IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 June 2024, accepted 24 June 2024, date of publication 28 June 2024, date of current version 8 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3420425

RESEARCH ARTICLE

On the Attractive and Repulsive Forces of Generalized Stochastic Neighbor Embedding With Alpha-Divergence

HSIN-YI LIN[1], (Member, IEEE), HUAN-HSIN TSENG[2], AND JEN-TZUNG CHIEN[3], (Senior Member, IEEE)

[1]Department of Mathematics and Computer Science, Seton Hall University, South Orange, NJ 07079, USA
[2]Computational Science Initiative, Brookhaven National Laboratory, Upton, NY 11973, USA
[3]Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan

Corresponding author: Hsin-Yi Lin (hsinyi.lin@shu.edu)

ABSTRACT Stochastic neighbor embedding (SNE) performs nonlinear transformation from high-dimensional observation space to low-dimensional latent space which preserves neighbor affinities. Data pairs in latent space tend to be crowded due to the dimensionality reduction. To mitigate the crowding problem, certain characteristics are favorable in the design of the SNE setting. This study presents a fundamental analysis of SNE that not only generalizes the previous SNEs but also provides a systematic way to understand the intrinsic properties. From the perspective of theoretical connection, we are able to conceive a new generalized SNE (g-SNE) by introducing a regularized power-law distribution with the α-divergence for manifold learning. The proposed method generalizes and incorporates various favorable features for the clustering process. In addition, the proposed method provides high flexibility, admitting tailored realizations to properly reflect the similarity between original and dimension-reduced samples. Experiments are performed to analyze the proposed method, and its effectiveness is demonstrated with several learning tasks.

INDEX TERMS α-divergence, clustering, dimensionality reduction, data visualization, manifold learning, stochastic neighbor embedding.

I. INTRODUCTION

With the advancement of modern technologies, vast data collected from diverse sources like social media, sensors, and digital images introduces the challenge of high dimensionality, which complicates data analysis and model training processes. Dimensionality reduction techniques effectively address these problems by condensing data into a lower-dimensional space, preserving crucial information while eliminating redundancies. This does not only accelerate computational tasks but also improves data interpretability [1], [2], aiding in pattern discovery to enhance model performance of a learned machine. For example, an optimal graph-based dimensionality reduction was applied to enhance and improve the performance for semi-supervised learning [3]. In another example, dimensionality reduction was used to manage the complexities of varying domain characteristics, while equilibrium distribution was leveraged for domain adaptive learning [4]. Solutions to dimensionality reduction are pivotal for efficient data analysis and visualization across diverse domains [5].

Dimensionality Reduction algorithms range from *linear* transformations, such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), to *nonlinear* mappings, such as Locally Linear Embedding (LLE) [6] and Stochastic Neighbor Embedding (SNE) [7], [8], [9], [10] which are regarded as nonparametric mappings. A parametric mapping based on deep neural network [11], [12] was learned to handle the unseen data in manifold learning [13], [14], while SNE has been extensively developed for probabilistic

The associate editor coordinating the review of this manuscript and approving it for publication was Qilian Liang.

dimensionality reduction and data visualization. Based on previous works, several methods with geometric insights have been introduced into generalized manifold settings. An extension of PCA to Symmetric Positive Definite (SPD) matrix manifolds was developed in [15]. HoroPCA [16], on the other hand, provides a PCA fitting for data lives on a hyperbolic space. More recently, an adapted SNE on the non-Euclidean hyperbolic space was introduced in [17]. Uniform Manifold Approximation and Projection (UMAP) [18] was proposed with a similar core idea as SNE, where similarity matrices in the high and low-dimensional space are compared. One key difference between SNE and UMAP is the optimization measurement. In fact, SNE frequently applies the Kullback-Leibler (KL) divergence, while UMAP applies the cross-entropy.

SNEs [7], [19] transform the data space into a probabilistic space and endow probability distributions on data points. The KL divergence is a typical measurement for the distributions in the high-dimensional space and that of the low-dimensional space [20], [21]. However, there is no clear analysis regarding how divergences and underlying probability functions mitigate the crowding problem where the pairwise distances in low-dimensional latent space do not fully manifest those in the high-dimensional observation space. Certain attempts by replacing the low-dimensional distribution as a heavy-tailed one [9] such as the Student-t distribution [8] were proven to be empirically helpful. Later on, it was pointed out that the power-law distribution (p-SNE) can further help with the clustering separation [22], from which this study is considerably extended.

This work intends to generalize SNE in a broader setting by integrating favorable characteristics. Initiated from a solid mathematical formulation, we propose a new generalized SNE that relaxes the usage of similarity measures, distribution functions, and divergences. Particularly, this study proposes employing a regularized power-law distribution of a symmetric target similarity on the general Riemannian structures. The proposed method is developed for a flexible manifold learning based on α-divergence to align the distributions in high-dimensions and low-dimensions. An additional parameter α admits extra degrees of freedom for flexibility. It is noticed that when $\alpha = -1$ and $\alpha = 0$, they correspond to the KL divergence and the Hellinger distance, respectively. With this generalized framework, we are allowed to explore the effect of attractive and repulsive forces and demonstrate how the low-dimensional neighbor representation is characterized. In this framework, it is found that the clustering performance is improved to a wide-separated visual representation, as shown in the experimental results.

Even with the aforementioned benefits, SNE can be computationally expensive, especially as the size of the dataset grows. This is due to the complexity of calculating pairwise probabilities and the iterative nature of minimizing the cost function. Although the proposed method inherits the limitations from the SNE framework, this weakness can be mitigated by GPU acceleration. This work focuses on comprehending the underlying mechanism of SNE clustering and generalizing it with favorable characteristics. The contributions of this work are summarized as follows:

1) This work investigates the fundamental mechanism of SNEs by analyzing individual fundamental principles and general properties. Moreover, the connection to attractive and repulsive forces is discussed as seen in Sec. IV, which yields insights into how SNE can be enhanced.
2) A new generalized SNE is proposed by the extended formulation to provide flexibility with favorable features where visual representations are illustrated as seen in Section III.
3) Our analyses demonstrate the influence of divergence measures in a loss function. By relaxing the typical Kullback-Leibler divergence into α-divergence, the proposed method is equipped with various capacities for realizing data similarity, as addressed in Sec. IV-A.

II. RELATED WORKS

SNE has several variants that adopt different types of similarity distributions in high-dimensional and low-dimensional space. The original SNE [7] describes the similarity distributions with conditional probabilities using the form of the Gaussian distribution

$$p_{j|i} = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)}{\sum_{k\neq i}\exp\left(-\|\mathbf{x}_i - \mathbf{x}_k\|^2\right)} \tag{1}$$

$$q_{j|i} = \frac{\exp\left(-\|\mathbf{y}_i - \mathbf{y}_j\|^2\right)}{\sum_{k\neq i}\exp\left(-\|\mathbf{y}_i - \mathbf{y}_k\|^2\right)} \tag{2}$$

where $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$ denote samples in the high-dimensional and the low-dimensional space, respectively. $p_{j|i}$ and $q_{j|i}$ denote the conditional probability distributions given $\mathbf{x}_i$ and $\mathbf{y}_i$ of the high-dimensional and low-dimensional spaces, respectively. The symmetric SNE [23], on the other hand, employs the joint probabilities with the Gaussian distribution with variance parameter.

Following the symmetric assumption, t-SNE [8] was then proposed to deploy the Student-t distribution of the following symmetric joint probability form for the low-dimensional space:

$$q_{ij} = \frac{\left(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2\right)^{-1}}{\sum_{k\neq l}\left(1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2\right)^{-1}}. \tag{3}$$

On the other hand, the p-SNE [22] applied the power-law distribution in the following form for low-dimensional space

$$q_{ij} = \begin{cases} c/\|\mathbf{y}_i - \mathbf{y}_j\|^2 & \text{if } \|\mathbf{y}_i - \mathbf{y}_j\| \geq r_0, \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

with a cut-off radius $r_0 > 0$. A spherical version of SNE, called s-SNE, was also introduced to visualize the hyperspectral data in [24], where the target geometry was set to be

the non-Euclidean d-dimensional unit sphere $\mathbb{S}^d$. The discrete exit distribution [25] was then adopted as

$$q_{ij} = \frac{\|\mathbf{y}_j - \rho\mathbf{y}_i\|^{-d}}{\sum_{k\neq l}\|\mathbf{y}_k - \rho\mathbf{y}_l\|^{-d}} \tag{5}$$

for any latent pair $\mathbf{y}_i, \mathbf{y}_j \in \mathbb{S}^{d-1}$ with a constant $\rho \in [0, 1)$.

In fact, various SNEs can be considered as special realizations of the method we develop next in Sec. III under the general Riemannian geometry with the arbitrary inner products (called *Riemannian metrics*) to provide the *weighted* similarity measures [26]. In deep learning terms, they correspond to the *attention* (weights) of the projected data. The proposed method derived from the Riemannian framework will then contain rich information in embedding structures.

Typically, SNEs were implemented by minimizing the KL divergence between high and low-dimensional distributions of data points. Other divergence measures such as β, γ, or Rényi may also be considered as well [27], [28]. Notably, one may also attempt to symmetrize the divergence measure between P and Q by combining two KL loss functions [29].

As divergence functions were studied and explored for their important role in general learning theories and optimization implementations, different classes of divergence functions can be incorporated into the SNE framework. α-divergence [30], [31] is a broader convex divergence function that includes the KL divergence as a special case. A study for a further generalized convex divergence measure can be found in [32].

This work extensively generalizes SNE and sets up a mathematical foundation that maximally captures the geometric features for clustering performance. An additionally proposed distribution in the study possesses both the heavy-tail property and the quick-growing behavior near the origin. There are two parameters η and β to adjust the probability decay rate and concentration at the origin. On the other hand, α-divergence is utilized to tune the clustering effect with a convex parameter α. The proposed generalization naturally includes the t-SNE as a special case while having a similar yet smoother behavior near the origin as the p-SNE. The comparison of the previous methods to the proposed framework will be discussed in Sec. IV.

III. GENERALIZED STOCHASTIC NEIGHBOR EMBEDDING

A. GENERAL FRAMEWORK

This study generalizes the SNE from the Riemannian geometry perspective. Let $(\mathcal{M}, h)$ and $(\mathcal{N}, g)$ be the Riemannian manifolds representing the *high-dimensional space* of dimension D and the *low-dimensional space* of dimension d, respectively. Given a set of N original (high-dimensional) data points $\mathcal{X} = \{\mathbf{x}_i \in \mathcal{M} | i = 1, \ldots, N\}$, the goal of SNE is to find a low-dimensional representation $\mathcal{Y} = \{\mathbf{y}_i \in \mathcal{N} | i = 1, \ldots, N\}$ such that $\mathbf{y}_i$ preserves the *pairwise similarity* of $\mathbf{x}_i$ in the latent space $\mathcal{N}$. The pairwise similarities in $\mathcal{M}, \mathcal{N}$ are typically modeled by the conditional probability $p_{j|i}$ of finding $\mathbf{x}_j$ as a neighbor of $\mathbf{x}_i$ and the conditional probability $q_{j|i}$ of having $\mathbf{y}_j$ as a neighbor of $\mathbf{y}_i$, where the exact definition of $p_{j|i}$ and $q_{j|i}$ will depend on the *distance measure* given by the Riemannian metric later in Eq. (7).

$$\begin{array}{ccc} (\mathcal{M}, \mathcal{X}) & \xrightarrow{\quad\varphi\quad} & (\mathcal{N}, \mathcal{Y}) \\ \Big\downarrow \xi & & \Big\downarrow \tau \\ (\Sigma, \mathcal{F}_\Sigma, P) & \xrightarrow[\text{SNE}]{\eta} & (\Omega, \mathcal{F}_\Omega, Q) \end{array}$$

FIGURE 1. A dimension reduction map φ from the high-dimensional space $\mathcal{M}$ to the low-dimensional space $\mathcal{N}$. The process of stochastic neighbor embedding is depicted by dash lines.

