Spectral Clustering with Perturbed Data

Ling Huang
Intel Research
ling.huang@intel.com

Donghui Yan
UC Berkeley
dhyan@stat.berkeley.edu

Michael I. Jordan
UC Berkeley
jordan@cs.berkeley.edu

Nina Taft
Intel Research
nina.taft@intel.com

Abstract

Spectral clustering is useful for a wide-ranging set of applications in areas such as biological data analysis, image processing and data mining. However, the computational and/or communication resources required by the method in processing large-scale data are often prohibitively high, and practitioners are often required to perturb the original data in various ways (quantization, downsampling, etc) before invoking a spectral algorithm. In this paper, we use stochastic perturbation theory to study the effects of data perturbation on the performance of spectral clustering. We show that the error under perturbation of spectral clustering is closely related to the perturbation of the eigenvectors of the Laplacian matrix. From this result we derive approximate upper bounds on the clustering error. We show that this bound is tight empirically across a wide range of problems, suggesting that it can be used in practical settings to determine the amount of data reduction allowed in order to meet a specification of permitted loss in clustering performance.

1 Introduction

A critical problem in machine learning is that of scaling: Algorithms should be effective computationally and statistically as various dimensions of a problem are scaled. One general tool for approaching large-scale problems is that of clustering or partitioning, in essence an appeal to the principle of divide-and-conquer. However, while the output of a clustering algorithm may yield a set of smaller-scale problems that may be easier to tackle, clustering algorithms can themselves be complex, and large-scale clustering often requires the kinds of preprocessing steps that are invoked for other machine learning algorithms [1], including proto-clustering steps such as quantization, downsampling and compression. Such preprocessing steps also arise in the distributed sensing and distributed computing setting, where communication and storage limitations may preclude transmitting the original data to centralized processors.

A number of recent works have begun to tackle the issue of determining the tradeoffs that arise under various "perturbations" of data, including quantization and downsampling [2, 3, 4]. Most of these analyses have been undertaken in the context of well-studied domains such as classification, regression and density estimation, for which there are existing statistical analyses of the effect of noise on performance. Although extrinsic noise differs conceptually from perturbations to data imposed by a data analyst to cope with resource limitations, the mathematical issues arising in the two cases are similar and the analyses of noise have provided a basis for the study of the tradeoffs arising from perturbations.

In this paper we focus on spectral clustering, a class of clustering methods that are based on eigendecompositions of affinity, dissimilarity or kernel matrices [5, 6, 7, 8]. These algorithms often outperform traditional clustering algorithms such as the K-means algorithm or hierarchical clustering. To date, however, their impact on real-world, large-scale problems has been limited; in particular, a distributed or "in-network" version of spectral clustering has not yet appeared. Moreover, there has been little work on the statistical analysis of spectral clustering, and thus there is little theory to guide the design of distributed algorithms. There is an existing literature on numerical techniques for

Procedure SpectralClustering $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$
Input: n data samples $\{\mathbf{x}_i\}_{i=1}^n, \mathbf{x}_i \in \mathbb{R}^d$
Output: Bipartition S and $\bar{S}$ of the input data

1. Compute the similarity matrix K:
   $K_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_k^2}\right), \forall \mathbf{x}_i, \mathbf{x}_j$
2. Compute the diagonal degree matrix D:
   $D_i = \sum_{j=1}^n K_{ij}$
3. Compute the normalized Laplacian matrix:
   $L = I - D^{-1}K$
4. Find the second eigenvector $\mathbf{v}_2$ of L
5. Obtain the two partitions using $\mathbf{v}_2$:
6. $S = \{[i] : v_{2i} > 0\}, \quad \bar{S} = \{[i] : v_{2i} \leq 0\}$

Figure 1: A spectral bipartitioning algorithm.

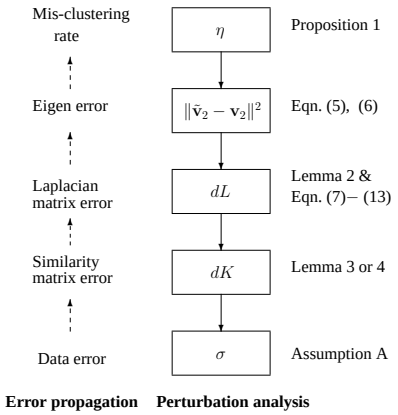

Figure 2: Perturbation analysis: from clustering error to data perturbation error.

scaling spectral clustering (including downsampling [9, 10] and the relaxation of precision requirements for the eigenvector computation [7]), but this literature does not provide end-to-end, practical bounds on error rates as a function of data perturbations.

In this paper we present the first end-to-end analysis of the effect of data perturbations on spectral clustering. Our focus is quantization, but our analysis is general and can be used to treat other kinds of data perturbation. Indeed, given that our approach is based on treating perturbations as random variables, we believe that our methods will also prove useful in developing statistical analyses of spectral clustering (although that is not our focus in this paper).

The paper is organized as follows. In Section 2, we provide a brief introduction to spectral clustering. Section 3 contains the main results of the paper; specifically we introduce the mis-clustering rate η, and present upper bounds on η due to data perturbations. In Section 4, we present an empirical evaluation of our analyses. Finally, in Section 5 we present our conclusions.

