Generalized t-SNE Through the Lens of Information Geometry

MASANARI KIMURA
Ridge-i Inc., Tokyo 100-0004, Japan
e-mail: m.kimura@ridge-i.com

I. INTRODUCTION
In recent years, with the increasing complexity of data, a large number of statistical analysis and machine learning approaches have become increasingly important. In particular, manifold learning [1], [2], dimension reduction [3], [4], and visualization [5], [6] are useful paradigms to help understand complex data.

Stochastic Neighbor Embedding (SNE) [7] is one of the powerful frameworks to achieve these objectives. SNE compresses the dimension so that the proximity of points in the original space is as close as possible to the proximity of the points after compression. However, it is known that SNE suffers from computational instability due to KL-divergence. Many implementations of SNE with simple modifications, it still suffers from computational complexity due to its dependence on KL-divergence. Our goal is to extend t-SNE in a natural way by the framework of information geometry. Our generalized t-SNE can outperform the original t-SNE with a well-chosen set of parameters. Furthermore, the experimental results for MNIST, Fashion MNIST and COIL-20, show that our generalized t-SNE outperforms the original t-SNE.

ABSTRACT t-SNE (t-distributed Stochastic Neighbor Embedding) is known to be one of the very powerful tools for dimensionality reduction and data visualization. By adopting the student’s t-distribution in the original SNE (Stochastic Neighbor Embedding), t-SNE achieves faster and more stable learning. However, t-SNE still poses computational complexity due to its dependence on KL-divergence.

INDEX TERMS Machine learning, visualization, clustering.

This theoretical framework allows us to interpret observations using the tools of differential geometry on a Riemannian manifold constructed by a set of probability distributions.

II. NOTATIONS AND PRELIMINARIES
Let \( \mathcal{X} \subset \mathbb{R}^d \) be the \( d \)-dimensional input space. The goal of manifold learning or dimensionality reduction is to obtain a map \( f : \mathcal{X} \rightarrow \mathcal{Y} \), where \( \mathcal{Y} \subset \mathbb{R}^s \) is the output space and \( s \ll d \).

For input data points \( X = \{x_i\}_{i=1}^n \) of sample size \( n \in \mathbb{N} \), t-SNE [8] starts by computing a joint probability distribution over all pairs of data points \( \{x_i, x_j\}_{1 \leq i \neq j \leq n} \), represented by a symmetric matrix \( P = (p_{ij})_{1 \leq i, j \leq n} \), where \( p_{ii} = 0 \) as

\[
p_{ij} := \frac{p_{ij} + p_{ji}}{2n} \tag{1}
\]

\[
p_{ji} := \frac{\exp\left(\|x_i - x_j\|^2/2\tau_i^2\right)}{\sum_{\ell \in \{1, 2, ..., n\} \setminus \{i\}} \exp\left(-\|x_i - x_\ell\|^2/2\tau_i^2\right)} \tag{2}
\]

Here \( \tau_i \) are tuning parameters, which are usually determined based on a certain perplexity measure and a simple binary search [7], [11]. Similarly, in a \( s \)-dimensional map \( \{y_i\}_{1 \leq i \leq n} \subset \mathbb{R}^s \), define the joint probability distribution over all pairs \( \{y_i, y_j\}_{1 \leq i \neq j \leq n} \) through a symmetric matrix \( Q = (q_{ij})_{1 \leq i, j \leq n} \), where \( q_{ii} = 0 \) as

\[
q_{ij} := \frac{1 + \|y_i - y_j\|^2}{\sum_{k, \ell \in \{1, 2, ..., n\} \setminus \{i, j\}} (1 + \|y_k - y_\ell\|^2)^{-1}} \tag{3}
\]

\( P \) and \( Q \) are similarity matrices summarizing the pairwise distances of the high dimensional data points \( \{x_i\}_{1 \leq i \leq n} \), and...
the low dimensional map \( \{y_i\}_{1 \leq i \leq n} \). Then t-SNE aims to find \( \{y_i\}_{1 \leq i \leq n} \) in \( \mathbb{R}^d \) that minimizes the KL-divergence [12] between \( P \) and \( Q \):

\[
(y_1, \ldots, y_n) = \arg \min_{y_1 \cdots y_n} D_{KL}[P\|Q].
\]

(4)

Here, KL-divergence is defined as follows:

