=T) = \bar{\lambda}^{(2)}$. It was shown that [18–21] in the linear-response regime, the excess work—work done above the difference in thermodynamic potentials—along this path can be computed as

$$P \propto \int_0^T \frac{d\bar{\lambda}^{\mathrm{T}}}{dt} g(\bar{\lambda}) \frac{d\bar{\lambda}}{dt} dt, \tag{14}$$

where the elements of the friction tensor g are given by [19]

$$g_{ij}(\bar{\lambda}) = \int_0^\infty \langle \delta \bar{Y}_i(0) \delta \bar{Y}_j(\tau) \rangle_{\bar{\lambda}} d\tau. \tag{15}$$

In Eq. (15), $\delta \bar{Y}_i = \bar{Y}_i - \langle \bar{Y}_i \rangle$ where $\langle \bar{Y}_i \rangle$ is the ensemble average value of the extensive variable $\bar{Y}_i$ when the intensive variables are fixed at $\bar{\lambda}(t)$. We note that a similar derivation exists for transforming two nonequilibrium steady-state (NESS) distributions [22]. However, NESS distributions cannot be expressed in the parametric Gibbs-Boltzmann form and therefore we do not pursue that direction here.

The friction tensor depends on the dynamics on the state space $\{a\}$ at a fixed $\bar{\lambda}$. When the transition rate matrix $\kappa_{a\to b}(\bar{\lambda})$ is provided, the friction tensor can be computed in a straightforward manner (see Appendix C for a description of the tensor calculations). What are reasonable choices for the dynamics? We want an "equilibrium" (detailed balanced) transition rate matrix that is constrained to reproduce the Gibbs-Boltzmann distribution $q(\bar{\lambda})$. Though this is not a requirement, we may also like the dynamics to be *local*; states that are *closer* to each other should have higher transition rates and vice versa for states that are farther apart. One way to incorporate the information about the underlying geometry is to require that the rates penalize transitions between geometrically "distant" states a and b. A simple transition rate matrix that satisfies these properties is the one that maximizes the path entropy [23]:

$$\kappa_{a\to b}(\bar{\lambda}) \propto \sqrt{\frac{q_b(\bar{\lambda})}{q_a(\bar{\lambda})}} \exp\left(-\frac{d(a,b)^2}{\varepsilon}\right). \tag{16}$$

Another choice is the so-called Glauber dynamics [24]:

$$\kappa_{a\to b}(\bar{\lambda}) \propto \frac{q_b(\bar{\lambda})}{q_a(\bar{\lambda}) + q_b(\bar{\lambda})} \exp\left(-\frac{d(a,b)^2}{\varepsilon}\right). \tag{17}$$

In Eqs. (16) and (17), $d(a,b)$ is a measure of separation between states a and b (for example, Euclidean distance) and $\varepsilon > 0$ plays the role of an inverse diffusion constant. We stress that any other choice of the dynamics will define a well-behaved friction tensor as long as the dynamics is reversible and reproduces the stationary distribution $q(\bar{\lambda})$.

When the dynamics is fast, the friction coefficient reduces (up to a proportionality) to the Fisher information matrix [18–21], which in the case of Gibbs-Boltzmann distributions is the matrix of fluctuations [25]:

$$g_{ij}(\bar{\lambda}) = \langle \bar{Y}_i \bar{Y}_j \rangle_{\bar{\lambda}} - \langle \bar{Y}_i \rangle_{\bar{\lambda}} \langle \bar{Y}_j \rangle_{\bar{\lambda}}. \tag{18}$$

If we assume that the rate of change of $\bar{\lambda}$ along a trajectory is kept constant, the paths that minimize excess work are also the paths that minimize the geodesic distance [18–21]. Hence, the length of the path of minimum excess work between two distributions, described by $\bar{\lambda}_1$ and $\bar{\lambda}_2$ respectively, also defines a metric distance between them. We note that the Fisher information matrix is invariant to a permutation of the indices. Therefore, if the Fisher information matrix is used instead of the friction tensor, the geodesic distances will not take into account the geometry of the state space.