2 Spectral clustering and data perturbation

2.1 Background on spectral clustering algorithms

Given a set of data points $\{\mathbf{x}_i\}_{i=1}^n, \mathbf{x}_i \in \mathbb{R}^{1\times d}$ and some notion of similarity between all pairs of data points $\mathbf{x}_i$ and $\mathbf{x}_j$, spectral clustering attempts to divide the data points into groups such that points in the same group are similar and points in different groups are dissimilar. The point of departure of a spectral clustering algorithm is a weighted *similarity graph* $G(V, E)$, where the vertices correspond to data points and the weights correspond to the pairwise similarities. Based on this weighted graph, spectral clustering algorithms form the graph Laplacian and compute an eigendecomposition of this Laplacian [5, 6, 7]. While some algorithms use multiple eigenvectors and find a k-way clustering directly, the most widely studied algorithms form a bipartitioning of the data by thresholding the second eigenvector of the Laplacian (the eigenvector with the second smallest eigenvalue). Larger numbers of clusters are found by applying the bipartitioning algorithm recursively. We present a specific example of a spectral bipartitioning algorithm in Fig. 1.

2.2 Input data perturbation

Let the data matrix $X \in \mathbb{R}^{n\times d}$ be formed by stacking n data samples in rows. To this data matrix we assume that perturbation W is applied, such that we obtain a perturbed version $\tilde{X}$ of the original data X. We assume that a spectral clustering algorithm is applied to $\tilde{X}$ and we wish to compare the results of this clustering with respect to the spectral clustering of X. This analysis captures a number of data perturbation methods, including data filtering, quantization, lossy compression and synopsis-based data approximation [11]. The multi-scale clustering algorithms that use "representative" samples to approximate the original data can be treated using our analysis as well [12].

3 Mis-clustering rate and effects of data perturbation

Let K and L be the similarity and Laplacian matrix on the original data X, and let $\tilde{K}$ and $\tilde{L}$ be those on the perturbed data. We define the *mis-clustering rate* η as the proportion of samples that have different cluster memberships when computed on the two different versions of the data, X and $\tilde{X}$. We wish to bound η in terms of the "magnitude" of the error matrix $W = \tilde{X} - X$, which we now define. We make the following general stochastic assumption on the error matrix W:

A. All elements of the error matrix W are i.i.d. random variables with zero mean, bounded variance σ^2 and bounded fourth central moment μ^4; and are independent of X.

Remark. (i) Note that we do not make i.i.d. assumptions on the elements of the similarity matrix; rather, our assumption refers to the input data only. (ii) This assumption is distribution free, and captures a wide variety of practical data collection and quantization schemes. (iii) Certain data perturbation schemes may not satisfy the independence assumption. We have not yet conducted an analysis of the robustness of our bounds to lack of independence, but in our empirical work we have found that the bounds are robust to relatively small amounts of correlation.

We aim to produce practically useful bounds on η in terms of σ and the data matrix X. The bounds should be reasonably tight so that in practice they could be used to determine the degree of perturbation σ given a desired level of clustering performance, or to provide a clustering error guarantee on the original data even though we have access only to its approximate version.

Fig. 2 outlines the steps in our theoretical analysis. Briefly, when we perturb the input data (e.g., by filtering, quantization or compression), we introduce a perturbation W to the data which is quantified by σ^2. This induces an error $dK := \tilde{K} - K$ in the similarity matrix, and in turn an error $dL := \tilde{L} - L$ in the Laplacian matrix. This further yields an error in the second eigenvector of the Laplacian matrix, which results in mis-clustering error. Overall, we establish an analytical relationship between the mis-clustering rate η and the data perturbation error σ^2, i.e., a functional relationship of the form $\eta = f(\sigma^2)$, where f is usually monotonically increasing. Our goal is to allow practitioners to specify a mis-clustering rate η^*, and by inverting this relationship, to determine the right magnitude of the perturbation σ^* allowed. That is, our work can provide a practical method to determine the tradeoff between data perturbation and the loss of clustering accuracy due to the use of $\tilde{X}$ instead of X. When the data perturbation can be related to computational or communications savings, then our analysis yields a practical characterization of the overall resource/accuracy tradeoff.

Practical Applications Consider in particular a clustering task in a distributed networking system that allows an application to specify a desired clustering error C^* on the distributed data (which is not available to the coordinator). Through a communication protocol similar to that in [4], the coordinator (e.g., network operation center) gets access to the perturbed data $\tilde{X}$ for spectral clustering. The coordinator can compute a clustering error bound C using our method. By setting $C \leq C^*$, it determines the tolerable data perturbation error σ^* and instructs distributed devices to use appropriate numbers of bits to quantize their data. Thus we can provide guarantees on the achieved error, $C \leq C^*$, with respect to the original distributed data even with access only to the perturbed data.