**Definition 1 (Kullback–Leibler Divergence [12]):** The Kullback–Leibler divergence or KL-divergence \( D_{KL} : \mathcal{P} \times \mathcal{P} \to [0, \infty) \) is defined between two Radon–Nikodym densities \( p \) and \( q \) of \( \mu \)-absolutely continuous probability measures by

\[
D_{KL}[p\|q] := \int_{\mathcal{X}} p \ln \frac{p}{q} d\mu(x),
\]

(5)

where \( \mathcal{P} := \{p(x) \geq 0 (\forall x \in \mathcal{X}) \mid \int_{\mathcal{X}} p(x) d\mu = 1\} \).

KL-divergence is a measure of the difference between two probability distributions in statistics and information theory [13], [14]. For continuous distributions, the KL-divergence is known to have computational difficulty. To be more specific, if \( q \) takes a small value relative to \( p \), the value of \( D_{KL}[p\|q] \) may diverge to infinity. The simplest idea to avoid this is to use very small \( \epsilon > 0 \) and modify \( D_{KL}[p\|q] \) as follows:

\[
D_{KL}^+ [p\|q] := \int_{\mathcal{X}} p \ln \frac{p}{q + \epsilon} d\mu(x).
\]

(6)

However, since \( q + \epsilon \) no longer satisfies the condition for a probability measure: \( \int_{\mathcal{X}} (q + \epsilon) d\mu(x) \neq 1 \), such an extension is unnatural. This computational difficulty is a critical issue in the implementation of t-SNE. For example, a gap occurs between theoretical analysis and implementation regarding t-SNE, making it difficult to describe numerical observations analytically.

As a more natural way to stabilize KL-divergence, the following skew divergences have been proposed:

**Definition 2 (Skew Divergence [15], [16]):** The skew divergence \( D^{(\lambda)} : \mathcal{P} \times \mathcal{P} \to [0, \infty) \) is defined between two Radon–Nikodym densities \( p \) and \( q \) of \( \mu \)-absolutely continuous probability measures by

\[
D^{(\lambda)} [p\|q] := \int_{\mathcal{X}} p \ln \frac{p}{(\lambda p + (1-\lambda)q)} d\mu(x),
\]

(7)

where \( \lambda \in (0, 1] \).

**A. EINSTEIN SUMMATION CONVENTION**

In the rest of this paper, Einstein summation convention [17] will be assumed, so that summation will be automatically taken over indices repeated twice in the term, e.g., \( a_i b^i = \sum_i a_i b^i \).

**B. CONTRACTION NOTATION FOR PARTIAL DERIVATIVES**

The variables in the partial derivative are represented by subscripts when it is clear from the context, e.g., \( \partial_i = \frac{\partial}{\partial x_i} \).

**III. INFORMATION GEOMETRY OF t-SNE**

We now consider the interpretation and extension of t-SNE through the lens of information geometry [9], [10].

We assume that \( P \) and \( Q \) are parameterized by unknown \( \theta \) and \( \xi \), that is, we assume two statistical models

\[
\mathcal{I} := \{P_{\theta} \subset \mathcal{P} \mid \theta = [\theta_1, \ldots, \theta_t] \in \Theta \subset \mathbb{R}^t\},
\]

(8)

\[
\mathcal{M} := \{Q_{\xi} \subset \mathcal{P} \mid \xi = [\xi_1, \ldots, \xi_u] \in \Xi \subset \mathbb{R}^u\}.
\]

(9)

We also denote \( (\mathcal{I}, g_{ij}^\mathcal{I}) \) as a statistical model equipped with the Riemannian metric \( g_{ij}^\mathcal{I} \). In particular, let \( g_{ij} = \lambda \) be the Fisher–Rao metric, which is the Riemannian metric induced from the Fisher information matrix [18]. Then, t-SNE can be regarded as learning a mapping from an input manifold \( (\mathcal{I}, g_{ij}^\mathcal{I}) \) to an output manifold \( (\mathcal{M}, g_{ij}^\mathcal{M}) \), \( f : \mathcal{I} \mapsto \mathcal{M} \).

We begin with an overview of the geometric concept of statistical manifolds.

