

Theoretical Foundations of t-SNE for Visualizing High-Dimensional Clustered Data

T. Tony Cai

*Department of Statistics and Data Science
University of Pennsylvania
Philadelphia, PA 19104, USA*

TCAI@WHARTON.UPENN.EDU

Rong Ma

*Department of Statistics
Stanford University
Stanford, CA 94305, USA*

RONGM@STANFORD.EDU

Editor: Ji Zhu

Abstract

This paper investigates the theoretical foundations of the t-distributed stochastic neighbor embedding (t-SNE) algorithm, a popular nonlinear dimension reduction and data visualization method. A novel theoretical framework for the analysis of t-SNE based on the gradient descent approach is presented. For the early exaggeration stage of t-SNE, we show its asymptotic equivalence to power iterations based on the underlying graph Laplacian, characterize its limiting behavior, and uncover its deep connection to Laplacian spectral clustering, and fundamental principles including early stopping as implicit regularization. The results explain the intrinsic mechanism and the empirical benefits of such a computational strategy. For the embedding stage of t-SNE, we characterize the kinematics of the low-dimensional map throughout the iterations, and identify an amplification phase, featuring the intercluster repulsion and the expansive behavior of the low-dimensional map, and a stabilization phase. The general theory explains the fast convergence rate and the exceptional empirical performance of t-SNE for visualizing clustered data, brings forth the interpretations of the t-SNE visualizations, and provides theoretical guidance for applying t-SNE and selecting its tuning parameters in various applications.

Keywords: Clustering; Data visualization; Foundation of data science; Nonlinear dimension reduction; t-SNE

1. Introduction

Data visualization is critically important for understanding and interpreting the structure of large datasets, and has been recognized as one of the fundamental topics in data science (Donoho, 2017). A collection of machine learning algorithms for data visualization and dimension reduction have been developed. Among them, the t-distributed stochastic neighbor embedding (t-SNE) algorithm, proposed by van der Maaten and Hinton (2008), is arguably one of the most popular methods and a state-of-art technique for a wide range of applications (Wang et al., 2021).

Specifically, t-SNE is an iterative algorithm for visualizing high-dimensional data by mapping the data points to a two- or finite-dimensional space. It creates a single map that

reveals the intrinsic structures in a high-dimensional dataset, including trends, patterns, and outliers, through a nonlinear dimension reduction technique. In the past decade, the original t-SNE algorithm, along with its many variants (for example, Yang et al. (2009); Carreira-Perpinán (2010); Xie et al. (2011); van der Maaten (2014); Gisbrecht et al. (2015); Pezzotti et al. (2016); Im et al. (2018); Linderman et al. (2019); Chatzimparmpas et al. (2020)), has made profound impact to the practice of scientific research, including genetics (Platzer, 2013), molecular biology (Olivon et al., 2018), single-cell transcriptomics (Kobak and Berens, 2019), computer vision (Cheng et al., 2015) and astrophysics (Traven et al., 2017). In particular, the extraordinary performance of t-SNE for visualizing high-dimensional data with intrinsic clusters has been widely acknowledged (van der Maaten, 2014; Kobak and Berens, 2019).

Compared to the extensive literature on the computational and numerical aspects of t-SNE, there is a paucity of fundamental results about its theoretical foundations (see Section 1.3 for a brief overview). The lack of theoretical understanding and justifications profoundly limits the users’ interpretation of the results as well as the potentials for further improvement of the method.

This paper aims to investigate the theoretical foundations of t-SNE. Specifically, we present a novel framework for the analysis of t-SNE, provide theoretical justifications for its competence in dimension reduction and visualizing clustered data, and uncover the fundamental principles underlying its exceptional empirical performance.

1.1 Basic t-SNE Algorithm

Let $\{X_i\}_{1 \leq i \leq n}$ be a set of p -dimensional data points. t-SNE 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 $\mathbf{P} = (p_{ij})_{1 \leq i, j \leq n} \in \mathbb{R}^{n \times n}$, where $p_{ii} = 0$ for all $1 \leq i \leq n$, and for $i \neq j$,

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n} \quad \text{with} \quad p_{j|i} = \frac{\exp(-\|X_i - X_j\|_2^2 / 2\tau_i^2)}{\sum_{\ell \in \{1, 2, \dots, n\} \setminus \{i\}} \exp(-\|X_i - X_\ell\|_2^2 / 2\tau_i^2)}. \quad (1)$$

Here τ_i are tuning parameters, which are usually determined based on a certain perplexity measure and a binary search strategy (Hinton and Roweis, 2002; van der Maaten and Hinton, 2008). Similarly, for a two-dimensional¹ map $\{y_i\}_{1 \leq i \leq n} \subset \mathbb{R}^2$, we define the joint probability distribution over all pairs $\{(y_i, y_j)\}_{1 \leq i \neq j \leq n}$ through a symmetric matrix $\mathbf{Q} = (q_{ij})_{1 \leq i, j \leq n}$ where $q_{ii} = 0$ for all $1 \leq i \leq n$ and for $i \neq j$,

$$q_{ij} = \frac{(1 + \|y_i - y_j\|_2^2)^{-1}}{\sum_{\ell, s \in \{1, 2, \dots, n\}, \ell \neq s} (1 + \|y_\ell - y_s\|_2^2)^{-1}}. \quad (2)$$

Intuitively, \mathbf{P} and \mathbf{Q} are similarity matrices summarizing the pairwise distances of the high-dimensional data points $\{X_i\}_{1 \leq i \leq n}$, and the two-dimensional map $\{y_i\}_{1 \leq i \leq n}$, respectively. Then t-SNE aims to find $\{y_i\}_{1 \leq i \leq n}$ that minimizes the KL-divergence between \mathbf{P} and \mathbf{Q} , that is,

$$(y_1, \dots, y_n) = \arg \min_{y_1, \dots, y_n} D_{KL}(\mathbf{P}, \mathbf{Q}) = \arg \min_{y_1, \dots, y_n} \sum_{\substack{i, j \in \{1, 2, \dots, n\} \\ i \neq j}} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \quad (3)$$

1. Throughout, we focus on the two-dimensional embedding for ease of presentation. However, all the theoretical results obtained in this work holds for any finite constant embedding dimension.

Many algorithms have been proposed to solve this optimization problem. The most widely used algorithm was proposed in van der Maaten and Hinton (2008), which draws on a variant of gradient descent algorithm, with an updating equation

$$y_i^{(k+1)} = y_i^{(k)} + hD_i^{(k)} + m^{(k+1)}(y_i^{(k)} - y_i^{(k-1)}), \quad \text{for } i = 1, \dots, n, \quad (4)$$

where $h \in \mathbb{R}_+$ is a prespecified step size parameter, $D_i^{(k)} = 4 \sum_{1 \leq j \leq n, j \neq i} (y_j^{(k)} - y_i^{(k)}) S_{ij}^{(k)} \in \mathbb{R}^2$ is the gradient term corresponding to y_i , with $S_{ij}^{(k)} = (p_{ij} - q_{ij}^{(k)}) / (1 + \|y_i^{(k)} - y_j^{(k)}\|_2^2) \in \mathbb{R}$, and $m^{(k)} \in \mathbb{R}_+$ is a prespecified momentum parameter. The algorithm starts with an initialization $y_i^{(0)} = y_i^{(-1)}$ for $i \in \{1, 2, \dots, n\}$, drawn independently from a uniform distribution on $[-0.01, 0.01]^2$, or from $N(0, \delta^2 I)$ for some small $\delta > 0$.

As indicated by van der Maaten and Hinton (2008), the inclusion of the momentum term $m^{(k+1)}(y_i^{(k)} - y_i^{(k-1)})$ in (4) is mainly to speed up the convergence and to reduce the risk of getting stuck in a local minimum. In this paper, for simplicity and generality we focus on the basic version of the t-SNE algorithm based on the simple gradient descent, with the updating equation

$$y_i^{(k+1)} = y_i^{(k)} + hD_i^{(k)}, \quad \text{for } i = 1, \dots, n. \quad (5)$$

In van der Maaten and Hinton (2008) and van der Maaten (2014), the recommended total number of iterations is 1000, while the step size h is initially set as 400 or 800, and is updated at each iteration by an adaptive learning rate scheme of Jacobs (1988).

The standard gradient descent algorithm as in (5) suffers from a slow convergence rate and even non-convergence in some applications. As an amelioration, van der Maaten and Hinton (2008) proposed an *early exaggeration* technique, applied to the initial stages of the optimization, that helps create patterns in the visualization and speed up the convergence. Such a computational strategy has been standard in practical use. In fact, most of the current software implementations of t-SNE are based on an early exaggeration stage followed by an embedding stage that iterates a certain gradient descent algorithm. In our setting, these two stages can be summarized as follows.

Early exaggeration stage. For the first $K_0 > 0$ iterations, the p_{ij} 's in the gradient term $D_i^{(k)}$ are multiplied by some exaggeration parameter $\alpha > 0$, so the updating equation for this early exaggeration stage becomes

$$y_i^{(k+1)} = y_i^{(k)} + h \sum_{1 \leq j \leq n, j \neq i} (y_j^{(k)} - y_i^{(k)}) S_{ij}^{(k)}(\alpha), \quad i = 1, \dots, n, \quad (6)$$

where $S_{ij}^{(k)}(\alpha) = (\alpha p_{ij} - q_{ij}^{(k)}) / (1 + \|y_i^{(k)} - y_j^{(k)}\|_2^2) \in \mathbb{R}$, and the factor 4 in $D_i^{(k)}$ is absorbed into the step size parameter h . We refer to this first stage of the t-SNE algorithm as the *early exaggeration stage*.

In van der Maaten and Hinton (2008), the authors choose $\alpha = 4$ and $K_0 = 50$ for the early exaggeration stage, whereas later in van der Maaten (2014), it is recommended that $\alpha = 12$ and $K_0 = 250$. In particular, it is empirically observed that, the early exaggeration technique enables t-SNE to find a better global structure in the early stages of the optimization by creating very tight clusters of points that easily move around in the embedding

space (van der Maaten, 2014); this observation is later supported by some pioneering theoretical investigations (see Section 1.3). Nevertheless, there are interesting questions to be answered concerning (i) the underlying principles and mechanism behind such a computational strategy, (ii) the limit behavior of the low-dimensional map, (iii) how sensitive is the performance of t-SNE with respect to the choice of tuning parameters (α, h, K_0), and (iv) how to efficiently determine these parameters to achieve the best empirical performance.

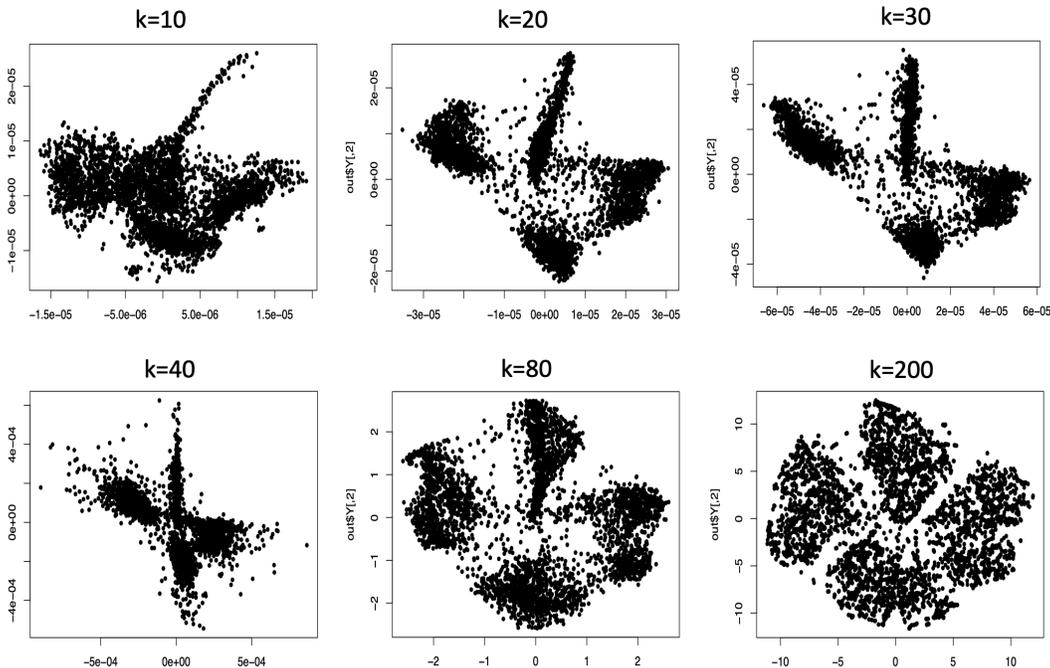


Figure 1: An illustration of the t-SNE iterations that visualize samples from the MNIST dataset (Section 5). Each sample corresponds an image of handwritten digit “2,” “4,” “6,” or “8.” The visualizations are obtained using the `Rtsne` function in the R package `Rtsne`, by selecting the exact t-SNE mode (`theta=0`, `pca=F`), dropping the momentum terms (`momentum=0`, `final_momentum = 0`), and setting `perplexity=30` (default), $\alpha = 12$ (default), $h = 200$ (default) in (6), and $K_0 = 40$. The first three plots (top row) correspond to the early exaggeration stage, while the last three plots (bottom row) correspond to the embedding stage.

Embedding stage. After the early exaggeration stage, the exaggeration parameter α is dropped and the original iterative algorithm (5) is carried out till attaining a prespecified number of steps. We refer to this second stage as the *embedding stage*. The final output is a two-dimensional map $\{y_i^{(K_1)}\}_{1 \leq i \leq n}$, commonly treated as a low-dimensional embedding of the original data $\{X_i\}_{1 \leq i \leq n}$, expected to preserve its intrinsic geometric structures.

In addition to data visualization, t-SNE is sometimes also used as an intermediate step for clustering, signal detection, among many other purposes. In particular, it has been

observed that, when applied to high-dimensional clustered data, t-SNE tends to produce a visualization with more separated clusters, which are often in good agreement with the clusters found by a dedicated clustering algorithm (Kobak and Berens, 2019). See Figure 1 for an example of data visualization using such a basic t-SNE algorithm.

1.2 Main Results and Our Contribution

A formal theoretical framework is introduced for the analysis of t-SNE that relies on a joint statistical and computational analysis. The key contribution of the present work can be summarized as follows:

- We rigorously establish the asymptotic equivalence between the early exaggeration stage and power iterations. Our theory unveils novel properties such as the implicit regularization effect and the necessity of early stopping in the early exaggeration stage for weakly clustered data.
- We characterize the behavior of t-SNE iterations at the embedding stage by identifying an amplification phase along with its intercluster repulsion and expansion phenomena, and a stabilization phase of this stage.
- We give the theoretical guidance for initialization and selecting the tuning parameters at both stages in a flexible and data-adaptive manner.
- We provide practical advice on applying t-SNE and interpreting the t-SNE visualizations of high-dimensional clustered data.

The main results can be explained in more detail from three perspectives.

Early exaggeration stage. Through a discrete-time analysis (Sections 2.1 and 2.2), we establish the asymptotic equivalence between the early exaggeration stage and power iterations based on the underlying graph Laplacian associated with the high-dimensional data, providing a spectral-graphical interpretation of the algorithm. We show the implicit spectral clustering mechanism underlying this stage, which explains the adaptivity and flexibility of t-SNE for visualizing clustered data without specifying the number of clusters. Specifically, for the cases where $\{X_i\}_{1 \leq i \leq n}$ are approximately clustered into R groups, we make the key observation that the coordinates of $\{y_i^{(k)}\}_{1 \leq i \leq n}$ converge to the R -dimensional Laplacian null space, leading to a limiting embedding where the elements of $\{y_i^{(k)}\}_{1 \leq i \leq n}$ are well-clustered according to their true cluster membership. On the other hand, through a continuous-time analysis (Section 2.3), we study the underlying gradient flow and uncover an implicit regularization effect depending on the number of iterations. In particular, our analysis implies that when dealing with noisy and approximately clustered data, one should stop early in the early exaggeration stage to avoid “overshooting.” These results justify the empirical observations about the benefits of the early exaggeration technique in creating cluster structures and speeding up the algorithm. For more details about comparison with the existing results, see Section 1.3 and the discussions after Corollaries 7 in Section 2.

Embedding stage. We provide a mechanical interpretation of the algorithm by characterizing the kinematics of the low-dimensional map at each iteration. Specifically, in

Section 3 we identify an amplification phase within the embedding stage, featuring the local intercluster repulsion (Theorem 13) and the global expansive behavior (Theorem 15) of $\{y_i^{(k)}\}_{1 \leq i \leq n}$. In the former case, it is shown that the movement of each $y_i^{(k)}$ to $y_i^{(k+1)}$ is jointly determined by the repulsive forces pointing toward $y_i^{(k)}$ from each of the other clusters (Figure 2), that amounts to increasing spaces between the existing clusters; in the latter case, it is shown the diameter of $\{y_i^{(k)}\}_{1 \leq i \leq n}$ may strictly increase after each iteration. We observe that, following the amplification phase, there is a stabilization phase where $\{y_i^{(k)}\}_{1 \leq i \leq n}$ is locally adjusted to achieve a finer embedding of $\{X_i\}_{1 \leq i \leq n}$. These results together explain the fast convergence rate and the exceptional empirical performance of t-SNE for visualizing clustered data. The articulation of these phenomena also leads to useful practical guidances. See below and Remark 16 in Section 3 for more details.

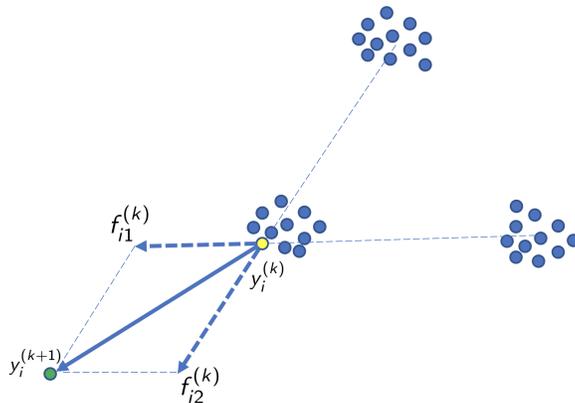


Figure 2: Illustration of the intercluster repulsion where the original data $\{X_i\}_{1 \leq i \leq n}$ have three clusters. The position of $y_i^{(k+1)}$ is jointly determined by $y_i^{(k)}$ and two repulsive forces $f_{i1}^{(k)}$ and $f_{i2}^{(k)}$ pushing $y_i^{(k)}$ away from the other two clusters.

Practical implications. The general theory brings forth the interpretations of the t-SNE output, and provides theoretical guidance for selecting tuning parameters and for initialization. In Section 4 we illustrate the general theory on two examples of high-dimensional clustered data, one generated from a Gaussian mixture model, and another from a noisy nested sphere model. We also analyze in Section 5 a real-world dataset to further demonstrate the practical implications of our theory. In particular, our analysis allows for a wider spectrum of tuning parameters (Figure 10 and Equation (39)) and initialization procedures than those considered in previous theoretical works (Arora et al., 2018; Linderman and Steinerberger, 2019). Moreover, our theoretical results support the state-of-art practice (Kobak and Berens, 2019; Kobak and Linderman, 2021), but also lead to novel insights (e.g., the first item below) that has been unknown to our knowledge. In the following, we summarize our general advice on applying t-SNE to potentially clustered data:

- For weakly clustered data, one may adopt the early exaggeration technique, but needs to stop early (for example, set $K_0 = \lfloor (\log n)^2 \rfloor$) to avoid overshooting – failure of stopping early may lead to false clustering; see Figure 5 below for an illustration.
- t-SNE visualization based on random initialization and early exaggeration is reliable in terms of cluster membership but not relative position of clusters. For example, the neighboring clusters in visualization may not be interpreted as neighboring clusters in the original data; see Figure 6 below for an illustration.
- Occasionally, false clustering may appear as an artifact of random initialization and intercluster repulsion. Therefore, it is helpful to run t-SNE multiple times to fully assess the effect of random initialization; see Figure 6 below for an illustration.
- For strongly clustered data, one can speed up the algorithm by replacing the early exaggeration stage by a simple spectral initialization² where $(\mathbf{y}_1^{(0)}, \mathbf{y}_2^{(0)})$ are the eigenvectors associated with the smallest two eigenvalues of $\mathbf{L}(\mathbf{P})$.

1.3 Related Work

The impressive empirical performance of t-SNE has recently attracted much theoretical interests. Lee and Verleysen (2011) investigated the benefits of the so-called shift-invariant similarities used in the stochastic neighbor embedding and its variants. Later on, they further identified two key properties of these visualization methods (Lee and Verleysen, 2014). In Shaham and Steinerberger (2017), a large family of methods including t-SNE as a special case were studied and shown to successfully map well-separated disjoint clusters from high dimensions to the real line so as to approximately preserve the clustering. In Arora et al. (2018), a theoretical framework was developed to formalize the notion of visualizing clustered data, which is used to analyze the early exaggeration stage of t-SNE, and to justify its high visualization quality. Linderman and Steinerberger (2019) showed that, in the early exaggeration stage of t-SNE, with properly chosen parameters α and h , a subset of the two-dimensional map belonging to the same cluster will shrink in diameter, suggesting well-clustered visualization following iterations. We note that connections between the early exaggeration stage and power iterations have been pointed out in Arora et al. (2018) and Linderman and Steinerberger (2019), but the discussions therein are mostly informal and heuristic. In contrast, we provide rigorous theoretical justification for such a connection, identify its condition and explicate its consequences. By extending the idea of t-SNE, Im et al. (2018) considered a class of methods with various loss functions based on the f -divergence, and theoretically assessed the performances of these methods based on a neighborhood-level precision-recall analysis. More recently, Zhang and Steinerberger (2021) proposed to view t-SNE as a force-based method which generates embeddings by balancing attractive and repulsive forces between data points. In particular, the limiting behavior of t-SNE was analyzed under a mean-field model where a single homogeneous cluster is present. At the empirical side, the recent works of Kobak and Berens (2019) and Kobak and Linderman (2021) summarize the state-of-art practice of using t-SNE to biological data.

2. An illustration is provided in Figure 4.1 of Linderman and Steinerberger (2019).

A comprehensive survey of existing data visualization methods and their properties can be found in Nonato and Aupetit (2018).

Despite these pioneering endeavors, the theoretical understanding of t-SNE is still limited. Many intriguing phenomena and important features that arise commonly in practice have not been well understood or properly explained. Moreover, it remains unclear how to properly interpret the t-SNE visualization and its potential artifacts. These important questions are carefully addressed in the current work for the case of clustered data. Compared to the existing works, the theoretical framework developed in our work leads to identification and explication of novel properties, phenomena, and important practical implications on t-SNE, as summarized at the beginning of Section 1.2.

1.4 Notation and Organization

For a vector $\mathbf{a} = (a_1, \dots, a_n)^\top \in \mathbb{R}^n$, we denote $\text{diag}(a_1, \dots, a_n) \in \mathbb{R}^{n \times n}$ as the diagonal matrix whose i -th diagonal entry is a_i , and define the ℓ_p norm $\|\mathbf{a}\|_p = (\sum_{i=1}^n a_i^p)^{1/p}$. For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n \times n}$, we define its Frobenius norm as $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n a_{ij}^2}$, its ℓ_∞ -norm as $\|\mathbf{A}\|_\infty = \max_{1 \leq i, j \leq n} |a_{ij}|$, and its spectral norm as $\|\mathbf{A}\| = \sup_{\|\mathbf{x}\|_2 \leq 1} \|\mathbf{A}\mathbf{x}\|_2$; we also denote $\mathbf{A}_{\cdot i} \in \mathbb{R}^n$ as its i -th column and $\mathbf{A}_i \in \mathbb{R}^n$ as its i -th row. Let $O(n, k) = \{\mathbf{V} \in \mathbb{R}^{n \times k} : \mathbf{V}^\top \mathbf{V} = \mathbf{I}_k\}$ be the set of all $n \times k$ orthonormal matrices and $O_n = O(n, n)$, the set of n -dimensional orthonormal matrices. For a rank r matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ with $1 \leq r \leq n$, its eigendecomposition is denoted as $\mathbf{A} = \mathbf{U}\mathbf{\Gamma}\mathbf{U}^\top$ where $\mathbf{U} \in O(n, r)$ with its columns being the eigenvectors, and $\mathbf{\Gamma} = \text{diag}(\lambda_1(\mathbf{A}), \lambda_2(\mathbf{A}), \dots, \lambda_r(\mathbf{A}))$ with $\lambda_{\min}(\mathbf{A}) = \lambda_1(\mathbf{A}) \leq \dots \leq \lambda_n(\mathbf{A}) = \lambda_{\max}(\mathbf{A})$ being the ordered eigenvalues of \mathbf{A} . For a smooth function $f(x)$, we denote $\dot{f}(x) = df(x)/dx$ and $\ddot{f}(x) = d^2f(x)/dx^2$. For any integer $n > 0$, we denote the set $[n] = \{1, 2, \dots, n\}$. For a finite set S , we denote its cardinality as $|S|$. For a subset $S \subseteq \mathbb{R}^n$, we define its diameter $\text{diam}(S) = \sup_{x, y \in S} \|x - y\|_2$. For sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = o(b_n)$ or $a_n \ll b_n$ if $\lim_n a_n/b_n = 0$, and write $a_n = O(b_n)$, $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if there exists a constant C such that $a_n \leq Cb_n$ for all n . We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. Throughout, C, C_1, C_2, \dots are universal constants, that can vary from line to line.

The rest of the paper is organized as follows. Section 2 presents the theoretical analysis for the early exaggeration stage of t-SNE. Section 3 analyzes the embedding stage. The general theory is then applied in Section 4 to two specific settings of model-based clustered data, one under a Gaussian mixture model and another under a noisy nested sphere model. Analysis of a real-world dataset is presented in Section 5. Section 6 discusses potential applications, extensions and other related problems. Proofs of our main results and supplementary figures are collected in Appendix A to F.

2. Analysis of the Early Exaggeration Stage

2.1 Asymptotic Graphical Interpretation and Localization

We start with a key observation that connects the updating equation (6) to some graph-related concepts. To this end, we introduce the following definition.

Definition 1 (Degree & Laplacian Operators) For a symmetric matrix $\mathbf{A} = (a_{ij})_{1 \leq i, j \leq n} \in \mathbb{R}^{n \times n}$, define the degree operator $\mathbf{D} : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}$ by $\mathbf{D}(\mathbf{A}) = \text{diag}(\sum_{i=1}^n a_{i1}, \dots, \sum_{i=1}^n a_{in})$, and the Laplacian operator $\mathbf{L} : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}$ by $\mathbf{L}(\mathbf{A}) = \mathbf{D}(\mathbf{A}) - \mathbf{A}$.