3.1 Upper bounding the mis-clustering rate

Little is currently known about the connection between clustering error and perturbations to the Laplacian matrix in the spectral clustering setting. [5] presented an upper bound for the clustering error, however this bound is usually quite loose and is not viable for practical applications. In this section we propose a new approach based on a water-filling argument that yields a tighter, practical bound. Let $\mathbf{v}_2$ and $\tilde{\mathbf{v}}_2$ be the unit-length second eigenvectors of L and $\tilde{L}$, respectively. We derive a relationship between the mis-clustering rate η and $\delta^2 := \|\tilde{\mathbf{v}}_2 - \mathbf{v}_2\|^2$.

The intuition behind our derivation is suggested in Fig. 3. Let a and b denote the sets of components in $\mathbf{v}_2$ corresponding to clusters of size k_1 and k_2, respectively, and similarly for a' and b' in the case of $\tilde{\mathbf{v}}_2$. If $\mathbf{v}_2$ is changed to $\tilde{\mathbf{v}}_2$ due to the perturbation, an incorrect clustering happens whenever a component of $\mathbf{v}_2$ in set a jumps to set b', denoted as $a \to b'$, or a component in set b jumps to set a', denoted as $b \to a'$. The key observation is that each flipping of cluster membership in either $a \to b'$

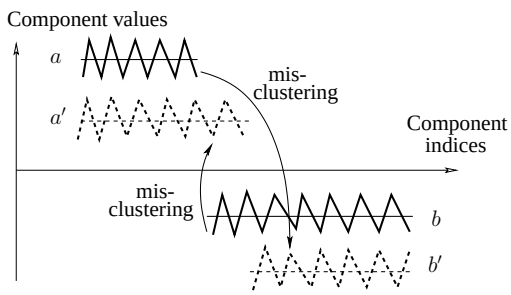

Figure 3: The second eigenvector $\mathbf{v}_2$ and its perturbed counterpart $\tilde{\mathbf{v}}_2$ (denoted by dashed lines).

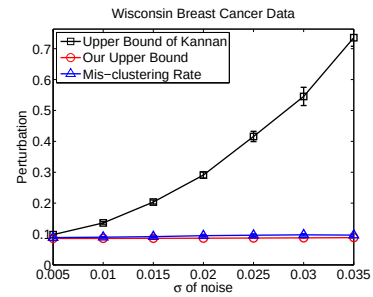

Figure 4: An example of the tightness of the upper bound for η in Eq. (1).

or $b \to a'$ contributes a fairly large amount to the value of δ^2, compared to the short-range drifts in $a \to a'$ or $b \to b'$. Given a fixed value of δ^2, the maximum possible number of flippings (i.e., missed clusterings) is therefore constrained, and this translates into an upper bound for η.

We make the following assumptions on the data X and its perturbation:

B1. The components of $\mathbf{v}_2$ form two clusters (with respect to the spectral bipartitioning algorithm in Fig. 1). The size of each cluster is comparable to n.

B2. The perturbation is small with the total number of mis-clusterings $m < \min(k_1, k_2)$, and the components of $\tilde{\mathbf{v}}_2$ form two clusters. The size of each cluster is comparable to n.

B3. The perturbation of individual components of $\mathbf{v}_2$ in each set of $a \to a'$, $a \to b'$, $b \to a'$ and $b \to b'$ have identical (not necessary independent) distributions with bounded second moments, respectively, and they are uncorrelated with the components in $\mathbf{v}_2$.

Our perturbation bound can now be stated as follows:

Proposition 1. *Under assumptions B1, B2 and B3, the mis-clustering rate η of the spectral bipartitioning algorithm under the perturbation satisfies $\eta \leq \delta^2 = \|\tilde{\mathbf{v}}_2 - \mathbf{v}_2\|^2$. If we further assume that all components of $\tilde{\mathbf{v}}_2 - \mathbf{v}_2$ are independent, then*

$$\eta \leq (1 + o_p(1))\mathrm{E}\|\tilde{\mathbf{v}}_2 - \mathbf{v}_2\|^2. \tag{1}$$

The proof of the proposition is provided in the Appendix.

Remarks. (i) Assumption B3 was motivated by our empirical work. Although it is difficult to establish general necessary and sufficient conditions for B3 to hold, in the Appendix we present some special cases that allow B3 to be verified a priori. It is also worth noting that B3 appears to hold (approximately) across a range of experiments presented in Section 4. (ii) If we assume piecewise constancy for $\mathbf{v}_2$, then we can relax the uncorrelated assumption in B3. (iii) Our bound has a different flavor than that obtained in [5]. Although the bound in Theorem 4.3 in [5] works for k-way clustering, it assumes a block-diagonal Laplacian matrix and requires the gap between the kth and $(k+1)^{th}$ eigenvalues to be greater than $1/2$, which is unrealistic in many data sets. In the setting of 2-way spectral clustering and a small perturbation, our bound is much tighter than that derived in [5]; see Fig. 4 in particular.