Following closely on the interpretations and definitions in [8], we know that comparing and aligning two sets of probability distributions $P = \{p_i\} = \{p_{j|i}\}$ and $Q = \{q_i\} = \{q_{j|i}\}$ will serve the purpose of *preserving data structures*. The process of matching P and Q is rendered by minimizing the (probability) divergences over samples as a cost function $\mathcal{L}$

$$\mathcal{L}(P, Q) = \sum_i \mathcal{D}_f(p_i \| q_i) = \sum_i \sum_j p_{ij} f\left(\frac{q_{ij}}{p_{ij}}\right) \tag{6}$$

where $p_{ij} := (p_{i|j} + p_{j|i})/2N$, $q_{ij} := (q_{i|j} + q_{j|i})/2N$ with sample size N are as defined in [33], and $\mathcal{D}_f(p_i \| q_i)$ is a divergence measure between probability measures p_i and q_i of samples $\mathbf{x}_i$ and $\mathbf{y}_i$, respectively. The specific form of p_{ij} and q_{ij} in our generalization is to be discussed in Eq. (10), (11).

A convex differentiable function $f : \mathbb{R}^+ \to \mathbb{R}$ is typically chosen with $f(1) = 0$. This objective evaluates how much two sets of distributions $p_i = \{p_{ij}\}$ and $q_i = \{q_{ij}\}$ agree to one another. Indeed, the loss function Eq. (6) attains the minimum value 0 when $p_{ij} = q_{ij}$ for the pairwise samples i, j.

As a result, the minimization of Eq. (6) acts as an implicit dimension reduction map $\varphi : \mathcal{M} \to \mathcal{N}$ found by an SNE to obtain the compressed samples $\mathcal{Y} = \varphi(\mathcal{X})$ in $\mathcal{N}$. Note that by considering the distribution of pairwise similarity, SNE transforms data $\mathcal{X}$ and $\mathcal{Y}$ into probability spaces $(\Sigma, \mathcal{F}_\Sigma, P)$ and $(\Omega, \mathcal{F}_\Omega, Q)$ with induced mapping ξ and τ respectively, as shown in Fig. 1, where $\mathcal{F}_\Sigma = 2^\Sigma$, $\mathcal{F}_\Omega = 2^\Omega$ denote their corresponding σ-algebras [34]. The learning process of SNE occurs in probability spaces by minimizing the divergence between two distributions. This mechanism is different from dimension reduction maps such as PCA and autoencoders [35], where a mapping φ is directly sought between data spaces $\mathcal{X}$ and $\mathcal{Y}$.

B. PROPOSED METHOD

Our formulation above contains three key factors that uniquely determine an SNE:

1) The *distance measure* between sample pairs $\{\mathbf{x}_i, \mathbf{x}_j\}$ on $\mathcal{M}$ and $\{\mathbf{y}_i, \mathbf{y}_j\}$ on $\mathcal{N}$,
2) The *distribution function* for conditional probabilities $p_{j|i}$ and $q_{j|i}$ (or joint probabilities p_{ij} and q_{ij}),

3) The *divergence measure* between probability spaces P and Q.

The proposed method, called the generalized SNE (*g*-SNE), extends these three factors to incorporate favorable features for clustering. The three extensions are explained in the following subsections:

1) DISTANCE MEASURES

Neighboring embedding is affected by the distance measure in the data space. A distance measure can stem from a Riemannian metric [36] that uniquely characterizes the geometry of a space, such as the shape of a sphere or a torus. Given Riemannian metrics $h : T\mathcal{M} \times T\mathcal{M} \to \mathbb{R}$ on $\mathcal{M}$ and $g : T\mathcal{N} \times T\mathcal{N} \to \mathbb{R}$ on $\mathcal{N}$, intrinsic *distances* on $\mathcal{M}$ and $\mathcal{N}$ are induced by

$$\begin{aligned} d_{ij} &:= d(\mathbf{x}_i, \mathbf{x}_j) := \sqrt{h(\mathbf{x}_i - \mathbf{x}_j, \mathbf{x}_i - \mathbf{x}_j)} \\ r_{ij} &:= r(\mathbf{y}_i, \mathbf{y}_j) := \sqrt{g(\mathbf{y}_i - \mathbf{y}_j, \mathbf{y}_i - \mathbf{y}_j)} \end{aligned} \tag{7}$$

where $T\mathcal{M}$, $T\mathcal{N}$ denote the tangent bundles of $\mathcal{M}$, $\mathcal{N}$, and certain *parallel transport* structures are assumed for simplifications. Consequently, Eq. (7) results in a general form for distance measures

$$d_{ij} = \sqrt{\sum_{k=1}^{D}\sum_{l=1}^{D} h_{kl}(x_{ik} - x_{jk})(x_{il} - x_{jl})} \tag{8}$$

$$r_{ij} = \sqrt{\sum_{k=1}^{d}\sum_{l=1}^{d} g_{kl}(y_{ik} - y_{jk})(y_{il} - y_{jl})} \tag{9}$$

where $\{h_{kl}, g_{kl}\}$ are *dynamical weights* at the coordinate $\{x_{ik}, x_{jl}, y_{ik}, y_{jl}\}$ of vectors $\{\mathbf{x}_i, \mathbf{x}_j, \mathbf{y}_i, \mathbf{y}_j\}$. Again, D and d are the dimensions of $\mathcal{M}$ and $\mathcal{N}$, respectively. Indeed, given a local coordinate or a basis $\{\mathbf{v}_k\}_{k=1}^{d}$ on $\mathcal{N}$, a latent point can be expanded by $\mathbf{y}_i = \sum_k y_{ik}\mathbf{v}_k$ with coefficients $y_{ik} \in \mathbb{R}$. Under the expansion, one has $g(\mathbf{y}_i, \mathbf{y}_j) = \sum_k \sum_l g_{kl} y_{ik} y_{jl}$ with $g_{kl} := g(\mathbf{v}_k, \mathbf{v}_l)(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ to give Eq. (9). Since each g_{kl} is a real-valued function of $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$, it can be regarded as the dynamical weighting at samples $\mathbf{y}_1, \ldots, \mathbf{y}_N$. The regular Euclidean distance $\|\mathbf{y}_i - \mathbf{y}_j\|^2 = \sum_k (y_{ik} - y_{jk})^2$ is then seen as a special case when $g_{kl} = \delta_{kl}$, a Kronecker delta function. The Mahalanobis distance [33] becomes another special realization when g_{kl} are all constants. In the term of deep learning, g_{kl} is an *attention weight* [37]. Similar computations can be carried out to yield attention weights h_{kl} in Eq. (8).

2) DISTRIBUTION FUNCTIONS

With the generalized distance measures $\{d_{ij}\}$ and $\{r_{ij}\}$ defined by Eq. (8), and (9), respectively, we may further generalize the distribution functions $P : \mathcal{F}_\Sigma \to [0, 1]$ and $Q : \mathcal{F}_\Omega \to [0, 1]$ by

$$P(\mathbf{x}_i, \mathbf{x}_j) = p_{ij} = \frac{p(d_{ij})}{\sum_{k \neq l} p(d_{kl})} \tag{10}$$

$$Q(\mathbf{y}_i, \mathbf{y}_j) = q_{ij} = \frac{q(r_{ij})}{\sum_{k \neq l} q(r_{kl})} \tag{11}$$

where $p : \mathbb{R}^+ \to \mathbb{R}^+$ and $q : \mathbb{R}^+ \to \mathbb{R}^+$ are non-negative decreasing functions, *i.e.*, $\dot{p}(d) \leq 0$ and $\dot{q}(r) \leq 0$ and the conditions $P(\mathbf{x}_i, \mathbf{x}_i) = 0$ and $Q(\mathbf{y}_i, \mathbf{y}_i) = 0$ are met. Distribution $P(\mathbf{x}_i, \mathbf{x}_j)$ characterizes the probability of having sample $\mathbf{x}_j$ as a neighbor of $\mathbf{x}_i$, similarly for $Q(\mathbf{y}_i, \mathbf{y}_j)$ in a way to reflect the similarity or affinity between data points.

This is usually a key measure in an SNE that shapes up the physical behavior of the latent space. In this regard, the original data distribution $p(d_{ij})$ is assumed to be Gaussian,

$$p(d_{ij}) = \exp\left(-d_{ij}^2\right). \tag{12}$$

We propose a *regularized power-law distribution* in $\mathcal{N}$ by letting

$$q(r_{ij}) = \frac{1}{\eta + r_{ij}^{\beta}} \tag{13}$$

with two parameters $\eta > 0$ and $\beta > 0$. The distribution is heavy-tailed for all $\beta > 0$, which controls the decay rate, while η connects to the steepness of how r_{ij} approaches zero. This distribution is proposed to mitigate the crowding problem by merging the beneficial characteristics of clustering effects in t-SNE [8] and p-SNE [22]. Indeed, our extension includes the Student-t distribution in Eq. (3) as a special case when $\eta = 1$ and $\beta = 2$, and the power-law distribution in Eq. (4) as $\eta \to 0$, $\beta = 2$.

3) DIVERGENCE MEASURES

To obtain the generality in comparing probability distributions, the α-divergence is employed in Eq. (6) as a measurement between P and Q by defining

$$f_\alpha(t) = \frac{4}{1 - \alpha^2}\left[\frac{1 - \alpha}{2} + \left(\frac{1 + \alpha}{2}\right)t - t^{\frac{1+\alpha}{2}}\right] \tag{14}$$

where $t > 0$ and $\alpha \in \mathbb{R}$ is a convexity parameter. Eq. (14) was derived by considering the 1-parameter α-connection family where the corresponding statistical interpretations have been explored and addressed in [38].

The α-divergence is differentiable and convex with $f''_\alpha(t) = t^{(\alpha-3)/2} \geq 0$. These nice properties allow this framework to explicitly calculate and analyze the SNE optimization process. Furthermore, the α-divergence includes the KL divergence and the Hellinger distance as two special cases when $\alpha = -1$ and $\alpha = 0$, respectively. This variational nature of α offers flexibility when catering to diverse data.

IV. ATTRACTIVE AND REPULSIVE FORCES IN SNES

The optimal solution $\mathcal{Y}^* = (\mathbf{y}_1^*, \ldots, \mathbf{y}_N^*)$ of the objective function in Eq. (6) determines the final SNE clustering result. The minimization process $\{\mathbf{y}_i(t) \to \mathbf{y}_i(t+1)\}$ for each iteration t of sample i characterizes the SNE mechanism. As there is no closed-form for the *dynamical process* $\mathbf{y}_i(t) \xrightarrow{t\to\infty} \mathbf{y}_i^*$, the gradient descent (GD) algorithm is commonly utilized to estimate the optimal solution. Consequently, the behavior of gradient values $\frac{\partial \mathcal{L}}{\partial \mathbf{y}_i}$ can be used to analyze and compare different types of SNEs.

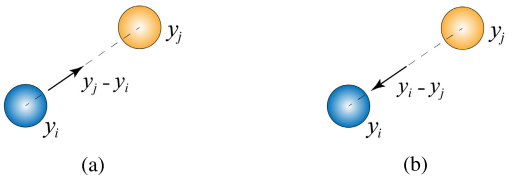


FIGURE 2. The directions of (a) an attractive force and (b) a repulsive force on particle $\mathbf{y}_i$ due to another particle $\mathbf{y}_j$ in the latent space.