V. LEARNING ISING MODEL FROM DATA

As a test case, we show that TMI can infer the energy landscape of an Ising model from samples of the Ising model distributions. To that end, we consider a nearest-neighbor Ising model with $n_s = 8$ spins arranged as shown in Fig. 1(a). Each spin σ can take the values 1 or -1. The probability of observing any spin configuration $\bar{\sigma}(a)$ is given by

$$p(\bar{\sigma}(a)) = \frac{1}{Z(H,J)} \exp[-HE_{\mathrm{mag}}(a) - JE_{\mathrm{int}}(a)], \tag{19}$$

where

$$E_{\mathrm{mag}}(a) = \sum_i \sigma(a)_i, \quad \text{and} \tag{20}$$

$$E_{\mathrm{int}}(a) = \sum_{i\,\mathrm{nn}\,j} \sigma(a)_i \sigma(a)_j. \tag{21}$$

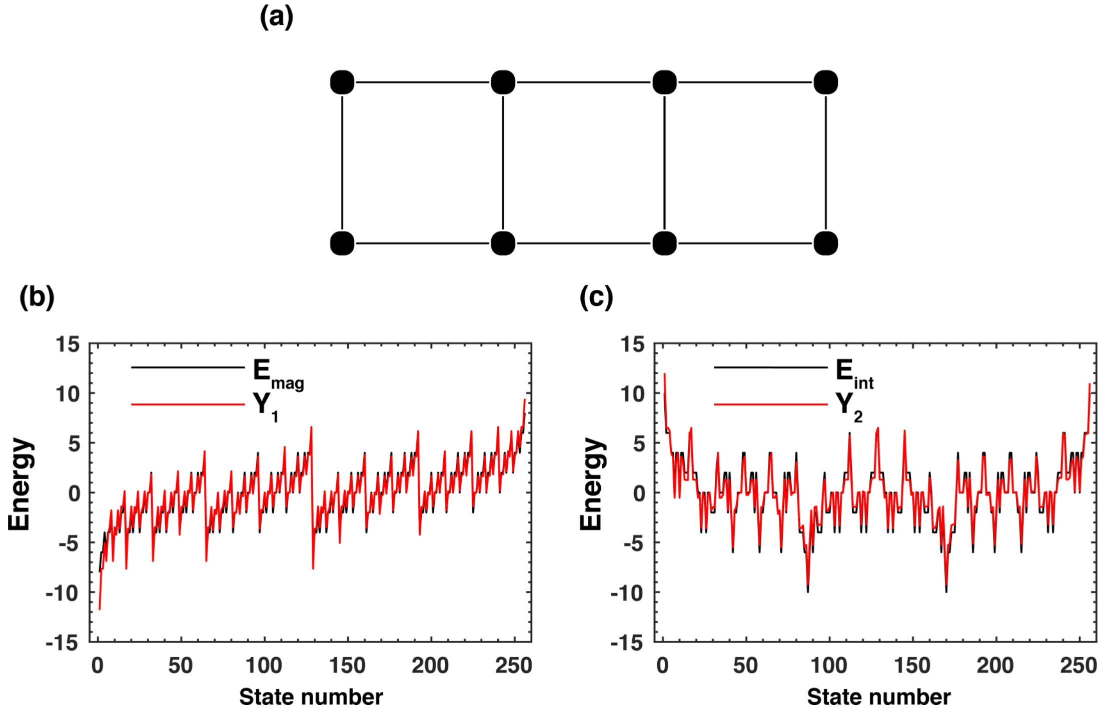

FIG. 1. (a) the connectivity graph of a eight-spin Ising model, (b) inferred extensive variable $\bar{Y}_1$ (red) compared to the true extensive variable E_{mag} (black), and (c) inferred extensive variable $\bar{Y}_2$ (red) compared to the true extensive variable E_{int}.

In Eq. (21), the summation is taken over the nearest neighbors of the graph shown in Fig. 1(a) and $Z(H, J)$ is the partition function.

We randomly sampled 50 pairs of H and J values from a uniform distribution where $H \in [-1, 1]$ and $J \in [-1, 1]$ and generated 50 Ising model distributions (see Appendix D for a discussion on the random sampling). Next, we approximated these input distributions using TMI with $K = 2$ extensive variables $\bar{Y}_1$ and $\bar{Y}_2$. We simultaneously inferred 50 pairs of intensive variables representing each of the 50 distributions.

As noted above, multiplication by a matrix $Y \to Y \times \mathbf{B}$ and $\Lambda \to \Lambda \times (\mathbf{B}^{-1})^{\mathrm{T}}$ does not change TMI predictions. Thus, in order to directly compare TMI predictions with the ground truth, we need to reorient the TMI-inferred variables. To that end, we find a matrix $\mathbf{B}$ such that (1) $\bar{Y}_1$ and $\bar{Y}_2$ have the same dot product as the vectors $\bar{E}_{\text{int}}$ and $\bar{E}_{\text{mag}}$ and (2) $\bar{Y}_1$ is orthogonal to $\bar{E}_{\text{int}}$. In Figs. 1(b) and 1(c) we show that the reoriented extensive variables $\bar{Y}_1$ and $\bar{Y}_2$ closely approximate the the true extensive variables E_{mag} and E_{int} respectively only from 50 sampled distributions. We note that constraints such as symmetry were not imposed when inferring the extensive variables.