3.2 Perturbation on the second eigenvector of Laplacian matrix

We now turn to the relationship between the perturbation of eigenvectors with that of its matrix. One approach is to simply draw on the classical domain of matrix perturbation theory; in particular, applying Theorem V.2.8 from [13], we have the following bound on the (small) perturbation of the second eigenvector:

$$\|\tilde{\mathbf{v}}_2 - \mathbf{v}_2\| \leq \frac{\|4dL\|_F}{\nu - \sqrt{2}\|dL\|_F}, \tag{2}$$

where ν is the gap between the second and the third eigenvalue. However, in our experimental evaluation we found that ν can be quite small in many data sets (e.g., ν ranges from 0.009 to 0.2 in

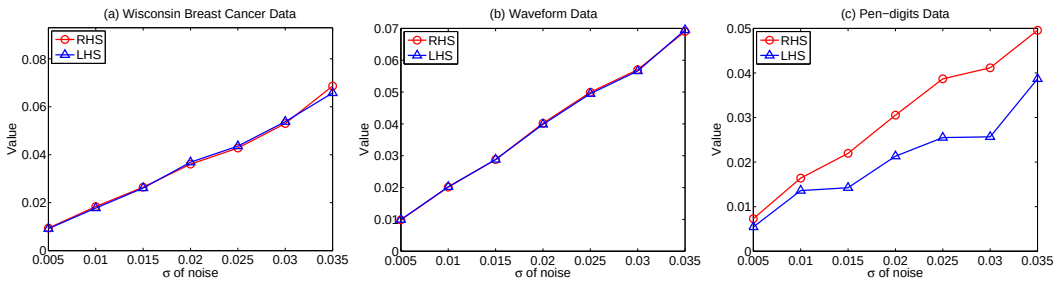

Figure 5: Experimental examples of the fidelity of the approximation in Eq. (5). We add i.i.d. zero mean Gaussian noise to the input data with different σ, and we see that the right-hand side (RHS) of (5) approximately upper bounds the left-hand side (LHS).

a large portion of data sets we used in our experiments), and in these cases the right-hand side of (2) can be quite large even for a small perturbation (or can be negative for a large perturbation). Thus the bound given by (2) is often not useful in practical applications.

To derive a more practically useful bound, we begin with a well-known first-order Taylor expansion to compute the perturbation on the second eigenvector of a Laplacian matrix as follows:

$$\begin{aligned}\tilde{\mathbf{v}}_2 - \mathbf{v}_2 &= \sum_{j=1, j\neq 2}^{n} \frac{\mathbf{v}_j^T dL \mathbf{v}_2}{\lambda_2 - \lambda_j}\mathbf{v}_j + O(dL^2) \approx \sum_{j=1, j\neq 2}^{n} \frac{\mathbf{v}_j}{\lambda_2 - \lambda_j} \sum_{p=1}^{n}\sum_{q=1}^{n} v_{pj} v_{q2} dL_{pq} \\ &= \sum_{p=1}^{n}\left[\left(\sum_{q=1}^{n} v_{q2} dL_{pq}\right)\cdot\left(\sum_{j=1, j\neq 2}^{n} \frac{v_{pj}\cdot \mathbf{v}_j}{\lambda_2 - \lambda_j}\right)\right] = \sum_{p=1}^{n} \beta_p \mathbf{u}_p, \end{aligned} \tag{3}$$

where $\beta_p = \sum_{q=1}^{n} v_{q2} dL_{pq}$ is a random variable determined by the effect of the perturbation on the Laplacian matrix L, and the vector $\mathbf{u}_p = \sum_{j=1, j\neq 2}^{n}\left(\frac{v_{pj}\mathbf{v}_j}{\lambda_2 - \lambda_j}\right)$ is a constant determined by the eigendecomposition of the Laplacian matrix L. Then we have

$$\mathrm{E}\|\tilde{\mathbf{v}}_2 - \mathbf{v}_2\|^2 \approx \mathrm{E}\left\|\sum_{p=1}^{n}\beta_p \mathbf{u}_p\right\|^2 = \sum_{p=1}^{n}\mathrm{E}\|\beta_p \mathbf{u}_p\|^2 + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathrm{E}\left(\beta_i \mathbf{u}_i \cdot \beta_j \mathbf{u}_j^T\right). \tag{4}$$

In our experimental work we have found that for $i \neq j$, $\beta_i \mathbf{u}_i$ is either very weakly correlated with $\beta_j \mathbf{u}_j$ (i.e., the total sum of all cross terms is typically one or two orders of magnitude less than that of squared term), or negatively correlated with $\beta_j \mathbf{u}_j$ (i.e., the total sum of all cross terms is less than zero). This empirical evidence suggests the following approximate bound:

$$\mathrm{E}\|\tilde{\mathbf{v}}_2 - \mathbf{v}_2\|^2 \lesssim \sum_{p=1}^{n}\mathrm{E}\beta_p^2 \cdot \|\mathbf{u}_p\|^2. \tag{5}$$

Examples of the fidelity of this approximation for particular data sets are shown in Fig. 5.

Finally, $\mathrm{E}\beta_p^2$ is related to dL_{pq}, and can be upper bounded by

$$\mathrm{E}\beta_p^2 = \mathrm{E}\left(\sum_{q=1}^{n} v_{q2} dL_{pq}\right)^2 \leq \sum_{i=1}^{n}\sum_{j=1}^{n}\left[v_{i2}v_{j2}\cdot \mathrm{E}\left(dL_{pi}\right)\mathrm{E}\left(dL_{pj}\right) + |v_{i2}v_{j2}|\sigma_{pi}\sigma_{pj}\right], \tag{6}$$

where σ_{pi} is the variance of dL_{pi}.