In fact, the gradient vectors can be connected to the *attractive* and *repulsive* forces from the perspective of mechanics. By depicting the loss function in Eq. (6) as a force field, the neighbor embedding behavior in $\mathcal{N}$ can be illustrated as the interaction between particles.

A mechanical system of N particles in $\mathcal{N} = \mathbb{R}^d$ with mass m and position $\mathbf{y}_i(t) \in \mathcal{N}$ subject to the potential energy $V = V(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ has the Lagrangian [39]

$$\mathscr{L}(\mathbf{y}_1, \ldots, \mathbf{y}_N, \dot{\mathbf{y}}_1, \ldots, \dot{\mathbf{y}}_N) = T - V \tag{15}$$

where $\dot{\mathbf{y}}_i$ is the velocity and $T = \frac{m}{2}\sum_{i=1}^{N}\|\dot{\mathbf{y}}_i\|^2$ denotes the total kinetic energy. Note that the Lagrangian $\mathscr{L}$ is to be distinguished with the loss function $\mathcal{L}$. The equation of motion for N particles Eq. (15) follows from the Euler-Lagrange equation $\frac{\partial \mathscr{L}}{\partial \mathbf{y}_i} = \frac{d}{dt}\frac{\partial \mathscr{L}}{\partial \dot{\mathbf{y}}_i}$ to give

$$F_i = m\ddot{\mathbf{y}}_i = -\frac{\partial V}{\partial \mathbf{y}_i} \tag{16}$$

which indicates the force exerted on particle $\mathbf{y}_i$ as $\ddot{\mathbf{y}}_i$ denotes the acceleration of $\mathbf{y}_i$.

SNE fits into this perspective when the learning objective $\mathcal{L}$ is identified with potential V. Put $V = \mathcal{L}$, one has

$$F_i = -\frac{\partial V}{\partial \mathbf{y}_i} = -\frac{\partial \mathcal{L}}{\partial \mathbf{y}_i}. \tag{17}$$

Then the loss *gradient* $\frac{\partial \mathcal{L}}{\partial \mathbf{y}_i}$ is legitimately regarded as the *force* on $\mathbf{y}_i$, by a negative sign. The existence of force naturally leads to the motion of latent variables $\mathbf{y}_i$, and therefore, the optimization process $\mathbf{y}_i(t) \xrightarrow{t\to\infty} \mathbf{y}_i^*$ to be discussed.

With the perspective of mechanics, we first examine a special case to see how the neighbor embedding process is transformed into force interactions. Under the KL divergence and $\mathcal{M} = \mathbb{R}^D, \mathcal{N} = \mathbb{R}^d$, Eqs. (6), (10), (11), and (17) yields

$$F_i = -2\sum_j p_{ij}\frac{\dot{q}(r_{ij})}{q(r_{ij})}\left(\frac{\mathbf{y}_j - \mathbf{y}_i}{r_{ij}}\right) + \frac{2}{Z}\sum_j \dot{q}(r_{ij})\left(\frac{\mathbf{y}_j - \mathbf{y}_i}{r_{ij}}\right), \tag{18}$$

where $Z = \sum_{k\neq l} q(r_{kl})$ is the partition function for normalization. Here $(\mathbf{y}_j - \mathbf{y}_i)/r_{ij}$ denotes the unit vector of attractive force. It is found that attractive force is greater than repulsive force whenever $p_{ij} > q_{ij}$ and vice versa so that the system seeks for the *equilibrium* when $p_{ij} \equiv q_{ij}$, which meets our intuition for the manifold learning [40], [41]. Notably, the joint distribution p_{ij} of original data $\{\mathbf{x}_i, \mathbf{x}_j\}$ only affects the *attractive force* rather than repulsive force. Repulsive force behaves like a global effect.

For the proposed generalizations in Sec. III-B, the gradient can be analytically computed by Eqs. (6), (8), (9), (10), (11), and (17) in a form of

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{y}_i} = {} & \frac{2}{Z}\sum_j \dot{q}(r_{ij})f'\left(\frac{q_{ij}}{p_{ij}}\right)\left(\frac{\mathbf{y}_i - \mathbf{y}_j}{r_{ij}}\right) \\ & - \frac{2}{Z^2}\left[\sum_{k\neq l} f'\left(\frac{q_{kl}}{p_{kl}}\right)q(r_{kl})\right]\left[\sum_j \dot{q}(r_{ij})\left(\frac{\mathbf{y}_i - \mathbf{y}_j}{r_{ij}}\right)\right] \\ & + \frac{1}{2Z}\sum_{k\neq l}\dot{q}(r_{kl})f'\left(\frac{q_{kl}}{p_{kl}}\right)\frac{(\nabla_{\mathbf{y}_i} r)(\mathbf{y}_k, \mathbf{y}_l)}{r_{kl}} \\ & - \frac{1}{2Z^2}\left[\sum_{m\neq n} f'\left(\frac{q_{mn}}{p_{mn}}\right)q(r_{mn})\right] \\ & \left[\sum_{k\neq l}\dot{q}(r_{kl})\frac{(\nabla_{\mathbf{y}_i} r)(\mathbf{y}_k, \mathbf{y}_l)}{r_{kl}}\right]\end{aligned} \tag{19}$$

where $f'(t) = \frac{d}{dt}f(t)$ and $Z = \sum_{k\neq l} q(r_{kl})$ and

$$(\nabla_{\mathbf{y}_i} r)(\mathbf{y}_k, \mathbf{y}_l) = \sum_{m\neq n}\left(\frac{\partial g_{mn}}{\partial \mathbf{y}_i}\right)(y_{km} - y_{lm})(y_{kn} - y_{ln}).$$

In the usual implementation where the Euclidean inner product as the Kronecker delta function $g_{nm} = \delta_{nm}$ is considered, Eq. (19) can be simplified as

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{y}_i} = {} & \frac{2}{Z}\sum_j \dot{q}(r_{ij})f'\left(\frac{q_{ij}}{p_{ij}}\right)\left(\frac{\mathbf{y}_i - \mathbf{y}_j}{r_{ij}}\right) \\ & - \frac{2}{Z^2}\left[\sum_{k\neq l} f'\left(\frac{q_{kl}}{p_{kl}}\right)q(r_{kl})\right]\left[\sum_j \dot{q}(r_{ij})\left(\frac{\mathbf{y}_i - \mathbf{y}_j}{r_{ij}}\right)\right].\end{aligned} \tag{20}$$

The effect of divergence measures can then be observed succinctly here. Furthermore, using α-divergence as given in Eq. (14), the gradient in Eq. (20) becomes

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{y}_i} = {} & \frac{4}{(1-\alpha)Z}\sum_j \dot{q}(r_{ij})\left(1 - \left(\frac{q_{ij}}{p_{ij}}\right)^{\frac{\alpha-1}{2}}\right)\left(\frac{\mathbf{y}_i - \mathbf{y}_j}{r_{ij}}\right) \\ & - \frac{4}{(1-\alpha)Z^2}\left[\sum_{k\neq l}\left(1 - \left(\frac{q_{kl}}{p_{kl}}\right)^{\frac{\alpha-1}{2}}\right)q(r_{kl})\right] \\ & \times \left[\sum_j \dot{q}(r_{ij})\left(\frac{\mathbf{y}_i - \mathbf{y}_j}{r_{ij}}\right)\right].\end{aligned} \tag{21}$$

It can be seen that the force composition can be complicated in general. However, when the special case using KL-divergence $\alpha = -1$ is adopted, the resulting g-SNE reveals that

$$\frac{\partial \mathcal{L}}{\partial \mathbf{y}_i} = 2\sum_j p_{ij}\frac{\dot{q}(r_{ij})}{q(r_{ij})}\left(\frac{\mathbf{y}_j - \mathbf{y}_i}{r_{ij}}\right) - \frac{2}{Z}\sum_j \dot{q}(r_{ij})\left(\frac{\mathbf{y}_j - \mathbf{y}_i}{r_{ij}}\right). \tag{22}$$

TABLE 1. Comparison of the attractive and repulsive forces of various SNEs, when considering the special case of $g_{nm} = \delta_{nm}$ and $\alpha = -1$.

	SNE	t-SNE	p-SNE	g-SNE
$q(r_{ij})$	$e^{-r_{ij}^2}$	$(1+r_{ij}^2)^{-1}$	$r_{ij}^{-\beta}$	$(\eta + r_{ij}^{\beta})^{-1}$
attractive force	$4p_{ij}r_{ij}$	$\frac{4p_{ij}r_{ij}}{1+r_{ij}^2}$	$\frac{2\beta p_{ij}}{r_{ij}}$	$\frac{2\beta p_{ij} r_{ij}^{\beta-1}}{\eta + r_{ij}^{\beta}}$
repulsive force	$\frac{1}{Z}4r_{ij}e^{-r_{ij}^2}$	$\frac{1}{Z}\frac{4r_{ij}}{(1+r_{ij}^2)^2}$	$\frac{1}{Z}\frac{2\beta}{r_{ij}^{\beta+1}}$	$\frac{1}{Z}\frac{2\beta r_{ij}^{\beta-1}}{(\eta + r_{ij}^{\beta})^2}$

A few more properties can be observed here. First, in Eq. (22), the first term is the *attractive force*, and the second term is the *repulsive force* since $\dot{q}(r_{ij}) < 0$. Second, *only* attractive force is coupled to p_{ij}. Namely, only attraction is related to the raw data affinity p_{ij} while the repulsion is universal, which does *not* concern data $\{\mathbf{x}_i\}$ at all. Third, Eq. (22) depicts that the attraction is proportional to the *decay rate* of the target similarity $\dot{q}_{ij}$ since p_{ij} is fixed. In fact, it is determined by $\dot{q}(r_{ij})/q(r_{ij}) = \frac{d}{dr}\log q(r_{ij}) < 0$, which is generally weak in magnitude since $\log q(r_{ij})$ is involved. Therefore, qualitatively choosing a *fast-decaying* target similarity will increase the latent particle force for separation. Fourth, the proposed g-SNE in Eq. (19) is not restricted to the α-divergence. The other divergences can also be applied, such as the χ^2-divergence by letting $f(t) = (t-1)^2$.