VI. ANALYSIS OF HANDWRITTEN DIGITS

Next, we illustrate the application of TMI using the MNIST dataset [26]. The dataset represents handwritten digits between 0 and 9. We randomly selected 500 digits from the set of all 6s and 9s from MNIST. The digits were represented as a 28 × 28 array of positive numbers. Each data point was normalized and treated as a distribution represented by a 784-dimensional probability vector. Given that there were two types of digits, we set out to infer $K = 2$ sample-independent extensive variables and 500 × 2 intensive variables. In Figs. 2(a) and 2(b) we show the two inferred extensive variables $\bar{Y}_1$ and $\bar{Y}_2$. Notably, TMI correctly identifies two extensive variables that correspond to a generic digit 9 and a generic digit 6 respectively. These represent the two inferred potential-energy minima in the data.

Moreover, as shown in panel (c), the two digits can also be classified by two different regions of the space of intensive variables; 9s are characterized by a high λ_1 and a low λ_2 while 6s are characterized by a low λ_1 and a high λ_2. Importantly, the Fisher-Rao metric on the space of intensive variables defines a notion of distance between the distributions as well as the "number of points" in any given volume element [27]. The heat map in panel (c) represents the logarithm of the volume element given by the square root of the determinant of the Fisher information matrix. It is clear that the reduced dimensional space is highly inhomogeneous; the same small change in λ_1 and λ_2 may have very different effects on the resulting distributions depending on the region of the space. The Fisher-Rao metric allows us to construct geodesics between pairs of data point. As shown in Fig. 2(c), the geodesic (dashed black line) between a 6 (top left) and a 9 (bottom right) is substantially different than the straight line (dashed red line).

VII. TMI PERFORMANCE IN DATA RECONSTRUCTION AND CLASSIFICATION

Mathematically, TMI is most closely related to non-negative matrix factorization (NMF) [1,2]. Therefore, we compared the performance of TMI with NMF. While TMI represents the thermodynamic potential of any state as a matrix product, NMF approximates the probabilities

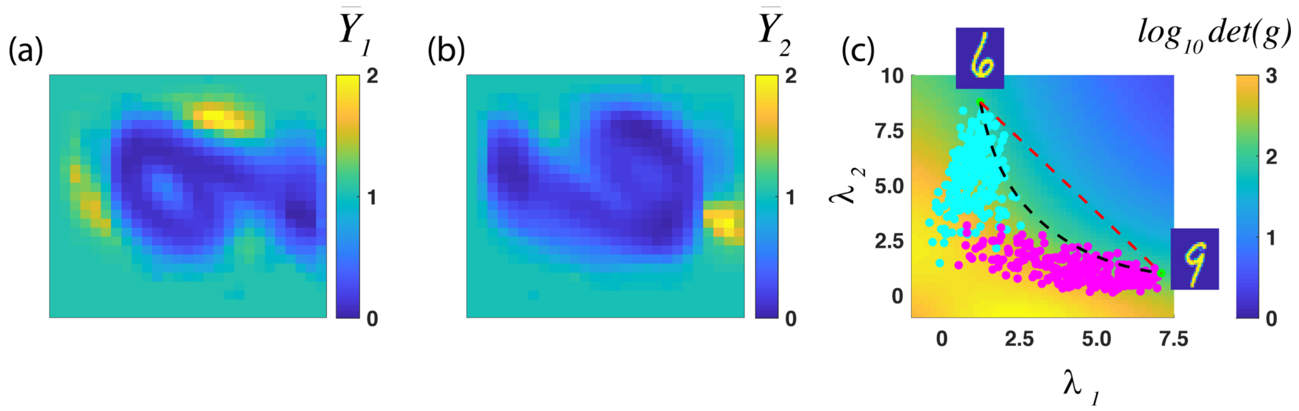

FIG. 2. (a) A heat map of the inferred extensive variable $\bar{Y}_1$ representing a generic "9." (b) A heat map of the inferred extensive variable $\bar{Y}_2$ representing a generic "6." (c) A scatter plot of the intensive bath parameters of the 500 data points. The data labeled "6" are colored cyan while the data labeled "9" are colored magenta. The heat map represents the volume element (square root of the determinant of the metric tensor). The dashed red line is a straight line transformation between two data points shown at the top left and bottom right. The dashed black line is the geodesic computed using the Fisher-Rao metric.