We define $\mathbf{S}_\alpha^{(k)} = (S_{ij}^{(k)}(\alpha))_{1 \leq i, j \leq n} \in \mathbb{R}^{n \times n}$ with $S_{ii}^{(k)}(\alpha) \equiv 0$ for all $i \in [n]$. Then we can rewrite the updating equation (6) using the matrix form as

$$\mathbf{y}_\ell^{(k+1)} = [\mathbf{I}_n - h\mathbf{L}(\mathbf{S}_\alpha^{(k)})]\mathbf{y}_\ell^{(k)}, \quad \ell = 1, 2, \quad (7)$$

where $\mathbf{I}_n \in \mathbb{R}^{n \times n}$ is the identity matrix, and $\mathbf{y}_\ell^{(k)} \in \mathbb{R}^n$ consists of the ℓ -th coordinates of $\{y_i^{(k)}\}_{1 \leq i \leq n}$. As a consequence, for each iteration k , if we treat the symmetric matrix $\mathbf{S}_\alpha^{(k)}$ as the adjacency matrix of a weighted graph $G^{(k)}$ with n nodes that summarizes the pairwise relationships between n data points $\{X_i\}_{1 \leq i \leq n}$, Equation (7) has an interpretation that links to the Laplacian matrix of such a weighted graph.

To better understand the meaning and the properties of the underlying graph $G^{(k)}$ that evolve over iterations, we take a closer look at its adjacency matrix $\mathbf{S}_\alpha^{(k)}$. In particular, one should keep in mind that in common applications of t-SNE, the early exaggeration stage has the following empirical features: (i) moderate or relatively large values of the exaggeration parameter α (default 12 in the R package `Rtsne`), (ii) local initializations $\{y_i^{(0)}\}_{1 \leq i \leq n}$ around the origin (see Section 1.1), and (iii) relative small diameters $\text{diam}(\{y_i^{(k)}\}_{1 \leq i \leq n})$ over the iterations (Figure 1).

Our next result shows that, these empirical features of t-SNE have deep connections to the asymptotic behavior of the evolving underlying graphs and their adjacency matrices $\{\mathbf{S}_\alpha^{(k)}\}_{k \geq 1}$ in the large sample limit (as $n \rightarrow \infty$).

Theorem 2 (Asymptotic Graphical Interpretation) Recall that $\mathbf{P} = (p_{ij})_{1 \leq i, j \leq n}$ is defined in (1) and denote $\eta^{(k)} = [\text{diam}(\{y_i^{(k)}\}_{1 \leq i \leq n})]^2$. Then for any $i, j \in [n]$ with $i \neq j$, and each $k \geq 1$ such that $\eta^{(k)} < 1$, we have

$$\left| S_{ij}^{(k)}(\alpha) - \alpha p_{ij} + \frac{1}{n(n-1)} \right| \leq \alpha p_{ij} \eta^{(k)} + \frac{2\eta^{(k)}}{n(n-1)(1-\eta^{(k)})}. \quad (8)$$

Consequently, if we denote $\mathbf{1}_n = (1, \dots, 1)^\top \in \mathbb{R}^n$, and $\mathbf{H}_n = \frac{1}{n(n-1)}(\mathbf{1}_n \mathbf{1}_n^\top - \mathbf{I}_n)$, then for each $k \geq 1$, as long as $(\eta^{(k)}, \alpha)$ satisfies

$$\eta^{(k)} \ll \frac{\|\mathbf{P}\|}{n\|\mathbf{P}\|_\infty}, \quad \alpha \gg \frac{1}{n\|\mathbf{P}\|}, \quad \text{as } n \rightarrow \infty, \quad (9)$$

we have

$$\lim_{n \rightarrow \infty} \frac{\|\mathbf{S}_\alpha^{(k)} - (\alpha\mathbf{P} - \mathbf{H}_n)\|}{\|\alpha\mathbf{P} - \mathbf{H}_n\|} = 0. \quad (10)$$

The above theorem implies that, for large n , as long as the diameter of $\{y_i^{(k)}\}_{1 \leq i \leq n}$ remains sufficiently small and the exaggeration parameter α sufficiently large, the adjacency

matrix $\mathbf{S}_\alpha^{(k)}$ behaves almost like a fixed matrix $\alpha\mathbf{P} - \mathbf{H}_n$ across the iterations. In other words, we may treat the updating equation (7) as an approximately linear equation

$$\mathbf{y}_\ell^{(k+1)} \approx [\mathbf{I}_n - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)]\mathbf{y}_\ell^{(k)}, \quad \ell = 1, 2, \quad (11)$$

where the linear operator $\mathbf{I}_n - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)$ only relies on the Laplacian of a fixed weighted graph whose adjacency matrix is given by the scaled and shifted similarity matrix $\alpha\mathbf{P} - \mathbf{H}_n$. This essentially opens the door to our key result on the asymptotic equivalence between the early exaggeration stage and power iterations.

Before we formally present such a result, we need to first point out an important phenomenon concerning the global behavior of the low-dimensional map at the early exaggeration stage. Specifically, we make the following assumptions on the initialization and the tuning parameters (α, h, k) :

(I1) $\{y_i^{(0)}\}_{1 \leq i \leq n}$ satisfies $\min_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2 > 0$, and $\max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_\infty = O(1)$ as $n \rightarrow \infty$; and

(T1) the parameters (α, h, k) satisfy $k(nh\alpha\|\mathbf{P}\|_\infty + h/n) = O(1)$ as $n \rightarrow \infty$.

Intuitively, Condition (I1) says that the initialization $\{y_i^{(0)}\}_{1 \leq i \leq n}$ should not be simply all zeros or unbounded, whereas the condition (T1) – as a consequence of (8) – essentially requires the cumulative deviations of $h\mathbf{L}(\mathbf{S}_\alpha^{(k)})$ from $h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)$ to be bounded. Our next result shows that, under these assumptions, the diameter of $\{y_i^{(k)}\}_{1 \leq i \leq n}$ may not increase throughout the iterations, so the embedding remains localized within the initial range.

Proposition 3 (Localization) *Suppose (I1) and (T1) hold. We have*

$$\text{diam}(\{y_i^{(k+1)}\}_{1 \leq i \leq n}) \leq C \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_\infty, \quad (12)$$

for some universal constant $C > 0$.

The above proposition confirmed the globally localized and non-expansive behavior of $\{y_i^{(k)}\}_{1 \leq i \leq n}$ over the early exaggeration stage observed in practice (Figure 1). Concerning Theorem 2, it tells us the step-specific condition $\eta^{(k)} \ll \|\mathbf{P}\|/(n\|\mathbf{P}\|_\infty)$ therein can be generalized to all finite k 's as long as the initialization $\{y_i^{(0)}\}_{1 \leq i \leq n}$ is concentrated around 0, that is, $\max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_\infty^2 \ll \|\mathbf{P}\|/(n\|\mathbf{P}\|_\infty)$. Furthermore, when (α, h) are chosen such that the step-wise deviation diminishes (i.e., $r_n = nh\alpha\|\mathbf{P}\|_\infty + h/n \rightarrow 0$), Proposition 3 indicates that (10) may remain true for even larger numbers of iterations as long as $k = O(r_n^{-1})$.

2.2 Asymptotic Power Iterations, Implicit Spectral Clustering and Early Stopping

With the above graphical interpretation of the updating equation (7) in mind, we now present our key result concerning the asymptotic equivalence between the early exaggeration stage and a power method based on the Laplacian matrix $\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)$. In particular, we make the following assumptions on the initialization and the tuning parameters:

- (I2) $\{y_i^{(0)}\}_{1 \leq i \leq n}$ satisfies $\max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_\infty^2 = o(\|\mathbf{P}\|/(n\|\mathbf{P}\|_\infty))$ as $n \rightarrow \infty$; and
 (T1.D) The parameters (α, h, k) satisfy $\alpha \gg (n\|\mathbf{P}\|)^{-1}$ and $k(nh\alpha\|\mathbf{P}\|_\infty + h/n) = o(1)$ as $n \rightarrow \infty$.

Condition (I2) follows from the discussion subsequent to Proposition 3, which along with Condition (T1.D), which is analogous to but stronger than (T1), ensures the conditions for Theorem 2 and Proposition 3 to hold simultaneously.

Theorem 4 (Asymptotic power iterations) *Under Conditions (I1) (I2) and (T1.D), we have (10) and (12) hold, and so does the asymptotic equivalence*

$$\lim_{n \rightarrow \infty} \frac{\|\mathbf{y}_\ell^{(k)} - [\mathbf{I}_n - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)]^k \mathbf{y}_\ell^{(0)}\|_2}{\|\mathbf{y}_\ell^{(0)}\|_2} = 0. \quad (13)$$

The above theorem suggests that each step of the early exaggeration stage may be treated as a power method in the sense that

$$\mathbf{y}_\ell^{(k+1)} \approx [\mathbf{I}_n - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)]^k \mathbf{y}_\ell^{(0)}. \quad (14)$$

The normalization by $\|\mathbf{y}_\ell^{(0)}\|_2$ in (13) makes sure the result to be scale-invariant to the initialization. It is well-known that, for a fixed matrix $\mathbf{G} \in \mathbb{R}^{n \times n}$ with 1 as its unique largest eigenvalue in magnitude, the power iteration $\mathbf{y}^{(k)} = \mathbf{G}^k \mathbf{y}^{(0)}$ converges to the associated eigenvector as $k \rightarrow \infty$. As a result, when treated as an approximate power method, the early exaggeration stage of t-SNE essentially aims to find the direction of the leading eigenvector(s) of the matrix $\mathbf{I}_n - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)$, which, as will be shown shortly, is actually equivalent to finding the eigenvector(s) associated with the smallest eigenvalue of the graph Laplacian $\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)$, or the null space of $\mathbf{L}(\mathbf{P})$.

Led by these observations, our next results concern the limiting behavior of the low-dimensional map $\{y_i^{(k)}\}_{1 \leq i \leq n}$ as the number of iterations $k \rightarrow \infty$. Note that any Laplacian matrix has an eigenvalue 0 associated with a trivial eigenvector $n^{-1/2}\mathbf{1}$. Given the affinity (14) between t-SNE and the power method, we start by showing that, the linear operator $[\mathbf{I}_n - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)]^k$ would converge eventually to a projection operator associated with the null space of the Laplacian $\mathbf{L}(\mathbf{P})$. In particular, we let $R \geq 1$ be the dimension of the null space of the Laplacian $\mathbf{L}(\mathbf{P}) \in \mathbb{R}^{n \times n}$; and assume

- (T2) the parameters (α, h) satisfies $\kappa < h\lambda_{R+1}(\mathbf{L}(\alpha\mathbf{P})) \leq h\lambda_n(\mathbf{L}(\alpha\mathbf{P})) < 1$ for some constant $\kappa \in (0, 1)$.

This assumption corresponds to the so-called ‘‘eigengap’’ condition in the random matrix literature, which gives the signal strength requirements for the recovery of the eigenvalues/eigenvectors, and, in the meantime, the conditions for the tuning parameters.

Theorem 5 (Convergence of power iterations) *Let $\mathbf{U} \in O(n, R)$ such that its columns consist of an orthogonal basis for the null space of $\mathbf{L}(\mathbf{P})$. Suppose $kh = o(n)$ and (T2) hold. Then, we have*

$$\lim_{k \rightarrow \infty} \frac{\|[\mathbf{I}_n - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)]^k \mathbf{y} - \mathbf{U}\mathbf{U}^\top \mathbf{y}\|_2}{\|\mathbf{y}\|_2} = 0. \quad (15)$$

Combining Theorems 4 and 5, we know that for sufficiently large n and k , the t-SNE iterations $\mathbf{y}_\ell^{(k)}$ may converge to the projection of the initial vectors $\mathbf{y}_\ell^{(0)}$ into the null space of the Laplacian $\mathbf{L}(\mathbf{P})$, that is

$$\mathbf{y}_\ell^{(k)} \approx \mathbf{U}\mathbf{U}^\top \mathbf{y}_\ell^{(0)}, \quad \ell \in [2]. \quad (16)$$

Now, to better understand the above theorem and its implications on the limiting behavior of t-SNE applied to clustered data, we study the null space of a special class of Laplacian matrices, corresponding to the family of weighted graphs consisting of $R \geq 2$ connected components. In fact, when the original data $\{X_i\}_{1 \leq i \leq n}$ are well-clustered and τ_i 's are appropriately chosen, the family of disconnected weighted graphs arise naturally since their adjacency matrices are good approximations of \mathbf{P} based on these data (Balakrishnan et al., 2011). We illustrate this point further in Section 4. In the following, we say a symmetric adjacency matrix \mathbf{P} is ‘‘well-conditioned’’ if its associated weighted graph has $R \geq 2$ connected components. Our next result characterizes the Laplacian null space corresponding to these disconnected weighted graphs.

Proposition 6 (Laplacian null space) *Suppose $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric and well conditioned. Then the smallest eigenvalue of the Laplacian $\mathbf{L}(\mathbf{A})$ is 0 and has multiplicity R , and the associated eigen subspace is spanned by $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_R\}$ where for each $r \in \{1, \dots, R\}$,*

$$[\boldsymbol{\theta}_r]_j = \begin{cases} 1/\sqrt{n_r} & \text{if the } j\text{-th node belongs to the } r\text{-th component} \\ 0 & \text{otherwise} \end{cases},$$

and n_r is the number of nodes in the r -th connected component. In particular, up to possible permutation of coordinates, any vector \mathbf{u} in the null space of $\mathbf{L}(\mathbf{A})$ can be expressed as

$$\mathbf{u} = \frac{a_1}{\sqrt{n_1}} \begin{bmatrix} \mathbf{1}_{n_1} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} + \frac{a_2}{\sqrt{n_2}} \begin{bmatrix} \mathbf{0} \\ \mathbf{1}_{n_2} \\ \vdots \\ \mathbf{0} \end{bmatrix} + \dots + \frac{a_R}{\sqrt{n_R}} \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{1}_{n_R} \end{bmatrix}, \quad (17)$$

for some $a_1, \dots, a_R \in \mathbb{R}$.

From the above proposition, for a well-conditioned matrix, the components of any \mathbf{u} in the Laplacian null space has at most R distinct values, and whenever $|\{a_1, \dots, a_R\}| = R$, the coordinates share the same value if and only if the corresponding nodes fall in the same connected component, i.e., the same cluster. Combining (16) and (17), one can see that, for strongly clustered data, the output from the early exaggeration stage essentially converges to the eigenvectors associated with the Laplacian null space. This leads to our fourth practical advice at the end of Section 1.2.

We now generalize the analysis to the setting where the data $\{X_i\}_{1 \leq i \leq n}$ is only weakly clustered in the sense that there exists a well-conditioned symmetric matrix \mathbf{P}^* close to \mathbf{P} under properly chosen $\{\tau_i\}$, and the underlying graph associated with \mathbf{P} may not be necessarily disconnected. More specifically, we assume

(T2.D) there exists a symmetric and well-conditioned matrix $\mathbf{P}^* \in \mathbb{R}^{n \times n}$ satisfying (T2) and is sufficiently close to \mathbf{P} in the sense that $kh\alpha \|\mathbf{L}(\mathbf{P}^* - \mathbf{P})\| = o(1)$.

For a given \mathbf{P} satisfying (T2.D), let n_r with $r \in [R]$ be the size of the r -th connected component in the graph associated with \mathbf{P}^* . Our next theorem obtains the implicit spectral clustering and early stopping properties of the early exaggeration stage.

Theorem 7 (Implicit clustering and early stopping) *Suppose the similarity \mathbf{P} and the tuning parameters (α, h, k) satisfy (T1.D) and (T2.D), and the initialization satisfies (I1) and (I2). Then there exists some permutation matrix $O \in \mathbb{R}^{n \times n}$ such that, for $\ell \in [2]$,*

$$\lim_{(k,n) \rightarrow \infty} \frac{\|\mathbf{y}_\ell^{(k)} - O\mathbf{z}_\ell\|_2}{\|\mathbf{y}_\ell^{(0)}\|_2} = 0, \quad (18)$$

where

$$\mathbf{z}_\ell = \left(\underbrace{z_{\ell 1}, \dots, z_{\ell 1}}_{n_1}, \underbrace{z_{\ell 2}, \dots, z_{\ell 2}}_{n_2}, \dots, \underbrace{z_{\ell R}, \dots, z_{\ell R}}_{n_R} \right)^\top \in \mathbb{R}^n, \quad (19)$$

and $z_{\ell r} = \boldsymbol{\theta}_r^\top \mathbf{y}_\ell^{(0)} / \sqrt{n_r}$ for $r \in [R]$.

Theorem 7 describes the limiting behavior of the low-dimensional map $\{y_i^{(k)}\}_{1 \leq i \leq n}$ as $(n, k) \rightarrow \infty$, when the original data is approximately clustered. Specifically, elements from $\{y_i^{(k)}\}_{1 \leq i \leq n}$ associated with a connected component of the underlying graph would converge cluster-wise towards a few points on \mathbb{R}^2 . In particular, Theorem 7 suggests that, at the end of the early exaggeration stage, although the samples belonging to the same underlying cluster tend to be clustered together in the t-SNE embeddings, the cluster centers of the t-SNE embeddings only rely on the initialization, rather than the actual positions of the underlying clusters. Therefore, if *initialized randomly and noninformatively*, the t-SNE embeddings at the end of the early exaggeration stage tend to preserve only the local structures (i.e., the closeness of the samples from the same cluster) but not the global structures (i.e., the relative positions of different clusters) of the original data (Kobak and Berens, 2019; Kobak and Linderman, 2021). This observation, as illustrated in Figure 6, leads to our second practical advice at the end of Section 1.2.

Our theory refines and improves the existing works such as Linderman and Steinerberger (2019) and Arora et al. (2018) in various aspects. Firstly, our theoretical framework formalizes and explains the asymptotic equivalence between the early exaggeration stage and the power iterations. The theory provides a precise description of the limiting behavior of the low-dimensional map and the theoretical conditions. Secondly, unlike the previous works where only one particular initialization and relatively limited range of tuning parameters were considered, our analysis yields general conditions and allows for more flexible choices of the initialization procedures and tuning parameters. Finally, our analysis unveils the need of stopping early in the early exaggeration stage for weakly clustered data, which is a novel feature. Specifically, both Conditions (T1.D) and (T2.D) allow $k \rightarrow \infty$ but in a controlled manner – whenever $\|\mathbf{L}(\mathbf{P}^* - \mathbf{P})\| \neq 0$, there is a data dependent upper bound on the iteration number

$$k \ll \frac{1}{h\alpha\|\mathbf{L}(\mathbf{P}^* - \mathbf{P})\|},$$

which becomes more stringent for weakly clustered data (i.e., $\|\mathbf{L}(\mathbf{P}^* - \mathbf{P})\|$ not too small). Such a phenomenon is also observed empirically (Figure 5), where failing to stop early would lead to false clustering.

2.3 Gradient Flow and Implicit Regularization

For $\ell \in \{1, 2\}$, let $\{\tilde{\mathbf{y}}_\ell^{(k)}\}_{k \geq 0}$ be the sequence defined by the power iterations $\tilde{\mathbf{y}}_\ell^{(k)} = [\mathbf{I}_n - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)]^k \mathbf{y}_\ell^{(0)}$. Theorem 4 shows that $\{\tilde{\mathbf{y}}_\ell^{(k)}\}_{k \geq 0}$ well approximates the t-SNE iterations $\{\mathbf{y}_\ell^{(k)}\}_{k \geq 0}$ in the large sample limit. The sequence $\{\tilde{\mathbf{y}}_\ell^{(k)}\}_{k \geq 0}$ admits the updating equation

$$\tilde{\mathbf{y}}_\ell^{(k+1)} = \tilde{\mathbf{y}}_\ell^{(k)} - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)\tilde{\mathbf{y}}_\ell^{(k)}, \quad k \geq 0, \quad (20)$$

with an initial value $\tilde{\mathbf{y}}_\ell^{(0)} = \mathbf{y}_\ell^{(0)}$. Treating Equation (20) as an auxiliary gradient descent algorithm to the original algorithm (7), a continuous-time analysis can be developed accordingly, which yields interesting insights about the t-SNE iterations $\{\mathbf{y}_\ell^{(k)}\}_{k \geq 0}$.

We begin by modeling $\{\tilde{\mathbf{y}}_\ell^{(k)}\}_{k \geq 0}$ by a smooth curve $\mathbf{Y}_\ell(t)$ with the Ansatz $\tilde{\mathbf{y}}_\ell^{(k)} \approx \mathbf{Y}_\ell(kh)$. Define a step function $\mathbf{y}_{\ell,h}(t) = \tilde{\mathbf{y}}_\ell^{(k)}$ for $kh \leq t < (k+1)h$, and as $h \rightarrow 0$, $\mathbf{y}_{\ell,h}(t)$ approaches $\mathbf{Y}_\ell(t)$ satisfying

$$\dot{\mathbf{Y}}_\ell(t) = \mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)\mathbf{Y}_\ell(t), \quad (21)$$

with the initial value $\mathbf{Y}_\ell(0) = \tilde{\mathbf{y}}_\ell^{(0)} = \mathbf{y}_\ell^{(0)}$. The above first-order differential equation (21) is usually referred as the gradient flow associated with the power iteration sequence $\{\tilde{\mathbf{y}}_\ell^{(k)}\}_{k \geq 0}$, whose limiting behavior can be studied through the step function $\mathbf{y}_{\ell,h}(t)$. The following theorem provides a non-asymptotic uniform upper bound on the deviation of $\mathbf{y}_{\ell,h}(t)$ from $\mathbf{Y}_\ell(t)$ over $t \in [0, T]$ and that of $\tilde{\mathbf{y}}_\ell^{(k)}$ from $\mathbf{Y}_\ell(kh)$ over $k \leq T/h$.

Proposition 8 (Gradient flow) *For $\ell = 1, 2$, and any given $T > 0$, we have*

$$\sup_{t \in [0, T]} \frac{\|\mathbf{y}_{\ell,h}(t) - \mathbf{Y}_\ell(t)\|_2}{\|\mathbf{Y}_\ell(t)\|_2} \leq Th \|\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)\|^2, \quad (22)$$

where $\mathbf{y}_{\ell,h}(t)$ is the continuous-time step process of $\{\tilde{\mathbf{y}}_\ell^{(k)}\}$ generated by (20), and $\mathbf{Y}_\ell(t)$ is the solution to the ordinary differential equation (21). As a consequence, for $t = hk$, if $kh^2 \|\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)\|^2 \rightarrow 0$ as $n \rightarrow \infty$, then for $\ell \in \{1, 2\}$,

$$\lim_{(n,k) \rightarrow \infty} \frac{\|\tilde{\mathbf{y}}_\ell^{(k)} - \mathbf{Y}_\ell(hk)\|_2}{\|\mathbf{y}_\ell^{(0)}\|_2} = 0. \quad (23)$$

Combining Theorems 4 and 8, we obtain the approximation $\mathbf{y}_\ell^{(k)} \approx \mathbf{Y}_\ell(kh)$ over a range of $k \geq 0$, for properly chosen parameters (α, h, k) and initialization. Consequently, the properties of the solution path $\mathbf{Y}_\ell(t)$ may provide important insights on the behavior of the t-SNE iterations at the early exaggeration stage. We start by stating the following proposition concerning the explicit expression of $\mathbf{Y}_\ell(t)$.

Proposition 9 (Solution path) *For $\ell \in \{1, 2\}$, the first-order linear differential equation (21) with initial value $\mathbf{Y}_\ell(0) = \mathbf{y}_\ell^{(0)}$ has the unique solution $\mathbf{Y}_\ell(t) = \exp(-t\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n))\mathbf{y}_\ell^{(0)}$, where $\exp(\cdot)$ is the matrix exponential defined as $\exp(\mathbf{A}) = \sum_{k=0}^{\infty} \frac{1}{k!} \mathbf{A}^k$. In particular,*

suppose $\mathbf{L}(\mathbf{P})$ have the eigendecomposition $\mathbf{L}(\mathbf{P}) = \sum_{i=1}^n \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$ where $0 = \lambda_1 \leq \dots \leq \lambda_n$ and $\mathbf{u}_1 = n^{-1/2} \mathbf{1}_n$. Then we also have

$$\mathbf{Y}_\ell(t) = (\mathbf{u}_1^\top \mathbf{y}_\ell^{(0)}) \mathbf{u}_1 + \sum_{i=2}^n e^{-t(\alpha\lambda_i - \frac{1}{n-1})} (\mathbf{u}_i^\top \mathbf{y}_\ell^{(0)}) \mathbf{u}_i. \quad (24)$$

Several important observations about the solution path $\mathbf{Y}_\ell(t)$ can be made. Firstly, by Proposition 9, for $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ where $m \in [n]$ is the largest integer such that $\alpha\lambda_m \leq \frac{1}{n-1}$, we have

$$\lim_{t \rightarrow \infty} \mathbf{Y}_\ell(t) \in \text{span}(\{\mathbf{u}_1, \dots, \mathbf{u}_m\}). \quad (25)$$

This can be treated as a continuous version of the limiting behavior of the power iterations obtained in Theorem 5: under the conditions of Theorem 5, we have $\alpha\lambda_m \leq \frac{1}{n-1}$ for all $m \in [R]$ but $\alpha\lambda_{R+1} > \frac{1}{n-1}$, so that (25) implies that $\mathbf{Y}_\ell(t)$ converges to the null space of $\mathbf{L}(\mathbf{P})$. Secondly, as long as $t = O(n)$, by orthogonality of $\{\mathbf{u}_i\}$, we have

$$\|\mathbf{Y}_\ell(t)\|_2^2 \lesssim \sum_{i=1}^n e^{-2t\alpha\lambda_i} (\mathbf{u}_i^\top \mathbf{y}_\ell^{(0)})^2. \quad (26)$$

The right-hand side is monotonically nonincreasing in t . Hence, the average distance of the rows in $(\mathbf{Y}_1(hk), \mathbf{Y}_2(hk))$ to the origin remains non-expansive over the iterations and is bounded up to a constant by that of $(\mathbf{Y}_1(0), \mathbf{Y}_2(0))$. This result echos Proposition 3 based on the discrete-time analysis.