**Definition 3 (Christoffel Symbols):** Let \( g \) be a Riemannian metric, particularly the Fisher information matrix, then the Christoffel symbols are given by

\[
\Gamma_{ij,k} = \frac{1}{2} \left( \partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij} \right).
\]

(10)

**Definition 4 (Levi-Civita Connection):** Let \( g \) be a Fisher–Riemannian metric. Levi-Civita connection \( \nabla^{(0)} \) is defined by

\[
\nabla^{(0)}(\partial_i, \partial_k) = \Gamma_{ij,k}.
\]

(11)

**Definition 5 (\( \nabla^{(1)} \)-Connection):** Let \( g \) be the Fisher–Riemannian metric, which is a 2-covariant tensor. Then, the \( \nabla^{(1)} \)-connection is defined by

\[
g \left( \nabla^{(1)}_{\partial_i} \partial_j, \partial_k \right) = \nabla^{(1)}_{\partial_i} \left( \partial_j \ln p \partial_k \ln p \right).
\]

(12)

It can also be expressed equivalently by explicitly writing as the Christoffel coefficients

\[
\nabla^{(1)}_{\partial_i} \left( \partial_j \ln p \partial_k \ln p \right).
\]

(13)

**Definition 6 (\( \nabla^{(1)} \)-Connection):** Let \( g \) be the Fisher–Riemannian metric, which is a 2-covariant tensor. Then, the \( \nabla^{(1)} \)-connection is defined by

\[
g \left( \nabla^{(1)}_{\partial_i} \partial_j, \partial_k \right) = \nabla^{(1)}_{\partial_i} \left( \partial_j \partial_k \ln p \partial_j \ln p \right)
\]

(14)

More generally, the following \( \nabla^{(\alpha)} \)-connection can be defined.

**Definition 7 (\( \nabla^{(\alpha)} \)-Connection):** For \( \alpha \in \mathbb{R} \), the \( \nabla^{(\alpha)} \)-connection is defined as

\[
\nabla^{(\alpha)} = \frac{1 + \alpha}{2} \nabla^{(1)} + \frac{1 - \alpha}{2} \nabla^{(-1)}.
\]

(15)

**Proposition 1:** The components \( \Gamma^{(\alpha)}_{ij,k} \) can be written as

\[
\Gamma^{(\alpha)}_{ij,k} = E \left( \partial_i \partial_j \ln p + \frac{1 - \alpha}{2} \partial_i \ln p \partial_j \ln p \right) d_k \ln p.
\]

(16)
The KL-divergence, which is a special case with \( \alpha = 1 \), induces the linear connection \( \nabla^{(1)} \). More generally, the \( \alpha \)-divergence with \( \alpha \in \mathbb{R} \) induces the \( \nabla^{(\alpha)} \)-connection.

**Definition 8 (\( \alpha \)-Divergence [19]):** Let \( \alpha \in \mathbb{R} \) be a real parameter. The \( \alpha \)-divergence between two probability distributions \( p \) and \( q \) is defined as
\[
D_{\alpha}[p||q] = \frac{4}{1 - \alpha^2} \left(1 - \sum_{i} p_i^{\frac{1+\alpha}{\alpha}} q_i^{\frac{1-\alpha}{\alpha}}\right).
\]  
(17)

Let \( M = \mathcal{P} \cup \mathcal{M} \). The mapping from \( P_\theta \) to \( Q_\theta \) can be regarded as choosing the shortest curve connecting \( P_\theta \in M \) and \( Q_\theta \in M \). Such a curve can be formulated as a geodesic, defined as follows.

**Definition 9 (\( \alpha \)-Representation [20]):** For some positive measure \( m_i^{-\alpha} \), the coordinate system \( \theta = (\theta^i) \) derived from the \( \alpha \)-divergence is
\[
\theta^i = m_i^{\frac{1-\alpha}{\alpha}} = f_\alpha(m_i),
\]  
(18)
where \( f_\alpha \) is defined as
\[
f_\alpha(x) = \begin{cases} x^{\frac{1-\alpha}{\alpha}} & (\alpha \neq 1), \\ \ln x & (\alpha = 1). \end{cases}
\]  
(19)

**Definition 10 (\( \alpha \)-Geodesic [10]):** The \( \alpha \)-geodesic connecting two probability vectors \( p \) and \( q \) is defined as
\[
m_f^{(\alpha,\lambda)}(p, q) = c(\lambda) f_\alpha^{-1}\left(1 - \lambda)f_\alpha(p) + \lambda f_\alpha(q)\right),
\]  
(20)
where \( c(\lambda) \) is the normalizing factor.