A. COMPARISON BETWEEN SNES

We can utilize Eq. (21) as a comprehensive tool to compare the following four realizations of SNE.

$$F_i^{\text{sne}} = 4\sum_j \left[p_{ij}r_{ij} - \frac{1}{Z}r_{ij}\exp(-r_{ij}^2)\right]\left(\frac{\mathbf{y}_j - \mathbf{y}_i}{r_{ij}}\right) \quad (23)$$

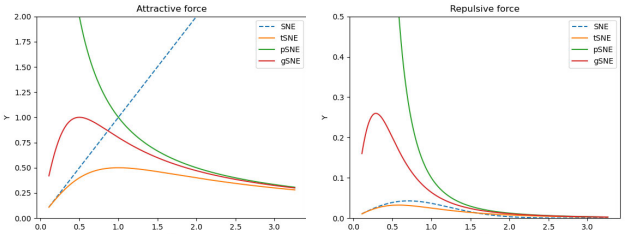
$$F_i^{t\text{-sne}} = 4\sum_j \left[\frac{p_{ij}r_{ij}}{1+r_{ij}^2} - \frac{1}{Z}\frac{r_{ij}}{(1+r_{ij}^2)^2}\right]\left(\frac{\mathbf{y}_j - \mathbf{y}_i}{r_{ij}}\right) \quad (24)$$

$$F_i^{p\text{-sne}} = 2\sum_j \left[\frac{\beta p_{ij}}{r_{ij}} - \frac{1}{Z}\frac{\beta}{r_{ij}^{\beta+1}}\right]\left(\frac{\mathbf{y}_j - \mathbf{y}_i}{r_{ij}}\right) \quad (25)$$

$$F_i^{g\text{-sne}} = 2\sum_j \left[\frac{p_{ij}\beta r_{ij}^{\beta-1}}{\eta + r_{ij}^{\beta}} - \frac{1}{Z}\frac{\beta r_{ij}^{\beta-1}}{(\eta + r_{ij}^{\beta})^2}\right]\left(\frac{\mathbf{y}_j - \mathbf{y}_i}{r_{ij}}\right). \quad (26)$$

B. DISCUSSION ON THE EXTENSION OF DISTRIBUTIONS

The calculations provide comparisons between SNEs from the perspective of forces. Table 1 contrasts the analytical forms of attractive and repulsive forces among four SNEs. Observing the repulsive force of SNE and t-SNE, it is found that the repulsive force of t-SNE has polynomial decays while r_{ij} increases, contrasting the exponential decays of the traditional SNE. The persistent repulsive force leads to improved cluster separation while reaching the equilibrium. From the perspective of attractive and repulsive forces, we see



(a) (b)

FIGURE 3. Magnitudes of (a) the attractive forces and (b) the repulsive forces versus latent distance r_{ij} (horizontal axis) for SNE, t-SNE, p-SNE, and g-SNE with $\beta = 2$, $\eta = 0.25$.

that t-SNE mitigates the crowding problem by increasing repulsion to those lower-dimensional particles $\mathcal{Y} = \{\mathbf{y}_i\}$.

p-SNE and g-SNE both possess polynomial decay as t-SNE to keep favorable characteristics for clustering. In addition, the hyperparameter β adjusts the strength of polynomial decay in p-SNE and g-SNE. When $\beta = 2$, the decay rate of p-SNE and g-SNE are the same as that of t-SNE. As β decreases, the decay rate of the repulsive force decreases, which results in stronger repulsive forces for large r_{ij}.

The behavior of attractive force, on the other hand, is different among t-SNE, p-SNE, and g-SNE. As r_{ij} decreases, the attractive force by t-SNE converges to zero, while p-SNE diverges to infinity. A strong attractive force may encourage the clustering effect, but the divergence may be harmful for reaching equilibrium and could result in instability. The design of g-SNE possesses the convergence around small r_{ij} for stability and, at the same time, introduces the parameter η to control the power of attractive forces. Due to the strong-force motions that lead to widely separated particles of different affinities, the crowding problem in g-SNE can be substantially mitigated.

The choices of hyperparameters for data visualization can be assisted with the understanding from the perspective of the forces. Figures 9 and 10 show the clustering effect of the hyperparameters β and η. Ideal clustering is expected to attract high-similarity samples (small r_{ij}) and repulse those with low similarities (large r_{ij}). The hyperparameter β controls the decay rate of repulsive forces as r_{ij} increases. As β increases, the repulsive force decays faster, which results in weaker repulsive forces between low-similarity samples. On the other hand, η connects to the strength of forces for sample pairs. As η decreases, the strength increases, which leads to better cluster separations.

C. DISCUSSION ON THE EXTENSION OF DIVERGENCE MEASURES

As smaller (or more negative) α in divergence measure Eq. (14) yields higher costs $\mathcal{L}_\alpha(P, Q)$ as shown in Table 3, the target probability Q is forced to get closer to the original data distribution P to result in loose clustering results such as the right most subfigures in Figs. 12 and 13. This illuminates

that total recovery of P by Q may not always be desired in the compressed dimension. Instead, allowing Q to deviate from P slightly provides wiggle room for latent samples to mitigate the crowding problem. The experimental results from this study suggest that a slight probability distortion improves the clustering effects in lower dimensions.

Considering the cases of larger α, the sensitivity to probability measure is then weakened to reflect flexibility on the latent points $\mathbf{y}_i$ of q_{ij} so that attractive and repulsive forces have more freedom to move latent particles without causing considerable variation in the loss function. This result justifies why larger α provides wider separation and more compact clustering. Consequently, one can tune α to adjust the divergence measure and sensitivity of similarity for the dataset to be learned in an unsupervised manner. In an illuminating example in Fig. 14(a), one may adjust proper α to unravel the entangled map in SNE with $\alpha = -1$ (Fig. 4(a)) such that this method has the potential to further reduce the crowding problem like t-SNE. Thus, from this example, one observes that the crowding problem does not solely result from the probability distribution but also from the divergence measure.

V. CONNECTION WITH SPHERICAL SNE

Our discussion below explores the connection between the spherical SNE (denoted by s-SNE) [24] and the g-SNE. The s-SNE maps the Gaussian probability $P = \{p_{ij}\}$ of flat geometry $\mathcal{M} = \mathbb{R}^D$ in Eq. (1) to the exit distribution $Q = \{q_{ij}\}$ of Eq. (5) on a sphere $\mathcal{N} = \mathbb{S}^{d-1}$.

The latent map $\mathcal{Y} = \{\mathbf{y}_i \in \mathbb{S}^{d-1}\}$ is obtained by minimizing the KL divergence

$$\mathcal{L}_{s\text{-sne}} = \sum_i \sum_j p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) + \sum_i \lambda_i \left(1 - \|\mathbf{y}_i\|^2\right) \quad (27)$$

where the spherical constraint is imposed by Lagrange multiplier $\lambda_i \in \mathbb{R}$. Then, the gradient of Eq. (27) is given by

$$\begin{aligned}\frac{\partial \mathcal{L}_{s\text{-sne}}}{\partial \mathbf{y}_i} &= d \sum_j p_{ij} \left(\frac{\mathbf{y}_i - \rho \mathbf{y}_j}{\|\mathbf{y}_i - \rho \mathbf{y}_j\|^2} + \rho \frac{\rho \mathbf{y}_i - \mathbf{y}_j}{\|\mathbf{y}_j - \rho \mathbf{y}_i\|^2}\right) \\ &\quad + d \left(\rho \sum_j q_{ij} \frac{\mathbf{y}_j - \rho \mathbf{y}_i}{\|\mathbf{y}_j - \rho \mathbf{y}_i\|^2} - \sum_j q_{ji} \frac{\mathbf{y}_i - \rho \mathbf{y}_j}{\|\mathbf{y}_i - \rho \mathbf{y}_j\|^2}\right) \\ &\quad - 2\lambda_i \mathbf{y}_i. \end{aligned} \quad (28)$$

The detailed derivation from Eq. (27) to (28) can be found in Appendix A.

We note that when $\rho \to 1$, we let $\rho = 1 - \epsilon$ with $\epsilon \to 0^+$ and the first two terms on the right-hand-side of Eq. (28) can be decomposed by Taylor series with respect to ϵ at reasonable distance $r_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\| \gg \epsilon$

$$\begin{aligned}\frac{\mathbf{y}_i - \rho \mathbf{y}_j}{||\mathbf{y}_i - \rho \mathbf{y}_j||^2} &= \frac{1}{r_{ij}^2} \left[(\mathbf{y}_i - \mathbf{y}_j) + \epsilon \mathbf{y}_j - \frac{2\epsilon}{r_{ij}^2} \right. \\ &\quad \left. \times \langle \mathbf{y}_j, \mathbf{y}_i - \mathbf{y}_j \rangle (\mathbf{y}_i - \mathbf{y}_j)\right] + \mathcal{O}(\epsilon^2) \end{aligned} \quad (29)$$

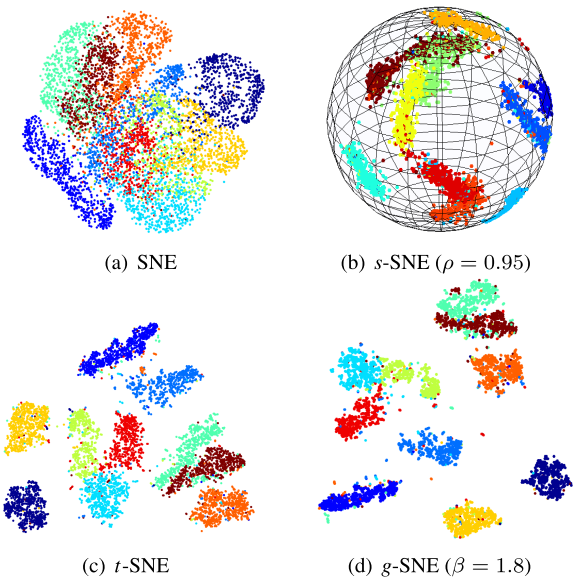


FIGURE 4. A visualization of MNIST digits in four SNEs. The proposed g-SNE shows a map with tighter and more widespread grouping.

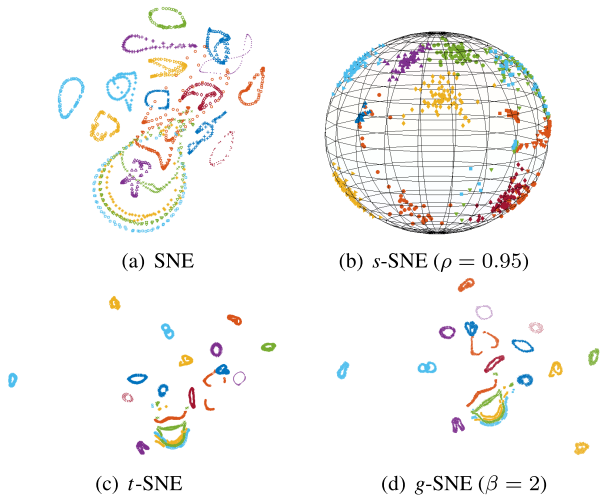


FIGURE 5. A visualization of COIL-20 dataset in four SNEs.

$$\begin{aligned}\rho \frac{\rho \mathbf{y}_i - \mathbf{y}_j}{||\mathbf{y}_j - \rho \mathbf{y}_i||^2} &= \frac{1-\epsilon}{r_{ij}^2} \left[(\mathbf{y}_i - \mathbf{y}_j) - \epsilon \mathbf{y}_i + \frac{2\epsilon}{r_{ij}^2} \right. \\ &\quad \left. \times \langle \mathbf{y}_i, \mathbf{y}_i - \mathbf{y}_j \rangle (\mathbf{y}_i - \mathbf{y}_j)\right] + \mathcal{O}(\epsilon^3) \end{aligned} \quad (30)$$

where $\mathcal{O}(\epsilon^k)$ denotes the k^{th}-order term of ϵ and beyond. The other two terms of Eq. (28) can be similarly decomposed so that collectively we have

$$\begin{aligned} F_i^{s\text{-sne}} &= -\frac{\partial \mathcal{L}_{s\text{-sne}}}{\partial \mathbf{y}_i} \\ &= 2\, d \sum_j \frac{1}{r_{ij}} \left(p_{ij} - q_{ij}\right) \left(\frac{\mathbf{y}_j - \mathbf{y}_i}{r_{ij}}\right) + 2\lambda_i \mathbf{y}_i + \mathcal{O}(\epsilon^1). \end{aligned} \quad (31)$$

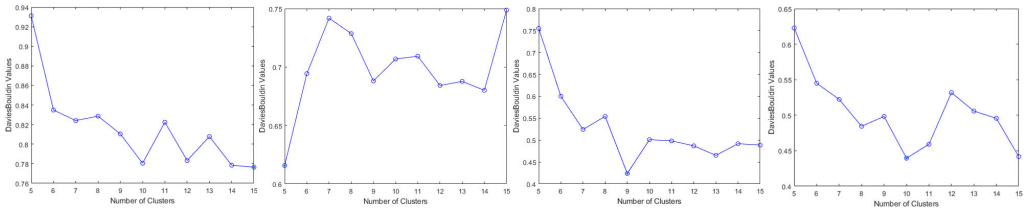


FIGURE 6. Davies-Bouldin indices of SNE, *s*-SNE, *t*-SNE and *g*-SNE (from left to right) in MNIST corresponding to Fig. 4. The true class number is 10.