The third and more insightful observation from (24) is its implications on the finite-time behavior of the original t-SNE sequence $\{\mathbf{y}_\ell^{(k)}\}_{k \geq 1}$, which complements our discrete-time analysis. Specifically, for finite $t > 0$, the coefficient of the i -th basis \mathbf{u}_i in $\mathbf{Y}_\ell(t)$ is proportional to $e^{-t\alpha\lambda_i}$, which is nonincreasing in λ_i . Consequently, (24) implies that, in the early steps of the iterations, the t-SNE algorithm imposes an *implicit regularization* effect on the low-dimensional map $\{y_i^{(k)}\}_{1 \leq i \leq n}$, in the sense that

$$\mathbf{y}_\ell^{(k)} \approx n^{-1} (\mathbf{1}_n^\top \mathbf{y}_\ell^{(0)}) \mathbf{1}_n + \sum_{i=2}^n e^{-kh(\alpha\lambda_i - \frac{1}{n-1})} (\mathbf{u}_i^\top \mathbf{y}_\ell^{(0)}) \mathbf{u}_i. \quad (27)$$

Comparing to the limit (25) or (16), during the early steps of the iterations, $\mathbf{y}_\ell^{(k)}$ is regularized as a conical sum of all the eigenvector basis $\{\mathbf{u}_i\}_{1 \leq i \leq n}$, with larger weights on the eigenvectors \mathbf{u}_i corresponding to the smaller eigenvalues of $\mathbf{L}(\mathbf{P})$, and smaller weights on those corresponding to the larger eigenvalues of $\mathbf{L}(\mathbf{P})$. As the iteration goes, the contributions from the less informative eigenvectors with larger eigenvalues λ_i such that $\alpha\lambda_i > \frac{1}{n-1}$ decrease exponentially in k , whereas the contributions from the more informative eigenvectors with smaller eigenvalues λ_i such that $\alpha\lambda_i < \frac{1}{n-1}$ increase with k .

Importantly, the inclusion of all the eigenvectors helps to better summarize the cluster information in the original data and to avoid convergence to the trivial eigenvector $n^{-1/2} \mathbf{1}_n$. Indeed, the convergence (25) by itself may not lead to a cluster structure in the limit, as in many applications with weakly clustered data, the graph corresponding to \mathbf{P} may be simply connected under finite samples, so that the null space $\text{span}(\{\mathbf{u}_1, \dots, \mathbf{u}_m\})$ is effectively the one-dimensional space spanned by $n^{-1/2} \mathbf{1}$ alone. However, as our next theorem

shows, the benefit of the implicit regularization, brought about by stopping early at the exaggeration stage, can be seen in the creation of desirable clusters in $\{y_i^{(k)}\}_{1 \leq i \leq n}$ for weakly clustered data with approximately block-structured \mathbf{P} . In particular, we make the following assumptions analogous to (T1.D) and (T2.D) in the discrete-time analysis.

(T1.C) the parameters (α, h, t) satisfy $\alpha \gg [n\lambda_{R+1}(\mathbf{L}(\mathbf{P}))]^{-1}$ and $t = o(n)$ as $n \rightarrow \infty$;

(T2.C) there exists a symmetric and well-conditioned matrix $\mathbf{P}^* \in \mathbb{R}^{n \times n}$ such that $\lambda_{R+1}(\mathbf{L}(\mathbf{P}^*)) \gg \max\{(t\alpha)^{-1}, \|\mathbf{L}(\mathbf{P}^* - \mathbf{P})\|\}$, and $t\alpha\|\mathbf{L}(\mathbf{P}^* - \mathbf{P})\| = o(1)$ as $n \rightarrow \infty$.

Similar to the previous conditions, Conditions (T1.C) and (T2.C) concerns the approximate block structure of \mathbf{P} , and ensures sufficient exaggeration and early stopping of the iterations.

Theorem 10 (Implicit regularization, clustering and early stopping) *Under Conditions (I1), (T1.C) and (T2.C), let $\mathbf{U}_0 \in O(n, R)$ such that its columns span the null space of \mathbf{P}^* . Then, we have*

$$\lim_{n \rightarrow \infty} \frac{\|\mathbf{Y}_\ell(t) - \mathbf{U}_0 \mathbf{U}_0^\top \mathbf{Y}_\ell(t)\|_2}{\|\mathbf{Y}_\ell(0)\|_2} = 0, \quad \ell \in \{1, 2\}, \quad (28)$$

and, for \mathbf{z}_ℓ defined in Theorem 7, there exists a permutation matrix $O \in \mathbb{R}^{n \times n}$ such that

$$\lim_{n \rightarrow \infty} \frac{\|\mathbf{Y}_\ell(t) - O \mathbf{z}_\ell\|_2}{\|\mathbf{Y}_\ell(0)\|_2} = 0, \quad \ell \in \{1, 2\}. \quad (29)$$

An immediate consequence of the above theorem is the following corollary, which arrives at the same conclusion as Theorem 7 through a different route.

Corollary 11 *Suppose the conditions of Theorems 4 and 10 hold with $t = hk$, and $k\alpha^2 h^2 \|\mathbf{L}(\mathbf{P} - \mathbf{H}_n)\|^2 \rightarrow 0$. Then the conclusion of Theorem 7 holds.*

The above theorems provide a deeper theoretical explanation of the need of stopping early at the exaggeration stage. On the one hand, the number of iterations should be sufficiently large so that $\{y_i^{(k)}\}_{1 \leq i \leq n}$ moves away from the initialization and is sufficiently close to a subspace where the underlying cluster information is properly stored. On the other hand, the iterations should be also stopped early for weakly clustered data to avoid “overshooting,” that is, convergence to the null space of the superficial Laplacian $\mathbf{L}(\mathbf{P})$, which may only include the non-informative trivial eigenvector $n^{-1/2} \mathbf{1}_n$ (Figure 5 right).

3. Analysis of the Embedding Stage

We have shown in Section 2 that the iterations in the early exaggeration stage essentially create clusters in the low-dimensional map $\{y_i^{(k)}\}_{1 \leq i \leq n}$, that agree with those underlying $\{X_i\}_{1 \leq i \leq n}$. However, as indicated by Proposition 3, so far the low-dimensional map is concentrated and localized around zero, which may not be ideal for visualization purpose. In addition, by Theorem 7, much information about $\{X_i\}_{1 \leq i \leq n}$ other than its cluster membership are not reflected by the low-dimensional map. In this section, we show that, after

transition to the embedding stage, the t-SNE iterations (5) essentially start by amplifying and refining the existing cluster structures in the low-dimensional map and then aim at a proper embedding of the original data.

We show that, starting from the embedding stage, the diameter of the low-dimensional map $\{y_i^{(k)}\}_{1 \leq i \leq n}$ grows fast and they move in clusters as inherited from the early exaggeration stage. Importantly, over the iterations, the elements of $\{y_i^{(k)}\}_{1 \leq i \leq n}$ belonging to different clusters would in general move away from each other, resulting to an enlarged visualization with more separated clusters. We refer these iteration steps presenting such a drastically expansive, and intercluster-repulsive behavior of $\{y_i^{(k)}\}_{1 \leq i \leq n}$ as the *amplification phase* of the embedding stage. We also show that, after a certain point, the conditions for the fast expansion phenomenon no longer hold, which likely causes the change of behavior, into a new phase which we refer as the *stabilization phase*. This is in line with the empirical observation (Figure 3) that, after a few fast expansive iterations in the amplification phase, the speed of expansion/amplification gradually reduces towards zero, and in the stabilization phase the diameter only increases very slowly with the iterations.

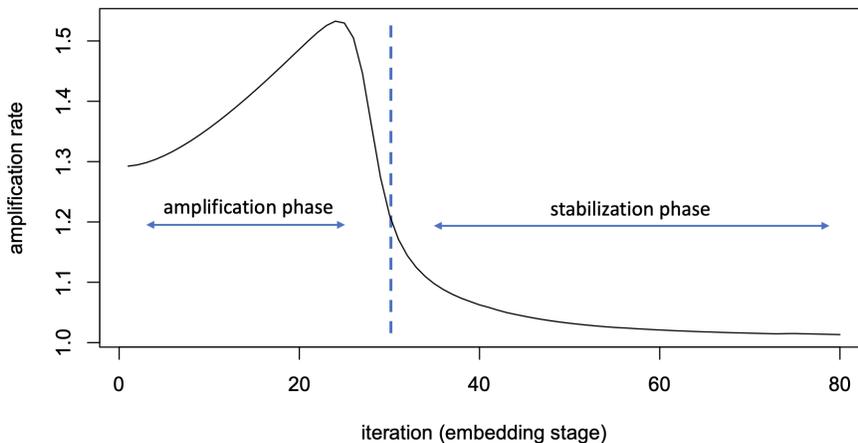


Figure 3: An illustration of the two phase of the embedding stage based on the 1600 MNIST samples described in Section 5. The iterations are counted from the beginning of the embedding stage, and the amplification rate is the ratio between the diameters of two consecutive embeddings.

Recall that the updating equation at the embedding stage is

$$y_i^{(k+1)} = y_i^{(k)} + h' \sum_{j \neq i} S_{ij}^{(k)} (y_j^{(k)} - y_i^{(k)}), \quad (30)$$

where h' is the step size that may not be identical to the one in the early exaggeration stage. To understand the behavior of t-SNE at this stage, we start with the following proposition characterizing the matrix $\mathbf{S}^{(k)} = (S_{ij}^{(k)})_{1 \leq i, j \leq n}$ over the amplification phase.

Proposition 12 *For any integer k , if $\text{diam}(\{y_i^{(k)}\}_{1 \leq i \leq n}) = o(1)$ as $n \rightarrow \infty$, then, for any $i, j \in [n]$ such that $i \neq j$,*

1. if $\lim_{n \rightarrow \infty} n^2 p_{ij} = 0$, it holds that $S_{ij}^{(k)} = -\frac{1+O(\eta^{(k)})}{n(n-1)}$ as $n \rightarrow \infty$; and
2. if $\lim_{n \rightarrow \infty} n^2 p_{ij} \geq c$ for some constant $c > 0$, it holds that $|S_{ij}^{(k)}| \asymp p_{ij}$ as $n \rightarrow \infty$.

Roughly speaking, Proposition 12 says that over the amplification phase, the matrix $\mathbf{S}^{(k)} = (S_{ij}^{(k)})_{1 \leq i, j \leq n}$ essentially has two types of entries, determined by the magnitude of the corresponding entries in \mathbf{P} . Specifically, $S_{ij}^{(k)}$ is negative with magnitude n^{-2} if p_{ij} is much smaller than n^{-2} , and otherwise $S_{ij}^{(k)}$ has the same magnitude as p_{ij} . This observation leads to the next theorem, which provides important insights on the updating equation (30) by partitioning the contributions of $\{y_i^{(k)}\}_{1 \leq i \leq n}$ to an updated $y_i^{(k+1)}$ into a few major components, each corresponding to a distinct cluster in the original data. To this end, we consider again the similarity matrix \mathbf{P} that is only approximately well-conditioned, as characterized by the following assumption.

(T2.E) There exists a symmetric and well-conditioned matrix $\mathbf{P}^* \in \mathbb{R}^{n \times n}$ satisfying (T2.D), $\lim_{n \rightarrow \infty} n^2 \|\mathbf{P} - \mathbf{P}^*\|_\infty = 0$, and $\lim_{n \rightarrow \infty} \frac{nr}{n} \rightarrow \gamma_r \in (0, 1)$ for each $r \in [R]$.

The existence of well-conditioned \mathbf{P}^* induces an equivalence class on $[n]$ characterizing the underlying cluster membership. Specifically, for any $i, j \in [n]$, we denote $i \sim j$ if and only if the i -th node and the j -th node belong to the same graph component. Therefore, we have the partition $[n] = \cup_{r \in [R]} H_r$ for mutually disjoint sets $\{H_r\}_{1 \leq r \leq R}$, with H_r corresponding to the r -th equivalence class.

Next, we make assumptions on the initialization and parameters (α, h, K_0) in the early exaggeration stage, where $K_0 = K_0(n) \rightarrow \infty$ is the total number of iterations in that stage. Specifically, we assume

(I3) the initialization is chosen such that $\|\mathbf{y}_1^{(0)}\|_2 \asymp \|\mathbf{y}_2^{(0)}\|_2$, $\max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_\infty = o(n^{-1/2})$ as $n \rightarrow \infty$, and there exists some constant $C > 1$ such that \mathbf{z}_ℓ defined in Theorem 7 satisfies $C^{-1} \leq n|z_{\ell i} - z_{\ell j}| / \|\mathbf{y}_\ell^{(0)}\|_2 \leq C$ for any $i, j \in [R]$ such that $i \neq j$, and $\ell \in \{1, 2\}$; and

(T1.E) the parameters (α, h, K_0) in (6) satisfy (T1.D), and, for $R_n = (1 - \kappa)^{K_0} + hK_0[(\alpha n \|\mathbf{P}\|_\infty + 1/n) \cdot \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_\infty^2 + \alpha \|\mathbf{L}(\mathbf{P}^* - \mathbf{P})\|]$, we have $nR_n(1 + n^2 \|\mathbf{P}^*\|_\infty) = o(1)$ as $n \rightarrow \infty$.

Condition (I3) is mild as it can be satisfied with high probability by a straightforward random initialization procedure, to be presented shortly. Condition (T1.E) is analogous to but slightly stronger than (T1.D), by requiring a smaller cumulative approximation error R_n between $\mathbf{L}(\mathbf{S}_\alpha^{(k)})$ and $\mathbf{L}(\alpha \mathbf{P}^* - \mathbf{H}_n)$, that is, more distinct clusters in $\{X_i\}_{1 \leq i \leq n}$.

Finally, for the parameters (h', K_1) in embedding stage, where $K_1 = K_1(n)$ is the number of iterations within the amplification phase, we make the following assumption that controls the cumulative approximation error in $\mathbf{S}^{(k)}$ as suggested by Proposition 12.

(T3.E) $\text{diam}(\{y_i^{(K_0+K_1)}\}_{1 \leq i \leq n}) = o(1)$, and the parameter h' in (30) satisfies $K_1 h' (n \|\mathbf{P}^*\|_\infty + 1/n) = O(1)$ as $n \rightarrow \infty$.

Theorem 13 (Intercluster repulsion) *Under Conditions (T1.E) (T2.E) (T3.E) and (I3), for each $K_0 \leq k \leq K_0 + K_1$ and any $i \in [n]$, we have*

$$y_i^{(k+1)} = y_i^{(k)} + \sum_{r \in [R] \setminus r_0} f_{ir}^{(k)} + \epsilon_i^{(k)}, \quad (31)$$

where $r_0 \in [R]$ such that $i \in H_{r_0}$, $\lim_{n \rightarrow \infty} \|\epsilon_i^{(k)}\|_2 / \|f_{ir}^{(k)}\|_2 = 0$ for all $r \in [R] \setminus r_0$, and

$$f_{ir}^{(k)} = \frac{h'|H_r|}{n(n-1)} \left(y_i^{(k)} - \frac{1}{|H_r|} \sum_{j \in H_r} y_j^{(k)} \right) \in \mathbb{R}^2.$$

In addition, we have

$$\sup_{K_0 \leq k \leq K_0 + K_1} \max_{(i,j): i \sim j} \|y_i^{(k)} - y_j^{(k)}\|_2 \ll n^{-1} (\|\mathbf{y}_1^{(0)}\|_2 + \|\mathbf{y}_2^{(0)}\|_2), \quad (32)$$

and

$$\inf_{K_0 \leq k \leq K_0 + K_1} \min_{(i,j): i \not\sim j} \|y_i^{(k)} - y_j^{(k)}\|_2 \gtrsim n^{-1} (\|\mathbf{y}_1^{(0)}\|_2 + \|\mathbf{y}_2^{(0)}\|_2). \quad (33)$$

A few remarks about the above theorem are in order. Firstly, Conditions (T1.E) (T2.E) and (I3) concerning the initialization, parameter selection and the number of iterations in the early exaggerations are not only compatible but also sufficient for the previous results, including Theorems 4 and 7. This suggests the above intercluster repulsive phenomenon at the embedding stage actually relies on the properties of the outputs from the early exaggeration stage, again yielding the necessity of the early exaggeration, or equivalent techniques. Secondly, as indicated by the next theorem, Condition (I3) on the initialization can be satisfied by the following simple local random initialization procedure.

Theorem 14 (Random initialization) *For any sequence $\sigma_n \rightarrow 0$ as $n \rightarrow \infty$, let $\mathbf{y}_\ell^{(0)} = \sigma_n \mathbf{g}_\ell / \|\mathbf{g}_\ell\|_2$, where $\mathbf{g}_\ell \in \mathbb{R}^n$ for $\ell \in [2]$ is independently generated from a standard multivariate normal distribution. Then $\{y_i^{(0)}\}_{1 \leq i \leq n}$ satisfies Condition (I3) with probability at least $1 - \delta$ for some sufficiently small constant $\delta > 0$.*

Thirdly, the above theorem provides a precise characterization of the kinematics of each y_i^k during the iterations, and its reliance on the data points $\{y_i^{(k-1)}\}$ in the previous step, as well as the cluster structure inherited from the early exaggeration stage. Specifically, $f_{ir}^{(k)}$ summarizes the contributions from the points $\{y_i^{(k)}\}_{i \in H_r}$ in the r -th cluster to the new point $y_i^{(k+1)}$. The theorem implies that, at the amplification phase, the behavior of $\{y_i^{(k)}\}_{1 \leq i \leq n}$ is mainly driven by the relative positions of the R clusters produced in the early exaggeration stage: for each point, a vector sum of the *repulsive forces* coming from all the other clusters at their current positions determine the direction and distance of its movement of each point in this iteration (Figure 2). As a consequence, after each iteration, the diameter of $\{y_i^{(k)}\}_{1 \leq i \leq n}$ would increase, till the end of the amplification phase, that is, when Condition (T3.E), or more specifically, $\text{diam}(\{y_i^{(k)}\}_{1 \leq i \leq n}) = o(1)$ no longer holds. This process improves the visualization quality by making the clusters more distinct and separated (Figure 1 with $k = 40$ and 80).

Our next result confirms the intuition that the diameter of $\{y_i^{(k)}\}_{1 \leq i \leq n}$ is bound to increase after each iteration in the amplification phase of the embedding stage.

Theorem 15 (Expansion) *Suppose the conditions of Theorem 13 hold. If in addition $\|\mathbf{P}^*\|_\infty \lesssim n^{-2}$, then for any $k \in \{K_0, K_0 + 1, \dots, K_1\}$, we have*

$$\text{diam}(\{y_i^{(k+1)}\}_{1 \leq i \leq n}) > \text{diam}(\{y_i^{(k)}\}_{1 \leq i \leq n}), \quad (34)$$

where $\text{diam}(\{y_i^{(k+1)}\}_{1 \leq i \leq n}) - \text{diam}(\{y_i^{(k)}\}_{1 \leq i \leq n}) \gtrsim \frac{h'}{n^2} \min_{\ell=1,2} \|\mathbf{y}_\ell^{(0)}\|_2$.

Once the diameter of $\{y_i^{(k)}\}_{1 \leq i \leq n}$ exceeds certain threshold, that is, when $\text{diam}(\{y_i^{(k)}\}_{1 \leq i \leq n})$ is at least of constant order, we arrive at the final stabilization phase. In this phase, the condition of Proposition 12 is violated, and, unlike what is claimed in part one of Proposition 12, the entries of the matrix $\mathbf{S}^{(k)}$ corresponding to the smaller entries in \mathbf{P} , that is, p_{ij} 's with $p_{ij} \ll n^{-2}$, no longer remain an almost constant value $1/n(n-1)$. In particular, the sign of $S_{ij}^{(k)}$ would generally rely on the relative magnitudes between p_{ij} and q_{ij} .

We rewrite (30) as

$$y_i^{(k+1)} = y_i^{(k)} + h' \sum_{j \neq i} \frac{p_{ij} - q_{ij}^{(k)}}{1 + d_{ij}^{(k)}} (y_j^{(k)} - y_i^{(k)}). \quad (35)$$

In the stabilization phase, the new position for $y_i^{(k+1)}$ is determined by the starting point $y_i^{(k)}$, and the averaged contributions from each of the other data points $\{y_j^{(k)}\}_{j \neq i}$. The contribution from $y_j^{(k)}$ to $y_i^{(k+1)}$ is either in or against the direction of $(y_j^{(k)} - y_i^{(k)})$, depending on $\text{sign}(p_{ij} - q_{ij})$. If $\text{sign}(p_{ij} - q_{ij}) = -1$, or, the similarity between $y_i^{(k)}$ and $y_j^{(k)}$ as measured by q_{ij} is greater than the similarity between X_i and X_j as measured by p_{ij} , the contribution from $y_j^{(k)}$ to $y_i^{(k+1)}$ is in the direction of $y_i^{(k)} - y_j^{(k)}$, resulting to a repulsive force that enlarges the distance between $y_i^{(k+1)}$ and $y_j^{(k+1)}$ after the iteration. Similarly, if $\text{sign}(p_{ij} - q_{ij}) = 1$, it means the similarity between $y_i^{(k)}$ and $y_j^{(k)}$ is smaller than their counterparts in $\{X_i\}_{1 \leq i \leq n}$, so the contribution $y_j^{(k)}$ to $y_i^{(k+1)}$ is in the opposite direction $y_j^{(k)} - y_i^{(k)}$, resulting to an attractive force that reduces the distance between $y_i^{(k+1)}$ and $y_j^{(k+1)}$ after the iteration. The iterations over the stabilization phase aim to locally adjust the relative positions of the low-dimensional map to make the final visualization more reliable and faithful.

Remark 16 In practice, expansion and repulsion effects help make clusters identified from the early exaggeration step more salient in the final visualization, and possibly more informative in terms of the local structures within the clusters. This is especially helpful if two clusters are positioned too close to each other at the end of the early exaggeration stage, as an artifact of the random initialization (e.g., the middle column of Figure 5). Moreover, the intercluster repulsion phenomenon explains the occasional appearance of false clusters in the t-SNE visualization (Kobak and Linderman, 2021). Specifically, our theory indicates that false clustering may appear due to an incidental combination of overlapped clusters from the early exaggeration stage with random initialization, and the intercluster repulsion from the embedding stage (Figures 5 and 6). This leads to our third general advice on practice at the end of Section 1.2.

4. Application I: Visualizing Model-Based Clustered Data

In the previous sections, we established the theoretical properties for the basic t-SNE algorithm under general conditions on the parameters $(\alpha, h, h', K_1, K_2)$, the initialization, and the similarity matrix \mathbf{P} constructed from the original data. In this section, we apply our general theory in two concrete examples of clustered data, one generated from a Gaussian mixture model and another from a noisy nested sphere model.

4.1 Gaussian Mixture Model

Consider the Gaussian mixture model

$$X_i | z_i = r \sim N(\mu_r, \Sigma), \quad z_i \stackrel{i.i.d.}{\sim} \text{Multinomial}(\pi_1, \dots, \pi_R), \quad \text{for } i \in [n], \quad (36)$$

where $\mu_r \in \mathbb{R}^p$ and $\sum_{r=1}^R \pi_r = 1$. We make the following assumptions.

- (C1) The mixing proportions $\{\pi_r\}_{1 \leq r \leq R}$ satisfy $\min_r \pi_r \geq c > 0$.
- (C2) There exists some large constant $C' > 0$ such that $\rho^2 = \min_{1 \leq j \neq k \leq R} \|\mu_j - \mu_k\|_2^2 \geq C' \max\{p, \log n\}$.
- (C3) There exists some constant $C > 0$ such that the population covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$ satisfies $C^{-1} \leq \lambda_1(\Sigma) \leq \lambda_p(\Sigma) \leq C$ and $\text{tr}(\Sigma)/p \leq C$.

Under the above Gaussian mixture model, we obtain the following corollary that provides the conditions for the theoretical results presented in the previous sections.

Corollary 17 *Suppose Conditions (C1) (C2) and (C3) hold, and $\tau_i^2 \asymp \max\{p, \log n\}$. If $\alpha \gg 1$, $K_0 h = o(n)$, $h \alpha \asymp n$, $1 \ll K_0 \ll \exp\{\frac{\rho^2}{\max\{p, \log n\}}\}$ and $K_0 h \alpha \sigma_n^2 \log n = o(n^2)$, then Conditions (T1.D) and (T2.D) hold. If in addition $\log n \ll K_0 \ll n^{-1} \exp\{\frac{\rho^2}{\max\{p, \log n\}}\}$, $K_0 h \alpha \sigma_n^2 \log n = o(n)$, and $K_1 h' = O(n)$, then Conditions (T1.E) (T2.E) and (T3.E) hold.*

As a consequence, suitable choices of the tuning parameters $(\alpha, h, h', K_0, K_1)$ under the Gaussian mixture model can be determined efficiently. For example, if $\rho^2 \gtrsim \log n \cdot \max\{p, \log n\}$, one could choose $K_0 = \lfloor (\log n)^2 \rfloor$, $\sigma_n = (\log n)^{-2}$, $h = h' = n^\delta$ and $\alpha = n^{1-\delta}$ for any constant $\delta \in (0, 1)$. By Corollary 17, Conditions (T1.D) and (T2.D) hold, so the conclusions of Theorem 7 follows for $k = K_0$; meanwhile, Conditions (T1.E) (T2.E) and (T3.E) also hold, so the conclusions of Theorem 13 hold for each K_1 with $K_0 \leq K_1 \leq n^{1-\delta}$. Note that the above results apply to both low-dimensional settings where $p = o(n)$ and high-dimensional settings where $p \gtrsim n$.

To demonstrate the effectiveness of the theoretical guidance, we generate $n = 1500$ samples of dimension $p = 100$ from a Gaussian mixture model with $r = 6$, $\rho^2 = p$, $\Sigma = \mathbf{I}_p$, and the cluster proportion vector $(0.1, 0.1, 0.1, 0.15, 0.25, 0.3)$. We use the above tuning parameters with various $\delta \in \{1/2, 1/3\}$ and `perplexity=30` (default). The t-SNE embeddings at the end of the early exaggeration stage $k = K_0 = \lfloor (\log n)^2 \rfloor = 53$ and at $k = 1000$ are included in Figure 4 below and Figure 7 in Appendix F, confirming the theoretical predictions. Moreover, Figure 8 in Appendix F shows that when the above separation condition (C2) is slightly violated (e.g., $\rho^2 = p^{4/5}$), t-SNE is still able to visualize clusters, which demonstrates the robustness of t-SNE with respect to the separation condition.