Remark. Through Eqs. (5) and (6), we can bound the squared norm of the perturbation on the second eigenvector in expectation, which in turn bounds the mis-clustering rate. To compute the bound, we need to estimate the first two moments of dL, which we discuss next.

3.3 Perturbation on the Laplacian matrix

Let D be the diagonal matrix with $D_i = \sum_j K_{ij}$. We define the normalized Laplacian matrix as $L = I - D^{-1}K$. Letting $\Delta = \tilde{D} - D$ and $dK = \tilde{K} - K$, we have the following approximation for $dL = \tilde{L} - L$:

Lemma 2. *If perturbation dK is small compared to K, then*

$$dL = (1 + o(1))\,\Delta D^{-2}K - D^{-1}dK. \tag{7}$$

Then, element-wise, the first two moments of dL can be estimated as

$$\begin{aligned} \mathrm{E}(dL) &\approx \mathrm{E}(\Delta)D^{-2}K - D^{-1}\mathrm{E}(dK) && (8)\\ \mathrm{E}(dL^2) &\approx \mathrm{E}\left(\Delta D^{-2}K \circ \Delta D^{-2}K - 2D^{-1}dK \circ \Delta D^{-2}K + D^{-1}dK \circ D^{-1}dK\right) \\ &= \mathrm{E}\left(\Delta^2\right)D^{-4}K^2 + D^{-2}\mathrm{E}\left(dK^2\right) - 2\mathrm{E}(\Delta dK)D^{-3} \circ K, && (9) \end{aligned}$$

where $\circ$ denotes element-wise product. The quantities needed to estimate $\mathrm{E}(dL)$ and $\mathrm{E}(dL^2)$ can be obtained from moments and correlations among the elements of the similarity matrix $\tilde{K}_{ij}$. In particular, we have

$$\mathrm{E}(dK_{ij}) = \mathrm{E}\left(\tilde{K}_{ij}\right) - K_{ij}, \quad \mathrm{E}(dK_{ij})^2 = \mathrm{E}\tilde{K}_{ij}^2 - 2K_{ij}\mathrm{E}\left(\tilde{K}_{ij}\right) + K_{ij}^2 \tag{10}$$

$$\mathrm{E}\Delta_i = \mathrm{E}\tilde{D}_i - D_i, \quad \mathrm{E}\tilde{D}_i = \sum_{j=1}^{n}\mathrm{E}\left(\tilde{K}_{ij}\right), \quad \mathrm{E}\Delta_i^2 = \mathrm{E}\tilde{D}_i^2 - 2D_i \cdot \mathrm{E}\tilde{D}_i + D_i^2 \tag{11}$$

$$\mathrm{E}\tilde{D}_i^2 = \mathrm{E}\left(\sum_{j=1}^{n}\tilde{K}_{ij}\right)^2 = \sum_{j=1}^{n}\mathrm{E}\tilde{K}_{ij}^2 + 2\sum_{j=1}^{n}\sum_{q=j+1}^{n}\left(\mathrm{E}\tilde{K}_{ij}\mathrm{E}\tilde{K}_{iq} + \rho_{ijq}^k\sigma_{ij}^k\sigma_{iq}^k\right) \tag{12}$$

$$\begin{aligned} \mathrm{E}(\Delta dK)_{ij} &= \mathrm{E}(\tilde{D}_i - D_i)(\tilde{K}_{ij} - K_{ij}) = \mathrm{E}\left(\tilde{D}_i\tilde{K}_{ij}\right) - D_i\mathrm{E}\tilde{K}_{ij} - K_{ij}\mathrm{E}\Delta_i \\ &= \mathrm{E}\left[\tilde{K}_{ij}^2 + \tilde{K}_{ij}\left(\sum_{q=1,q\neq j}^{n}\tilde{K}_{iq}\right)\right] - D_i\mathrm{E}\tilde{K}_{ij} - K_{ij}\mathrm{E}\Delta_i \\ &= \mathrm{E}\tilde{K}_{ij}^2 + \sum_{q=1,q\neq j}^{n}\left(\mathrm{E}\tilde{K}_{ij}\mathrm{E}\tilde{K}_{iq} + \rho_{ijq}^k\sigma_{ij}^k\sigma_{iq}^k\right) - D_i\mathrm{E}\tilde{K}_{ij} - K_{ij}\mathrm{E}\Delta_i, \end{aligned} \tag{13}$$

where σ_{ij}^k is the standard deviation of $\tilde{K}_{ij}$ and $-1 \leq \rho_{ijq}^k \leq 1$ is the correlation coefficient between $\tilde{K}_{ij}$ and $\tilde{K}_{iq}$. Estimating all $\rho_{ijq}^k{}'s$ would require an intensive effort. For simplicity, we could set ρ_{ijq}^k to 1 in Eq. (12) and to -1 in Eq. (13), and obtain an upper bound for $\mathrm{E}(dL^2)$. This bound could optionally be tightened by using a simulation method to estimate the values of ρ_{ijq}^k. However, in our experimental work we have found that our results are insensitive to the values of ρ_{ijq}^k, and setting $\rho_{ijq}^k = 0.5$ usually achieves good results.