themselves as a matrix product. Briefly, in NMF, positive valued data is expressed as a product of two matrices:

$$\mathbf{x}_a^{(\alpha)} \approx q_a^{(\alpha)} = \sum_k \mathbf{l}_k^{(\alpha)} \mathbf{y}_{ka}. \tag{22}$$

The matrices $\mathbf{l}$ and $\mathbf{y}$ are determined by minimizing the Kullback-Leibler divergence between the data $\{\mathbf{x}_a^{(\alpha)}\}$ and the approximation $\{q_a^{(\alpha)}\}$ (L_2 norm minimization is possible as well). NMF is a widely used technique to model positive valued data as it leads to interpretable positive valued decomposition (see [28] for a review). We note that NMF-based decomposition of the data is a linear superposition of positive valued "feature vectors" $\bar{\mathbf{y}}_k$s with positive valued "coefficients" $\bar{\mathbf{l}}^{(\alpha)}$s. In contrast, TMI expresses the data as a multiplicative decomposition [see Eq. (1)].

To compare the ability of TMI and NMF to approximate the data, we chose four data sets of very different origins. The first was the MNIST dataset of handwritten digits [26]. From the MNIST dataset, we randomly selected 500 samples comprising digits from 0 to 9. As above, each digit was represented by a 28 × 28 array of pixel intensities which was normalized to 1. The second was the time series data collected on the gut microbiome of a human [29]. The microbiome data consisted of 318 samples collected approximately daily over a period of a year from the feces of one human individual. Each sample was represented by the relative abundances of 70 most abundant bacterial operational taxonomic units (OTUs). The third dataset comprised a "bag of words" description [30] of papers submitted to the Neural Information Processing Systems conference (downloaded from [31]). Each paper was represented as a collection of words wherein each word was assigned a frequency in each submitted article. The fourth dataset comprised 472 gray-scale images of human faces stored as an array of 19 × 19 pixels (the CBCL database of faces [32]) (see Appendix E for details of the data sets).

We approximated each of the data sets using TMI and NMF with several different values of K. For each K we compared the Kullback-Leibler divergence between the data $\{\mathbf{x}_a^{(\alpha)}\}$ and the reconstruction $\{q_a^{(\alpha)}\}$. As seen in Table I, TMI consistently performed better than NMF at reconstructing the data for every value of K. One possible reason behind this performance is that real data sets often have widely varying amplitudes. For example, the intensity of any given pixel in a set of images can vary substantially from image to image [33]. The multiplicative approximation using the intensive variables in TMI may be better suited to capture such variability compared to the linear superposition in NMF.

Next, we tested how TMI performed in data classification using the MNIST dataset. To that end, used the 500 MNIST digits as above and inferred intensive variables and extensive variables for a range of K values. We used these intensive variables and the known identities of the digits to train a support vector machine (SVM) classifier. Next, we randomly selected 2000 digits from the dataset and predicted their identities. Similarly, we fitted the same data with NMF and trained an SVM classifier with the same hyperparameters. The accuracy of the two identifications is shown in Table II.

TABLE I. Comparison of KL divergences between the data $\{\mathbf{x}_a^{(\alpha)}\}$ and the approximate reconstruction $\{q_a^{(\alpha)}\}$ using TMI and NMF respectively. K indicates the number of extensive variables used to model the data. The divergences are reported as an average per data point.

	MNIST		Microbiome		NIPS		CBCL	
K	TMI	NMF	TMI	NMF	TMI	NMF	TMI	NMF
1	0.89	0.92	0.29	0.30	0.24	0.24	0.033	0.036
2	0.78	0.82	0.18	0.20	0.22	0.23	0.022	0.028
3	0.68	0.74	0.14	0.16	0.21	0.22	0.019	0.022
4	0.59	0.68	0.11	0.14	0.21	0.21	0.018	0.019
5	0.53	0.64	0.09	0.12	0.20	0.21	0.016	0.017
10	0.34	0.50	0.05	0.07	0.18	0.19	0.010	0.012
20	0.17	0.38	0.02	0.04	0.15	0.18	0.005	0.007
40	0.07	0.26	0.01	0.01	0.12	0.16	0.002	0.003

TABLE II. Comparison of classification success rate (fraction) of randomly chosen handwritten digits from the MNIST dataset using TMI and NMF. The error bars are standard deviations estimated using 20 equal subsamples of the test set.