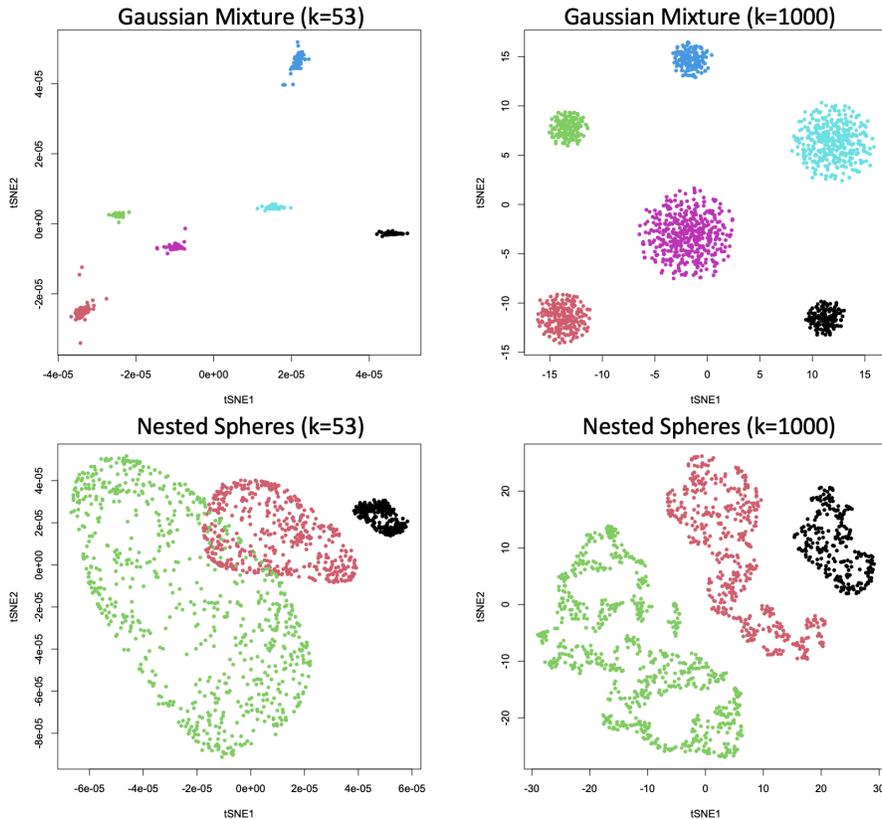


Figure 4: t-SNE visualizations of the model-generated samples as described in Section 4, using the theory-guided tuning parameters with $\delta = 1/3$ (see Figure 7 for similar results with $\delta = 1/2$). The left column shows outputs from the early exaggeration stage, whereas the right column are the corresponding final embeddings.

Remark 18 Arora et al. (2018) analyzed the early exaggeration stage of t-SNE based on a slightly different theoretical framework under the Gaussian mixture model with a mean separation condition $\rho \gtrsim p^{1/4}$, and under the mixture model of log-concave distributions with a separation condition $\rho \gtrsim p^{5/12}$. Compared to these results, our separation condition $\rho \gtrsim p^{1/2}$ in (C2) is strong, and it is unclear to us if such a restriction is intrinsic to our theoretical framework or an artifact from our proof strategy. Nevertheless, nailing down the sharp information threshold for t-SNE visualization is an important and fundamental problem – we plan to have a more systematic treatment of this in a subsequent work.

4.2 Noisy Nested Sphere Model

Consider the model of nested spheres with radial noise (Amini and Razaee, 2021), where for $i \in [n]$, we have

$$X_i = \mu_i + \frac{\mu_i}{\|\mu_i\|_2} \xi_i, \quad \xi_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2) \quad (37)$$

and

$$\mu_i | z_i = r \sim P_k, \quad z_i \stackrel{i.i.d.}{\sim} \text{Multinomial}(\pi_1, \dots, \pi_R), \quad (38)$$

with $\sum_{r=1}^R \pi_r = 1$ and $\{P_k\}$ being uniform distributions on nested spheres in \mathbb{R}^p of various radii $\rho_{\min} = \rho_1 < \rho_2 < \dots < \rho_R = \rho_{\max}$. We make the following assumptions concerning the separation distances between the underlying nested spheres.

(C4) There exists some γ such that $\max\{n^{-1}, \sigma^2 \rho_{\min}^{-2}\} \log n \ll \gamma \ll 1$ and $\max_{r \in [R-1]} \frac{\rho_r}{\rho_{r+1}} \ll 1 - C\sqrt{\gamma \log \gamma}$ for some sufficiently large constant $C > 0$.

(C5) There exists some small constant $c > 0$ such that $c \min |\rho_{r+1} - \rho_r| \geq \sigma \sqrt{\log n}$.

In Condition (C4), the separation distance is characterized by the ratio ρ_r/ρ_{r+1} whereas in Condition (C5) the distance is characterized by the difference $\rho_{r+1} - \rho_r$. The following corollary provides a sufficient condition for the results presented in the previous sections.

Corollary 19 *Suppose Assumptions (C1) (C4) and (C5) hold, and $\tau_i^2 \asymp \gamma \rho_{z_i}^2$. If $K_0 h = o(n)$, $\alpha h = O(\gamma n)$, $h \alpha \lambda_{R+1}(\mathbf{L}(\mathbf{P}^*)) \geq \kappa$ for some constant $\kappa \in (0, 1)$, $K_0 \gg 1$, $K_0 h (\alpha/\gamma + 1) \sigma_n^2 \log n = o(n^2)$, and $\log \frac{K_0 h \alpha}{n} \ll \gamma^{-1} (1 - \max_{r \in [R-1]} \frac{\rho_r}{\rho_{r+1}})^2 + \log \gamma$, then Conditions (T1.D) and (T2.D) hold. If in addition $K_0 \gg \log n$, $K_0 h (\alpha/\gamma + 1) \sigma_n^2 \log n = o(n)$, $\log K_0 h \alpha \ll \gamma^{-1} (1 - \max_{r \in [R-1]} \frac{\rho_r}{\rho_{r+1}})^2 + \log \gamma$, and $K_1 h' = O(\gamma n)$, then Conditions (T1.E) (T2.E) and (T3.E) hold.*

Again, suitable choices of the tuning parameters $(\alpha, h, h', K_0, K_1)$ under the noisy nested sphere model can be determined efficiently. For example, let's consider the case where $\rho_{r+1} - \rho_r = \Delta$ for all $r \in [R-1]$. Specifically, suppose there exists some small constant $c > 0$ such that $\Delta \geq c \rho_R$, and that $\gamma = c(\log n)^{-1}$ satisfies (C4) and $\lambda_{R+1}(\mathbf{L}(\mathbf{P}^*)) \gtrsim \frac{1}{\gamma n}$ in probability. Then, by Corollary 19, the desired visualization properties such as those in Theorems 7 and 13 would hold with high probability, as long as we choose $K_0 = \lfloor (\log n)^2 \rfloor$, $K_1 \leq n^{1-\delta}/\log n$, $\sigma_n = (\log n)^{-2}$, $h = h' = n^\delta$ and $\alpha = \gamma n^{1-\delta}$ for any constant $\delta \in (0, 1)$. Figures 4 and 7 in the Appendix show the t-SNE embeddings of $n = 1500$ samples of dimension $p = 50$, at the end of the early exaggeration stage $k = K_0 = \lfloor (\log n)^2 \rfloor = 53$ and at $k = 1000$, generated from Model (37) with $r = 3, \sigma = 1, (\rho_1, \rho_2, \rho_3) = (10, 25, 50)$ and cluster proportion $(0.17, 0.33, 0.5)$. For the tuning parameters, the above analytical values with $\gamma = 0.5$ and $\delta \in \{1/3, 1/2\}$ are used. As a result, clusters of three nested spheres are visible in all t-SNE embeddings, confirming our theoretical predictions.

5. Application II: Visualizing Real-World Clustered Data

Finally, we demonstrate our theory by applying t-SNE to the MNIST³ dataset, which contains images of hand-written digits. Specifically, we focus on $n = 4N = 1600$ images of hand-written digits “2,” “4,” “6” and “8,” with each digit having $N = 400$ images. Each image contains 28×28 pixels and was treated as a 784-dimensional vector. Based on our theoretical analysis, we set the tuning parameters

$$\alpha = n^{1-\delta}, \quad h = h' = n^\delta, \quad K_0 = \lfloor (\log n)^2 \rfloor \quad (39)$$

3. <http://yann.lecun.com/exdb/mnist/>

with $\delta = 2/3$. Again, we use the default perplexity ($=30$), leading to an approximate block matrix \mathbf{P} , with block structure corresponding to the cluster membership (Figure 9).

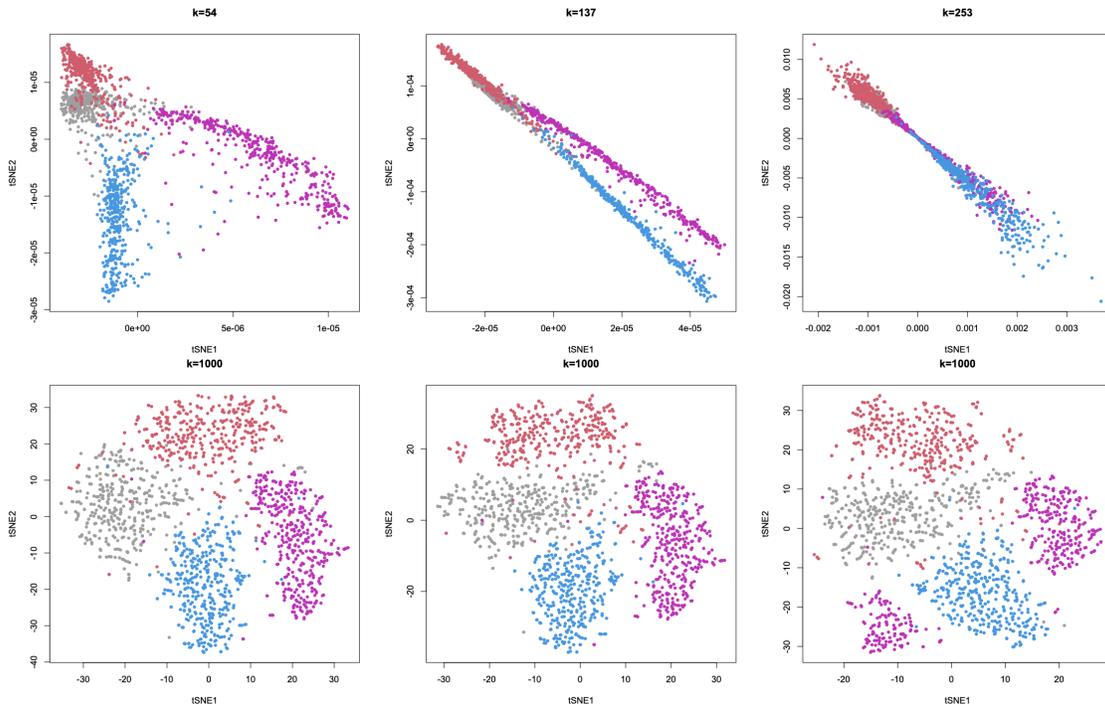


Figure 5: Illustration of t-SNE embeddings of 1600 MNIST samples at the end of embedding stage (bottom row), and their corresponding outputs from the early exaggeration stage (top row). Different columns have identical initializations and tuning parameters, but distinct number of iterations for the early exaggeration stage. The colors of the dots indicate the underlying four clusters.

To demonstrate the necessity of stopping early at the early exaggeration stage, in Figure 5, we show the t-SNE embeddings at the end of embedding stage (bottom row), and their corresponding outputs from the early exaggeration stage (top row). Different columns have identical initialization and tuning parameters, but distinct numbers of iterations for the early exaggeration stage, namely $K_0 = \lfloor (\log n)^2 \rfloor = 54$ (left), $K_0 = \lfloor n^{2/3} \rfloor = 137$ (middle), and $K_0 = \lfloor n^{3/4} \rfloor = 253$ (right). Comparing the top three plots in Figure 5, we can clearly see that when K_0 far exceeds our theory-guided value $\lfloor (\log n)^2 \rfloor$, the cluster patterns is no longer visible, which, in the case of $K_0 = 253$, led to false clustering in the final visualization. Moreover, the middle column of Figure 5 also demonstrates the importance of the embedding stage, especially its underlying intercluster repulsion and expansion effects, as to making the cluster patterns more salient in the final visualization.

Next we assess the effects and artifacts of the random initialization. In Figure 6, we fix all the tuning parameters as in (39) and generate t-SNE visualizations from three different random initializations. Comparing the first two plots, we observe that the relative positions of the clusters vary with the initialization. For example, the purple cluster and red cluster

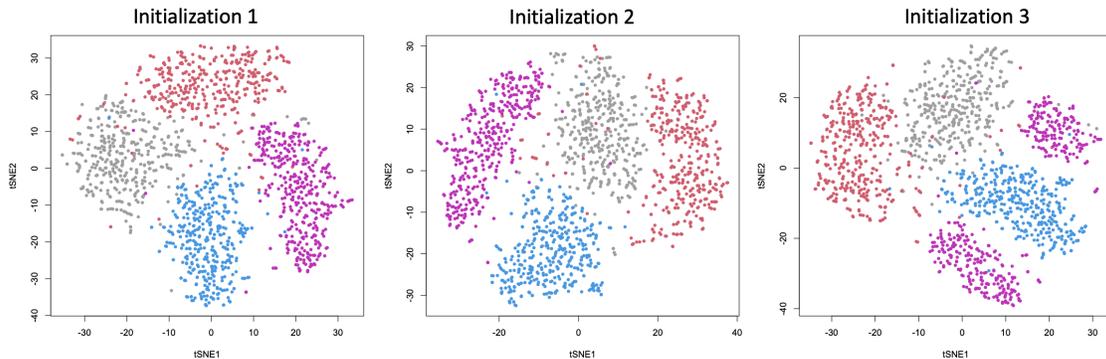


Figure 6: t-SNE visualizations of 1600 MNIST samples based on three different random initializations and identical tuning parameters in (39).

are neighbors in the left panel but not in the middle panel. This echoes our theoretical prediction (discussion after Theorem 7) and justifies our second practical advice in Section 1.2. On the other hand, on the right panel of Figure 6, we find that even with a proper choice of the tuning parameters, false clustering may still appear as an artifact of the random initialization (cf. Remark 16 and the third practical advice in Section 1.2).

Finally, we point out that our theory-guided values for the tuning parameters are flexible, robust and adaptive to the sample size. For example, in Figure 10, we present three more visualizations of $n = 2400$ ($N = 600$ for each digit) MNIST samples, by using the tuning parameters in (39) with various $\delta \in \{1/3, 1/2, 2/3\}$, and an identical random initialization. The cluster patterns are visible and similar in all the cases, showing the effectiveness of our tuning parameters and the insensitivity to the choice of δ .

6. Discussion

The present paper provides theoretical foundations of t-SNE for visualizing clustered data and obtains insights about its theoretical properties and interpretations. We believe that some of the conditions may be relaxed by adopting more advanced technical tools. For example, the current analysis of the early exaggeration stage relies on the well-celebrated Davis-Kahan matrix perturbation inequality (cf. Section B.3), which may be further improved by leveraging advanced results in Random Matrix Theory, such as Benaych-Georges and Nadakuditi (2012) and Bao et al. (2021).

There are still many interesting questions that remain to be explored. For instance, what is the limiting behavior of the low-dimensional map $\{y_i^{(k)}\}_{1 \leq i \leq n}$ towards the end of the embedding stage, after transition to the stabilization phase? How to interpret the local structure within a cluster (DePavia and Steinerberger, 2020; Robinson and Pierce-Hoffman, 2020)? How many iterations are needed for the embedding stage? How to determine the bandwidth $\{\tau_i\}$ in a data-driven and adaptive manner (Ding and Ma, 2022)? The present work is a first step towards answering these important questions.

Moreover, our theoretical framework is generic and can be generalized to study other algorithms that are closely related to or share similar features with t-SNE. For example, in addition to the variants of t-SNE mentioned in Section 1, many dimension reduction and data visualization methods, such as multidimensional scaling (Kruskal, 1978), kernel principal component analysis (Schölkopf et al., 1997), and Laplacian eigenmap (Belkin and Niyogi, 2003), start with a similarity matrix summarizing the pairwise distances within a dataset, and then proceed by either explicitly or implicitly exploiting the spectral properties of the similarity matrix. In this connection, the general ideas behind our theoretical analysis, such as identifying the underlying structured graph and properties of its adjacency or Laplacian matrix (Sections 2.1 and 2.2), studying the gradient flow associated with the discrete algorithm (Section 2.3), and the mechanical/kinematic view of the updating equation (Section 3), can be adopted to uncover the underlying mechanism and the properties of these methods.

It is also interesting to explore the fundamental limit for data visualization and dimension reduction. For example, what are the necessary conditions for the data $\{X_i\}_{1 \leq i \leq n}$ to guarantee the existence of a low-dimensional map $\{y_i\}_{1 \leq i \leq n}$ being a metric embedding of it? Whether t-SNE has to sacrifice some global structures in order to locally embed the data well (Chari et al., 2021)? These problems are left for future investigation.

Acknowledgement

The authors are grateful to the editors and four anonymous referees for their comments and suggestions which have significantly improved the results and presentation of the paper. The research of Tony Cai was supported in part by NSF grant DMS-2015259 and NIH grant R01-GM129781. The research of Rong Ma was supported by Professor David Donoho at Stanford University. This work was partially completed while Rong Ma was a PhD candidate in Biostatistics at the University of Pennsylvania. Rong Ma would like to thank Mingyao Li for introducing the subject, and Michaël Aupetit, David Donoho, Jeyong Lee, Stefan Steinerberger, Yiqiao Zhong and James Zou for helpful discussions.

References

- Arash A Amini and Zahra S Razaee. Concentration of kernel matrices with application to kernel spectral clustering. *The Annals of Statistics*, 49(1):531–556, 2021.
- Sanjeev Arora, Wei Hu, and Pravesh K Kothari. An analysis of the t-SNE algorithm for data visualization. In *Conference on Learning Theory*, pages 1455–1462. PMLR, 2018.
- Sivaraman Balakrishnan, Min Xu, Akshay Krishnamurthy, and Aarti Singh. Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems*, pages 954–962, 2011.
- Zhigang Bao, Xiucan Ding, and Ke Wang. Singular vector and singular subspace distribution for the matrix denoising model. *The Annals of Statistics*, 49(1):370–392, 2021.
- Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

- Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.
- John Charles Butcher. *Numerical methods for ordinary differential equations*. John Wiley & Sons, 2008.
- T Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89, 2018.
- Miguel A Carreira-Perpinán. The elastic embedding algorithm for dimensionality reduction. In *ICML*, volume 10, pages 167–174, 2010.
- Tara Chari, JoeYta Banerjee, and Lior Pachter. The specious art of single-cell genomics. *BioRxiv*, 2021.
- Angelos Chatzimpampas, Rafael M Martins, and Andreas Kerren. t-viSNE: Interactive assessment and interpretation of t-SNE projections. *IEEE Transactions on Visualization and Computer Graphics*, 26(8):2696–2714, 2020.
- Jian Cheng, Haijun Liu, Feng Wang, Hongsheng Li, and Ce Zhu. Silhouette analysis for human action recognition based on supervised temporal t-SNE and incremental learning. *IEEE Transactions on Image Processing*, 24(10):3203–3217, 2015.
- Adela DePavia and Stefan Steinerberger. Spectral clustering revisited: Information hidden in the fiedler vector. *arXiv preprint arXiv:2003.09969*, 2020.
- Xiucui Ding and Rong Ma. Learning low-dimensional nonlinear structures from high-dimensional noisy data: An integral operator approach. *arXiv preprint arXiv:2203.00126*, 2022.
- David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.
- Andrej Gisbrecht, Alexander Schulz, and Barbara Hammer. Parametric nonlinear dimensionality reduction using kernel t-SNE. *Neurocomputing*, 147:71–82, 2015.
- Geoffrey Hinton and Sam T Roweis. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*, volume 15, pages 833–840, 2002.
- Daniel Jiwoong Im, Nakul Verma, and Kristin Branson. Stochastic neighbor embedding under f-divergences. *arXiv preprint arXiv:1811.01247*, 2018.
- Robert A Jacobs. Increased rates of convergence through learning rate adaptation. *Neural Networks*, 1(4):295–307, 1988.
- Dmitry Kobak and Philipp Berens. The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, 10(1):1–14, 2019.

- Dmitry Kobak and George C Linderman. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology*, 39(2):156–157, 2021.
- Joseph B Kruskal. *Multidimensional Scaling*. Number 11. Sage, 1978.
- John A Lee and Michel Verleysen. Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants. *Procedia Computer Science*, 4: 538–547, 2011.
- John A Lee and Michel Verleysen. Two key properties of dimensionality reduction methods. In *2014 IEEE symposium on computational intelligence and data mining (CIDM)*, pages 163–170. IEEE, 2014.
- George C Linderman and Stefan Steinerberger. Clustering with t-SNE, provably. *SIAM Journal on Mathematics of Data Science*, 1(2):313–332, 2019.
- George C Linderman, Manas Rachh, Jeremy G Hoskins, Stefan Steinerberger, and Yuval Kluger. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature Methods*, 16(3):243–245, 2019.
- Anne Marsden. Eigenvalues of the laplacian and their relationship to the connectedness of a graph. *University of Chicago, REU*, 2013.
- Luis Gustavo Nonato and Michael Aupetit. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2650–2673, 2018.
- Florent Olivon, Nicolas Elie, Gwendal Grelier, Fanny Roussi, Marc Litaudon, and David Touboul. Metgem software for the generation of molecular networks based on the t-SNE algorithm. *Analytical Chemistry*, 90(23):13900–13908, 2018.
- Nicola Pezzotti, Boudewijn PF Lelieveldt, Laurens Van Der Maaten, Thomas Höllt, Elmar Eisemann, and Anna Vilanova. Approximated and user steerable tSNE for progressive visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 23(7): 1739–1752, 2016.
- Alexander Platzer. Visualization of snps with t-SNE. *PloS One*, 8(2):e56883, 2013.
- Isaac Robinson and Emma Pierce-Hoffman. Tree-sne: Hierarchical clustering and visualization using t-sne. *arXiv preprint arXiv:2002.05687*, 2020.
- Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.
- Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, pages 583–588. Springer, 1997.
- Uri Shaham and Stefan Steinerberger. Stochastic neighbor embedding separates well-separated clusters. *arXiv preprint arXiv:1702.02670*, 2017.

- Gregor Traven, Gal Matijević, Tomaz Zwitter, M Žerjal, Janez Kos, Martin Asplund, Joss Bland-Hawthorn, Andrew R Casey, Gayandhi De Silva, Kenneth Freeman, et al. The galah survey: classification and diagnostics with t-SNE reduction of spectral information. *The Astrophysical Journal Supplement Series*, 228(2):24, 2017.
- Laurens van der Maaten. Accelerating t-SNE using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014.
- Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.
- Richard S Varga. *Geršgorin and his circles*, volume 36. Springer Science & Business Media, 2010.
- Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, 22: 1–73, 2021.
- Bo Xie, Yang Mu, Dacheng Tao, and Kaiqi Huang. m-SNE: Multiview stochastic neighbor embedding. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(4):1088–1096, 2011.
- Zhirong Yang, Irwin King, Zenglin Xu, and Erkki Oja. Heavy-tailed symmetric stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*, volume 22, pages 2169–2177, 2009.
- Yulan Zhang and Stefan Steinerberger. t-SNE, forceful colorings and mean field limits. *arXiv preprint arXiv:2102.13009*, 2021.

Appendix A. Discrete-Time Analysis of the Early Exaggeration Stage

A.1 Proof of Theorem 2

For simplicity, we ignore the superscript (k) in $y_i^{(k)}, q_{ij}^{(k)}, \eta^{(k)}$ and $S_{ij}^{(k)}(\alpha)$. Since $S_{ij}(\alpha) = \frac{\alpha p_{ij} - q_{ij}}{1 + \|y_i - y_j\|_2^2}$, if we denote $d_{ij} = \|y_i - y_j\|_2^2$, we have

$$\begin{aligned} \left| S_{ij}(\alpha) - \alpha p_{ij} + \frac{1}{n(n-1)} \right| &\leq \left| \alpha p_{ij} - \frac{\alpha p_{ij}}{1 + d_{ij}} \right| + \left| \frac{q_{ij}}{1 + d_{ij}} - \frac{1}{n(n-1)} \right| \\ &= d_{ij} \alpha p_{ij} + \left| \frac{1}{Z(1 + d_{ij})^2} - \frac{1}{n(n-1)} \right|, \end{aligned}$$

where

$$Z = \sum_{i \neq j} (1 + d_{ij})^{-1} = \sum_{i \neq j} \left(1 - \frac{d_{ij}}{1 + d_{ij}} \right) = n(n-1) - \sum_{i \neq j} \frac{d_{ij}}{1 + d_{ij}} \equiv n(n-1) - \Delta.$$

Now since $\Delta \leq n(n-1) \min\{\eta, 1\} = n(n-1)\eta$, we have

$$\begin{aligned} \left| \frac{1}{Z(1 + d_{ij})^2} - \frac{1}{n(n-1)} \right| &= \left| \frac{(1 + d_{ij})^{-2} n(n-1) - n(n-1) + \Delta}{n^2(n-1)^2 - n(n-1)\Delta} \right| \\ &\leq \frac{|[(1 + d_{ij})^{-2} - 1]n(n-1) + \Delta|}{n^2(n-1)^2(1 - \eta)} \leq \frac{2\eta}{n(n-1)(1 - \eta)} \end{aligned} \quad (40)$$

Hence

$$\left| S_{ij}(\alpha) - \alpha p_{ij} + \frac{1}{n(n-1)} \right| \leq \eta \alpha p_{ij} + \frac{2\eta}{n(n-1)(1 - \eta)}.$$

For the second statement, we note that

$$\|\mathbf{H}_n\| \leq \frac{1}{n-1} + \frac{1}{n(n-1)} \lesssim \frac{1}{n}. \quad (41)$$

Then as long as $\|\alpha \mathbf{P}\| \gg \frac{1}{n}$, we have $\|\alpha \mathbf{P} - \mathbf{H}_n\| \geq \|\alpha \mathbf{P}\| - \frac{1}{n} \asymp \|\alpha \mathbf{P}\|$. Therefore, under the condition that $\eta \ll \frac{\|\mathbf{P}\|}{n\|\mathbf{P}\|_\infty} \leq 1$,

$$\frac{\|\mathbf{S}_\alpha - (\alpha \mathbf{P} - \mathbf{H}_n)\|}{\|\alpha \mathbf{P} - \mathbf{H}_n\|} \lesssim \frac{n\|\mathbf{S}_\alpha - (\alpha \mathbf{P} - \mathbf{H}_n)\|_\infty}{\|\alpha \mathbf{P}\|} \lesssim \frac{\alpha n\|\mathbf{P}\|_\infty \eta + \eta/n}{\|\alpha \mathbf{P}\|}$$

Then, the first term $\eta n\|\mathbf{P}\|_\infty/\|\mathbf{P}\| \rightarrow 0$ as $\eta \ll \|\mathbf{P}\|/(n\|\mathbf{P}\|_\infty)$; the second term $\frac{\eta}{n\alpha\|\mathbf{P}\|} \lesssim \frac{1}{n\alpha\|\mathbf{P}\|} \rightarrow 0$ as $\alpha \gg (n\|\mathbf{P}\|)^{-1}$.