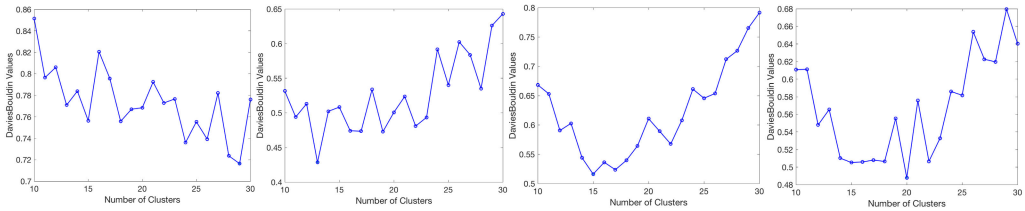


FIGURE 7. Davies-Bouldin indices of SNE, *s*-SNE, *t*-SNE and *g*-SNE (from left to right) in COIL-20 corresponding to Fig. 5. The true class number is 20.

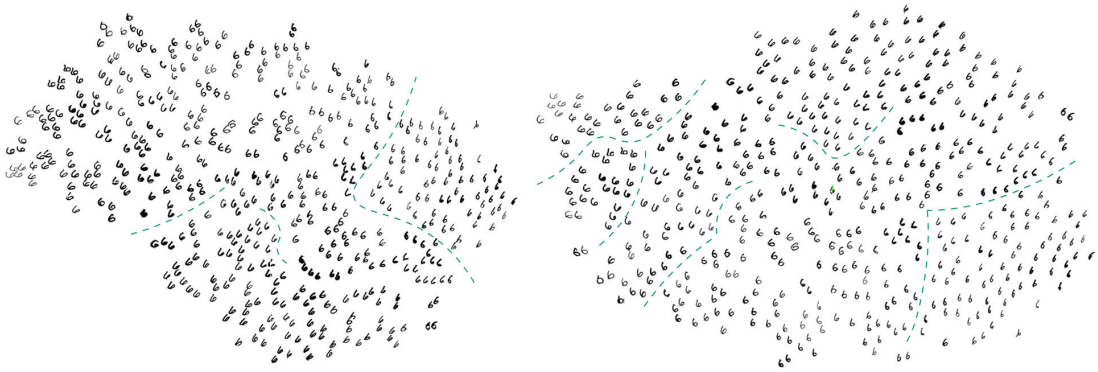


FIGURE 8. [Left] *t*-SNE v.s. [right] *g*-SNE ($\beta = 2.2$) for clustering of 500 MNIST images of digit "6" only. Auxiliary green lines show the natural separation of different writing styles. It is noticed in *g*-SNE, for instance, that the curly "6" are mostly confined at the upper-left corner, while the slanted ones are on the lower-right corner.

Comparing Eq. (31) with Eq. (25), we recognize that the leading term of the force (gradient) of *s*-SNE and *p*-SNE are *identical* when latent samples are close. This shows that *s*-SNE has a similar gradient force behavior as the *p*-SNE, *i.e.,* *g*-SNE of $\eta \to 0$.

VI. EXPERIMENTS

A. EXPERIMENTAL SETUP

Experiments are conducted to investigate the effect of dimensionality reduction on three datasets: the MNIST handwritten digits [42], the COIL-20 objects [43], and the Olivetti faces [8]. The MNIST dataset has 60,000 training images of 10 handwritten digits, each of which is an image of 28×28 pixels. The COIL-20 dataset contains 1,440 images of 20 different objects, each of which was snapshotted in 72 angles equally sampled among 360°. The image size is 128×128. The Olivetti faces dataset comprises 400 images from 40 distinct persons, each of which has 10 facial variations from viewpoints or expressions. The image size is 64×64.

The proposed method is implemented by PyTorch with GPU support available at https://github.com/hsinyilin19/generalizedSNE. During the SNE learning, the PCA whitening process is applied to all datasets to reduce to 50 dimensions such that some noises are suppressed with most data structures retained. Subsequently, different SNEs are performed to reduce the dimension from $D = 50$ to $d = 2$ for visualization. Latent variables $\mathbf{y}_i$ are randomly initialized in a bounded disk centered at zero. In MNIST, 6,000 images among 10 digits were randomly selected for evaluation. Different classes of digits, objects, and faces are shown in different colors. The experiments compare several SNE methods including SNE [7], *t*-SNE [8], *s*-SNE [24] and the proposed *g*-SNE with the realization using α-divergence.

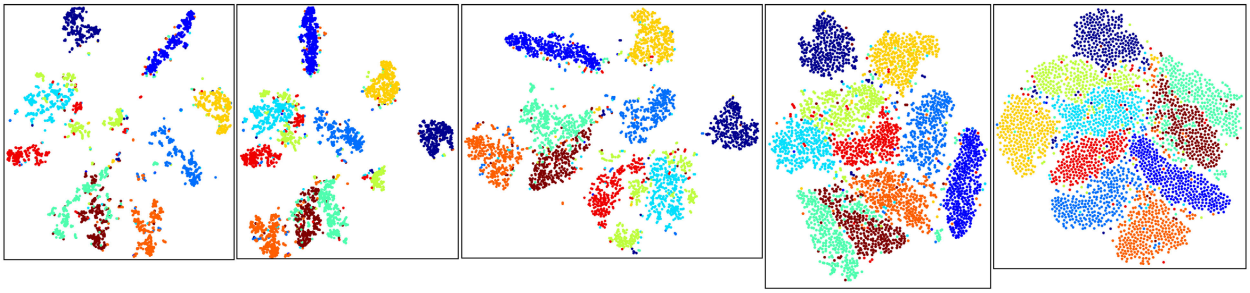


FIGURE 9. Comparison of g-SNE on $\beta = 1.5, 1.8, 2, 2.5, 3$ (from left to right) in MNIST with fixed divergence $\alpha = -1$.

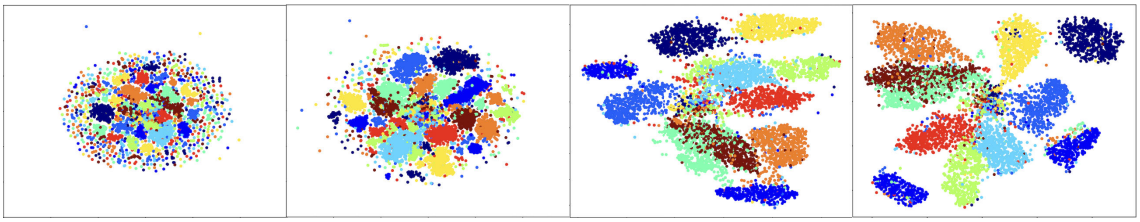


FIGURE 10. Comparison of g-SNE on $\eta = 5, 1, 0.1, 0.075$ (from left to right) on MNIST with fixed $\alpha = -0.5$ and $\beta = 3$.

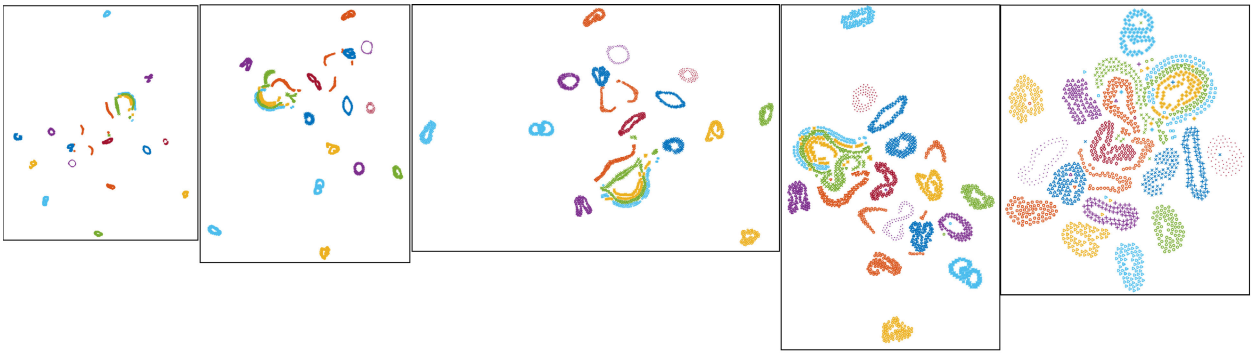


FIGURE 11. Comparison of g-SNE on $\beta = 1.5, 1.8, 2, 2.5, 3$ (from left to right) in COIL-20 with fixed divergence $\alpha = -1$.

B. EXPERIMENTAL RESULTS

1) COMPARISON OF DIFFERENT SNES

Fig. 4 displays the latent maps of different SNEs on MNIST under KL divergence where the crowding problem is accompanied by dimensionality reduction from $D = 50$ to $d = 2$. The SNE (as shown in Fig. 4(a)) obviously suffers more from the problem where 10 digits in different colors can be confusing in a two-dimensional representation. As one uses the other SNEs with stronger attractive and repulsive forces, such as t-SNE, the clustering effect for visualization becomes more clear. Fig. 4(c) on MNIST and Fig. 5(c) on COIL-20 show this property.

For s-SNE, the clustering effect significantly improves when $\rho \to 1$ as in Fig. 4(b) and 5(b). As analyzed above, s-SNE is locally indistinguishable due to the property of power-law behavior.

In the experiments, the proposed g-SNE is observed to result in a clear separation between different groups in MNIST and COIL-20, Fig. 4(d) and 5(d), respectively, when compared with t-SNE and s-SNE. Distinct objects are also expelled far apart from one another due to the strong forces resulting from the distribution function. Ideally, each object in COIL-20 should shrink to a tiny circle as depicted by t-SNE and g-SNE in Fig. 5, while the SNE, on the contrary, has multiple objects overlapped and non-separated in Fig. 5(a).

Two scores are used to evaluate the performance of a dimension reduction map. The clustering performance may be measured by the Davies-Bouldin index (DBI) [44] as shown in Fig. 6 and 7, where the lowest DBI is given by g-SNE to approximate the true number of classes. The grouping accuracy is measured by the generalization error of one nearest neighbor (1)-NN) with 10-fold cross-validation applied. Table 2 indicates that g-SNE and t-SNE outperform

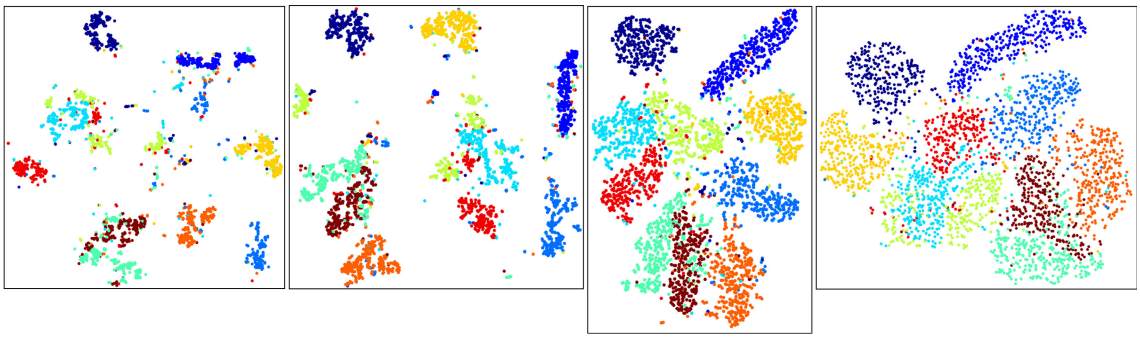


FIGURE 12. Comparison of SNE with the Student-t distribution under different divergences $\alpha = 0, -0.5, -1.5, -2$ (from left to right) on MNIST. $\alpha = -1$ (KL divergence) was given in Fig. 4(b).