K	TMI	NMF
5	0.62 ± 0.05	0.49 ± 0.04
10	0.75 ± 0.04	0.63 ± 0.05
15	0.77 ± 0.03	0.67 ± 0.05
20	0.80 ± 0.04	0.69 ± 0.05

Similar to its ability to fit the data accurately, TMI also performs significantly better than NMF at classifying the data.

VIII. DISCUSSION

The manifold assumption [4], commonly invoked in modern data analysis, posits that high dimensional data is governed by a few parameters and as a result can be represented by a lower dimensional manifold residing in the higher dimension. Several manifold inference methods such as diffusion maps [3], Laplacian eigenmaps [4], isomaps [5], tSNE (t-stochastic neighborhood embedding) [6], and UMAP (uniform manifold approximation and projection) [7] have been developed to approximately reconstruct these manifolds from the data.

While the manifold-based methods achieve dimensionality reduction, unlike other approaches such as principal component analysis (PCA) or non-negative matrix factorization (NMF) [1,2], they cannot obtain an approximate reconstruction of the original data using lower dimensional "features." At the same time, these methods do not obtain an analytical description of the manifold but only approximate it using a nonlinear embedding of the data points in the lower dimensional space. As a result, analytical manipulations such as computation of geodesics and volume elements are not possible.

We presented TMI, an approach rooted in statistical physics to embed positive valued high dimensional data points in lower dimensions. TMI possesses advantages of both manifold approximation methods as well as matrix-based dimensionality reduction methods. (1) Similar to matrix-based methods such as PCA, SVD, and NMF, TMI can approximate data using lower dimensional features. Notably, similar to NMF, this decomposition is positive valued [see Eq. (1)] and thus interpretable. Moreover, given the multiplicative nature of the decomposition, TMI appears to be better suited to model real data compared to NMF. (2) Similar to manifold approximation methods, TMI can infer an approximate lower dimensional manifold on which the data resides. Importantly, unlike previously developed methods (discussed above), TMI defines an analytically tractable and readily computable Riemannian manifold (with an associated distance metric) in the lower dimension. This in turn allows us to compute geodesics and volume elements in the reduced dimensional description.

While TMI outperformed NMF in modeling and classifying data, in the current implementation, TMI was slower than NMF. Therefore, in the future, it will be important to optimize the numerical algorithms in TMI. Similarly, the calculation of the geodesic can be time consuming given that it requires solving boundary value nonlinear differential equations. However, numerically efficient techniques have been developed [21,34,35] which will be more useful in situations when using $K > 2$ extensive variables. Another potential way to avoid solving the nonlinear differential equations is to rely on the observation that the geodesics pass through the data rich regions of the $\bar{\lambda}$ space. Consequently, we can potentially approximate the geodesic as the shortest path on a graph connecting the data points themselves.

ACKNOWLEDGMENTS

We would like to thank Dr. S. Saxena for numerous discussions about the topic and Dr. M. Rachh for useful comments on the manuscript. We would like to thank Mayar Shahin for pointing out a mistake in the calculations. We would like to thank University of Florida for its generous support.

APPENDIX A: COST FUNCTION IS CONVEX IN λs and Ys

In this section, we show that the (1) cost function in Eq. (6) is convex with respect to $\lambda_k^{(\alpha)}\,\forall\, k \in [1, K]$ when all the other λs and all the Ys are fixed and (2) the cost function is convex in $Y_{ka}\,\forall\, a \in [1, d]$ when λs and all other Ys are fixed.

The double derivative of the cost function for a fixed α is given by Eq. (A1):

$$\frac{\partial^2 C}{\partial \lambda_k^{(\alpha)} \lambda_j^{(\alpha)}} = (\langle Y_{ka} Y_{ja}\rangle_\alpha - \langle Y_{ka}\rangle_\alpha \langle Y_{ja}\rangle_\alpha). \tag{A1}$$

The matrix in Eq. (A1) is a covariance matrix. Given that covariance matrices are non-negative, the Hessian matrix in Eq. (A1) is non-negative as well.