Remark. Eqs. (8)–(13) allow us to estimate (i.e., to upper bound) the first two moments of dL using those of dK, which are computed using Eq. (15) or (16) in Section 3.4.

3.4 Perturbation on the similarity matrix

The similarity matrix $\tilde{K}$ on perturbed data $\tilde{X}$ is

$$\tilde{K}_{ij} = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j + \epsilon_i - \epsilon_j||^2}{2\sigma_k^2}\right), \tag{14}$$

where σ_k is the kernel bandwidth. Then, given data X, the first two moments of $dK_{ij} = \tilde{K}_{ij} - K_{ij}$, the error in the similarity matrix, can be determined by one of the following lemmas.

Lemma 3. *Given X, if all components of ϵ_i and ϵ_j are i.i.d. Gaussian $N(0, \sigma^2)$, then*

$$\mathrm{E}\left(\tilde{K}_{ij}\right) = M_{ij}\left(-\frac{\sigma^2}{\sigma_k^2}\right), \quad \mathrm{E}\left(\tilde{K}_{ij}^2\right) = M_{ij}\left(-\frac{2\sigma^2}{\sigma_k^2}\right), \tag{15}$$

where $M_{ij}(t) = \left[\exp\left(\frac{\lambda_{ij}t}{1-2t}\right)/(1-2t)^{d/2}\right]$, and $\lambda_{ij} = \left(||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma^2\right)$.

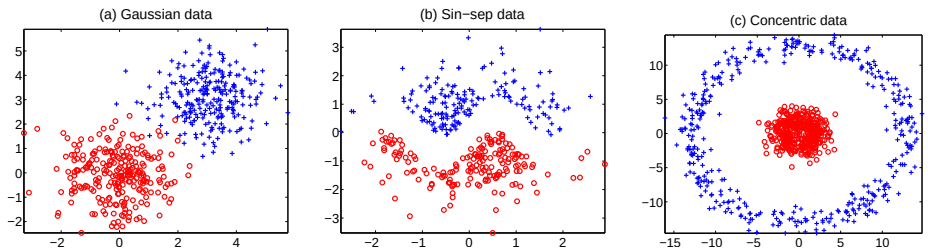

Figure 6: Synthetic data sets illustrated in two dimensions.

Lemma 4. *Under Assumption* A*, given* X *and for large values of the dimension* d*, the first two moments of* $\tilde{K}_{ij}$ *can be computed approximately as follows:*

$$\mathrm{E}\left(\tilde{K}_{ij}\right) = M_{ij}\left(-\frac{1}{2\sigma_k^2}\right), \quad \mathrm{E}\left(\tilde{K}_{ij}^2\right) = M_{ij}\left(-\frac{1}{\sigma_k^2}\right), \tag{16}$$

where $M_{ij}(t) = \exp\left[\left(\lambda_{ij} + 2d\sigma^2\right)t + \left(d\mu^4 + d\sigma^4 + 4\sigma^2\lambda_{ij}^2\right)t^2\right]$*, and* $\lambda_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||^2$.

Remark. (i) Given data perturbation error σ, kernel bandwidth σ_k and data X, the first two moments of dK_{ij} can be estimated directly using (15) or (16). (ii) Through Eqs. (1)–(16), we have established a relationship between the mis-clustering rate η and the data perturbation magnitude σ. By inverting this relationship (e.g., using binary search), we can determine a σ^* for a given η^*.

4 Evaluation

In this section we present an empirical evaluation of our analysis on 3 synthetic data sets (see Fig. 6) and 6 real data sets from the UCI repository [14]. The data domains are diverse, including image, medicine, agriculture, etc., and the different data sets impose different difficulty levels on the underlying spectral clustering algorithm, demonstrating the wide applicability of our analysis.

In the experiments, we use data quantization as the perturbation scheme to evaluate the upper bound provided by our analysis on the clustering error. Fig. 7 plots the mis-clustering rate and the upper bound for data sets subject to varying degrees of quantization. As expected, the mis-clustering rate increases as one decreases the number of quantization bits. We find that the error bounds are correct and remarkably tight, which validate the assumptions we make in the analysis. It is also interesting to note that even when using as few as 3-4 bits, the clustering degrades very little in both real error and as assessed by our bound. The effectiveness of our bound should allow the practitioner to determine the right amount of quantization given a permitted loss in clustering performance.

5 Conclusion

In this paper, we proposed a theoretical analysis of the clustering error for spectral clustering in the face of stochastic perturbations. Our experimental evaluation has provided support for the assumptions made in the analysis, showing that the bound is tight under conditions of practical interest. We believe that our work, which provides an analytical relationship between the mis-clustering rate and the variance of the perturbation, constitutes a critical step towards enabling a large class of applications that seek to perform clustering of objects, machines, data, etc in a distributed environment. Many networks are bandwidth constrained, and our methods can guide the process of data thinning so as to limit the amount of data transmitted through the network for the purpose of clustering.