**IV. GEODESICAL SKEW t-SNE**

Using the notion of geodesics, skew divergence is generalized in terms of information geometry as follows.

**Definition 11 (\( \alpha \)-Geodesical Skew divergence [21]):** The \( \alpha \)-geodesical skew divergence \( D^{(\alpha,\lambda)}_{GS} \) is defined between two Radon–Nikodym densities \( p \) and \( q \) of \( \mu \)-absolutely continuous probability measures by:
\[
D^{(\alpha,\lambda)}_{GS}[p||q] := \int_{\mathcal{P}} p \ln \frac{p}{m_f^{(\alpha,\lambda)}(p, q)} d\mu(x),
\]  
(21)
where \( \alpha \in \mathbb{R} \) and \( \lambda \in (0, 1] \).

By considering a procedure to minimize the \( D^{(\alpha,\lambda)}_{GS}[p||q] \) in t-SNE, we can derive a skew t-SNE:
\[
(y_1, \ldots, y_n) = \arg \min_{y_1, \ldots, y_n} D^{(\alpha,\lambda)}_{GS}[P||Q].
\]  
(22)

The update rule of \( y_i \) by the gradient descent is given as
\[
y_{i}^{t+1} = y_{i}^{t} - \eta \frac{\partial D^{(\alpha,\lambda)}_{GS}[P||Q]}{\partial y_{i}} + \alpha(t)(y_{i}^{t} - y_{i}^{t-1}),
\]  
(23)
where \( t, \eta \) and \( \alpha(t) \) are respectively the iteration, the learning rate and the momentum of iteration \( t \). Gradient descent strategies do not always converge to the global optimal solution, but they are guaranteed to converge to the saddle point under suitable continuity [22].

**A. GEOMETRY OF THE EMBEDDED MANIFOLDS**

We illustrate the geometric advantages of our generalization of t-SNE. From the definition of \( \alpha \)-geodesical skew divergences, we can see that the generalized t-SNE obtains embeddings along the geodesics associated with the parameters. This suggests that generalized t-SNE may be able to select the optimal parameters from a bundle of curves, whereas the existing KL-divergence based t-SNE is restricted to a single curve. For a more formal discussion, we introduce the notion of flatness of a manifold.

**Definition 12:** The curvature of the linear connection \( \nabla \) is given by
\[
R_{ijk} = \partial_i \Gamma_{jk} - \partial_j \Gamma_{ik} + \Gamma_{il} \Gamma_{jk} - \Gamma_{jl} \Gamma_{ik}.
\]  
(24)

Similarly, let \( R^{(\alpha)} \) be the curvature tensor in the \( \alpha \)-coordinate system. A manifold is said to be \( \alpha \)-flat if \( R^{(\alpha)} = 0 \). Since curvature is a local property that depends on the coordinate system, which curves connect manifolds with the shortest distance also depends on the properties of the manifold. Figure 1 shows a conceptual graph to help with intuitive understanding.

Original t-SNE assumes the student’s t-distribution with 1 degree of freedom, but for the sake of general discussion, we consider the following version with \( r \) degrees of freedom:
\[
\mathcal{M} = \left\{ q(z) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{\pi r} \Gamma\left(\frac{r}{2}\right)} \left(1 + \frac{u}{r}(z-u)\right)^{-\frac{r+1}{2}} \right\},
\]  
(25)
where \( u \) is location parameter and \( v \) is scale parameter and \( r \) is degrees of freedom. Is known that the Fisher information matrix of \( \mathcal{M} \) is given as follows:
\[
(g_{ij}) = \begin{pmatrix} \frac{r+1}{v^2(r+3)} & 0 \\ 0 & \frac{2r}{v^2(r+3)} \end{pmatrix}
\]  
(26)