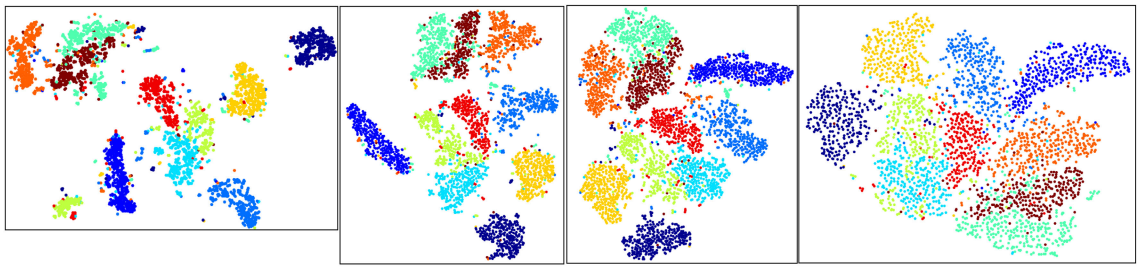
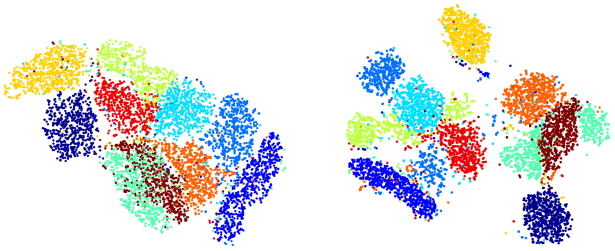


FIGURE 13. Comparison of g-SNE under different divergences $\alpha = -0.8, -1.2, -1.5, -2$ (from left to right) on MNIST with fixed $\beta = 2$.



(a) $\alpha = -0.85$ (b) $\alpha = -0.7$

FIGURE 14. Illustration of symmetric SNE with different α may unravel entangled samples (6,000 MNIST images) in the latent space. Fig. 4(a) of $\alpha = -1$ (KL divergence) should be included for comparison here.

SNE and s-SNE in different datasets and conditions. The proposed g-SNE performs better than t-SNE for most cases in Table 2.

We further conduct a separate experiment on single-digit images of MNIST to visualize the clustering effect. Fig. 8 illustrates the map of 500 samples with digit "6" only, where g-SNE ($\beta = 2.2$) is used to plot against t-SNE. One notices that there exist spatial divisions that naturally distinguish several hand-writing styles of "6", as shown by auxiliary green lines. This is also believed to be a consequence of strong forces that attract (*resp.* expel) similar (*resp.* distinct) handwriting. However, such a phenomenon is less clear in the case of t-SNE. Next, we further analyze the effect of g-SNE.

TABLE 2. Comparison of error rates by using 1-NN classifier with 10-fold cross-validation. $\rho = 0.95$ and $\beta = 2$ was used for s-SNE and g-SNE, respectively.

	MNIST $N = 1K$	MNIST $N = 6K$	MNIST $N = 14K$	COIL-20	Olivetti faces
SNE	36.7%	30.9%	28.1%	13.0%	44.5%
s-SNE	16.2%	13.9%	6.7%	7.5%	8.5%
t-SNE	11.6%	4.8 %	3.9 %	**0.2** %	5.3%
g-SNE	**11.1**%	**4.4**%	**3.8**%	0.3%	**4.5**%

2) EVALUATION OF G-SNE

By varying the exponent β of the target similarity $q(r_{ij}) = 1/(\eta + r_{ij}^{\beta})$, one derives different embeddings. As shown in Fig. 9 on MNIST and Fig. 11 on COIL-20, smaller β gives tighter clustering while larger β shows bigger blobs but yields consistent grouping effect. This can be understood from Eq. (26) since the attractive force increases linearly with β for fixed p_{ij} and r_{ij} while the repulsion has exponential growth (or decay) when two latent particles $\mathbf{y}_i$ and $\mathbf{y}_j$ are close $r_{ij} \to 0$ (or far apart $r_{ij} \gg 1$). Consequently, slightly dis-similar objects tend to be expelled easily under large β, making it hard to form compact clustering. Empirical range $\beta \in [1.5, 3]$ typically yields stable numerical convergence. Slight differences exist for different datasets.

The effect of parameter η can be observed from Fig. 10. As η decreases, the cluster separation effect grows stronger. Fig. 10 shows the distinction between the t-SNE ($\eta = 1$) and the p-SNE ($\eta \to 0$). It was observed that the introduction of

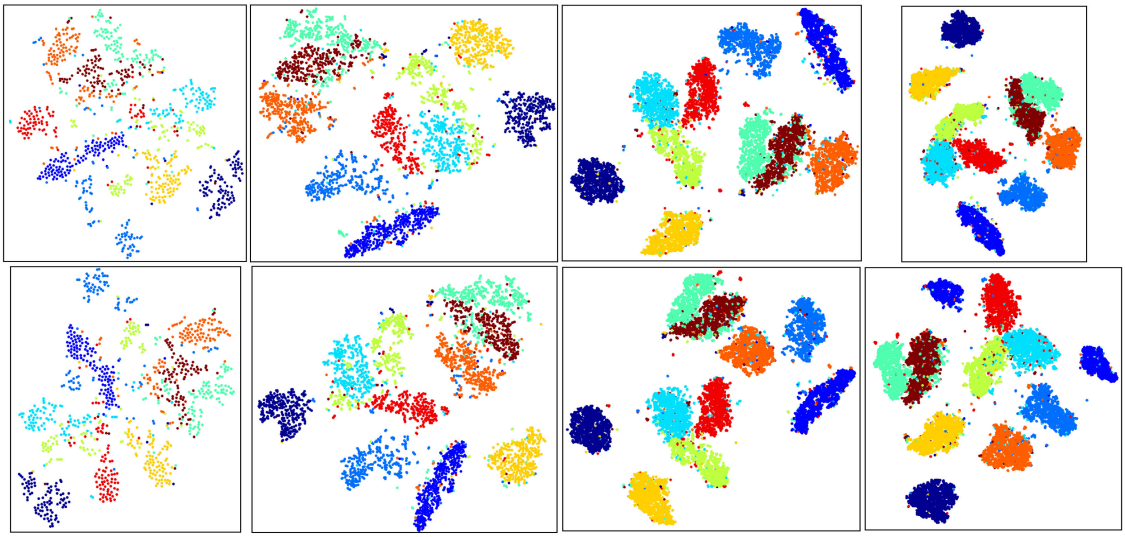


FIGURE 15. *t*-SNE (upper 4 figures) v.s. *g*-SNE (lower 4 figures) under various sample sizes in MNIST, $N = 1K, 3K, 10K, 14K$ (from left to right) with $\alpha = -1$ (KL divergence).

η grants the g-SNE more stability than the p-SNE. Increased stability of g-SNE enlarges the viable range of parameters such as β, and thus more flexibility can be ensured.

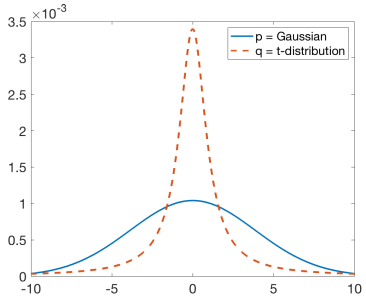
3) EVALUATION OF DIVERGENCE MEASURES

To observe how divergences affect the dimensionality reduction behavior, we utilize the proposed g-SNE with different α-divergence to view the corresponding change. Fig. 12 demonstrates the effect of different α on t-SNE, while Fig. 13 shows how α adjusts the g-SNE map. Although theoretically positive $\alpha > 0$ is possible, our empirical study shows that $\alpha \leq 0$ is more numerically viable.

Both Fig. 12 and 13 suggest that larger α achieves better separation between distinct clusters. A qualitative explanation is that smaller (more negative) α requires higher conformity between two probabilities p_{ij} and q_{ij}. Thus the latent distribution q_{ij} of either t-SNE or g-SNE is forced to approach the original p_{ij} in order to minimize the loss function Eq. (6), where in this case p_{ij} is the Gaussian with wider support. Since a smaller α is sensitive to the difference of distributions, it leads to wide-spread clusters as in the case of $\alpha = -2$. A numerical illustration in Table 3 exemplifies this assertion.

On the other hand, larger α may weaken the sensitivity to the difference of p_{ij} and q_{ij} so that attractive and repulsive forces have more freedom to move latent particles $\mathbf{y}_i$ without increasing loss values. Thus, more flexibility can be acquired to derive wider separation and compact clustering under larger α. Consequently, tuning both α and β allows us to adjust the divergence measure and sensitivity of similarity depending on tasks. An illuminating example Fig. 14 shows that adjusting a proper α unravels an entangled map from Fig. 4(a) to Fig. 14(a) and (b) under the *same* SNE. This suggests that the other SNE methods also have the potential to reduce the crowding problem, just like t-SNE. This example also reveals that the crowding problem may be alleviated from the perspective of a probability distribution or a divergence measure.

TABLE 3. Numerical values of $\mathcal{L}_\alpha(P, Q)$ computed by Eq. (6), (14) with an example $p(r_{ij}) = \exp(-r_{ij}^2/30)$, $q(r_{ij}) = 1/(1 + r_{ij}^2)$. Observe that $\mathcal{L}_\alpha(P, Q)$ increases with decreasing α, where $\alpha = 0, -1$ corresponds to the Hellinger and the KL divergence, respectively.



α	$\mathcal{L}_\alpha(P, Q)$
0.0	0.33971
-0.5	0.34289
-1.0	0.35003
-1.5	0.36145
-2.0	0.37746
-2.5	0.39862
-3.0	0.42565
-3.5	0.45949
-4.0	0.50132

(a) t-SNE (b) g-SNE

FIGURE 16. Visualization of Olivetti faces using *t*-SNE and the proposed *g*-SNE with $\alpha = -0.8$ and $\beta = 2.2$.

FIGURE 17. Visualization of Olivetti faces represented by g-SNE corresponding to Fig. 16(b).

4) EVALUATION OF VARIOUS SAMPLE SIZES

We also study the effect of g-SNE under various sample sizes using MNIST. Fig. 15 shows that g-SNE driven by the strong forces in general provides good clustering effects even when samples are few (*e.g.*, $N = 1K$). As the sample size grows, reduction maps of both t-SNE and g-SNE acquire better representation, such as lower classification error and lower DBI. However, it is also known that for large datasets, SNE techniques soon become numerically intense due to the complexity $O(N^2)$ for a dataset of N samples. Some possible modifications of SNE for large data were via the use of random walk-based similarities [8] or adaptations of the quad-tree approximation [45].

In the final, we compare the visualization of Olivetti faces using t-SNE and the g-SNE with $\alpha = -0.8$ and $\beta = 2.2$ in Fig. 16. The face images corresponding to the proposed method are shown in Fig. 17, where the faces of distinct persons are clearly separated.