Next, we look at the Hessian with respect to the Y_{ka}s for a fixed k when λs and other Ys are fixed. We have the derivative

$$\frac{\partial C}{\partial Y_{ka}} = -\sum_\alpha \lambda_k^{(\alpha)} q_a^{(\alpha)} + \sum_\alpha \mathbf{x}_a^{(\alpha)} \lambda_k^{(\alpha)}, \tag{A2}$$

$$\frac{\partial^2 C}{\partial Y_{ka} Y_{kb}} = -\sum_\alpha \lambda_k^{(\alpha)} \frac{\partial q_a^{(\alpha)}}{\partial Y_{kb}}. \tag{A3}$$

We have the derivative

$$\frac{\partial q_a^{(\alpha)}}{\partial Y_{kb}} = \frac{\partial}{\partial Y_{kb}} \frac{f_a^{(\alpha)}}{Z^{(\alpha)}}, \tag{A4}$$

where

$$f_a^{(\alpha)} = \exp\left(-\sum_k \lambda_k^{(\alpha)} Y_{ka}\right). \tag{A5}$$

From Eq. (A4), we have

$$\frac{\partial q_a^{(\alpha)}}{\partial Y_{kb}} = \frac{1}{Z^{(\alpha)}} \frac{\partial f_a^{(\alpha)}}{\partial Y_{kb}} - q_a^{(\alpha)} \frac{\partial \ln Z^{(\alpha)}}{\partial Y_{kb}}. \tag{A6}$$

We have

$$\frac{\partial \ln Z^{(\alpha)}}{\partial Y_{kb}} = -\lambda_k^{(\alpha)} q_b^{(\alpha)} \tag{A7}$$

and

$$\frac{\partial f_a^{(\alpha)}}{\partial Y_{kb}} = -\delta_{ab} \lambda_k^{(\alpha)} f_a^{(\alpha)}. \tag{A8}$$

Putting everything together, we have

$$\frac{\partial q_a^{(\alpha)}}{\partial Y_{kb}} = -\delta_{ab} q_a^{(\alpha)}\lambda_k^{(\alpha)} + q_a^{(\alpha)}\lambda_k^{(\alpha)} q_b^{(\alpha)} \tag{A9}$$

$$= q_a^{(\alpha)}\lambda_k^{(\alpha)}\big(-\delta_{ab} + q_b^{(\alpha)}\big). \tag{A10}$$

Thus, the elements of the Hessian are given by

$$\frac{\partial^2 C}{\partial Y_{ka} Y_{kb}} = -\sum_{\alpha}\big(\lambda_k^{(\alpha)}\big)^2 q_a^{(\alpha)}\big(q_b^{(\alpha)} - \delta_{ab}\big)$$

$$= \delta_{ab}\left(\sum_{\alpha}\big(\lambda_k^{(\alpha)}\big)^2 q_a^{(\alpha)}\right) - \sum_{\alpha}\big(\lambda_k^{(\alpha)}\big)^2 q_a^{(\alpha)} q_b^{(\alpha)}. \tag{A11}$$

Given that $q_a^{(\alpha)} > q_a^{(\alpha)} q_b^{(\alpha)}$, the sum of off-diagonal entries in the Hessian matrix is smaller than the diagonal entry; according to Gershgorin's disk theorem, the Hessian matrix in Eq. (A11) will be positive semidefinite.

APPENDIX B: TMI FOR UN-NORMALIZED DATA

The cost function for un-normalized distributions can be written as [36]

$$C = \sum_{a,\alpha}\left(x_a^{(\alpha)}\ln\frac{x_a^{(\alpha)}}{q_a^{(\alpha)}} - x_a^{(\alpha)} + q_a^{(\alpha)}\right), \tag{B1}$$

where

$$q_a^{(\alpha)} = \exp\left(-\sum_{k=1}^{K}\lambda_k^{(\alpha)} Y_{ka}\right) \tag{B2}$$

is the un-normalized distribution and $\{x_a^{(\alpha)}\}$ is the un-normalized positive valued data. We rewrite C after dropping terms that do not depend on λs and Ys:

$$C = \sum_{a,\alpha,k} x_a^{(\alpha)}\lambda_k^{(\alpha)} Y_{ka} + \sum_{a,\alpha} q_a^{(\alpha)}. \tag{B3}$$

We differentiate Eq. (B3) with respect to $\lambda_k^{(\alpha)}$ and set the derivative to zero:

$$\sum_a x_a^{(\alpha)} Y_{ka} = \sum_a q_a^{(\alpha)} Y_{ka}. \tag{B4}$$

Notably, Eq. (B4) is identical to the normalized version [see Eq. (7)]. Similarly, we differentiate with respect to Ys and set the gradient to zero:

$$\sum_{\alpha} x_a^{(\alpha)}\lambda k^{(\alpha)} = \sum_{\alpha} q_a^{(\alpha)}\lambda_k^{(\alpha)}. \tag{B5}$$

Similar to Eq. (B4), Eq. (B5) is identical to Eq. (8). This shows that TMI can be implemented with both normalized and un-normalized data.