References

[1] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," in *Advances in Neural Information Processing Systems 20*, 2007.

[2] A. Silberstein, G. P. A. Gelfand, K. Munagala, and J. Yang, "Suppression and failures in sensor networks: A Bayesian approach," in *Proceedings of VLDB*, 2007.

[3] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Nonparametric decentralized detection using kernel methods," *IEEE Transactions on Signal Processing*, vol. 53, no. 11, pp. 4053–4066, 2005.

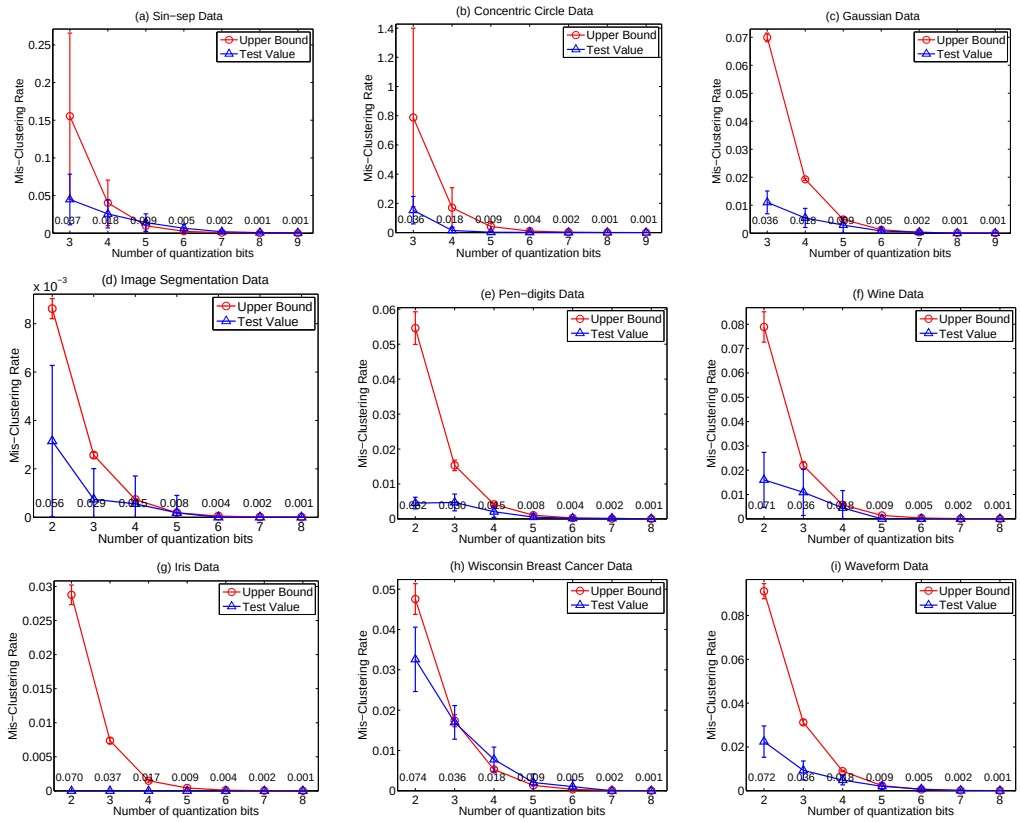

Figure 7: Upper bounds of clustering error on approximate data obtained from quantization as a function of the number of bits. (a–c) Simulated data sets (1000 sample size, 2, 2, 10 features, respectively); (d) Statlog image segmentation data (2310 sample size, 19 features); (e) Handwritten digits data (10992 sample size, 16 features); (f) Wine data (178 sample size, 13 features); (g) Iris data (150 sample size, 4 features). (h) Wisconsin breast cancer data (569 sample size, 30 features); (i) Waveform data (5000 sample size, 21 features). The x-axis shows the number of quantization bits and (above the axis) the corresponding data perturbation error σ. Error bars are derived from 25 replications. In the experiments, all data values are normalized in range $[0, 1]$. For data sets with more than two clusters, we choose two of them for the experiments.

[4] L. Huang, X. Nguyen, M. Garofalakis, A. D. Joseph, M. I. Jordan, and N. Taft, "In-network PCA and anomaly detection," in *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[5] R. Kannan, S. Vempala, and A. Vetta, "On clusterings: Good, bad and spectral," *Journal of the ACM*, vol. 51, no. 3, pp. 497–515, 2004.

[6] A. Y. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems (NIPS)*, 2002.

[7] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[8] U. von Luxburg, M. Belkin, and O. Bousquet, "Consistency of spectral clustering," *Annals of Statistics*, vol. 36, no. 2, pp. 555–586, 2008.

[9] P. Drineas and M. W. Mahoney, "On the Nyström method for approximating a Gram matrix for improved kernel-based learning," in *Proceedings of COLT*, 2005, pp. 323–337.

[10] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nyström method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, 2004.

[11] G. Cormode and M. Garofalakis, "Sketching streams through the net: Distributed approximate query tracking," in *Proceedings of VLDB*, 2005, pp. 13–24.