VII. CONCLUSION

This study developed a generalized stochastic neighbor embedding approach. By setting off from general conditions on Riemannian manifolds $(\mathcal{M}, h, d)$ and $(\mathcal{N}, g, r)$ equipped with flexible probability distributions $\{p_{ij}\}$ and $\{q_{ij}\}$ and an arbitrary α-divergence function $\mathcal{D}_{f_\alpha}$, a comprehensive analysis for the stochastic neighbor embedding was derived. The generalized SNE not only unified several SNEs but also revealed intrinsic properties of dimensionality reductions. Moreover, a connection between mathematics and physics was established. The mechanism accounting for the clustering effect was fully determined by the geometric elements given above. The attractive and repulsive force correspon-

dence characterized the dynamics of latent variables as mechanical particles in the lower dimension.

By analyzing our general formulation, a new method g-SNE was further conceived to incorporate favorable characteristics for flexibility and proposed to enhance the clustering performance. Extensive experiments were further conducted to investigate the effects of g-SNE. Overall, observations showed that the proposed generalizations provided higher flexibility by appropriate combinations of (α, β, η), and the proposed g-SNE offered strengthened clustering results.

APPENDIX A THE DERIVATION FROM EQ. (27) TO (28)

Lemma 1: For any vectors $\mathbf{y}_i, \mathbf{y}_j, \mathbf{y}_k \in \mathbb{R}^s$, *we have* $\frac{\partial}{\partial \mathbf{y}_i}\langle \mathbf{y}_j, \mathbf{y}_k\rangle := \nabla_{\mathbf{y}_i}\langle \mathbf{y}_j, \mathbf{y}_k\rangle = \delta_{ij}\mathbf{y}_k + \delta_{ik}\mathbf{y}_j$.

Proof: Let $\mathbf{y}_i = \sum_\alpha y_{i\alpha}\,\mathbf{e}_\alpha$ where $\{\mathbf{e}_\alpha \in \mathbb{R}^s\}$ is a chosen coordinate basis and $y_{i\alpha} \in \mathbb{R}^1$ are the corresponding coefficients. Similarly, we expand $\mathbf{y}_j = \sum_\beta y_{j\beta}\,\mathbf{e}_\beta$ and $\mathbf{y}_k = \sum_\gamma y_{k\gamma}\,\mathbf{e}_\gamma$ in terms of the basis with $y_{j\beta}, y_{k\gamma} \in \mathbb{R}^1$. Since the gradient is defined by $\nabla_{\mathbf{y}_i} = \sum_\alpha \mathbf{e}_\alpha \frac{\partial}{\partial y_{i\alpha}}$, we have

$$\nabla_{\mathbf{y}_i}\langle \mathbf{y}_j, \mathbf{y}_k\rangle := \sum_\alpha \mathbf{e}_\alpha \frac{\partial}{\partial y_{i\alpha}}\langle \mathbf{y}_j, \mathbf{y}_k\rangle \tag{32}$$

$$= \sum_\alpha \mathbf{e}_\alpha \frac{\partial}{\partial y_{i\alpha}}\left(\sum_\beta y_{j\beta}\, y_{k\beta}\right) \tag{33}$$

$$= \sum_\alpha \mathbf{e}_\alpha \left(\sum_\beta \left(\delta_{ij}\,\delta_{\alpha\beta}\, y_{k\beta} + y_{j\beta}\,\delta_{ik}\delta_{\alpha\beta}\right)\right) \tag{34}$$

$$= \sum_\alpha \mathbf{e}_\alpha \left(\delta_{ij} y_{k\alpha} + \delta_{ik} y_{j\alpha}\right) = \delta_{ij}\mathbf{y}_k + \delta_{ik}\mathbf{y}_j \tag{35}$$

where $\delta_{ij}, \delta_{\alpha\beta}$, etc denote the standard Kronecker delta functions and the usual 1-dimensional calculus is applied from (33) to (34). □

Lemma 2: $\nabla_{\mathbf{y}_i}||\mathbf{y}_k - \rho\,\mathbf{y}_j||^2 = 2\left(\mathbf{y}_k - \rho\,\mathbf{y}_j\right)\left(\delta_{ik} - \rho\,\delta_{ij}\right)$.

proof: As $\nabla_{\mathbf{y}_i}||\mathbf{y}_k - \rho\,\mathbf{y}_j||^2 = \nabla_{\mathbf{y}_i}\langle \mathbf{y}_k - \rho\,\mathbf{y}_j, \mathbf{y}_k - \rho\,\mathbf{y}_j\rangle$, similar calculations to Lemma 1 lead to the conclusion. □

Lemma 3: For a differentiable real-valued function $f : \mathbb{R}^+ \to \mathbb{R}$, *we have* $\nabla_{\mathbf{y}_i} f(||\mathbf{y}_k - \rho\,\mathbf{y}_j||^2) = 2\,f'(||\mathbf{y}_k - \rho\,\mathbf{y}_j||)\left(\mathbf{y}_k - \rho\,\mathbf{y}_j\right)\left(\delta_{ik} - \rho\,\delta_{ij}\right)$.

Proof: Let $z = ||\mathbf{y}_k - \rho\,\mathbf{y}_j||^2$ and

$$\nabla_{\mathbf{y}_i} f(||\mathbf{y}_k - \rho\,\mathbf{y}_j||^2) = \sum_\alpha \mathbf{e}_\alpha \frac{\partial}{\partial y_{i\alpha}} f(z)$$

$$= \sum_\alpha \mathbf{e}_\alpha \left.\frac{\partial f(z)}{\partial z}\right|_{z=||\mathbf{y}_k-\rho\,\mathbf{y}_j||^2} \cdot \frac{\partial z}{\partial y_{i\alpha}}$$

$$= \left.\frac{\partial f(z)}{\partial z}\right|_{z=||\mathbf{y}_k-\rho\,\mathbf{y}_j||^2} \cdot \underbrace{\left(\sum_\alpha \mathbf{e}_\alpha \frac{\partial z}{\partial y_{i\alpha}}\right)}_{\nabla_{\mathbf{y}_i}||\mathbf{y}_k-\rho\,\mathbf{y}_j||^2}$$

□

Lemma 4: $\nabla_{\mathbf{y}_i}||\mathbf{y}_k - \rho\,\mathbf{y}_j||^{-d} = -\frac{d(\mathbf{y}_k-\rho\,\mathbf{y}_j)}{||\mathbf{y}_k-\rho\,\mathbf{y}_j||^{d+2}}\left(\delta_{ik} - \rho\,\delta_{ij}\right)$.

Proof: Apply Lemma 3 with a special case $f(z) = z^{-d/2}$. □

Lemma 5: $\nabla_{\mathbf{y}_i}\left(\sum_{m\neq n}||\mathbf{y}_m - \rho\,\mathbf{y}_n||^{-d}\right) = -d\sum_{m\neq i}||\mathbf{y}_i - \rho\,\mathbf{y}_m||^{-d-2}(\mathbf{y}_i - \rho\,\mathbf{y}_m) + d\rho\sum_{m\neq i}||\mathbf{y}_m - \rho\mathbf{y}_i||^{-d-2}(\mathbf{y}_m - \rho\mathbf{y}_i)$.

Proof: Apply Lemmas 2, 3 and 4 to obtain the result. □

Next, we compute $\frac{\partial}{\partial \mathbf{y}_i}\mathcal{L}_{s\text{-sne}}$ with $\mathcal{L}_{s\text{-sne}}$ defined by Eq. (27). Since $\sum_{jk} p_{jk}\log p_{jk}$ has no relation with the lower dimension coordinate $\mathbf{y}$, i.e. $\frac{\partial}{\partial \mathbf{y}_i}\left(\sum_{jk} p_{jk}\log p_{jk}\right) \equiv 0$, we have

$$\frac{\partial}{\partial \mathbf{y}_i}\mathcal{L}_{s\text{-sne}}$$

$$= -\nabla_{\mathbf{y}_i}\left(\sum_{jk} p_{jk}\log q_{jk}\right) - 2\lambda_i\,\mathbf{y}_i \tag{36}$$

$$= -\sum_{jk}\frac{p_{jk}}{q_{jk}}\left(\nabla_{\mathbf{y}_i} q_{jk}\right) - 2\lambda_i\,\mathbf{y}_i \tag{37}$$

$$= -\sum_{jk}\frac{p_{jk}}{q_{jk}}\left(\frac{\nabla_{\mathbf{y}_i}||\mathbf{y}_k - \rho\,\mathbf{y}_j||^{-d}}{\sum_{m\neq n}||\mathbf{y}_m - \rho\,\mathbf{y}_n||^{-d}}\right. \tag{38}$$

$$\left. - \frac{||\mathbf{y}_k - \rho\,\mathbf{y}_j||^{-d}\nabla_{\mathbf{y}_i}\left(\sum_{m\neq n}||\mathbf{y}_m - \rho\,\mathbf{y}_n||^{-d}\right)}{\left(\sum_{m\neq n}||\mathbf{y}_m - \rho\,\mathbf{y}_n||^{-d}\right)^2}\right) - 2\lambda_i\,\mathbf{y}_i \tag{39}$$

where the definition of q_{jk} in Eq. (5) is used to obtain the last equality. Then, two major terms in Eq. (38),(39) can be computed separately, where the first term in Eq. (38),

$$-\sum_{jk}\frac{p_{jk}}{q_{jk}}\left(\frac{\nabla_{\mathbf{y}_i}||\mathbf{y}_k - \rho\,\mathbf{y}_j||^{-d}}{\sum_{m\neq n}||\mathbf{y}_m - \rho\,\mathbf{y}_n||^{-d}}\right) \tag{40}$$

$$= -\sum_{jk}\frac{p_{jk}}{q_{jk}}\frac{-d\,||\mathbf{y}_k - \rho\,\mathbf{y}_j||^{-d-2}\left(\mathbf{y}_k - \rho\,\mathbf{y}_j\right)\left(\delta_{ik} - \rho\,\delta_{ij}\right)}{\sum_{m\neq n}||\mathbf{y}_m - \rho\,\mathbf{y}_n||^{-d}} \tag{41}$$

$$= d\sum_j \frac{p_{ji}}{q_{ji}}\frac{||\mathbf{y}_i - \rho\,\mathbf{y}_j||^{-d-2}\left(\mathbf{y}_i - \rho\,\mathbf{y}_j\right)}{\sum_{m\neq n}||\mathbf{y}_m - \rho\mathbf{y}_n||^{-d}} \tag{42}$$

$$-d\rho\sum_k \frac{p_{ik}}{q_{ik}}\frac{||\mathbf{y}_k - \rho\,\mathbf{y}_i||^{-d-2}\left(\mathbf{y}_k - \rho\,\mathbf{y}_i\right)}{\sum_{m\neq n}||\mathbf{y}_m - \rho\mathbf{y}_n||^{-d}} \tag{43}$$

$$= d\sum_j \frac{p_{ji}\,\cancel{q_{ji}}\left(\mathbf{y}_i - \rho\,\mathbf{y}_j\right)}{\cancel{q_{ji}}\,||\mathbf{y}_i - \rho\,\mathbf{y}_j||^2} - d\rho\sum_k \frac{p_{ik}\,\cancel{q_{ik}}\cdot\left(\mathbf{y}_k - \rho\,\mathbf{y}_i\right)}{\cancel{q_{ik}}\,||\mathbf{y}_k - \rho\,\mathbf{y}_i||^2} \tag{44}$$

$$= d\sum_j p_{ij}\left(\frac{\mathbf{y}_i - \rho\,\mathbf{y}_j}{||\mathbf{y}_i - \rho\,\mathbf{y}_j||^2} - \rho\frac{\mathbf{y}_j - \rho\,\mathbf{y}_i}{||\mathbf{y}_j - \rho\,\mathbf{y}_i||^2}\right) \tag{45}$$