**Proposition 2:** The \( \alpha \)-curvature tensor of the \( \mathcal{M} \) is given as follows.
\[
R^{(\alpha)}_{1212} = \frac{-(r+1)((r+5)^2 - \alpha^2(r-1)^2)}{v^4(r+3)(r+5)^2}.
\]  
(27)
Proof: We have
\[
\Gamma_{112} = \Gamma_{112}^k = \frac{r + 1}{v^3(r + 3)} , \quad \Gamma_{121} = \Gamma_{121}^k = -\frac{r + 1}{v^3(r + 3)} ,
\]
\[
\Gamma_{222} = \Gamma_{222}^k = -\frac{2r}{v^3(r + 3)} , \quad \Gamma_{122} = \Gamma_{122}^k = 0 .
\]
Let \( \ell = \ln p \). We also have
\[
T_{112} = \mathbb{E}[\partial_1 \ell \partial_1 \ell \partial_2 \ell] = \frac{2(r + 1)(r - 1)}{v^3(r + 3)(r + 5)} ,
\]
\[
T_{222} = \mathbb{E}[\partial \ell \partial \ell \partial \ell] = \frac{8r(r - 1)}{v^3(r + 3)(r + 5)} ,
\]
\[
T_{111} = T_{122} = 0 .
\]
By combining them,
\[
\Gamma_{112}^{(o)} = \Gamma_{112} - \frac{\alpha}{2} T_{112} = \frac{(r + 1)(r + 5) - \alpha(r - 1)}{v^3(r + 3)(r + 5)} ,
\]
\[
\Gamma_{222}^{(o)} = \Gamma_{222} - \frac{\alpha}{2} T_{222} = \frac{2r[(r + 5) + 2\alpha(r - 1)]}{v^3(r + 3)(r + 5)} .
\]
TABLE 1. The $Q_{\text{local}}$ of the scale-independent quality criteria [28]. By choosing the parameters well, we can see that our generalized t-SNE outperforms the original t-SNE.

|          | MNIST [23] | Fashion-MNIST [24] | COIL-20 [25] |
|----------|------------|---------------------|--------------|
| $\lambda = 1$ (t-SNE) | $0.6735 (\pm 0.102)$ | $0.6359 (\pm 0.158)$ | $0.8534 (\pm 0.087)$ |
| $\alpha = 1, \lambda = 0.25$ | $0.6782 (\pm 0.099)$ | $0.6282 (\pm 0.166)$ | $0.8628 (\pm 0.094)$ |
| $\alpha = 1, \lambda = 0.5$ | $0.6819 (\pm 0.113)$ | $0.6552 (\pm 0.157)$ | $0.8601 (\pm 0.092)$ |
| $\alpha = 1, \lambda = 0.75$ | $0.6834 (\pm 0.121)$ | $0.6171 (\pm 0.159)$ | $0.8589 (\pm 0.090)$ |
| $\alpha = 0, \lambda = 0.25$ | $0.6903 (\pm 0.119)$ | $0.6287 (\pm 0.162)$ | $0.8426 (\pm 0.093)$ |
| $\alpha = 0, \lambda = 0.5$ | $0.6928 (\pm 0.116)$ | $0.6114 (\pm 0.148)$ | $0.8588 (\pm 0.087)$ |
| $\alpha = 0, \lambda = 0.75$ | $0.6975 (\pm 0.110)$ | $0.6398 (\pm 0.149)$ | $0.8243 (\pm 0.089)$ |
| $\alpha = 1, \lambda = 0.5$ | $0.6917 (\pm 0.098)$ | $0.6724 (\pm 0.148)$ | $0.8488 (\pm 0.091)$ |
| $\alpha = 1, \lambda = 0.75$ | $0.6866 (\pm 0.096)$ | $0.6382 (\pm 0.157)$ | $0.8523 (\pm 0.093)$ |
| $\alpha = 2, \lambda = 0.25$ | $0.6726 (\pm 0.112)$ | $0.6441 (\pm 0.153)$ | $0.8570 (\pm 0.090)$ |
| $\alpha = 2, \lambda = 0.5$ | $0.6718 (\pm 0.109)$ | $0.6293 (\pm 0.150)$ | $0.8692 (\pm 0.088)$ |
| $\alpha = 2, \lambda = 0.75$ | $0.6709 (\pm 0.122)$ | $0.6239 (\pm 0.146)$ | $0.8724 (\pm 0.087)$ |