A.2 Proof of Proposition 3

Note that $y_{\ell i}^{(k+1)} \leq \|[\mathbf{I} - h\mathbf{L}(\mathbf{S}_\alpha^{(k)})]_i\|_1 \|y_\ell^{(k)}\|_\infty$ for any $k \geq 0$, where

$$\|[\mathbf{I} - h\mathbf{L}(\mathbf{S}_\alpha^{(k)})]_i\|_1 = \left| 1 - h \sum_{j=1}^n S_{ij}^{(k)}(\alpha) \right| + h \sum_{j \neq i} |S_{ij}^{(k)}(\alpha)| \leq 1 + 2h \sum_{j=1}^n |S_{ij}^{(k)}(\alpha)|.$$

For the last term, we have

$$h \sum_{j=1}^n |S_{ij}^{(k)}(\alpha)| \leq hn \|\mathbf{S}_\alpha^{(k)}\|_\infty \leq nh(\alpha \|\mathbf{P}\|_\infty + \|\mathbf{Q}^{(k)}\|_\infty) \leq nh\alpha \|\mathbf{P}\|_\infty + \frac{h(1 + \eta^{(k)})}{n-1}$$

where the last inequality follows from (40), that is, $\|\mathbf{Q}^{(k)}\|_\infty \leq 1/Z$, $Z \geq n(n-1)/(1 + \eta^{(k)})$, so that $\|\mathbf{Q}^{(k)}\|_\infty \leq (1 + \eta^{(k)})/n(n-1)$. Then we have

$$y_{\ell i}^{(k+1)} \leq \left(1 + 2nh\alpha \|\mathbf{P}\|_\infty + \frac{2h(1 + \eta^{(k)})}{n-1}\right) \|\mathbf{y}_\ell^{(k)}\|_\infty,$$

or

$$\|\mathbf{y}_\ell^{(k+1)}\|_\infty \leq \left(1 + 2nh\alpha \|\mathbf{P}\|_\infty + \frac{2h(1 + \eta^{(k)})}{n-1}\right) \|\mathbf{y}_\ell^{(k)}\|_\infty.$$

Whenever $\eta^{(k)}$ and $\max\{\|\mathbf{y}_1^{(k)}\|_\infty, \|\mathbf{y}_2^{(k)}\|_\infty\}$ are bounded by an absolute constant, by setting $r_n = nh\alpha \|\mathbf{P}\|_\infty + \frac{h}{n}$ and assuming $r_n = O(1)$ (by Condition (T1)), we have

$$\|\mathbf{y}_\ell^{(k+1)}\|_\infty \leq (1 + Cr_n) \|\mathbf{y}_\ell^{(k)}\|_\infty, \quad (42)$$

and

$$\begin{aligned} \eta^{(k+1)} &\leq 4 \max_{i \in [n], \ell \in [2]} |y_{i\ell}^{(k)}|^2 \leq 8 \max\{\|\mathbf{y}_1^{(k+1)}\|_\infty^2, \|\mathbf{y}_2^{(k+1)}\|_\infty^2\} \\ &\leq 8(1 + Cr_n) \max\{\|\mathbf{y}_1^{(k)}\|_\infty^2, \|\mathbf{y}_2^{(k)}\|_\infty^2\} = O(1) \end{aligned}$$

In other words, we have shown that for any k such that $\eta^{(k)}$ and $\max\{\|\mathbf{y}_1^{(k)}\|_\infty, \|\mathbf{y}_2^{(k)}\|_\infty\}$ are bounded, then (42) holds, and $\eta^{(k+1)}$ and $\max\{\|\mathbf{y}_1^{(k+1)}\|_\infty, \|\mathbf{y}_2^{(k+1)}\|_\infty\}$ are bounded.

Now Condition (II) says that $\max\{\|\mathbf{y}_1^{(0)}\|_\infty, \|\mathbf{y}_2^{(0)}\|_\infty\} = O(1)$ and $\eta^{(0)} \leq 4 \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_\infty^2 = O(1)$. By induction, we know that $\eta^{(k)}$ and $\max\{\|\mathbf{y}_1^{(k+1)}\|_\infty, \|\mathbf{y}_2^{(k+1)}\|_\infty\}$ are bounded and (42) holds for all $k \geq 1$. Applying (42) iteratively, we have for any $k \geq 1$,

$$\|\mathbf{y}_\ell^{(k)}\|_\infty \leq (1 + Cr_n)^k \|\mathbf{y}_\ell^{(0)}\|_\infty. \quad (43)$$

Therefore, as long as $k = k(n)$ such that $kr_n = O(1)$ (by Condition (T1)), we have $\|\mathbf{y}_\ell^{(k)}\|_\infty / \|\mathbf{y}_\ell^{(0)}\|_\infty = O(1)$, or

$$\frac{\text{diam}(\{y_i^{(k)}\}_{1 \leq i \leq n})}{\max_{i \in [n], \ell \in [2]} |y_{i\ell}^{(0)}|} \leq \frac{\max_{i \in [n], \ell \in [2]} |y_{i\ell}^{(k)}|}{\max_{i \in [n], \ell \in [2]} |y_{i\ell}^{(0)}|} = O(1).$$

A.3 Proof of Theorem 4

The results concerning (10) and (12) follows directly from Theorem 2 and Proposition 3. To see that (13) holds, we first prove the following proposition.

Proposition 20 Let $\mathbf{E}_\alpha^{(k)} = \mathbf{S}_\alpha^{(k)} - (\alpha\mathbf{P} - \mathbf{H}_n)$ and $\zeta = \sup_{k \geq 0} \|\mathbf{L}(\mathbf{E}_\alpha^{(k)})\|$. Suppose the initialization satisfies $\|\mathbf{y}_\ell^{(0)}\| \neq 0$ for $\ell = 1, 2$, and (α, h, K) satisfies $h\|\mathbf{L}(\alpha\mathbf{P})\| < 2$, $Kh\zeta = O(1)$ and $Kh = O(n)$ as $n \rightarrow \infty$. Then for $\ell \in \{1, 2\}$, it holds that

$$\sup_{1 \leq k \leq K} \frac{\|\mathbf{y}_\ell^{(k)} - [\mathbf{I} - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)]^k \mathbf{y}_\ell^{(0)}\|_2}{\|\mathbf{y}_\ell^{(0)}\|_2} = O(Kh\zeta). \quad (44)$$

Consequently, for (α, h, k) such that $h\|\mathbf{L}(\alpha\mathbf{P})\| < 2$, $kh\zeta = o(1)$ and $kh = O(n)$, we have

$$\lim_{n \rightarrow \infty} \frac{\|\mathbf{y}_\ell^{(k)} - [\mathbf{I} - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)]^k \mathbf{y}_\ell^{(0)}\|_2}{\|\mathbf{y}_\ell^{(0)}\|_2} = 0. \quad (45)$$

Proof By linearity of the Laplacian operator, we have

$$\mathbf{L}(\mathbf{E}_\alpha^{(k)}) = \mathbf{L}(\mathbf{S}_\alpha^{(k)}) - \mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n) = \mathbf{L}(\mathbf{S}_\alpha^{(k)}) - \mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)$$

For $\ell = 1, 2$, the updating equation $\mathbf{y}_\ell^{(k+1)} = (\mathbf{I} - h\mathbf{L}(\mathbf{S}_\alpha^{(k)}))\mathbf{y}_\ell^{(k)}$ can be written as

$$\begin{aligned} \mathbf{y}_\ell^{(k+1)} &= (\mathbf{I} - h\mathbf{L}(\mathbf{S}_\alpha^{(k)}))(\mathbf{I} - h\mathbf{L}(\mathbf{S}_\alpha^{(k-1)})) \dots (\mathbf{I} - h\mathbf{L}(\mathbf{S}_\alpha^{(0)}))\mathbf{y}_\ell^{(0)} \\ &= (\mathbf{I} - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n) - h\mathbf{L}(\mathbf{E}_\alpha^{(k)}))(\mathbf{I} - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n) - \mathbf{L}(\mathbf{E}_\alpha^{(k-1)})) \dots \\ &\quad \times (\mathbf{I} - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n) - \mathbf{L}(\mathbf{E}_\alpha^{(0)}))\mathbf{y}_\ell^{(0)} \\ &= (\mathbf{I} - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n))^{k+1} \mathbf{y}_\ell^{(0)} + \boldsymbol{\epsilon}^{(k)}, \end{aligned}$$

where

$$\begin{aligned} \|\boldsymbol{\epsilon}^{(k)}\|_2 &\leq \|\mathbf{y}_\ell^{(0)}\|_2 \left[\binom{k+1}{1} h\zeta \lambda^k + \binom{k+1}{2} (h\zeta)^2 \lambda^{k-1} + \dots + \binom{k+1}{k+1} (h\zeta)^{k+1} \lambda^0 \right] \\ &\leq \|\mathbf{y}_\ell^{(0)}\|_2 [(h\zeta + \lambda)^{k+1} - \lambda^{k+1}] \\ &\leq \|\mathbf{y}_\ell^{(0)}\|_2 \lambda^{k+1} [(h\zeta/\lambda + 1)^{k+1} - 1] \end{aligned}$$

where $\lambda = \|\mathbf{I} - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)\|$ and $\zeta = \sup_{k \geq 0} \|\mathbf{L}(\mathbf{E}_\alpha^{(k)})\|$. We need the following lemma.

Lemma 21 If $\|h\mathbf{L}(\alpha\mathbf{P})\| < 2$, then $1 \leq \|\mathbf{I} - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)\| \leq 1 + \frac{h}{n-1}$.

The above lemma implies

$$\|\boldsymbol{\epsilon}^{(k)}\|_2 \leq \|\mathbf{y}_\ell^{(0)}\|_2 (1 + Ch/n)^{k+1} [(h\zeta + 1)^{k+1} - 1].$$

By the binomial identity,

$$\begin{aligned} (1 + x_n)^k &= 1 + kx_n + \frac{k(k-1)}{2} x_n^2 + \frac{k(k-1)(k-2)}{3!} x_n^3 + \dots + x_n^k \\ &\leq 1 + kx_n + \frac{k^2}{2} x_n^2 + \frac{k^3}{3!} x_n^3 + \dots + \frac{k^k}{k!} x_n^k \\ &\leq 1 + kx_n \left(1 + kx_n + \frac{k^2 x_n^2}{2!} + \dots + \frac{k^{k-1} x_n^{k-1}}{(k-1)!} \right) \\ &\leq 1 + kx_n e^{kx_n}. \end{aligned}$$

Then if we denote $\mathbf{u}_{R+1}, \dots, \mathbf{u}_n$ as the columns of \mathbf{U}_\perp , we have

$$(\mathbf{I} - h\mathbf{L}(\mathbf{P} - \mathbf{H}_n))^k = \sum_{i=1}^{R-1} \left(1 + \frac{h}{n-1}\right)^k \mathbf{u}_i \mathbf{u}_i^\top + n^{-1} \mathbf{1} \mathbf{1}^\top + \sum_{i=R+1}^n \left(1 + \frac{h}{n-1} - h\lambda_i\right)^k \mathbf{u}_i \mathbf{u}_i^\top.$$

Hence

$$\begin{aligned} & \left\| (\mathbf{I} - h\mathbf{L})^k \mathbf{y} - \left(1 + \frac{h}{n-1}\right)^k \sum_{i=1}^{R-1} \mathbf{u}_i \mathbf{u}_i^\top \mathbf{y} - n^{-1} \mathbf{1} \mathbf{1}^\top \mathbf{y} \right\|_2 = \left\| \sum_{i=R+1}^n \left(1 + \frac{h}{n-1} - h\lambda_i\right)^k \mathbf{u}_i \mathbf{u}_i^\top \mathbf{y} \right\|_2 \\ & = \sqrt{\sum_{i=R+1}^n \left(1 + \frac{h}{n-1} - h\lambda_i\right)^{2k} (\mathbf{u}_i^\top \mathbf{y})^2} \leq \left(1 + \frac{h}{n-1} - h\lambda_{R+1}\right)^k \|\mathbf{y}\|_2. \end{aligned}$$

The final result follows by noting that $1 + \frac{h}{n-1} - h\lambda_{R+1} < 1 - \kappa/2 < 1$.

A.5 Proof of Proposition 6

Firstly, since \mathbf{A} is nonnegative, by the Geršgorin circle theorem (Varga, 2010), $\mathbf{L}(\mathbf{A})$ is positive semi-definite. For any $\boldsymbol{\theta}_r$, $r \in \{1, \dots, R\}$, it holds that $\mathbf{L}(\mathbf{P})\boldsymbol{\theta}_r = 0$. It follows that $\{\boldsymbol{\theta}_r\}_{r=1}^R$ is a set of eigenvectors corresponding to the smallest eigenvalue 0. In addition, since the graph corresponding to the weighted adjacency matrix \mathbf{P} has R connected components, by the spectral property of the Laplacian matrix (see, for example, Theorem 3.10 of Marsden (2013)), the null space of $\mathbf{L}(\mathbf{P})$ has dimension R . This implies that the eigenvalue 0 of $\mathbf{L}(\mathbf{P})$ has multiplicity R . Lastly, as $\{\boldsymbol{\theta}_k\}_{k=1}^K$ are linearly independent, the eigen subspace associated with the eigenvalue 0 is spanned by $\{\boldsymbol{\theta}_r\}_{r=1}^R$.

A.6 Proof of Theorem 7

Let $\mathbf{E}'_\alpha^{(k)} = \mathbf{S}_\alpha^{(k)} - \alpha\mathbf{P}^* + \mathbf{H}_n$ and $\zeta' = \sup_{k \geq 1} \|\mathbf{L}(\mathbf{E}'_\alpha^{(k)})\|$. Then similar arguments as in the proof of Theorem 4 imply that $\boldsymbol{\epsilon}'^{(k)} = \mathbf{y}_\ell^{(k+1)} - (\mathbf{I} - h\mathbf{L}(\alpha\mathbf{P}^* - \mathbf{H}_n))^{k+1} \mathbf{y}_\ell^{(0)}$ satisfies

$$\frac{\|\boldsymbol{\epsilon}'^{(k)}\|_2}{\|\mathbf{y}_\ell^{(0)}\|_2} \leq (1 + Ch/n)^{k+1} [(1 + h\zeta')^{k+1} - 1].$$

As a result, as long as $h\zeta' \ll 1/k$, we have

$$\lim_{(n,k) \rightarrow \infty} \frac{\|\mathbf{y}_\ell^{(k)} - [\mathbf{I} - h\mathbf{L}(\alpha\mathbf{P}^* - \mathbf{H}_n)]^k \mathbf{y}_\ell^{(0)}\|_2}{\|\mathbf{y}_\ell^{(0)}\|_2} = 0.$$

Now by the inequality $\|\mathbf{L}(\mathbf{A})\| \leq \|\mathbf{D}(\mathbf{A})\| + \|\mathbf{A}\| \leq 2n\|\mathbf{A}\|_\infty$, and the bounded initialization, we have

$$\begin{aligned} \|\mathbf{L}(\mathbf{E}'_\alpha^{(k)})\| & \leq \|\mathbf{L}(\mathbf{S}_\alpha^{(k)} - \alpha\mathbf{P} + \mathbf{H}_n)\| + \alpha\|\mathbf{L}(\mathbf{P}^* - \mathbf{P})\| \\ & \leq 2n\|\mathbf{S}_\alpha^{(k)} - \alpha\mathbf{P} + \mathbf{H}_n\|_\infty + \alpha\|\mathbf{L}(\mathbf{P}^* - \mathbf{P})\| \\ & \leq (\alpha n\|\mathbf{P}\|_\infty + 1/n) \cdot \max_{i \in [n], \ell \in [2]} |y_{i\ell}^{(0)}|^2 + \alpha\|\mathbf{L}(\mathbf{P}^* - \mathbf{P})\| \end{aligned} \quad (48)$$

where the last inequality follows from Proposition 2. Thus, the condition $h\zeta' \ll 1/k$ can be implied by

$$k[(nh\alpha\|\mathbf{P}\|_\infty + h/n) \cdot \max_{i \in [n], \ell \in [2]} |y_{i\ell}^{(0)}|^2 + \alpha h \|\mathbf{L}(\mathbf{P}^* - \mathbf{P})\|] = o(1),$$

which holds under Conditions (I2) (T1.D) and (T2.D). Now, if we further assume $\kappa < h\lambda_{R+1}(\mathbf{L}(\alpha\mathbf{P}^*)) \leq h\|\mathbf{L}(\alpha\mathbf{P}^*)\| \leq 1$ and $k \rightarrow \infty$, by Theorem 5, we also have

$$\lim_{(n,k) \rightarrow \infty} \frac{\|\mathbf{y}_\ell^{(k)} - \mathbf{U}\mathbf{U}^\top \mathbf{y}_\ell^{(0)}\|_2}{\|\mathbf{y}_\ell^{(0)}\|_2} = 0,$$

where the columns of $\mathbf{U} \in O(n, R-1)$ span the null space of $\mathbf{L}(\mathbf{P}^*)$. By Proposition 6, we know that the matrix $\mathbf{U}\mathbf{U}^\top \begin{bmatrix} \mathbf{y}_1^{(0)} & \mathbf{y}_2^{(0)} \end{bmatrix} \in \mathbb{R}^{n \times 2}$ has at most R distinct rows, and any two rows corresponding to the same graph component in G have the identical values. Then, the final results follow by setting $\{z_1, \dots, z_R\}$ such that $z_r = (z_{1r}, z_{2r})$ is the same as the rows in $\mathbf{U}\mathbf{U}^\top \begin{bmatrix} \mathbf{y}_1^{(0)} & \mathbf{y}_2^{(0)} \end{bmatrix}$ corresponding to the r -th graph component.

Appendix B. Continuous-Time Analysis of the Early Exaggeration Stage

B.1 Proof of Proposition 8

Note that the algorithm (20) is in fact the Euler scheme for solving the differential equation (21). We can apply the standard differential equation theory to obtain the global approximation error for the Euler scheme. By taking derivative on both sides of the differential equation (21), we have

$$\ddot{\mathbf{Y}}_\ell(t) = \mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)\dot{\mathbf{Y}}_\ell(t) = \mathbf{L}^2(\alpha\mathbf{P} - \mathbf{H}_n)\mathbf{Y}_\ell(t).$$

Since $\|\ddot{\mathbf{Y}}_\ell(t)\|_2 \leq \|\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)\|^2 \|\mathbf{Y}_\ell(t)\|_2$, by Theorem 212A of Butcher (2008), we have

$$\sup_{t \in [0, T]} \frac{\|\mathbf{y}_{\ell, h}(t) - \mathbf{Y}_\ell(t)\|_2}{\|\mathbf{Y}_\ell(t)\|_2} \leq Th \|\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)\|^2.$$

Consequently, for any $k \leq T/h$, if we set $t = kh \leq T$, then

$$\frac{\|\tilde{\mathbf{y}}_\ell^{(k)} - \mathbf{Y}_\ell(kh)\|_2}{\|\mathbf{Y}_\ell(kh)\|_2} = \frac{\|\mathbf{y}_{\ell, h}^{(t)} - \mathbf{Y}_\ell(t)\|_2}{\|\mathbf{Y}_\ell(t)\|_2} \leq Th \|\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)\|^2.$$

This proves the second statement of the theorem.

B.2 Proof of Proposition 9

By standard theory of ODE, we have $\mathbf{Y}_\ell(t) = \sum_{i=1}^n e^{-t\sigma_i} (\mathbf{u}_i^\top \mathbf{y}_\ell^{(0)}) \mathbf{u}_i$, where $\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n) = \sum_{i=1}^n \sigma_i \mathbf{u}_i \mathbf{u}_i^\top$ is the eigendecomposition of $\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)$. The final result follows from the fact that $\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n) = \mathbf{L}(\alpha\mathbf{P}) - \frac{1}{n-1} \mathbf{I} + \frac{1}{n(n-1)} \mathbf{1}\mathbf{1}^\top$, so that $\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)$ and $\mathbf{L}(\alpha\mathbf{P})$ share the same set of eigenvectors, and

$$\sigma_i = \alpha\lambda_i - \frac{1}{n-1}, \quad i \in \{2, \dots, n\},$$

and $\sigma_1 = \alpha\lambda_1$.

B.3 Proof of Theorem 10

Let $\mathbf{U}_0 \in O(n, R)$ be the matrix whose columns span the null space of $\mathbf{L}(\mathbf{P}^*)$, and the first column of \mathbf{U}_0 is $n^{-1/2}\mathbf{1}$. Let $\mathbf{U} \in O(n, R)$ be the matrix whose columns correspond to the smallest R eigenvalues of $\mathbf{L}(\mathbf{P})$. By standard Davis-Kahan matrix perturbation inequality, we have

$$\|\mathbf{U}_{0\perp}^\top \mathbf{U}\| \leq \frac{\|\mathbf{L}(\mathbf{E})\|}{\lambda_{R+1}(\mathbf{L}(\mathbf{P}^*))}, \quad (49)$$

where $\mathbf{E} = \mathbf{P} - \mathbf{P}^*$, and $\lambda_{R+1}(\mathbf{L}(\mathbf{P}^*))$ is the smallest nonzero eigenvalue of $\mathbf{L}(\mathbf{P}^*)$. Note that if we define $\boldsymbol{\beta}_t = (\beta_{t,1}, \dots, \beta_{t,n}) \in \mathbb{R}^n$ such that $\beta_{t,i} = e^{-t(\alpha\lambda_i - \frac{1}{n-1})}(\mathbf{u}_i^\top \mathbf{y}_\ell^{(0)})$ for $i = 2, \dots, n$ and $\beta_{t,1} = n^{-1/2}\mathbf{1}_n^\top \mathbf{y}_\ell^{(0)}$, we can write $\mathbf{Y}_\ell(t) = \mathbf{U}\boldsymbol{\beta}_{t,1:R} + \mathbf{U}_\perp\boldsymbol{\beta}_{t,(R+1):n}$. In particular, we have

$$\|\mathbf{Y}_\ell(t)\|_2 = \|\mathbf{U}\boldsymbol{\beta}_{t,1:R}\|_2 + \|\mathbf{U}_\perp\boldsymbol{\beta}_{t,(R+1):n}\|_2.$$

Since $\|\mathbf{Y}_\ell(t)\|_2 \lesssim \|\mathbf{Y}_\ell(0)\|_2$, it follows that

$$\begin{aligned} & \frac{\|\mathbf{U}_0\mathbf{U}_0^\top \mathbf{Y}_\ell(t) - \mathbf{Y}_\ell(t)\|_2}{\|\mathbf{Y}_\ell(0)\|_2} \\ &= \frac{\|\mathbf{U}_0\mathbf{U}_0^\top \mathbf{U}\boldsymbol{\beta}_{t,1:R} - \mathbf{U}\boldsymbol{\beta}_{t,1:R}\|_2}{\|\mathbf{Y}_\ell(t)\|_2} \frac{\|\mathbf{Y}_\ell(t)\|_2}{\|\mathbf{Y}_\ell(0)\|_2} + \frac{\|\mathbf{U}_0\mathbf{U}_0^\top \mathbf{U}_\perp\boldsymbol{\beta}_{t,(R+1):n} - \mathbf{U}_\perp\boldsymbol{\beta}_{t,(R+1):n}\|_2}{\|\mathbf{Y}_\ell(0)\|_2} \\ &\lesssim \|(\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{I})\mathbf{U}\mathbf{U}^\top\| \cdot \frac{\|\mathbf{U}\boldsymbol{\beta}_{t,1:R}\|_2}{\|\mathbf{U}\boldsymbol{\beta}_{t,1:R}\|_2 + \|\mathbf{U}_\perp\boldsymbol{\beta}_{t,(R+1):n}\|_2} + \|(\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{I})\mathbf{U}_\perp\mathbf{U}_\perp^\top\| \cdot \frac{\|\mathbf{U}_\perp\boldsymbol{\beta}_{t,(R+1):n}\|_2}{\|\mathbf{Y}_\ell(0)\|_2} \\ &\leq \|(\mathbf{U}_0)_\perp^\top \mathbf{U}\| + \|(\mathbf{U}_0)_\perp^\top \mathbf{U}_\perp\| \cdot \frac{\|\mathbf{U}_\perp\boldsymbol{\beta}_{t,(R+1):n}\|_2}{\|\mathbf{Y}_\ell(0)\|_2} \\ &\leq \|(\mathbf{U}_0)_\perp^\top \mathbf{U}\| + \frac{\|\mathbf{U}_\perp\boldsymbol{\beta}_{t,(R+1):n}\|_2}{\|\mathbf{Y}_\ell(0)\|_2} \\ &\leq \frac{\|\mathbf{L}(\mathbf{E})\|}{\lambda_{R+1}(\mathbf{L}(\mathbf{P}^*))} + \frac{e^{-t(\alpha\lambda_{R+1}(\mathbf{L}(\mathbf{P})) - \frac{1}{n-1})}\|\mathbf{U}_\perp\mathbf{y}_\ell^{(0)}\|_2}{\|\mathbf{y}_\ell^{(0)}\|_2} \\ &\leq \frac{\|\mathbf{L}(\mathbf{E})\|}{\lambda_{R+1}(\mathbf{L}(\mathbf{P}^*))} + e^{-t(\alpha\lambda_{R+1}(\mathbf{L}(\mathbf{P})) - \frac{1}{n-1})}. \end{aligned}$$

Therefore, whenever

$$\|\mathbf{L}(\mathbf{E})\| \ll \lambda_{R+1}(\mathbf{L}(\mathbf{P}^*)), \quad \alpha\lambda_{R+1}(\mathbf{L}(\mathbf{P})) \gg \frac{1}{n}, \quad t\alpha\lambda_{R+1}(\mathbf{L}(\mathbf{P})) \rightarrow \infty, \quad (50)$$

we have

$$\lim_{(t,n) \rightarrow \infty} \frac{\|\mathbf{U}_0\mathbf{U}_0^\top \mathbf{Y}_\ell(t) - \mathbf{Y}_\ell(t)\|_2}{\|\mathbf{Y}_\ell(0)\|_2} = 0.$$

On the other hand, note that $\mathbf{Y}_\ell(t) = \mathbf{U}\boldsymbol{\Gamma}_1(t)\mathbf{U}^\top \mathbf{y}_\ell^{(0)} + \mathbf{U}_\perp\boldsymbol{\Gamma}_2(t)\mathbf{U}_\perp^\top \mathbf{y}_\ell^{(0)}$, where $\boldsymbol{\Gamma}_1(t) = \text{diag}(1, \dots, e^{-t(\alpha\lambda_R(\mathbf{L}(\mathbf{P})) - \frac{1}{n-1})})$, and $\boldsymbol{\Gamma}_2(t) = \text{diag}(e^{-t(\alpha\lambda_{R+1}(\mathbf{L}(\mathbf{P})) - \frac{1}{n-1})}, \dots, e^{-t(\alpha\lambda_n(\mathbf{L}(\mathbf{P})) - \frac{1}{n-1})})$.