APPENDIX C: COMPUTING THE FRICTION TENSOR

Consider a transition rate matrix κ whose stationary distribution is given by $q_a(\bar{\lambda})$. The probability of being in state b at time t conditioned on being in state a at time $t = 0$ is given by K_{ab} where the matrix K is given by

$$K = \exp(\kappa\tau) = V\exp(\Lambda\tau)V^{-1}, \tag{C1}$$

where $V\Lambda V^{-1}$ is the diagonalization of κ. We can now express the friction tensor:

$$g_{ij}(\bar{\lambda}) = \int_0^{\infty}\langle\delta Y_i(0)\delta Y_j(\tau)\rangle_{\bar{\lambda}}\, d\tau \tag{C2}$$

$$= \int_0^{\infty}\left(\sum_{a,b} q_a\delta Y_{ia}\delta Y_{jb}K_{ab}\right)d\tau \tag{C3}$$

$$= \int_0^{\infty} C_i\exp(\Lambda\tau)D_j\, d\tau, \tag{C4}$$

where

$$C_i = (q\circ\delta\bar{Y}_i)^{\mathrm{T}}V \quad \text{and}\; D_j = V^{-1}\delta\bar{Y}_j, \tag{C5}$$

where $a\circ b$ is the Haddamard (elementwise) product. Thus, we have

$$g_{ij} = \int_0^{\infty}\sum_a C_{ia}D_{ja}\exp(\Lambda_a\tau)d\tau \tag{C6}$$

$$= -\sum_a\frac{C_{ia}D_{ja}}{\Lambda_a}, \tag{C7}$$

where the sum omits the zero eigenvalue.

APPENDIX D: FIGURE FOR ISING MODEL

Figure 3 shows 50 randomly sampled points in the space of intensive variables (H, J) used in the inference of the intensive and extensive variables in Fig. 1.

APPENDIX E: DATA FOR NMF/TMI COMPARISON AND IMPLEMENTATION OF NMF

1. Microbiome data

The microbiome data were obtained from David *et al.* [29]. Briefly, the data consisted of bacterial operational taxonomic

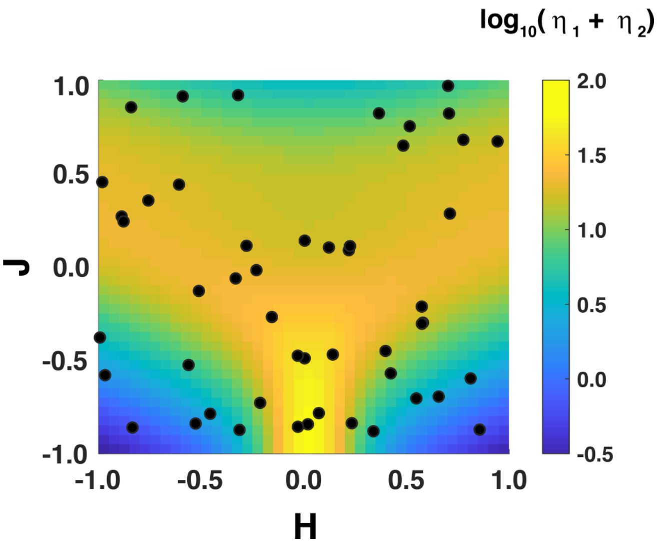

FIG. 3. Scatter plot of 50 randomly chosen H and J values between $[-11]$. The color represents the logarithm of the trace (sum of eigenvalues η_1 and η_2) of the Fisher information matrix of the Ising model [20].

unit (OTU) abundances collected over a period of a year. There were 318 samples; each sample comprised relative abundances of $\sim 8 \times 10^3$ OTUs. Based on our previous analysis [37], we discarded from the data OTUs whose average relative abundance was less than 0.1% as these abundances are likely to represent technical noise in data collection. The data on the remaining 70 high abundant OTUs were renormalized to relative fractions.

2. Bag of words data from NIPS conferences

The bag of words description [30] is a simple way to characterize text documents. Briefly, for a collection of documents, one first identifies all possible words. Next, the frequency of each word in each document is estimated. The document is then represented as a vector of frequencies, regardless of the order in which the words appear.

We downloaded the bag of words model of article submissions to the NIPS conference from the UCI machine learning repository [31]. From the data, we removed article submissions that were characterized by less than 1000 words and words that had less than 100 appearances across all articles. The resultant dataset had 1322 articles each represented by a normalized probability vector with 2753 entries.

3. Implementation of non-negative matrix factorization

We implemented a modified algorithm to learn the matrices $\mathbf{l}$ and $\mathbf{y}$ in Eq. (22). We followed the update algorithm that corresponds to minimization of Kullback-Leibler divergence between the data and the approximate representation [36]. To ensure normalization of the approximate reconstruction, in each iteration of $\mathbf{l}$, for each α, we scaled the vectors $\bar{\mathbf{l}}^{(\alpha)}$ such that the predictions $q^{(\alpha)}$ sum to 1.