where Lemma 4 is applied in Eq. (40) to get Eq. (41) and using definition in Eq. (5) to derive Eq. (44) from Eq. (43). Finally, by arranging the index in Eq. (44) and using the symmetry $p_{ij} = p_{ji}$, Eq. (45) yields the first two terms of Eq. (28). For the term in Eq. (39):

$$\sum_{jk}\frac{p_{jk}}{q_{jk}}\cdot\frac{||\mathbf{y}_k - \rho\,\mathbf{y}_j||^{-d}\cdot\nabla_{\mathbf{y}_i}\left(\sum_{m\neq n}||\mathbf{y}_m - \rho\,\mathbf{y}_n||^{-d}\right)}{\left(\sum_{m\neq n}||\mathbf{y}_m - \rho\,\mathbf{y}_n||^{-d}\right)^2} \tag{46}$$

$$= \sum_{jk} \frac{p_{jk}}{q_{jk}} \cdot \frac{q_{jk} \cdot \nabla_{\mathbf{y}_i} \left(\sum_{m \neq n} ||\mathbf{y}_m - \rho\, \mathbf{y}_n||^{-d} \right)}{\sum_{m \neq n} ||\mathbf{y}_m - \rho\, \mathbf{y}_n||^{-d}} \tag{47}$$

$$= \underbrace{\left(\sum_{jk} p_{jk} \right)}_{1} \cdot \frac{\nabla_{\mathbf{y}_i} \left(\sum_{m \neq n} ||\mathbf{y}_m - \rho\, \mathbf{y}_n||^{-d} \right)}{\sum_{m \neq n} ||\mathbf{y}_m - \rho\, \mathbf{y}_n||^{-d}} \tag{48}$$

$$= -d \frac{\sum_{m \neq i} ||\mathbf{y}_i - \rho\, \mathbf{y}_m||^{-d-2} \, (\mathbf{y}_i - \rho\, \mathbf{y}_m)}{\sum_{m \neq n} ||\mathbf{y}_m - \rho\, \mathbf{y}_n||^{-d}} \tag{49}$$

$$+ d\rho \frac{\sum_{m \neq i} ||\mathbf{y}_m - \rho\, \mathbf{y}_i||^{-d-2} \, (\mathbf{y}_m - \rho\, \mathbf{y}_i)}{\sum_{m \neq n} ||\mathbf{y}_m - \rho\, \mathbf{y}_n||^{-d}} \tag{50}$$

$$= -d \sum_{m \neq i} q_{mi} \frac{\mathbf{y}_i - \rho\, \mathbf{y}_m}{||\mathbf{y}_i - \rho\, \mathbf{y}_m||^2} + d\rho \sum_{m \neq i} q_{im} \frac{\mathbf{y}_m - \rho\, \mathbf{y}_i}{||\mathbf{y}_m - \rho\, \mathbf{y}_i||^2} \tag{51}$$

where again the definition of q_{jk} in Eq. (5) is applied in Eq. (46) to obtain Eq. (47). We notice that the summation over j, k in Eq. (47) is only on p_{jk}, no other terms involved. Consequently, by definition, the probability p_{jk} sums up to 1 to derive Eq. (48). Simply applying Lemma 5 on Eq. (48) gets to Eq. (50) and subsequently using the definition in Eq. (5) arrives at Eq. (51). Combining Eq. (45) and Eq. (51) concludes the proof of Eq. (28).

REFERENCES

[1] B. Hosseini and B. Hammer, "Interpretable discriminative dimensionality reduction and feature selection on the manifold," in *Machine Learning and Knowledge Discovery in Databases*. Cham, Switzerland: Springer, 1007, pp. 310–326.

[2] H. A. Chipman and H. Gu, "Interpretable dimension reduction," *J. Appl. Statist.*, vol. 32, no. 9, pp. 969–987, Nov. 2005.

[3] Z. Liu, Z. Lai, W. Ou, K. Zhang, and R. Zheng, "Structured optimal graph based sparse feature extraction for semi-supervised learning," *Signal Process.*, vol. 170, May 2020, Art. no. 107456.

[4] Z. Liu, T. Wang, F. Zhu, X. Chen, D. Pelusi, and A. V. Vasilakos, "Domain adaptive learning based on equilibrium distribution and dynamic subspace approximation," *Expert Syst. Appl.*, vol. 249, Sep. 2024, Art. no. 123673.

[5] L. van der Maaten, E. O. Postma, and J. van den Herik, "Dimensionality reduction: A comparative review," Tilburg Univ., Tilburg, The Netherlands, Tech. Rep. 5-2009, 2009.

[6] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[7] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, Dec. 2003, pp. 857–864.

[8] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[9] Z. Yang, I. King, Z. Xu, and E. Oja, "Heavy-tailed symmetric stochastic neighbor embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Vancouver, BC, Canada: Curran Associates, Dec. 2009, pp. 2169–2177.

[10] K. Bunte, S. Haase, M. Biehl, and T. Villmann, "Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences," *Neurocomputing*, vol. 90, pp. 23–45, Aug. 2012.

[11] M. R. Min, L. Maaten, Z. Yuan, A. J. Bonner, and Z. Zhang, "Deep supervised t-distributed embedding," in *Proc. Int. Conf. Mach. Learn.*, Haifa, Israel, Jun. 2010, pp. 791–798.

[12] J.-T. Chien and C.-H. Chen, "Deep discriminative manifold learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Shanghai, China, Mar. 2016, pp. 2672–2676.

[13] Y. Fan, J. Liu, P. Liu, Y. Du, W. Lan, and S. Wu, "Manifold learning with structured subspace for multi-label feature selection," *Pattern Recognit.*, vol. 120, Dec. 2021, Art. no. 108169.

[14] M. T. Islam, Z. Zhou, H. Ren, M. B. Khuzani, D. Kapp, J. Zou, L. Tian, J. C. Liao, and L. Xing, "Revealing hidden patterns in deep neural network feature space continuum via manifold learning," *Nature Commun.*, vol. 14, no. 1, p. 8506, Dec. 2023.

[15] I. Horev, F. Yger, and M. Sugiyama, "Geometry-aware principal component analysis for symmetric positive definite matrices," in *Proc. Asian Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, vol. 45, Hong Kong, G. Holmes and T.-Y. Liu, Eds., Nov. 2016, pp. 1–16. [Online]. Available: https://proceedings.mlr.press/v45/Horev15.html

[16] I. Chami, A. Gu, D. Nguyen, and C. Ré, "HoroPCA: Hyperbolic dimensionality reduction via horospherical projections," in *Proc. Int. Conf. Mach. Learn.*, 2021.

[17] Y. Guo, H. Guo, and S. X. Yu, "CO-SNE: Dimensionality reduction and visualization for hyperbolic data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11–20.

[18] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.

[19] J. N. Böhm, P. Berens, and D. Kobak, "Attraction-repulsion spectrum in neighbor embeddings," *J. Mach. Learn. Res.*, vol. 23, no. 95, pp. 1–32, 2022.

[20] X. Du, X. Zheng, X. Lu, and A. A. Doudkin, "Multisource remote sensing data classification with graph fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10062–10072, Dec. 2021.

[21] W. Chen, H. Wang, Y. Zhang, P. Deng, Z. Luo, and T. Li, "T-distributed stochastic neighbor embedding for co-representation learning," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 2, pp. 1–18, Apr. 2024.

[22] H.-H. Tseng, I. El Naqa, and J.-T. Chien, "Power-law stochastic neighbor embedding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Oct. 2017, pp. 2347–2351.

[23] J. Cook, I. Sutskever, A. Mnih, and G. E. Hinton, "Visualizing similarity data with a mixture of maps," in *Proc. Int. Conf. Artif. Intell. Statist.*, San Juan, Puerto Rico, Mar. 2007, pp. 67–74.

[24] D. Lunga and O. Ersoy, "Spherical stochastic neighbor embedding of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 857–871, Feb. 2013.

[25] S. Kato, "A distribution for a pair of unit vectors generated by Brownian motion," *Bernoulli*, vol. 15, no. 3, pp. 898–921, Aug. 2009.

[26] M. Belkin and P. Niyogi, "Semi-supervised learning on Riemannian manifolds," *Mach. Learn.*, vol. 56, nos. 1–3, pp. 209–239, Jul. 2004.

[27] Z. Yang, J. Peltonen, and S. Kaski, "Optimization equivalence of divergences improves neighbor embedding," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 460–468.

[28] K. Narayan, A. Punjani, and P. Abbeel, "Alpha-beta divergences discover micro and macro structures in data," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 796–804.

[29] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski, "Information retrieval perspective to nonlinear dimensionality reduction for data visualization," *J. Mach. Learn. Res.*, vol. 11, no. 2, pp. 451–490, 2010.

[30] H. Zhu and R. Rohwer, "Bayesian invariant measurements of generalization," *Neural Process. Lett.*, vol. 2, no. 6, pp. 28–31, Dec. 1995.

[31] J. Zhang, "Divergence function, duality, and convex analysis," *Neural Comput.*, vol. 16, no. 1, pp. 159–195, Jan. 2004.

[32] J.-T. Chien and H.-L. Hsieh, "Convex divergence ICA for blind source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 302–313, Jan. 2012.

[33] P. C. Mahalanobis, "On the generalized distance in statistics," *Sankhyā Indian J. Statist. A*, vol. 80, pp. S1–S7, 2008.

[34] R. Durrett, *Probability: Theory and Examples*. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[35] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

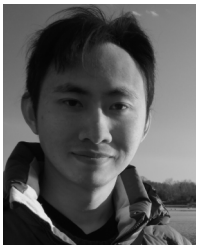
[36] J. Jost and J. Jost, *Riemannian Geometry and Geometric Analysis*, vol. 42005. Springer, 2008.

[37] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[38] S.-I. Amari, "Differential geometry of curved exponential families-curvatures and information loss," *Ann. Statist.*, vol. 10, no. 2, pp. 357–385, Jun. 1982.

[39] H. Goldstein, C. P. Poole, and J. L. Safko, *Classical Mechanics*. Reading, MA, USA: Addison-Wesley, 2002.
[40] H. Han, W. Li, J. Wang, G. Qin, and X. Qin, "Enhance explainability of manifold learning," *Neurocomputing*, vol. 500, pp. 877–895, Aug. 2022.
[41] X. Chen, R. Chen, Q. Wu, F. Nie, M. Yang, and R. Mao, "Semisupervised feature selection via structured manifold learning," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 5756–5766, Jul. 2022.
[42] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[43] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-20)," Tech. Rep. CUCS-005-96, 1996.
[44] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.
[45] Z. Yang, J. Peltonen, and S. Kaski, "Scalable optimization of neighbor embedding for visualization," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 127–135.

HSIN-YI LIN (Member, IEEE) received the Ph.D. degree in mathematics from the University of Maryland, College Park. She is currently an Assistant Professor of mathematics and computer science with Seton Hall University. Her research interests include partial differential equations and their application in modeling and machine learning algorithms.



HUAN-HSIN TSENG received the B.A. and Ph.D. degrees in mathematics and physics from National Tsing Hua University, Hsinchu, Taiwan. He is currently a Research Scientist of investigating quantum machine learning and deep learning algorithms with the Computational Science Initiative, Brookhaven National Laboratory. His research interests include the intersections of mathematics, physics, and state-of-the-art machine learning.

JEN-TZUNG CHIEN (Senior Member, IEEE) is currently a Professor in electronics and electrical engineering, and computer science with National Yang Ming Chiao Tung University, Taiwan. His research interests include machine learning, deep learning and Bayesian learning with applications on natural language processing, and computer vision.