We have

$$\begin{aligned}
 \frac{\|\mathbf{U}_0\mathbf{U}_0^\top\mathbf{Y}_\ell(t) - \mathbf{U}_0\mathbf{U}_0^\top\mathbf{Y}_\ell(0)\|_2}{\|\mathbf{Y}_\ell(0)\|_2} &= \frac{\|\mathbf{U}\mathbf{\Gamma}_1(t)\mathbf{U}^\top\mathbf{Y}_\ell(0) + \mathbf{U}_\perp\mathbf{\Gamma}_2(t)\mathbf{U}_\perp^\top\mathbf{Y}_\ell(0) - \mathbf{U}_0\mathbf{U}_0^\top\mathbf{Y}_\ell(0)\|_2}{\|\mathbf{Y}_\ell(0)\|_2} \\
 &\leq \frac{\|\mathbf{U}\mathbf{\Gamma}_1(t)\mathbf{U}^\top\mathbf{Y}_\ell(0) - \mathbf{U}_0\mathbf{U}_0^\top\mathbf{Y}_\ell(0)\|_2}{\|\mathbf{Y}_\ell(0)\|_2} + \frac{\|\mathbf{U}_\perp\mathbf{\Gamma}_2(t)\mathbf{U}_\perp^\top\mathbf{Y}_\ell(0)\|_2}{\|\mathbf{Y}_\ell(0)\|_2} \\
 &\leq \|\mathbf{U}(\mathbf{\Gamma}_1(t) - \mathbf{I}_R)\mathbf{U}^\top + \mathbf{U}\mathbf{U}^\top - \mathbf{U}_0\mathbf{U}_0^\top\| + \|\mathbf{U}_\perp\mathbf{\Gamma}_2(t)\mathbf{U}_\perp^\top\| \\
 &\leq \|\mathbf{\Gamma}_1(t) - \mathbf{I}_R\| + 2\|\mathbf{U}^\top\mathbf{U}_{0\perp}\| + \|\mathbf{\Gamma}_2(t)\| \\
 &\leq |e^{-t(\alpha\lambda_{R+1}(\mathbf{L}(\mathbf{P})) - \frac{1}{n-1})} - 1| + \frac{2\|\mathbf{L}(\mathbf{E})\|}{\lambda_{R+1}(\mathbf{L}(\mathbf{P}^*))} + e^{-t(\alpha\lambda_{R+1}(\mathbf{L}(\mathbf{P})) - \frac{1}{n-1})},
 \end{aligned}$$

where we used the inequality $\|\mathbf{U}\mathbf{U}^\top - \mathbf{U}_0\mathbf{U}_0^\top\| \leq 2\|\mathbf{U}^\top\mathbf{U}_{0\perp}\|$ from Lemma 1 of Cai and Zhang (2018). Hence, whenever

$$\|\mathbf{L}(\mathbf{E})\| \ll \lambda_{R+1}(\mathbf{L}(\mathbf{P}^*)), \quad \alpha\lambda_{R+1}(\mathbf{L}(\mathbf{P})) \gg \frac{1}{n}, \quad t\alpha\lambda_{R+1}(\mathbf{L}(\mathbf{P})) \rightarrow \infty, \quad t\left(\alpha\lambda_{R+1}(\mathbf{L}(\mathbf{P})) - \frac{1}{n-1}\right) \rightarrow 0, \quad (51)$$

we also have

$$\lim_{(n,t) \rightarrow \infty} \frac{\|\mathbf{U}_0\mathbf{U}_0^\top\mathbf{Y}_\ell(t) - \mathbf{U}_0\mathbf{U}_0^\top\mathbf{y}_\ell^{(0)}\|_2}{\|\mathbf{y}_\ell^{(0)}\|_2} = 0.$$

The second statement can be obtained by noticing that

$$\frac{\|\mathbf{Y}_\ell(t) - O\mathbf{z}_\ell\|_2}{\|\mathbf{Y}_\ell(0)\|_2} \leq \frac{\|\mathbf{Y}_\ell(t) - \mathbf{U}_0\mathbf{U}_0^\top\mathbf{Y}_\ell(t)\|_2}{\|\mathbf{Y}_\ell(0)\|_2} + \frac{\|\mathbf{U}_0\mathbf{U}_0^\top\mathbf{Y}_\ell(t) - \mathbf{U}_0\mathbf{U}_0^\top\mathbf{Y}_\ell(0)\|_2}{\|\mathbf{Y}_\ell(0)\|_2},$$

and that $\mathbf{U}_0\mathbf{U}_0^\top\mathbf{Y}_\ell(0) = O\mathbf{z}_\ell$. To see the above conditions hold under Conditions (I1) (T1.C) and (T2.C), we note that by Weyl's inequality, $\max_i |\lambda_i(\mathbf{L}(\mathbf{P})) - \lambda_i(\mathbf{L}(\mathbf{P}^*))| \leq \|\mathbf{L}(\mathbf{E})\|$. Since $\lambda_{R+1}(\mathbf{L}(\mathbf{P})) \geq \lambda_{R+1}(\mathbf{L}(\mathbf{P}^*)) - \|\mathbf{L}(\mathbf{E})\|$, the conditions in (50) can be implied by

$$\|\mathbf{L}(\mathbf{E})\| \ll \lambda_{R+1}(\mathbf{L}(\mathbf{P}^*)), \quad \alpha\lambda_{R+1}(\mathbf{L}(\mathbf{P})) \gg \frac{1}{n}, \quad t\alpha\lambda_{R+1}(\mathbf{L}(\mathbf{P}^*)) \rightarrow \infty.$$

On the other hand, the conditions in (51) can be implied by

$$\begin{aligned}
 \|\mathbf{L}(\mathbf{E})\| \ll \lambda_{R+1}(\mathbf{L}(\mathbf{P}^*)), \quad \alpha\lambda_{R+1}(\mathbf{L}(\mathbf{P})) \gg \frac{1}{n}, \quad t\alpha\lambda_{R+1}(\mathbf{L}(\mathbf{P}^*)) \rightarrow \infty, \\
 t\alpha\|\mathbf{L}(\mathbf{E})\| \rightarrow 0, \quad t/n \rightarrow 0.
 \end{aligned}$$

These are ensured by the conditions of the theorem.

Appendix C. Analysis of the Embedding Stage

C.1 Proof of Proposition 12

Note that $\frac{q_{ij}^{(k)}}{1+d_{ij}^{(k)}} = \frac{1}{Z^{(k)}(1+d_{ij}^{(k)})^2} = \frac{(1+d_{ij}^{(k)})^{-2}}{\sum_{i \neq j} (1+d_{ij}^{(k)})^{-1}}$, where $(1+\eta^{(k)})^{-2} \leq (1+d_{ij}^{(k)})^{-2} \leq 1$ for all $i \neq j$. It holds that

$$\begin{aligned}
 \left| \frac{q_{ij}^{(k)}/(1+d_{ij}^{(k)})}{1/n(n-1)} - 1 \right| &= \left| \frac{n(n-1)(1+d_{ij}^{(k)})^{-2}}{\sum_{i \neq j} (1+d_{ij}^{(k)})^{-1}} - 1 \right| \\
 &\leq \frac{|(1+d_{ij}^{(k)})^{-2}n(n-1) - \sum_{i \neq j} (1+d_{ij}^{(k)})^{-1}|}{\sum_{i \neq j} (1+d_{ij}^{(k)})^{-1}} \\
 &\leq \frac{|(1+d_{ij}^{(k)})^{-2}n(n-1) - \sum_{i \neq j} (1+d_{ij}^{(k)})^{-1}|}{n(n-1)/(1+\eta^{(k)})} \\
 &\leq \max \left\{ \frac{|n(n-1) - \sum_{i \neq j} (1+d_{ij}^{(k)})^{-1}|}{n(n-1)/(1+\eta^{(k)})}, \frac{|(1+\eta^{(k)})^{-2}n(n-1) - \sum_{i \neq j} (1+d_{ij}^{(k)})^{-1}|}{n(n-1)/(1+\eta^{(k)})} \right\} \\
 &\leq \max \left\{ \frac{|(1+\eta^{(k)})^{-2}n(n-1) - n(n-1)|}{n(n-1)/(1+\eta^{(k)})}, \frac{|n(n-1) - n(n-1)(1+\eta^{(k)})^{-1}|}{n(n-1)/(1+\eta^{(k)})}, \right. \\
 &\quad \left. \frac{|(1+\eta^{(k)})^{-2}n(n-1) - n(n-1)(1+\eta^{(k)})^{-1}|}{n(n-1)/(1+\eta^{(k)})} \right\},
 \end{aligned}$$

where the last two inequalities follows from the fact that $|a-b| \leq \max\{|c-b|, |d-b|\}$ for any $c < a < d$. Therefore, by $\eta^{(k)} = o(1)$, we have

$$\left| \frac{q_{ij}^{(k)}/(1+d_{ij}^{(k)})}{1/n(n-1)} - 1 \right| \lesssim |(1+\eta^{(k)})^{-2} - 1| + |(1+\eta^{(k)})^{-1} - 1| \lesssim \eta^{(k)}.$$

Now since $p_{ij} \geq 0$, we have $S_{ij}^{(k)} = \frac{p_{ij}-q_{ij}^{(k)}}{1+d_{ij}^{(k)}} \geq \frac{p_{ij}}{1+\eta} - \frac{1+O(\eta^{(k)})}{n(n-1)}$. Similarly, we can obtain

$S_{ij}^{(k)} = \frac{p_{ij}-q_{ij}^{(k)}}{1+d_{ij}^{(k)}} \leq p_{ij} - \frac{1+O(\eta^{(k)})}{n(n-1)}$. Hence, if $p_{ij} = o(n^{-2})$ and $\eta = o(1)$, we have

$$S_{ij}^{(k)} \asymp -\frac{1+O(\eta^{(k)})}{n(n-1)}.$$

This proves the first statement of the lemma. The second statement can be obtained from the similar argument.

C.2 Proof of Theorem 13

The proof is divided into two parts.

Case I: $k = K_0$. By (30), summing up the contribution of all the points with index $j \neq i$ such that $j \sim i$, we have

$$\begin{aligned} \left\| \sum_{j:j \sim i} S_{ij}^{(k)} (y_j^{(k)} - y_i^{(k)}) \right\|_2 &\lesssim (n^{-2} + \|\mathbf{P}^*\|_\infty + \|\mathbf{E}\|_\infty) \cdot \sum_{j:j \sim i} \|y_j^{(k)} - y_i^{(k)}\|_2 \\ &\lesssim \sqrt{n}(n^{-2} + \|\mathbf{P}^*\|_\infty) R_n (\|\mathbf{y}_1^{(0)}\|_2 + \|\mathbf{y}_2^{(0)}\|_2) \end{aligned}$$

where $R_n = (1 - \kappa)^{k+1} + hk((\alpha n \|\mathbf{P}\|_\infty + 1/n) \cdot \max_{i \in [n], \ell \in [2]} |y_{i\ell}^{(0)}|^2 + \alpha \|\mathbf{L}(\mathbf{P}^* - \mathbf{P})\|)$, the first inequality follows from Proposition 12, and the second inequality follows from the fact that,

$$\begin{aligned} \sum_{j:j \sim i} \|y_j^{(k)} - y_i^{(k)}\|_2 &\lesssim \sum_{\ell=1}^2 \|\mathbf{y}_\ell^{(k)} - O\mathbf{z}_\ell\|_1 \leq \sqrt{n} \sum_{\ell=1}^2 \|\mathbf{y}_\ell^{(k)} - O\mathbf{z}_\ell\|_2 \\ &\leq \sqrt{n} \sum_{\ell=1}^2 [\|\mathbf{y}_\ell^{(k)} - (\mathbf{I} - h\mathbf{L}(\alpha\mathbf{P}^* - \mathbf{H}_n))^k \mathbf{y}_\ell^{(0)}\|_2 + \|(\mathbf{I} - h\mathbf{L}(\alpha\mathbf{P}^* - \mathbf{H}_n))^k \mathbf{y}_\ell^{(0)} - O\mathbf{z}_\ell\|_2] \\ &\lesssim \sqrt{n} [(1 - \kappa)^{k+1} + hk \sup_{s < k} \|\mathbf{L}(\mathbf{S}_\alpha^{(s)} - \alpha\mathbf{P}^* + \mathbf{H}_n)\|] (\|\mathbf{y}_1^{(0)}\|_2 + \|\mathbf{y}_2^{(0)}\|_2) \end{aligned} \quad (52)$$

$$\begin{aligned} &\lesssim \sqrt{n} [(1 - \kappa)^{k+1} + hk((\alpha n \|\mathbf{P}\|_\infty + 1/n) \cdot \max_{i \in [n], \ell \in [2]} |y_{i\ell}^{(0)}|^2 + \alpha \|\mathbf{L}(\mathbf{P}^* - \mathbf{P})\|)] (\|\mathbf{y}_1^{(0)}\|_2 + \|\mathbf{y}_2^{(0)}\|_2) \\ &= \sqrt{n} R_n (\|\mathbf{y}_1^{(0)}\|_2 + \|\mathbf{y}_2^{(0)}\|_2) \end{aligned} \quad (53)$$

where (52) follow from Theorems 4 and 5, by assuming $\kappa + \frac{h}{n-1} < h\lambda_{R+1}(\mathbf{L}(\alpha\mathbf{P}^*)) \leq h\lambda_n(\mathbf{L}(\alpha\mathbf{P}^*)) \leq 1 + \frac{h}{n-1}$, and the last inequality follows from (48), by assuming $kh[(n\alpha\|\mathbf{P}\|_\infty + h1/n) \max_{i \in [n], \ell \in [2]} |y_{i\ell}^{(0)}|^2 + \alpha h \|\mathbf{L}(\mathbf{P}^* - \mathbf{P})\|] = o(1)$.

On the other hand, we consider the contribution of all the other points $j \not\sim i$. Suppose $i \in H_r$. Then for any $s \neq r$, by the fact that $S_{ij}^{(k)} = \frac{1+O(\eta^{(k)})}{n(n-1)}$ for $i \not\sim j$, we have

$$\begin{aligned} \left\| h' \sum_{j \in H_s} S_{ij}^{(k)} (y_j^{(k)} - y_i^{(k)}) - f_{is}^{(k)} \right\|_2 &= h' \left\| \sum_{j \in H_s} S_{ij}^{(k)} (y_j^{(k)} - y_i^{(k)}) - \frac{1}{n(n-1)} \sum_{j \in H_s} (y_i^{(k)} - y_j^{(k)}) \right\|_2 \\ &\lesssim \frac{h'\eta^{(k)}}{n^2} \left\| \sum_{j \in H_s} (y_i^{(k)} - y_j^{(k)}) \right\|_2 = \frac{h'\eta^{(k)}}{n^2} \|[\mathbf{t}^\top \mathbf{y}_1^{(k)}, \mathbf{t}^\top \mathbf{y}_2^{(k)}]\|_2 \leq \frac{h'n_s \eta^{(k)}}{n^2} \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(k)}\|_2, \end{aligned}$$

where $\mathbf{t} = (t_1, \dots, t_n)^\top$ such that $t_i = |H_s| = n_s$ and $t_j = -1\{j \in H_s\}$, and the last inequality follows from $\|\mathbf{t}\|_2 \lesssim n_s$. In particular, since $\|\mathbf{y}_\ell^{(k)}\|_2 \leq \|\mathbf{I} - h\mathbf{L}(\mathbf{S}_\alpha^{(k-1)})\| \cdot \|\mathbf{y}_\ell^{(k-1)}\|_2 \leq \|\mathbf{y}_\ell^{(k-1)}\|_2$, whenever $h\|\mathbf{S}_\alpha^{(k-1)}\| < 2$, or $\|\mathbf{I} - h\mathbf{L}(\mathbf{S}_\alpha^{(k-1)})\| \leq 1$, we have

$$\|\mathbf{y}_\ell^{(k)}\|_2 \leq \|\mathbf{y}_\ell^{(0)}\|_2,$$

whenever $\sup_{s < k} h\|\mathbf{S}_\alpha^{(s)}\| < 2$. Now since

$$\begin{aligned} \sup_{s < k} \|\mathbf{L}(\mathbf{S}_\alpha^{(s)})\| &\leq \sup_{s < k} \|\mathbf{L}(\mathbf{S}_\alpha^{(s)} - \alpha\mathbf{P}^* + \mathbf{H}_n)\| + \|\mathbf{L}(\alpha\mathbf{P}^* - \mathbf{H}_n)\| \\ &\leq (\alpha n \|\mathbf{P}\|_\infty + 1/n) \cdot \max_{i \in [n], \ell \in [2]} |y_{i\ell}^{(0)}|^2 + \alpha \|\mathbf{L}(\mathbf{P}^* - \mathbf{P})\| + \|\mathbf{L}(\alpha\mathbf{P}^* - \mathbf{H}_n)\|, \end{aligned}$$

it suffices to have $h\|\mathbf{L}(\alpha\mathbf{P}^* - \mathbf{H}_n)\| \leq c < 2$ for some constant c (the last inequality is ensured by $h\|\mathbf{L}(\alpha\mathbf{P}^* - \mathbf{H}_n)\| \leq h\|\mathbf{L}(\alpha\mathbf{P}^*)\| \leq 1 + h/n$ in (T1.E)). Hence, combining the previous arguments, we have

$$\begin{aligned} \|\epsilon_i\|_2 &\lesssim \left\| h' \sum_{j:j\sim i} S_{ij}^{(k)}(y_j^{(k)} - y_i^{(k)}) \right\|_2 + R \left\| h' \sum_{j \in H_s} S_{ij}^{(k)}(y_j^{(k)} - y_i^{(k)}) - f_{is}^{(k)} \right\|_2 \\ &\lesssim \left[\sqrt{n}h'(\|\mathbf{P}^*\|_\infty + n^{-2})R_n + \frac{h'\eta^{(k)}}{n} \right] \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2. \end{aligned}$$

On the other hand, define $D_{rs} = z_r - z_s$. Then

$$\begin{aligned} \left\| f_{is}^{(k)} - \frac{h'n_s}{n(n-1)} D_{rs} \right\|_2 &= \left\| \frac{h'n_s}{n(n-1)} \left(y_i^{(k)} - \frac{1}{n_s} \sum_{j \in H_s} y_j^{(k)} \right) - \frac{h'n_s}{n(n-1)} D_{rs} \right\|_2 \\ &= \frac{h'}{n(n-1)} \left\| \sum_{j \in H_s} (y_i^{(k)} - y_j^{(k)}) - n_s(z_r - z_s) \right\|_2 = \frac{h'}{n(n-1)} \left\| [\mathbf{t}^\top(\mathbf{y}_1^{(k)} - O\mathbf{z}_1), \mathbf{t}^\top(\mathbf{y}_2^{(k)} - O\mathbf{z}_2)] \right\|_2 \\ &\lesssim \frac{h'n_s}{n^2} \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(k)} - O\mathbf{z}_\ell\|_\infty \lesssim \frac{h'n_s}{n^2} \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(k)} - O\mathbf{z}_\ell\|_2 \lesssim \frac{h'n_s}{n^2} R_n \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2, \end{aligned}$$

For sufficiently large n , assuming $\|\mathbf{y}_1^{(0)}\|_2 \asymp \|\mathbf{y}_2^{(0)}\|_2$,

$$\frac{\|D_{rs}\|_2}{\max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2} \gtrsim \min_{\ell \in [2]} \frac{\min_{r \neq s} |z_{\ell r} - z_{\ell s}|}{\|\mathbf{y}_\ell^{(0)}\|_2} \geq \frac{c}{n}, \quad (54)$$

Consequently, if $nR_n = o(1)$,

$$\|f_{is}^{(k)}\|_2 \geq \frac{h'n_s}{n^2} \|D_{rs}\|_2 - \frac{h'n_s R_n}{n^2} \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2 \gtrsim \left[\frac{h'n_s}{n^3} - \frac{h'n_s R_n}{n^2} \right] \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2 \gtrsim \frac{h'n_s}{n^3} \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2.$$

If in addition $\frac{(\|\mathbf{P}^*\|_\infty + n^{-2})R_n n^{7/2}}{n_s} \rightarrow 0$ and $n^2\eta^{(k)}/n_s \rightarrow 0$ (this is implied by Proposition 3 and $\sqrt{n} \max_{i \in [n], \ell \in [2]} |y_{i\ell}^{(0)}| \rightarrow 0$), we have

$$\|\epsilon_i\|_2 / \|f_{is}^{(k)}\|_2 \lesssim \frac{[\sqrt{n}h'(\|\mathbf{P}^*\|_\infty + n^{-2})R_n + h'\eta^{(k)}/n]n^3}{h'n_s} = \frac{(\|\mathbf{P}^*\|_\infty + n^{-2})R_n n^{7/2}}{n_s} + n^2\eta^{(k)}/n_s = o(1).$$

This completes the proof for $k = K_0$.

Case II. $k > K_0$. In order to show that the above results still hold for $k > k_0$, we show that (i) as $n \rightarrow \infty$

$$\sup_{k_0 \leq k \leq K} \max_{(i,j): i \sim j} \|y_i^{(k)} - y_j^{(k)}\|_2 \lesssim V_n, \quad (55)$$

and (ii) as $n \rightarrow \infty$

$$\inf_{k_0 \leq k \leq K} \min_{(i,j): i \not\sim j} \|y_i^{(k)} - y_j^{(k)}\|_2 \gtrsim B_n. \quad (56)$$

For any $r, s \in \{1, \dots, R\}$ and $r \neq s$, we choose $c(r)$ and $c(s)$ such that $c(r) \in H_r$ and $c(s) \in H_s$, and define

$$D_{rs}^{(k)} = y_{c(r)}^{(k)} - y_{c(s)}^{(k)}.$$

In particular, by (55), as $n \rightarrow \infty$, the choices of specific $c(r)$ and $c(s)$ are unimportant. Now By (56), for each $k_0 \leq k \leq K$, we have

$$\begin{aligned} \|f_{is}^{(k)}\|_2 &= \left\| \frac{h'n_s}{n(n-1)} \left(y_i^{(k)} - \frac{1}{n_s} \sum_{j \in H_s} y_j^{(k)} \right) \right\|_2 = \frac{h'}{n(n-1)} \left\| \sum_{j \in H_s} (y_i^{(k)} - y_j^{(k)}) \right\|_2 \\ &\geq \frac{h'n_s}{n(n-1)} (\|y_i^{(k)} - y_{j^*}^{(k)}\|_2 - \max_{j_1, j_2 \in H_s} \|y_{j_1}^{(k)} - y_{j_2}^{(k)}\|_2) \gtrsim \frac{h'n_s}{n^2} (B_n - V_n). \end{aligned}$$

Hence,

$$\|f_{is}\|_2 \gtrsim \frac{h'n_s B_n}{n^2}, \quad (57)$$

whenever $B_n \gg V_n$. On the other hand, for the error term, we have

$$\|\epsilon_i\|_2 \lesssim \left\| h' \sum_{j: j \sim i} S_{ij}^{(k)} (y_j^{(k)} - y_i^{(k)}) \right\|_2 + R \left\| h' \sum_{j \in H_s} S_{ij}^{(k)} (y_j^{(k)} - y_i^{(k)}) - f_{is}^{(k)} \right\|_2,$$

where

$$\left\| h' \sum_{j: j \sim i} S_{ij}^{(k)} (y_j^{(k)} - y_i^{(k)}) \right\|_2 \lesssim h' (n^{-2} + \|\mathbf{P}^*\|_\infty) n_r \max_{i \sim j} \|y_i^{(k)} - y_j^{(k)}\|_2 \lesssim h' (n^{-2} + \|\mathbf{P}^*\|_\infty) n_r V_n,$$

and

$$\begin{aligned} \left\| h' \sum_{j \in H_s} S_{ij}^{(k)} (y_j^{(k)} - y_i^{(k)}) - f_{is}^{(k)} \right\|_2 &\leq h' \left\| \sum_{j \in H_s} S_{ij}^{(k)} (y_j^{(k)} - y_i^{(k)}) - \frac{1}{n(n-1)} \sum_{j \in H_s} (y_i^{(k)} - y_j^{(k)}) \right\|_2 \\ &\lesssim \frac{h'\eta^{(k)}}{n^2} \left\| \sum_{j \in H_s} (y_i^{(k)} - y_j^{(k)}) \right\|_2 \lesssim \frac{h'\eta^{(k)} n_s B_n}{n^2}, \end{aligned}$$

where we used the key inequality

$$\max_{i, j \in [n]} \|y_i^{(k)} - y_j^{(k)}\|_2 \lesssim B_n. \quad (58)$$

Hence, it follows that

$$\|\epsilon_i\|_2 \lesssim (\|\mathbf{P}^*\|_\infty + n^{-2}) n_r h' V_n + \frac{h'\eta^{(k)} n_s B_n}{n^2} \quad (59)$$

and

$$\frac{\|\epsilon_i\|_2}{\|f_{is}^{(k)}\|_2} \lesssim \frac{(\|\mathbf{P}^*\|_\infty + n^{-2}) V_n n^3 + \eta^{(k)} n_s B_n}{n_s B_n} = o(1) \quad (60)$$

whenever $(\|\mathbf{P}^*\|_\infty + n^{-2}) V_n n^2 / B_n \rightarrow 0$ and $\eta^{(k)} \rightarrow 0$. In fact, we will show in the next part that $V_n = R_n \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2$ and $B_n = n^{-1} \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2$, so these conditions become $\|\mathbf{P}^*\|_\infty R_n n^3 \rightarrow 0$ and $\eta^{(k)} = o(1)$, and both are true under (I3) and (T1.E).