[1] D. D. Lee and H. S. Seung, Nature (London) **401**, 788 (1999).
[2] T. Hofmann, in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann Publishers Inc., San Francisco, CA, 1999), pp. 289–296.
[3] R. R. Coifman and S. Lafon, Appl. Comput. Harmonic Anal. **21**, 5 (2006).
[4] M. Belkin and P. Niyogi, Neural Comput. **15**, 1373 (2003).
[5] M. Balasubramanian and E. L. Schwartz, Science **295**, 7 (2002).
[6] L. v. d. Maaten and G. Hinton, J. Machi. Learn. Res. **9**, 2579 (2008).
[7] L. McInnes, J. Healy, and J. Melville, arXiv:1802.03426.
[8] P. D. Dixit *et al.*, J. Chem. Phys. **148**, 010901 (2018).
[9] K. Ghosh, P. D. Dixit, L. Agozzino, and K. A. Dill, Annu. Rev. Phys. Chem. **71**, 213 (2020).
[10] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, Rep. Prog. Phys. **81**, 032601 (2018).
[11] P. D. Dixit, Biophys. J. **104**, 2743 (2013).
[12] P. D. Dixit, E. Lyashenko, M. Niepel, and D. Vitkup, Cell Syst. **10**, 204 (2020).
[13] C. Savin and G. Tkačik, Curr. Opin. Neurobiol. **46**, 120 (2017).
[14] W. Bialek *et al.*, Proc. Natl. Acad. Sci. USA **109**, 4786 (2012).
[15] O. K. Ernst, T. Bartol, T. Sejnowski, and E. Mjolsness, J. Chem. Phys. **149**, 034107 (2018).
[16] O. K. Ernst, T. M. Bartol, T. J. Sejnowski, and E. Mjolsness, Phys. Rev. E **99**, 063315 (2019).
[17] S. Saxena *et al.*, bioRxiv, 650093(2019).
[18] G. E. Crooks, Phys. Rev. Lett. **99**, 100602 (2007).
[19] D. A. Sivak and G. E. Crooks, Phys. Rev. Lett. **108**, 190602 (2012).
[20] G. M. Rotskoff and G. E. Crooks, Phys. Rev. E **92**, 060102(R) (2015).
[21] G. M. Rotskoff, G. E. Crooks, and E. Vanden-Eijnden, Phys. Rev. E **95**, 012148 (2017).
[22] D. Mandal and C. Jarzynski, J. Stat. Mech. (2016) 063204.
[23] P. D. Dixit, A. Jain, G. Stock, and K. A. Dill, J. Chem. Theory Comput. **11**, 5464 (2015).
[24] R. J. Glauber, J. Math. Phys. **4**, 294 (1963).
[25] A. Caticha, arXiv:0808.0012.
[26] Y. LeCun, C. Cortes, and C. Burges, AT&T Labs (online). Available: http://yann.lecun.com/exdb/mnist.
[27] A. Caticha, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MAXENT 2014)*, edited by A. Mohammad-Djafari and F. Barbaresco, AIP Conf. Proc. No. 1641 (AIP, Melville, NY, 2015), pp. 15–26.
[28] Y.-X. Wang and Y.-J. Zhang, IEEE Trans. Knowl. Data Eng. **25**, 1336 (2012).
[29] L. A. David *et al.*, Genome Biol. **15**, R89 (2014).
[30] Y. Zhang, R. Jin, and Z.-H. Zhou, Int. J. Mach. Learn. Cybern. **1**, 43 (2010).
[31] A. Asuncion and D. Newman, Uci machine learning repository (2007).
[32] H. A. Rowley, S. Baluja, and T. Kanade, IEEE Trans. Pattern Anal. Mach. Intell. **20**, 23 (1998).
[33] D. L. Ruderman and W. Bialek, in *Advances in Neural Information Processing Systems* (1994), pp. 551–558.
[34] M. Heymann and E. Vanden-Eijnden, Phys. Rev. Lett. **100**, 140601 (2008).
[35] M. Heymann and E. Vanden-Eijnden, Commun. Pure . Appl. Math. **61**, 1052 (2008).
[36] D. D. Lee and H. S. Seung, in *Advances in Neural Information Processing Systems* (2001), pp. 556–562.
[37] B. W. Ji *et al.*, Nat. Methods **16**, 731 (2019).