Thus

$$
\Gamma_{121}^{(\alpha)} = \Gamma_{121} - \frac{\alpha}{2} T_{121} = \frac{(r + 1)((r + 5) - \alpha(r - 1))}{v^3(r + 3)(r + 5)},
$$

$$
\Gamma_{111}^{(\alpha)} = \Gamma_{122} = 0.
$$

Thus

$$
R_{121}^{(\alpha)} = \left( \partial_1 \Gamma_{11}^{(\alpha)} - \partial_1 \Gamma_{21}^{(\alpha)} \right) g_{12} + \Gamma_{12}^{(\alpha)} \Gamma_{21}^{(\alpha)} - \Gamma_{21}^{(\alpha)} \Gamma_{11}^{(\alpha)} - \frac{(r + 1)((r + 5)^2 - \alpha^2(r - 5)^2)}{v^4(r + 3)(r + 5)^2}.
$$

Proposition 3: Embedding manifolds of the original t-SNE have the constant curvature $R_{121}^{(\alpha)} = -\frac{1}{2\pi}$.

Proof: By substituting $r = 1$, we can prove immediately.

V. EXPERIMENTAL RESULTS

In this section, we present several numerical experiments. In our experiments, we apply generalized t-SNE to the MNIST [23], Fashion-MNIST [24] and COIL-20 [25] datasets. Since the generalized t-SNE includes the original t-SNE by modifying the parameters, we study the behavior of the method by conducting experiments with multiple parameters. Throughout the entire experiment, we use perplexity = 30.

A. CLUSTERING PERFORMANCE

First, we investigate the clustering performance of generalized t-SNE. We apply $k$-Nearest Neighbors [26] to the dimensionally reduced representation of the raw data and observe the clustering performance.

Table 1 shows the evaluation of the embedding space by using the scale-independent quality criteria [16]:

$$
Q_{\text{local}} := \frac{1}{K_{\text{max}}} \sum_{K=1}^{K_{\text{max}}} \frac{1}{KN} \sum_{k,i \in F_K \times S_K} q_{ki},
$$

$F_K = \{1, \ldots, K\}$, $S = \{K + 1, \ldots, N - 1\}$, $Q = (q_{ki})_{1 \leq k,i \leq N-1}$ is the co-ranking matrix [27] and

$$
K_{\text{max}} := \arg \max \left\{ \left( \frac{1}{KN} \sum_{k,i \in F_K \times S_K} q_{ki} - \frac{K}{N-1} \right) \right\}.
$$

This metric is the rank-based scale-independent quality criteria for the embedded spaces, and the embedding space with high $Q_{\text{local}}$ is evaluated as having good properties.

From Table 1, we can see that the generalized t-SNE outperforms the original t-SNE with a suitable choice of parameters. In fact, a t-test between the parameter set when the best performance is achieved for each dataset and the performance of the original t-SNE yields $t_{\text{MNIST}} = -5.059$, $t_{\text{Fashion-MNIST}} = -5.331$, $t_{\text{COIL-20}} = -4.883$, indicating a significant difference in all experiments.

In addition, the visualization results for each dataset are shown in Figures 2, 3 and 4. From these figures, we can see that our generalized t-SNE converges faster than the original t-SNE with respect to the number of iterations.
observe that the behavior of the generalized t-SNE changes depending on the choice of parameters.

B. NUMERICAL ANALYSIS FOR THE CONVERGENCE

Figure 5 shows the numerical convergence analysis of t-SNE and skew t-SNE. In this experiment, we performed optimization by SGD [29] with $\lambda = 0.05$ on the MNIST dataset [23] and set the parameters of skew t-SNE to $\alpha \in \{0, 1, 2\}$ and $\lambda = 0.5$. From this experiment, we can see that skew t-SNE converges dramatically faster than the original t-SNE.

VI. RELATED WORKS

Many variants of t-SNE have been proposed due to the excellent experimental results [30]–[32]. Kernel t-SNE [33], [34] is an efficient extension of t-SNE to a parametric framework, which preserves the flexibility of original t-SNE, but enables explicit out-of-sample extensions. In addition, there are variants that leverage the power of GPUs to solve the computational complexity of t-SNE [35], [36].

Other research areas around t-SNE include its parameter tuning [11], [37]. As mentioned earlier, the original t-SNE has a large computational time complexity, so the automation of parameter tuning is very important.

VII. CONCLUSION AND DISCUSSION

In this paper, we generalized t-SNE from the perspective of information geometry. Our generalization is derived very naturally, and suggests the possibility of the existence of a better parameter set.