Proof of (55), (56) and (58). To show these two inequalities, we need to obtain a general iteration formula over the embedding stage. Note that for any $i, j \in [n]$,

$$\begin{aligned} \|y_i^{(k+1)} - y_j^{(k+1)}\|_2 &\leq \|y_i^{(k)} - y_j^{(k)}\|_2 + \left\| h' \sum_{m:m \sim i} S_{im}^{(k)}(y_m^{(k)} - y_i^{(k)}) \right\|_2 + \left\| h' \sum_{m:m \sim j} S_{jm}^{(k)}(y_m^{(k)} - y_j^{(k)}) \right\|_2 \\ &\quad + \left\| h' \sum_{m:m \sim i} S_{im}^{(k)}(y_m^{(k)} - y_i^{(k)}) - h' \sum_{m:m \sim j} S_{jm}^{(k)}(y_m^{(k)} - y_j^{(k)}) \right\|_2, \end{aligned}$$

where

$$\left\| h' \sum_{m:m \sim i} S_{im}^{(k)}(y_m^{(k)} - y_i^{(k)}) \right\|_2 \lesssim h'(\|\mathbf{P}^*\|_\infty + n^{-2})n_r \max_{i \sim j} \|y_i^{(k)} - y_j^{(k)}\|_2,$$

and, by Proposition 12,

$$\begin{aligned} &\left\| h' \sum_{m:m \sim i} S_{im}^{(k)}(y_m^{(k)} - y_i^{(k)}) - h' \sum_{m:m \sim j} S_{jm}^{(k)}(y_m^{(k)} - y_j^{(k)}) \right\|_2 \\ &\lesssim h'R \left\| \sum_{m \in H_s} [S_{im}^{(k)}(y_m^{(k)} - y_i^{(k)}) - S_{jm}^{(k)}(y_m^{(k)} - y_j^{(k)})] \right\|_2 \\ &\lesssim \frac{h'Rn_s}{n^2} \max_{i \sim j} \|y_j^{(k)} - y_i^{(k)}\|_2. \end{aligned}$$

Hence, we have the key iteration formula

$$\max_{i \sim j} \|y_i^{(k+1)} - y_j^{(k+1)}\|_2 \leq \left[1 + C \left(h'(\|\mathbf{P}^*\|_\infty + n^{-2})n_r + \frac{h'Rn_s}{n^2} \right) \right] \max_{i \sim j} \|y_i^{(k)} - y_j^{(k)}\|_2, \quad (61)$$

and

$$\max_{i,j \in [n]} \|y_i^{(k+1)} - y_j^{(k+1)}\|_2 \leq \left[1 + C \left(h'(\|\mathbf{P}^*\|_\infty + n^{-2})n_r + \frac{h'Rn_s}{n^2} \right) \right] \max_{i,j \in [n]} \|y_i^{(k)} - y_j^{(k)}\|_2, \quad (62)$$

Apply (61) and (62) iteratively, we have

$$\max_{i \sim j} \|y_i^{(k_0+k)} - y_j^{(k_0+k)}\|_2 \leq \left[1 + C \left(h'(\|\mathbf{P}^*\|_\infty + n^{-2})n_r + \frac{h'Rn_s}{n^2} \right) \right]^k \max_{i \sim j} \|y_i^{(k_0)} - y_j^{(k_0)}\|_2, \quad (63)$$

and

$$\max_{i,j \in [n]} \|y_i^{(k_0+k)} - y_j^{(k_0+k)}\|_2 \leq \left[1 + C \left(h'(\|\mathbf{P}^*\|_\infty + n^{-2})n_r + \frac{h'Rn_s}{n^2} \right) \right]^k \max_{i,j \in [n]} \|y_i^{(k_0)} - y_j^{(k_0)}\|_2, \quad (64)$$

Therefore, as long as $kh'((\|\mathbf{P}^*\|_\infty + n^{-2})n + \frac{R}{n}) = O(1)$, we have

$$\max_{i \sim j} \|y_i^{(k_0+k)} - y_j^{(k_0+k)}\|_2 \lesssim \max_{i \sim j} \|y_i^{(k_0)} - y_j^{(k_0)}\|_2, \quad (65)$$

and

$$\max_{i,j \in [n]} \|y_i^{(k_0+k)} - y_j^{(k_0+k)}\|_2 \lesssim \max_{i,j \in [n]} \|y_i^{(k_0)} - y_j^{(k_0)}\|_2, \quad (66)$$

Similarly, for $i \asymp j$, we also have

$$\begin{aligned} \|y_i^{(k+1)} - y_j^{(k+1)}\|_2 &\geq \|y_i^{(k)} - y_j^{(k)}\|_2 - \left\| h' \sum_{m:m \sim i} S_{im}^{(k)}(y_m^{(k)} - y_i^{(k)}) \right\|_2 - \left\| h' \sum_{m:m \sim j} S_{jm}^{(k)}(y_m^{(k)} - y_j^{(k)}) \right\|_2 \\ &\quad - \left\| h' \sum_{m:m \asymp i} S_{im}^{(k)}(y_m^{(k)} - y_i^{(k)}) - h' \sum_{m:m \asymp j} S_{jm}^{(k)}(y_m^{(k)} - y_j^{(k)}) \right\|_2. \end{aligned}$$

If $\max_{i \sim j} \|y_j^{(k)} - y_i^{(k)}\|_2 \lesssim \min_{i \asymp j} \|y_j^{(k)} - y_i^{(k)}\|_2$, we have

$$\|y_i^{(k+1)} - y_j^{(k+1)}\|_2 \geq \|y_i^{(k)} - y_j^{(k)}\|_2 - C \left(h'(\|\mathbf{P}^*\|_\infty + n^{-2})n_r + \frac{h'Rn_s}{n^2} \right) \min_{i \sim j} \|y_j^{(k)} - y_i^{(k)}\|_2, \quad (67)$$

or

$$\min_{i \sim j} \|y_i^{(k+1)} - y_j^{(k+1)}\|_2 \geq \left[1 - C \left(h'(\|\mathbf{P}^*\|_\infty + n^{-2})n_r + \frac{h'Rn_s}{n^2} \right) \right] \min_{i \sim j} \|y_j^{(k)} - y_i^{(k)}\|_2. \quad (68)$$

As long as $h'(\|\mathbf{P}^*\|_\infty + n^{-2})n + \frac{h'R}{n} \leq c$ for some small constant $c > 0$, we have

$$\min_{i \sim j} \|y_i^{(k+1)} - y_j^{(k+1)}\|_2 \gtrsim \min_{i \sim j} \|y_j^{(k)} - y_i^{(k)}\|_2 \gtrsim \max_{i \sim j} \|y_j^{(k)} - y_i^{(k)}\|_2 \gtrsim \max_{i \sim j} \|y_j^{(k+1)} - y_i^{(k+1)}\|_2.$$

Thus, if $\max_{i \sim j} \|y_j^{(K_0)} - y_i^{(K_0)}\|_2 \lesssim \min_{i \asymp j} \|y_j^{(K_0)} - y_i^{(K_0)}\|_2$, we can also apply (68) iteratively, to have

$$\min_{i \sim j} \|y_i^{(K_0+k)} - y_j^{(K_0+k)}\|_2 \geq \left[1 - C \left(h'(\|\mathbf{P}^*\|_\infty + n^{-2})n_r + \frac{h'Rn_s}{n^2} \right) \right]^k \min_{i \sim j} \|y_j^{(K_0)} - y_i^{(K_0)}\|_2. \quad (69)$$

Under the same condition that $kh'(\|\mathbf{P}^*\|_\infty + n^{-2})n + \frac{R}{n} = O(1)$, we have

$$\min_{i \sim j} \|y_i^{(K_0+k)} - y_j^{(K_0+k)}\|_2 \gtrsim \min_{i \asymp j} \|y_j^{(K_0)} - y_i^{(K_0)}\|_2. \quad (70)$$

By the above arguments, we only need to show that

$$\max_{i \sim j} \|y_i^{(k)} - y_j^{(k)}\|_2 \ll n^{-1} \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2, \quad (71)$$

$$\min_{i \sim j} \|y_i^{(k)} - y_j^{(k)}\|_2 \gtrsim n^{-1} \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2 \quad (72)$$

hold for $k = K_0$. Then it suffices to set $V_n = Rn \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2$ and $B_n = n^{-1} \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2$ and $B_n \gg V_n$ holds naturally.

To see (71), note that for $k = K_0$,

$$\max_{i \sim j} \|y_i^{(k)} - y_j^{(k)}\|_2 \lesssim \max_{i \sim j} |y_{\ell i}^{(k)} - y_{\ell j}^{(k)}| \lesssim \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(k)} - \mathbf{Oz}_\ell\|_\infty \lesssim Rn \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2,$$

which implies (71) and $V_n = Rn \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2$ by previous assumption $nRn = o(1)$.

To see (72), note that for $i \asymp j$ such that $i \in H_r$ and $j \in H_s$,

$$\|y_i^{(k)} - y_j^{(k)}\|_2 / \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2 \geq \|D_{rs}\|_2 / \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2 - \|y_i^{(k)} - y_j^{(k)} - D_{rs}\|_2 / \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2 \geq cn^{-1},$$

and that $B_n = n^{-1} \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2$, where the last inequality follows from (54) and

$$\|y_i^{(k)} - y_j^{(k)} - D_{rs}\|_2 / \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2 \lesssim \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(k)} - O\mathbf{z}_\ell\|_\infty / \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2 = o(n^{-1})$$

Finally, to see (58), based on a similar argument we have

$$\max_{i,j \in [n]} \|y_i^{(k)} - y_j^{(k)}\|_2 / \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2 \leq \|D_{rs}\|_2 / \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2 + \|y_i^{(k)} - y_j^{(k)} - D_{rs}\|_2 / \max_{\ell \in [2]} \|\mathbf{y}_\ell^{(0)}\|_2 \leq cn^{-1},$$

where $i \in H_s$ and $j \in H_r$.

C.3 Proof of Theorem 14

By construction, we have $\|\mathbf{y}_1^{(0)}\|_2 = \|\mathbf{y}_2^{(0)}\|_2 = \sigma_n$, and

$$\max_{i \in [n], \ell \in [2]} |y_{i\ell}^{(0)}| \lesssim \frac{\sigma_n \Phi(1-\delta)}{\sqrt{n}} = o(n^{-1/2}) \quad (73)$$

with probability at least $1 - \delta$. Finally, note that by Theorem 7,

$$z_{\ell r} = \boldsymbol{\theta}_r^\top \mathbf{y}_\ell^{(0)} / \sqrt{n_r} = \frac{1}{n_r} \sum_{i \in H_r} y_{i\ell}^{(0)}. \quad (74)$$

Then

$$n|z_{\ell i} - z_{\ell j}| / \|\mathbf{y}_\ell^{(0)}\|_2 = \sigma_n^{-1} n \left| \frac{1}{n_i} \sum_{k \in H_i} y_{k\ell}^{(0)} - \frac{1}{n_j} \sum_{k \in H_j} y_{k\ell}^{(0)} \right| = n \left| \frac{1}{n_i} \sum_{k \in H_i} g_{k\ell}^{(0)} - \frac{1}{n_j} \sum_{k \in H_j} g_{k\ell}^{(0)} \right| / \|\mathbf{g}_\ell\|_2.$$

Note that $\frac{\sqrt{n}}{n_i} \sum_{k \in H_i} g_{k\ell}^{(0)}$ and $\frac{\sqrt{n}}{n_j} \sum_{k \in H_j} g_{k\ell}^{(0)}$ are independent centered random variables with variances n^2/n_i^2 and n^2/n_j^2 respectively. There exist some constants (C, δ) such that

$$C^{-1} \leq \left| \frac{\sqrt{n}}{n_i} \sum_{k \in H_i} g_{k\ell}^{(0)} - \frac{\sqrt{n}}{n_j} \sum_{k \in H_j} g_{k\ell}^{(0)} \right| \leq C,$$

with probability at least $1 - \delta$. Now since $1/2 \leq \|\mathbf{g}_\ell\|_2^2/n \leq 2$ with probability at least $1 - n^{-c}$. Then by combining the above two results, for sufficiently large n , we have

$$\frac{1}{C\sqrt{2}} \leq n|z_{\ell i} - z_{\ell j}| / \|\mathbf{y}_\ell^{(0)}\|_2 \leq C\sqrt{2} \quad (75)$$

with probability at least $1 - 2\delta$. This proves the theorem.

C.4 Proof of Theorem 15

Define $i_0^{(k)} = \arg \max_i y_{i1}^{(k)}$ and $i_1^{(k)} = \arg \min_i y_{i1}^{(k)}$. For simplicity, we drop the superscript (k) in i_0 and i_1 when there is no risk of confusion. In general, it suffices to show that

$$y_{i_0\ell}^{(k+1)} > y_{i_0\ell}^{(k)}, \quad y_{i_1\ell}^{(k+1)} < y_{i_1\ell}^{(k)} \quad (76)$$

for $\ell = 1, 2$, at each iteration. Without loss of generality, we only show that $y_{i_01}^{(k+1)} > y_{i_01}^{(k)}$ as the proofs of the other results are the same. Note that

$$\begin{aligned} y_{i_01}^{(k+1)} &= y_{i_01}^{(k)} + h' \sum_{j:j \approx i_0} S_{i_0j}^{(k)} (y_{j1}^{(k)} - y_{i_01}^{(k)}) + h' \sum_{j:j \sim i_0} S_{i_0j}^{(k)} (y_{j1}^{(k)} - y_{i_01}^{(k)}) \\ &\equiv y_{i_01}^{(k)} + F_{i_01}^{(k)} + E_{i_01}^{(k)}. \end{aligned} \quad (77)$$

By definition of i_0 and Proposition 12, we have $y_{j1}^{(k)} - y_{i_01}^{(k)} < 0$ and $S_{i_0j}^{(k)} < 0$ for all $j \approx i$, so that $F_{i_01}^{(k)} > 0$. Moreover, since

$$F_{i_01}^{(k)} \gtrsim \frac{h'}{n} \min_{j \approx i} |y_{j1}^{(k)} - y_{i_01}^{(k)}|,$$

and

$$|E_{i_01}^{(k)}| \lesssim \max_{\substack{j \neq i_0 \\ j \sim i_0}} |y_{j1}^{(k)} - y_{i_01}^{(k)}| \cdot h' \sum_{\substack{j \neq i_0 \\ j \sim i_0}} S_{i_0j}^{(k)} \lesssim h'n \|\mathbf{P}^*\|_\infty \max_{i \sim j} |y_{j1}^{(k)} - y_{i_01}^{(k)}|,$$

by the same argument that leads to (55) and (56) in the proof of Theorem 13, we have $\max_{i \sim j} |y_{j1}^{(k)} - y_{i_01}^{(k)}| \ll \min_{j \approx i} |y_{j1}^{(k)} - y_{i_01}^{(k)}|$. Then, in equation (77), we have $F_{i_01}^{(k)} \gg |E_{i_01}^{(k)}|$ under the condition that $\|\mathbf{P}^*\|_\infty \lesssim n^{-2}$.

Appendix D. Analysis of Two Examples

D.1 Proofs of the Gaussian Mixture Model

For given $\{z_i\}_{1 \leq i \leq n}$, we define the equivalence relationship over $[n]$ such that $i \sim j$ whenever $z_i = z_j$. Thus, for given $\{X_i\}_{1 \leq i \leq n}$, we can define the symmetric matrix $\mathbf{P}^* = (p_{ij}^*) \in \mathbb{R}^{n \times n}$ such that $p_{ij}^* = p_{ij}$ if $i \sim j$, and $p_{ij}^* = 0$ otherwise. The following proposition concerns properties of the similarity matrix \mathbf{P} under the Gaussian mixture model.

Proposition 22 *Under conditions of Corollary 17, we have*

$$P\left(\|\mathbf{P}\|_\infty \lesssim \frac{1}{n^2}\right) \geq 1 - \frac{1}{n^c}, \quad (78)$$

and the following events

$$\mathcal{B}_1 = \left\{ \|\mathbf{L}(\mathbf{P} - \mathbf{P}^*)\| \lesssim \frac{1}{n} e^{-c\rho^2/p}, \|\mathbf{P} - \mathbf{P}^*\|_\infty \lesssim \frac{1}{n^2} e^{-c\rho^2/p} \right\}, \quad (79)$$

$$\mathcal{B}_2 = \left\{ \min_{i \sim j} p_{ij}^* \gtrsim \frac{1}{n^2}, \lambda_{R+1}(\mathbf{L}(\mathbf{P}^*)) \asymp \|\mathbf{L}(\mathbf{P}^*)\| \asymp \frac{1}{n} \right\}, \quad (80)$$

hold with probability at least $1 - n^{-c}$.

Proof Firstly, we define the Gaussian kernel matrix associated with the data points $\{X_i\}_{1 \leq i \leq n}$ as

$$\mathbf{K} = (K(X_i, X_j))_{1 \leq i, j \leq n}, \quad K(X_i, X_j) = \exp\left(-\frac{\|X_i - X_j\|_2^2}{2\tau_i^2}\right). \quad (81)$$

Let K_{ij} be the (i, j) -th entry of \mathbf{K} . By the definition of p_{ij} , for each pair $\{i, j\} \subset \{1, \dots, n\}$ such that $i \neq j$, we define the map $F_{ij} : \mathbb{R}_+^{n \times n} \rightarrow (0, 1)$ where

$$p_{ij} = F_{ij}(\mathbf{K}) = \frac{K_{ij}}{2n \sum_{\ell \neq i} K_{i\ell}} + \frac{K_{ji}}{2n \sum_{\ell \neq j} K_{j\ell}}.$$

To show (78), it suffices to show $\sum_{\ell \neq i} K_{i\ell} \gtrsim n$, with high probability, as by definition $K_{ij} \leq 1$. This is done in the following lemma.

Lemma 23 *Under conditions of Corollary 17, for any $i \sim j$,*

$$P(\min\{K_{ij}, \mathbb{E}K_{ij}\} \gtrsim c') \geq 1 - n^{-c}, \quad (82)$$

and, for any given $i \in \{1, \dots, n\}$,

$$P(|\{s \in \{1, \dots, n\} : s \sim i\}| \geq c'n) \geq 1 - e^{-cn}. \quad (83)$$

Next, to show (79), we only need to obtain an upper bound for $\max_{i \asymp j} K_{ij}$, or a lower bound for $\max_{i \asymp j} \|X_i - X_j\|_2^2$. We write $X_i = \mu_{z_i} + \Sigma^{1/2}W_i$ where $W_i \sim N(0, I_p)$ so that

$$\|X_i - X_j\|_2^2 = \|\mu_{z_i} - \mu_{z_j}\|_2^2 + (W_i - W_j)^\top \Sigma (W_i - W_j) + 2(\mu_{z_i} - \mu_{z_j})^\top \Sigma^{1/2}(W_i - W_j).$$

On the one hand, by the Hanson-Wright inequality (Rudelson and Vershynin, 2013), we have, for $t \gtrsim p$,

$$P(|(W_i - W_j)^\top \Sigma (W_i - W_j) - 2\text{tr}(\Sigma)| > t) \leq 2e^{-ct}.$$

On the other hand, standard concentration inequality for sub-Gaussian random variables indicates

$$P(|2(\mu_{z_i} - \mu_{z_j})^\top \Sigma^{1/2}(W_i - W_j)| > C\|\mu_{z_i} - \mu_{z_j}\|_2 \sqrt{t}) \leq 2e^{-ct}.$$

By choosing $t = C \max\{p, \log n\}$ in the above inequalities, we have

$$P(\|X_i - X_j\|_2^2 \geq [\|\mu_{z_i} - \mu_{z_j}\|_2 + C \max\{\sqrt{p}, \sqrt{\log n}\}]^2) \leq n^{-c}.$$

Now since $\rho^2 \gtrsim \max\{p, \log n\}$, we have $P(\|X_i - X_j\|_2^2 \geq \|\mu_{z_i} - \mu_{z_j}\|_2^2) \leq n^{-c}$. In other words, under the same event, we have

$$P\left(\max_{i \asymp j} K_{ij} \lesssim e^{-c\rho^2 / \max\{p, \log n\}}\right) \geq 1 - n^{-c}.$$

Finally, to show (80), we define the matrix $\mathbf{P}_0 \in \mathbb{R}^{n \times n}$ such that $p_{0,ij} = p_{ij}^*$ for $i \asymp j$, and $p_{0,ij} = F_{ij}(\mathbb{E}[\mathbf{K}|z_1, \dots, z_n])$ otherwise. Conditional on $\{z_i\}$, \mathbf{P}_0 is a block-wise constant matrix. Then $\lambda_{R+1}(\mathbf{L}(\mathbf{P}_0)) = \|\mathbf{L}(\mathbf{P}_0)\| \asymp n^{-1}$. To see this, we need the following lemma.

Lemma 24 *Under conditions of Corollary 17, for any pair $\{i, j\} \subset \{1, \dots, p\}$ such that $i \sim j$, we have*

$$p_{0,ij} \asymp \frac{1}{n^2}, \quad (84)$$

with probability at least $1 - e^{-cn}$.

The following lemma concerns the relation between \mathbf{P}^* and \mathbf{P}_0 .

Lemma 25 *Under conditions of Corollary 17, with probability at least $1 - n^{-c} - n^2 e^{-cn}$ for some constant $c > 0$, we have*

$$\|\mathbf{P}^* - \mathbf{P}_0\|_\infty \lesssim \frac{1}{n^2} \|\mathbf{K} - \mathbb{E}[\mathbf{K}|z_1, \dots, z_n]\|_\infty. \quad (85)$$

Moreover, with probability at least $1 - n^2(n^{-c} + e^{-cn} + 2e^{-t^2})$, we have

$$\|\mathbf{P}^* - \mathbf{P}_0\|_\infty \lesssim \frac{t}{n^2 \max\{\sqrt{p}, \sqrt{\log n}\}}. \quad (86)$$

Consequently, we have

$$P\left(\|\mathbf{P}^* - \mathbf{P}_0\|_\infty \lesssim \frac{\sqrt{\log n}}{n^2 \max\{\sqrt{p}, \sqrt{\log n}\}}\right) \geq 1 - n^{-c} \quad (87)$$

Finally, in order to show (80), it suffices to use Weyl's inequality $|\lambda_{R+1}(\mathbf{L}(\mathbf{P}_0)) - \lambda_{R+1}(\mathbf{L}(\mathbf{P}^*))| \leq \|\mathbf{L}(\mathbf{P}_0) - \mathbf{L}(\mathbf{P}^*)\|$. This proves the proposition. \blacksquare

The verification of (T1.D) and (T2.D) is straightforward. To check (T1.E) (T2.E) and (T3.E), we note that $\alpha h K_0 \|\mathbf{L}(\mathbf{P}^* - \mathbf{P})\| \lesssim K_0 e^{-c\rho^2/\max\{p, \log n\}}$. Hence, for (T1.E) to hold, we need $K_0 n e^{-c\rho^2/\max\{p, \log n\}} = o(1)$. Moreover, the condition $\text{diam}(\{y_i^{(K_0+K_1)}\}_{1 \leq i \leq n}) = o(1)$ follows from Theorem 15 and the condition $h'K_1 = O(n)$.

D.2 Proof of the Noisy Nested Sphere Model

Similarly, we define the symmetric matrix $\mathbf{P}^* = (p_{ij}^*) \in \mathbb{R}^{n \times n}$ such that $p_{ij}^* = p_{ij}$ if $i \sim j$, and $p_{ij}^* = 0$ otherwise. The following proposition concerns properties of the similarity matrix \mathbf{P} under the noisy nested sphere model.

Proposition 26 *Under the conditions of Corollary 19, we have*

$$P\left(\|\mathbf{P}\|_\infty \lesssim \frac{1}{\gamma n^2}\right) \geq 1 - \frac{1}{n^{c'}}, \quad (88)$$

$$P\left(\|\mathbf{P} - \mathbf{P}^*\|_\infty \lesssim \frac{1}{\gamma n^2} \exp\left\{-\min_{r \in [R-1]} \frac{c(\rho_{r+1} - \rho_r)^2}{\gamma \rho_{r+1}^2}\right\}\right) \geq 1 - \frac{1}{n^{c'}}, \quad (89)$$

$$P\left(\|\mathbf{L}(\mathbf{P} - \mathbf{P}^*)\| \lesssim \frac{1}{\gamma n} \exp\left\{-\min_{r \in [R-1]} \frac{c(\rho_{r+1} - \rho_r)^2}{\gamma \rho_{r+1}^2}\right\}\right) \geq 1 - \frac{1}{n^{c'}}. \quad (90)$$

Proof As in the proof of Proposition 22, we note that $p_{ij} = F_{ij}(\mathbf{K}) = \frac{K_{ij}}{2n \sum_{\ell \neq i} K_{i\ell}} + \frac{K_{ji}}{2n \sum_{\ell \neq j} K_{j\ell}}$. Then we need to find lower bound for $\sum_{\ell \neq j} K_{j\ell}$ and $\sum_{\ell \neq i} K_{i\ell}$, as $K_{ij} \leq 1$.

Lemma 27 *Under the conditions of Corollary 19, conditional on the event in (83), for any $i \in [1 : n]$*

$$P\left(\sum_{j:j \sim i} \min\{K_{ij}, \mathbb{E}K_{ij}\} \geq C\gamma n\right) \geq 1 - n^{-c}. \quad (91)$$

Now, for $i \asymp j$, we have

$$\begin{aligned} \|X_i - X_j\|_2 &= \left\| \mu_i - \mu_j + \frac{\mu_i}{\|\mu_i\|_2} \xi_i - \frac{\mu_j}{\|\mu_j\|_2} \xi_j \right\|_2 \\ &= \|\theta_i(\rho_{z_i} + \xi_i) - \theta_j(\rho_{z_j} + \xi_j)\|_2 \\ &\geq |(\rho_{z_i} + \xi_i) - (\rho_{z_j} + \xi_j)| \\ &\geq |\rho_{z_i} - \rho_{z_j}| - |\xi_i - \xi_j| \\ &\geq |\rho_{z_i} - \rho_{z_j}| - \sigma\sqrt{\log n} \end{aligned}$$

Now as long as $\min_{i \asymp j} |\rho_{z_i} - \rho_{z_j}| \geq C\sigma\sqrt{\log n}$ for some sufficiently large C , we have

$$\|X_i - X_j\|_2 \geq C|\rho_{z_i} - \rho_{z_j}|.$$

Hence, with probability at least $1 - n^{-c}$, we have

$$\max_{i \asymp j} (K_{ij} + K_{ji}) \lesssim \exp\left(-\frac{C(\rho_{z_i} - \rho_{z_j})^2}{\tau_i^2}\right) + \exp\left(-\frac{C(\rho_{z_i} - \rho_{z_j})^2}{\tau_j^2}\right) \quad (92)$$

$$\lesssim \exp\left(-\frac{C(\rho_{z_i} - \rho_{z_j})^2}{\max\{\tau_i^2, \tau_j^2\}}\right) \quad (93)$$

so that under the same event, for $i \asymp j$, we have

$$p_{ij} \lesssim \frac{K_{ij} + K_{ji}}{\gamma n^2} \lesssim \frac{1}{\gamma n^2} \exp\left(-\frac{C(\rho_{z_i} - \rho_{z_j})^2}{\max\{\tau_i^2, \tau_j^2\}}\right),$$

or (89). Finally, note that for $C' \leq C/c$,

$$\min_{i \asymp j} \frac{C(\rho_{z_i} - \rho_{z_j})^2}{\max\{\tau_i^2, \tau_j^2\}} = \min_{i \asymp j} \frac{C(\rho_{z_i} - \rho_{z_j})^2}{C'\gamma \max\{\rho_{z_i}^2, \rho_{z_j}^2\}} \geq \min_{r \in [R-1]} \frac{c(\rho_{r+1} - \rho_r)^2}{\gamma \rho_{r+1}^2}.$$

This along with the fact $\|\mathbf{L}(\mathbf{P}^* - \mathbf{P})\| \lesssim n\|\mathbf{P}^* - \mathbf{P}\|_\infty$ implies (90). \blacksquare

Now we check (T1.D) and (T2.D). Specifically, we need $K_0 \rightarrow \infty$, $K_0 h = o(n)$, $n h \alpha \|\mathbf{P}\|_\infty = O(1)$, $K_0 h \alpha \|\mathbf{L}(\mathbf{P} - \mathbf{P}^*)\| = o(1)$, $1/2 + h/n \leq h \alpha \lambda_{R+1}(\mathbf{L}(\mathbf{P}^*)) \leq h \alpha \|\mathbf{L}(\mathbf{P}^*)\| \leq 1 + h/n$ and $K_0 h (n \alpha \|\mathbf{P}\|_\infty + 1/n) \max_{i \in [n], \ell \in [2]} |y_{i\ell}^{(0)}|^2 = o(1)$. To have these conditions hold with probability at least $1 - n^{-c}$, by Proposition 26, we need

$$K_0 h = o(n), \quad \frac{h \alpha}{\gamma n} = O(1), \quad \frac{K_0 h \alpha}{\gamma n} \exp\left\{-\min_{r \in [R-1]} \frac{c(\rho_{r+1} - \rho_r)^2}{\gamma \rho_{r+1}^2}\right\} = o(1),$$

$$0.5 \leq h\alpha\lambda_{R+1}(\mathbf{L}(\mathbf{P})) \leq \frac{h\alpha}{\gamma n} \leq 1.5, \quad (94)$$

and

$$K_0 h(\alpha/\gamma + 1)\sigma_n^2 \log n = o(n^2), \quad K_0 \rightarrow \infty.$$

This proves the first statement. To show (T1.E) (T2.E) and (T3.E) hold, we only need to check the following conditions

$$n(1 - \kappa)^{K_0} + hK_0[(\alpha n\|\mathbf{P}^*\|_\infty + 1/n)\sigma_n^2 \log n + \alpha n\|\mathbf{L}(\mathbf{P}^* - \mathbf{P})\|](1 + n^2\|\mathbf{P}^*\|_\infty) = o(1),$$

and

$$K_1 h'(n\|\mathbf{P}^*\|_\infty + 1/n) = O(1), \quad n^2\|\mathbf{P} - \mathbf{P}^*\|_\infty = o(1).$$

Again, by Proposition 26, the above conditions hold with probability at least $1 - n^{-c}$ if $K_0 \gg \log n$,

$$n(1 - \kappa)^{K_0} + hK_0 \left[(\alpha/\gamma + 1)\sigma_n^2 \log n/n + \frac{\alpha}{\gamma} \exp \left\{ - \min_{r \in [R-1]} \frac{c(\rho_{r+1} - \rho_r)^2}{\gamma \rho_{r+1}^2} \right\} \right] (1 + 1/\gamma) = o(1),$$

and

$$K_1 h'(1/\gamma + 1) = O(n), \quad \frac{1}{\gamma} \exp \left\{ - \min_{r \in [R-1]} \frac{c(\rho_{r+1} - \rho_r)^2}{\gamma \rho_{r+1}^2} \right\} = o(1).$$

This completes the proof of the corollary.

Appendix E. Proof of Auxiliary Lemmas

E.1 Proof of Lemma 21

Note that $\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n) = \mathbf{L}(\alpha\mathbf{P}) - \mathbf{L}(\mathbf{H}_n) = \mathbf{L}(\alpha\mathbf{P}) + \frac{1}{n(n-1)}\mathbf{1}\mathbf{1}^\top - \frac{1}{n-1}\mathbf{I}_n$, and the Laplacian $\mathbf{L}(\alpha\mathbf{P})$ is positive semi-definite (as a result of the Geršgorin circle theorem (Varga, 2010) and that $\mathbf{L}(\alpha\mathbf{P})$ is a symmetric diagonally dominant matrix with real non-negative diagonal entries) and has the smallest eigenvalue $\lambda_1 = 0$ with an eigenvector $n^{-1/2}\mathbf{1}$. Then, if $\lambda_1 \leq \dots \leq \lambda_n$ are the eigenvalues of $\mathbf{L}(\alpha\mathbf{P})$, the eigenvalues of $\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)$ are $(\lambda_1, \lambda_2 - (n-1)^{-1}, \dots, \lambda_n - (n-1)^{-1})$. Consequently, the smallest eigenvalue of $h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)$ is $\min\{0, h(\lambda_2 - (n-1)^{-1})\} \in [-\frac{h}{n-1}, 0]$. On the other hand, we also have $\|h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)\| \leq \|h\mathbf{L}(\alpha\mathbf{P})\| < 2$. Then, it follows that $1 \leq \|\mathbf{I} - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)\| \leq 1 + \frac{h}{n-1}$.

E.2 Proof of Lemma 23

Proof of (82). On the one hand, if we let $\Sigma = U\Lambda U^\top$ be the eigen-decomposition of Σ , where $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$, then in light of (98) below,

$$A = \frac{1}{4}U\left(\frac{2}{\tau^2}\mathbf{I} + \Lambda^{-1}\right)U^\top = \frac{1}{4}U\text{diag}(2/\tau^2 + \lambda_1^{-1}, \dots, 2/\tau^2 + \lambda_p^{-1})U^\top, \quad (95)$$

and therefore

$$|2A| \cdot |2\Sigma| = \prod_{i=1}^p \left(\frac{1}{\tau^2} + \frac{1}{2\lambda_i} \right) (2\lambda_i) = \prod_{i=1}^p \left(\frac{2\lambda_i}{\tau^2} + 1 \right) \leq \left(\frac{2\lambda_1}{\tau^2} + 1 \right)^p \lesssim \left(\frac{C}{p} + 1 \right)^p \leq C'.$$

In other words, we have shown

$$\mathbb{E}K_{ij} \gtrsim c', \quad \text{if } i \sim j. \quad (96)$$

On the other hand, we show that

$$P\left(\frac{\|X_i - X_j\|_2^2}{2\tau^2} \leq C\right) \leq 1 - n^{-c}, \quad \text{if } i \sim j. \quad (97)$$

To see (97), note that $X_i - X_j \sim N(0, 2\Sigma)$, by the Hanson-Wright inequality (Rudelson and Vershynin, 2013), for $t \gtrsim \|\Sigma\|_F^2 / \|\Sigma\|_2 \asymp p$, we have $P(\|X_i - X_j\|_2^2 \geq C(\text{tr}(\Sigma) + t)) \leq e^{-ct/\|\Sigma\|_2}$. Whenever $\text{tr}(\Sigma) \lesssim p$ and $\|\Sigma\|_2 \leq C$, by setting $t \asymp \max\{p, \log n\}$, we have $P(\|X_i - X_j\|_2^2 \geq C(p + \log n)) \leq n^{-c}$. This implies (97) if we choose $\tau^2 \asymp \max\{p, \log n\}$. Combining (96) and (97), we have (82).

Proof of (83). It is equivalent to show that, for any $x_i \stackrel{i.i.d.}{\sim} \text{Bernoulli}(\eta)$ where $i = 1, \dots, n$ and $p \in (0, 1)$, with probability at least $1 - e^{-cn}$, we have $\sum_{i=1}^n x_i \geq c'n$, for some constants $c, c' > 0$. By Hoeffding's inequality, for any $t \geq 0$, $P(\frac{1}{n} \sum_{i=1}^n x_i - \eta \geq t) \leq e^{-2nt^2}$, which implies $P\left(\sum_{i=1}^n x_i \geq (1 - \eta)n\right) \leq e^{-2n}$, if we set $t = 1$. This proves (83).

E.3 Proof of Lemma 24

Let $\mu_i = \mathbb{E}X_i$ and $\mu_j = \mathbb{E}X_j$. Since $Z_{ij} = X_i - X_j \sim N(\mu_i - \mu_j, 2\Sigma)$, we have

$$\begin{aligned} \mathbb{E}K_{ij} &= \mathbb{E} \exp(-Z_{ij}^\top Z_{ij} / 2\tau^2) \\ &= \frac{1}{(2\pi)^{p/2} |2\Sigma|^{1/2}} \int \exp(-Z_{ij}^\top Z_{ij} / 2\tau^2) \exp(-(Z_{ij} - \mu_i + \mu_j)^\top \Sigma^{-1} (z_{ij} - \mu_i + \mu_j) / 4) dZ_{ij} \\ &= \frac{1}{(2\pi)^{p/2} |2\Sigma|^{1/2}} \int \exp(-Z_{ij}^\top Z_{ij} / 2\tau^2 - (Z_{ij} - \mu_i + \mu_j)^\top \Sigma^{-1} (Z_{ij} - \mu_i + \mu_j) / 4) dZ_{ij} \\ &= \frac{1}{(2\pi)^{p/2} |2\Sigma|^{1/2}} \int \exp(-Z_{ij}^\top (\frac{2}{\tau^2} \mathbf{I} + \Sigma^{-1}) Z_{ij} / 4 + Z_{ij}^\top \Sigma^{-1} \Delta_{ij} / 2 - \Delta_{ij}^\top \Sigma^{-1} \Delta_{ij} / 4) dZ_{ij} \\ &= \frac{1}{(2\pi)^{p/2} |2\Sigma|^{1/2}} \exp(b^\top A^{-1} b / 4 - \Delta_{ij}^\top \Sigma^{-1} \Delta_{ij} / 4) \int \exp(-(Z_{ij} - A^{-1} b)^\top A (Z_{ij} - A^{-1} b)) dZ_{ij} \\ &= \frac{1}{|2A|^{1/2} |2\Sigma|^{1/2}} \exp(b^\top A^{-1} b / 4 - \Delta_{ij}^\top \Sigma^{-1} \Delta_{ij} / 4) \end{aligned} \quad (98)$$

where $A = (\frac{2}{\tau^2} \mathbf{I} + \Sigma^{-1}) / 4$ and $b = \Sigma^{-1} \Delta_{ij} / 2$. The results follows by noting that $\Delta_{ij} = 0$.

On the one hand, by (95), we have

$$|2A| \cdot |2\Sigma| = \prod_{i=1}^p \left(\frac{1}{\tau^2} + \frac{1}{2\lambda_i} \right) (2\lambda_i) = \prod_{i=1}^p \left(\frac{2\lambda_i}{\tau^2} + 1 \right) \leq \left(\frac{2\lambda_1}{\tau^2} + 1 \right)^p \lesssim \left(\frac{C}{p} + 1 \right)^p \leq C'.$$

On the other hand,

$$|2A| \cdot |2\Sigma| = \prod_{i=1}^p \left(\frac{1}{\tau^2} + \frac{1}{2\lambda_i} \right) (2\lambda_i) = \prod_{i=1}^p \left(\frac{2\lambda_i}{\tau^2} + 1 \right) \geq 1.$$

This implies $\mathbb{E}K_{ij} \asymp 1$.

To obtain bounds for $p_{0,ij}$, it suffices to see that $P(\sum_{\ell \neq i} \mathbb{E}K_{\ell i} \gtrsim n) \geq 1 - e^{-cn}$, which follows from (83).

E.4 Proof of Lemma 25

We start with the proof of (85). For any $\{i, j\} \subset \{1, \dots, n\}$ such that $i \neq j$, we have

$$\frac{\partial F_{ij}(\mathbf{K})}{\partial K_{kl}} = \begin{cases} \frac{\sum_{s \notin \{i, j\}} K_{is}}{2n(\sum_{s \neq i} K_{is})^2} + \frac{\sum_{s \notin \{i, j\}} K_{js}}{2n(\sum_{s \neq j} K_{js})^2}, & \text{if } \{k, \ell\} = \{i, j\} \\ -\frac{K_{kl}}{2n(\sum_{s \neq i} K_{is})^2}, & \text{if } i \in \{k, \ell\}, j \notin \{k, \ell\} \\ -\frac{K_{kl}}{2n(\sum_{s \neq j} K_{js})^2}, & \text{if } j \in \{k, \ell\}, i \notin \{k, \ell\} \\ 0, & \text{otherwise} \end{cases}. \quad (99)$$

For any $\mathbf{K}_1, \mathbf{K}_2 \in \mathbb{R}_+^{n \times n}$ and any $\{i, j\} \subset \{1, \dots, p\}$ such that $i \neq j$, we have

$$F_{ij}(\mathbf{K}_1) - F_{ij}(\mathbf{K}_2) = \sum_{1 \leq k \neq \ell \leq n} \frac{\partial F_{ij}(K^*)}{\partial K_{kl}} ([\mathbf{K}_1]_{kl} - [\mathbf{K}_2]_{kl}) \leq \|\mathbf{K}_1 - \mathbf{K}_2\|_\infty \sum_{1 \leq k \neq \ell \leq n} \left| \frac{\partial F_{ij}(\mathbf{K}^*)}{\partial K_{kl}} \right|,$$

where $\mathbf{K}^* = t\mathbf{K}_1 + (1-t)\mathbf{K}_2$ for some $t \in (0, 1)$.

In the following, we show that for $\mathbf{K}_1 = \mathbf{K}$ and $\mathbf{K}_2 = \mathbb{E}[\mathbf{K}|z_1, \dots, z_n]$, for given $i \neq j$, it holds that

$$\left| \frac{\partial F_{ij}(\mathbf{K}^*)}{\partial K_{kl}} \right| \lesssim \frac{1}{n^2}, \quad \text{for } \{k, \ell\} = \{i, j\}, \quad (100)$$

and

$$\left| \frac{\partial F_{ij}(\mathbf{K}^*)}{\partial K_{kl}} \right| \lesssim \frac{1}{n^3}, \quad \text{for } j \in \{k, \ell\}, i \notin \{k, \ell\} \text{ or } i \in \{k, \ell\}, j \notin \{k, \ell\}, \quad (101)$$

with probability at least $1 - e^{-cp} - e^{-cn}$ for some constant $c > 0$, and conclude that

$$\sum_{1 \leq k \neq \ell \leq n} \left| \frac{\partial F_{ij}(\mathbf{K}^*)}{\partial K_{kl}} \right| \lesssim \frac{1}{n^2}, \quad (102)$$

under the same event. This along with a union bound argument leads to (85). To show (100), we note that

$$\left| \frac{\partial F_{ij}(\mathbf{K})}{\partial K_{ij}} \right| = \frac{\sum_{s \notin \{i, j\}} K_{is}}{2n(\sum_{s \neq i} K_{is})^2} + \frac{\sum_{s \notin \{i, j\}} K_{js}}{2n(\sum_{s \neq j} K_{js})^2} \leq \frac{1}{2n \sum_{s \neq i} K_{is}} + \frac{1}{2n \sum_{s \neq j} K_{js}}.$$

Note that by (82), we have

$$\frac{1}{\sum_{s \neq i} K_{is}^*} + \frac{1}{\sum_{s \neq j} K_{js}^*} \lesssim \frac{1}{n},$$

with probability at least $1 - e^{-cp} - e^{-cn}$ for some constant $c > 0$. Therefore (100) holds with high probability. On the other hand, to show (101), it suffices to see that, for $j \in \{k, \ell\}, i \notin \{k, \ell\}$,

$$\left| \frac{\partial F_{ij}(\mathbf{K}^*)}{\partial K_{kl}} \right| \lesssim \frac{1}{n(\sum_{s \neq i} K_{is}^*)^2} \lesssim \frac{1}{n^3}$$

with probability at least $1 - e^{-cp} - e^{-cn}$. This completes the proof of (85).

Next we prove (86). The proof follows directly from

$$P(\|\mathbf{K} - \mathbb{E}[\mathbf{K}|z_1, \dots, z_n]\|_\infty \geq t / \max\{\sqrt{p}, \sqrt{\log n}\}) \leq 2n^2 e^{-t^2}. \quad (103)$$

Note that, the kernel function $K(X_i, X_j) : \mathbb{R}^{2p} \rightarrow \mathbb{R}$ is L -Lipschitz with respect to the ℓ_2 norm on \mathbb{R}^{2p} , with $L = \frac{\sqrt{2}}{e\tau} \asymp \frac{1}{\max\{\sqrt{p}, \sqrt{\log n}\}}$. By the concentration inequality for Lipschitz continuous functions (see, e.g., Theorem 4 of Amini and Razaee (2021)), it holds that

$$P(|K(X_i, X_j) - \mathbb{E}K(X_i, X_j)| \geq Ct / \max\{\sqrt{p}, \sqrt{\log n}\}) \leq 2e^{-ct^2}.$$

By applying the union bound, we have

$$\begin{aligned} P(\|\mathbf{K} - \mathbb{E}[\mathbf{K}|z_1, \dots, z_n]\|_\infty \geq Ct/\sqrt{p}) &= P\left(\max_{1 \leq i \neq j \leq n} |K(X_i, X_j) - \mathbb{E}K(X_i, X_j)| \geq Ct / \max\{\sqrt{p}, \sqrt{\log n}\}\right) \\ &\leq 2n^2 e^{-ct^2}. \end{aligned}$$

This completes the proof.

E.5 Proof of Lemma 27

Suppose $\{z_i\}$ and $\{\mu_i\}$ are given. Let $S_i(\gamma) = \{j \in [n] : j \sim i, \frac{\mu_i^\top \mu_j}{\|\mu_i\|_2 \|\mu_j\|_2} \geq 1 - \gamma\}$ for some $\gamma \in (0, 1)$. We show that for some properly chosen $\{\tau_i\}$,

$$\min\{K_{ij}, \mathbb{E}K_{ij}\} \geq C, \quad \text{for any } j \in S_i(\gamma), \quad (104)$$

for some constant $C > 0$ with probability at least $1 - n^{-c}$. Apparently, this leads to $\sum_{j:j \sim i} \min\{K_{ij}, \mathbb{E}K_{ij}\} \geq \sum_{j \in S_i(\gamma)} \min\{K_{ij}, \mathbb{E}K_{ij}\} \geq C|S_i(\gamma)|$, and we only need to show that $|S_i(\gamma)| \geq \gamma n$ with the claimed probability. We first show (104). Since for any $j \in S_i(\gamma)$, if we denote $\theta_i = \mu_i / \|\mu_i\|_2$, we have

$$\begin{aligned} \|X_i - X_j\|_2^2 &\lesssim \|\theta_i - \theta_j\|_2^2 \rho_{z_i}^2 + |\xi_i + \xi_j|^2 \\ &\lesssim (1 - \theta_i^\top \theta_j) \rho_{z_i}^2 + \sigma^2 \log n \\ &\lesssim \gamma \rho_{z_i}^2 + \sigma^2 \log n \end{aligned}$$

with probability at least $1 - n^{-c}$. Here we used the tail bound $P(|\xi_i + \xi_j| \leq C\sigma\sqrt{\log n}) \geq 1 - n^{-c}$. In particular, if $\gamma \min_{r \in [R]} \rho_r^2 \gg \sigma^2 \log n$, the above argument leads to

$$P\left(\min\{K_{ij}, \mathbb{E}K_{ij}\} \gtrsim \exp\left\{-C \frac{\gamma \rho_{z_i}^2}{\tau_i^2}\right\}\right) \geq 1 - n^{-c}. \quad (105)$$

Therefore, if we choose τ_i such that $\tau_i^2 \gtrsim \gamma \rho_{z_i}^2$, we have (104).

Secondly, we obtain lower bound for $|S_i(\gamma)|$. Note that $\{\theta_i\}$ are uniformly drawn on \mathbb{S}^{p-1} . It follows from the spherical area formula that $|S_i(\gamma)|$ is a binomial random variable with distribution $\text{Bin}(n, \gamma/2)$, so that $P\left(|S_i| - \frac{n\gamma}{2} \lesssim t\sqrt{\gamma(1-\gamma)n}\right) \geq 1 - e^{-t^2}$. By choosing $t = C\sqrt{\log n}$, we have $P(|S_i| \geq n\gamma/2 - C\sqrt{\gamma n \log n}) \geq 1 - n^{-c}$. If $\gamma \gtrsim \log n/n$, we have $P(|S_i| \gtrsim n\gamma) \geq 1 - n^{-c}$. This proves the lemma.

Appendix F. Supplementary Figures

This section includes additional figures from the numerical studies presented in Sections 4 and 5. Specifically, Figure 7 contains the final t-SNE embeddings of the model-generated samples as described in Section 4, but with different tuning parameters where $\delta = 1/2$. Figure 8 shows that when the separation condition $\rho^2 \gg p$ is slightly violated, t-SNE is still able to visualize clusters from the Gaussian mixture model, which demonstrates the robustness of t-SNE with respect to the separation condition.

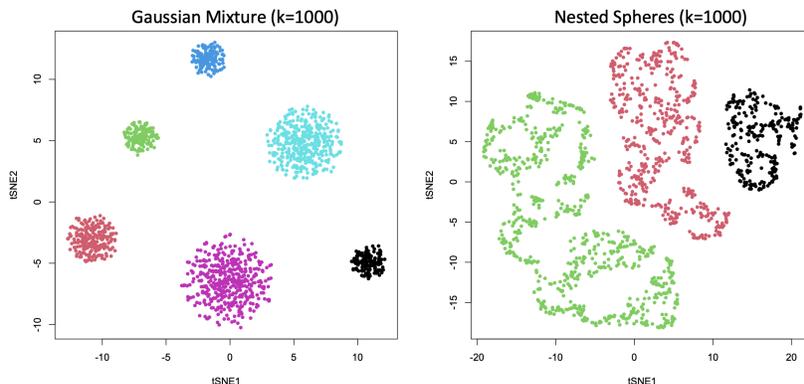


Figure 7: The final t-SNE embeddings of the model-generated samples as described in Section 4, using the tuning parameters with $\delta = 1/2$.

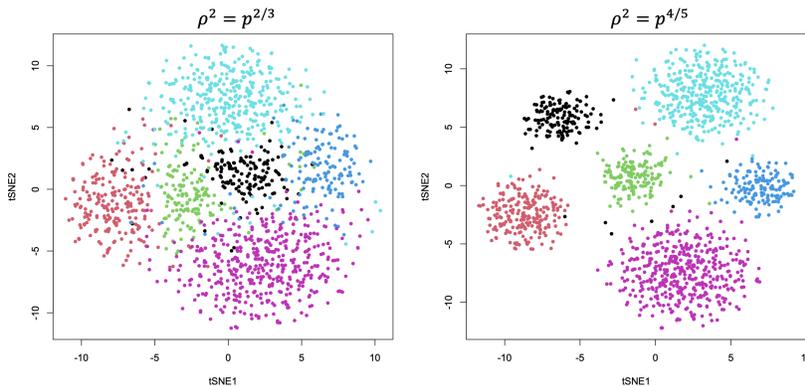


Figure 8: The final t-SNE embeddings of the samples generated from the Gaussian mixture model with separation $\rho^2 = p^{2/3}$ (left) and $\rho^2 = p^{4/5}$ (right), using the tuning parameters with $\delta = 1/3$.

Figure 9 is a heatmap of the similarity matrix \mathbf{P} for the $n = 1600$ MNIST samples corresponding to digits “2,” “4,” “6,” and “8,” analyzed in Section 5. It justifies our assumption on the approximate block structure on \mathbf{P} . Figure 10 contains t-SNE visualizations

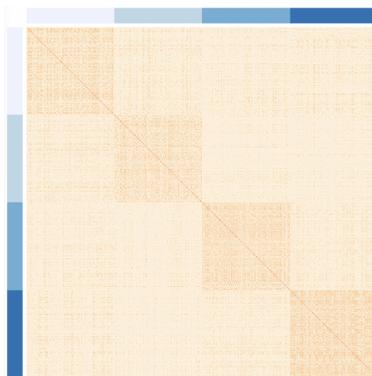


Figure 9: A heatmap of the similarity matrix \mathbf{P} for the $n = 1600$ MNIST samples corresponding to digits “2,” “4,” “6,” and “8,” analyzed in Section 5. The color bars represent the cluster labels of the columns and rows.

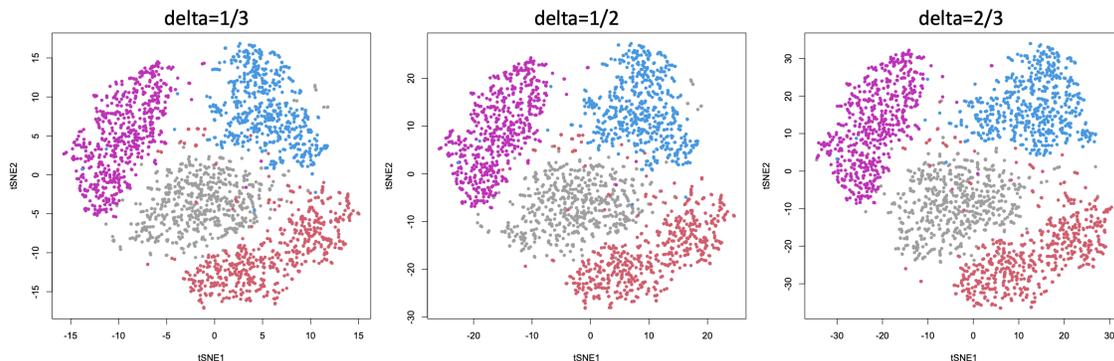


Figure 10: t-SNE visualizations of $n = 2400$ MNIST samples with an identical random initialization but different values of δ for the tuning parameters in (39).

of $n = 2400$ MNIST samples with an identical random initialization but different values of δ for the tuning parameters in (39). The similarity in the cluster patterns indicates robustness and flexibility of our theory-guided choices for the tuning parameters.