A. FUTURE WORKS

We have the following future studies:

- Deriving the convergence rate of skew t-SNE. In Section V-B, we show numerically that our skew t-SNE converges significantly faster than the original t-SNE. It is very important in practical applications to derive the convergence rate of the proposed method theoretically, for example, $O(N)$ and $O(N^2)$.

- Information criterion for parameter selection. As shown in Figure 6, generalized t-SNE produces different clustering results depending on the values of the two parameters. Therefore, some criteria for determining the parameters are very useful [38], [39].

- Additional experiments. For example, it would be useful to investigate the combination of our generalized t-SNE with other clustering methods other than k-means [40], [41].

REFERENCES

[1] L. Cayton, “Algorithms for manifold learning,” Univ. California, San Diego, CA, USA, Tech. Rep., 2005, p. 1, vol. 12, nos. 1–17.

[2] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen, “Image reconstruction by domain-transform manifold learning,” Nature, vol. 555, no. 7697, pp. 487–492, 2018.

[3] I. K. Fodor, “A survey of dimension reduction techniques,” Lawrence Livermore Nat. Lab., Livermore, CA, USA, Tech. Rep., 2002.

[4] L. McInnes, J. Healy, N. Saul, and L. Großberger, “UMAP: Uniform manifold approximation and projection,” J. Open Source Softw., vol. 3, no. 29, p. 861, Sep. 2018.

[5] C. D. Hansen and C. R. Johnson, Visualization Handbook. Amsterdam, The Netherlands: Elsevier, 2011.

[6] M. Aparicio and C. J. Costa, “Data visualization,” Commun. Des. Quart. Rev., vol. 3, no. 1, pp. 7–11, 2015.

[7] G. Hinton and S. Roweis, “Stochastic neighbor embedding,” in Proc. NIPS, vol. 15, 2002, pp. 833–840.

[8] L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, no. 11, pp. 2579–2605, 2008.

[9] S.-I. Amari and H. Nagaoka, Methods of Information Geometry, vol. 191. Providence, RI, USA: American Mathematical Society, 2000.

[10] S.-I. Amari, Information Geometry and Its Applications, vol. 194. Springer, 2016.

[11] Y. Cao and L. Wang, “Automatic selection of t-SNE perplexity,” 2017, arXiv:1708.03229. [Online]. Available: http://arxiv.org/abs/1708.03229

[12] S. Kullback and R. A. Leibler, “On information and sufficiency,” Ann. Math. Statist., vol. 22, no. 1, pp. 79–86, 1951.

[13] Y. Sakamoto, M. Ishiguro, and G. Kitagawa, Akaike Information Criterion Statistics, vol. 81, no. 10.5555. Dordrecht, The Netherlands: D. Reidel, 1986, p. 26853.

[14] J. Goldberger, S. Gordon, and H. Greenspan, “An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures,” in Proc. ICCV, vol. 3, 2003, pp. 487–493.

[15] J. Lin, “Divergence measures based on the Shannon entropy,” IEEE Trans. Inf. Theory, vol. 37, no. 1, pp. 145–151, Jan. 1991.

[16] L. Lee, “Measures of distributional similarity,” 2000, arXiv:cs/0001012. [Online]. Available: https://arxiv.org/abs/cs/0001012

[17] E. W. Weisstein. (2014). Einstein Summation. [Online]. Available: https://mathworld.wolfram.com/

[18] S.-I. Amari, Differential-Geometrical Methods in Statistics, vol. 28. Springer, 2012.

[19] S.-I. Amari, “Differential-geometrical methods in statistics,” Lect. Notes Statist., vol. 28, p. 1, 1995.

[20] N. Ay, J. Jost, H. Ván Lé, and L. Schwachhöfer, “Information geometry and sufficient statistics,” Probab. Theory Rel. Fields, vol. 162, nos. 1–2, pp. 327–364, Jun. 2015.

[21] M. Kimura and H. Hino, “α-geodesical skew divergence,” Entropy, vol. 23, no. 5, p. 528, Apr. 2021. [Online]. Available: https://www.mdpi.com/1099-4300/23/5/528

[22] R. M. Gower, N. Loizou, X. Qian, A. Salianbayev, E. Shulgin, and P. Richtárík, “SGD: General analysis and improved rates,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 5200–5209.

[23] Y. LeCun and C. Cortes. (2010). MNIST Handwritten Digit Database. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[24] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms,” 2017, arXiv:1708.07747. [Online]. Available: http://arxiv.org/abs/1708.07747

[25] S. Nene, S. K. Nayar, and H. Murase. (1996). Columbia Object Image Library (COIL-20). [Online]. Available: http://www.cs.columbia.edu/CAVE/software/softlib/coin-20.php

[26] L. E. Peterson, “K-nearest neighbor,” Scholarpedia, vol. 4, no. 2, p. 1883, 2009.

[27] J. A. Lee and M. Verleysen, “Rank-based quality assessment of nonlinear dimensionality reduction,” in Proc. ESANN, 2008, pp. 49–54.
[28] J. A. Lee and M. Verleysen, “Scale-independent quality criteria for dimensionality reduction.” *Pattern Recognit. Lett.*, vol. 31, no. 14, pp. 2248–2257, 2010.

[29] H. Robbins, “A stochastic approximation method,” *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 2007.

[30] B. Kang, D. G. García, J. Lijffijt, R. Santos-Rodríguez, and T. De Bie, “Conditional t-SNE: Complementary t-SNE embeddings through factorizing out prior information.” 2019, arXiv:1905.10086. [Online]. Available: http://arxiv.org/abs/1905.10086

[31] A. Chatzimparmpas, R. M. Martins, and A. Kerren, “t-viSNE: Interactive assessment and interpretation of t-SNE projections,” *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 8, pp. 2696–2714, Aug. 2020.

[32] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, “Stochastic neighbor embedding with Gaussian and student-t distributions: Tutorial and survey.” 2020, arXiv:2009.10301. [Online]. Available: http://arxiv.org/abs/2009.10301

[33] A. Gisbrecht, A. Schulz, and B. Hammer, “Parametric nonlinear dimensionality reduction using kernel t-SNE,” *Neurocomputing*, vol. 147, pp. 71–82, Jan. 2015.

[34] H. Zhang, P. Wang, X. Gao, Y. Qi, and H. Gao, “Out-of-sample data visualization using bi-kernel t-SNE,” *Inf. Visualizat.*, vol. 20, no. 1, pp. 20–34, Jan. 2021.

[35] D. M. Chan, R. Rao, F. Huang, and J. F. Canny, “t-SNE-CUDA: GPU-accelerated t-SNE and its applications to modern data,” in *Proc. 30th Int. Symp. Comput. Archit. High Perform. Comput. (SBAC-PAD)*, Sep. 2018, pp. 330–338.

[36] N. Pezzotti, J. Thijssen, A. Mordvintsev, T. Höllt, B. Van Lew, B. P. F. Lelieveldt, E. Eisemann, and A. Vilanova, “GPGPU linear complexity t-SNE optimization,” *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 1, pp. 1172–1181, Jan. 2020.

[37] A. C. Belkina, C. O. Ciccolella, R. Anno, R. Halpert, J. Spidlen, and J. E. Snyder-Cappione, “Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets,” *Nature Commun.*, vol. 10, no. 1, pp. 1–12, Dec. 2019.

[38] A. Soriano, L. Vergara, B. Ahmed, and A. Salazar, “Fusion of scores in a detection context based on alpha integration,” *Neural Comput.*, vol. 27, no. 9, pp. 1983–2010, 2015.

[39] G. Safont, A. Salazar, and L. Vergara, “Vector score alpha integration for classifier late fusion,” *Pattern Recognit. Lett.*, vol. 136, pp. 48–55, Aug. 2020.

[40] Y. Zhang, X. Bai, R. Fan, and Z. Wang, “Deviation-sparse fuzzy C-means with neighbor information constraint,” *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 1, pp. 185–199, Jan. 2019.

[41] Y. Tang, F. Ren, and W. Pedrycz, “Fuzzy C-means clustering through SSIM and patch for image segmentation,” *Appl. Soft Comput.*, vol. 87, Feb. 2020, Art. no. 105928.

**MASANARI KIMURA** was born in Aomori, Japan. He received the B.S. degree in library and information science from the University of Tsukuba, in 2018. He is currently pursuing the Ph.D. degree in statistics with The Graduate University for Advanced Studies. Since 2018, he has been working as a Research Engineer at Ridge-i Inc. Since 2019, he has been also working as an External Research Fellow with the National Institute of Advanced Industrial Science and Technology.