[12] D. Kushnir, M. Galun, and A. Brandt, "Fast multiscale clustering and manifold identification," *Pattern Recognition*, vol. 39, no. 10, pp. 1876–1891, 2006.

[13] G. W. Stewart and J. Guang Sun, *Matrix Perturbation Theory*. Academic Press, 1990.

[14] A. Asuncion and D. Newman, "UCI Machine Learning Repository, Department of Information and Computer Science," 2007, http://www.ics.uci.edu/ mlearn/MLRepository.html.

[15] D. Yan, "Some issues with dimensionality in statistical inference," Ph.D. dissertation, University of California, Berkeley, 2008.

[16] J. A. Rice, *Mathematical Statistics and Data Analysis*. Duxbury Press, 1995.

6 Appendix

In this Appendix we provide more detailed analyses and proofs that are omited in the main body of the paper due to space limitation.

Some heuristics on assumption B3 It is difficult to show that assumption B3 is valid under very general conditions. To gain insight, we here show its validity in the case where the original similarity matrix has the following block-diagonal structure (assuming there are two clusters with sizes p and q, respectively):

$$P_0 = \begin{bmatrix} 1_{p\times p} & 0_{p\times q} \\ 0_{q\times p} & 1_{q\times q} \end{bmatrix},$$

where $1_{p\times p}$, $1_{q\times q}$ denote matrices with all elements 1, and $0_{p\times q}$, $0_{q\times p}$ denote matrices with all elements 0. This is the case where points in the same cluster have perfect affinity and points from different clusters have no affinity.

We further assume the perturbed similarity matrix is given by

$$P_\epsilon = \begin{bmatrix} 1_{p\times p} & U(\tilde{0},\epsilon) \\ U(\tilde{0},\epsilon)^T & 1_{q\times q} \end{bmatrix},$$

where $U(\tilde{0},\epsilon)$ denotes a $p\times q$ matrix with elements generated i.i.d. according to some distribution $\mathbb{P}$ (e.g., uniform distribution) on interval $[0,\epsilon]$, with ϵ being a small constant. This model has been studied recently in [15], which obtains the following result for the unnormalized second eigenvector $\tilde{\mathbf{v}}_2$ of the Laplacian matrix of P_ϵ:

Proposition 5 *Assuming all elements ϵ_{ij} in matrix P_ϵ are i.i.d. uniform over interval $[0,\epsilon]$ for some constant $\epsilon = o(\frac{1}{p+q})$, and $\frac{p}{q}\to\alpha$ for some constant α when p and q grow. Then when p and q grow, the following holds*

$$\tilde{v}_{2k} = \begin{cases} 1-(1+\gamma)p^2[\frac{\epsilon_{k.}}{p+\epsilon_{k.}} - \mathbb{E}\frac{\epsilon_{1.}}{p+\epsilon_{1.}}] + R_{k,p,q} & k=1,...,p-1 \\ 1-(1+\gamma)(p-1)[\frac{\epsilon_{p.}}{p+\epsilon_{p.}} - \mathbb{E}\frac{\epsilon_{1.}}{p+\epsilon_{1.}}] + R_{p,p,q} & k=p \\ -\gamma+(1+\gamma)q^2[\frac{\epsilon_{.(k-p)}}{q+\epsilon_{.(k-p)}} - \mathbb{E}\frac{\epsilon_{.1}}{q+\epsilon_{.1}}] + R_{k,p,q} & k=p+1,...,p+q-1 \\ -\gamma+(1+\gamma)(q-1)[\frac{\epsilon_{.q}}{q+\epsilon_{.q}} - \mathbb{E}\frac{\epsilon_{.1}}{q+\epsilon_{.1}}] + R_{p+q,p,q} & k=p+q \end{cases}$$

where $\tilde{v}_{2k}$ denotes the kth component of eigenvector $\tilde{\mathbf{v}}_2$, γ is a constant, $\epsilon_{i.} = \sum_{j=1}^q \epsilon_{ij}, i=1,...,p$, and $\epsilon_{.j} = \sum_{i=1}^p \epsilon_{ij}, j=1,...,q$, $R_{k,p,q}$'s are remainders with $\max_{1\le k\le p+q}|R_{k,p,q}| = o_p(1)$.

From Proposition 5, we can easily see that excluding the pth and the $(p+q)^{th}$ components, all other components in $\tilde{\mathbf{v}}_2$ that belong to the same clusters follow the same distribution, up to a first order approximation (in the sense of the general matrix perturbation theory). Moreover, since the second eigenvector $\mathbf{v}_2$ of the Laplacian matrix of P_0 is piecewise constant, it immediately follows that the individual perturbations are uncorrelated with their initial values.

Proof of Proposition 1 We use the notation introduced earlier in Section 3.1, and introduce additional notation in Fig. 8. Recall that a and b denote the sets of elements in $\mathbf{v}_2$ corresponding to two clusters under consideration, and similarly for a' and b' in $\tilde{\mathbf{v}}_2$. Let m denote the total number of missed clusterings: $m = m_- + m_+$, where m_- denotes the number of cluster flippings of $a\to b'$, and m_+ the number of flippings of $b\to a'$. As an abuse of notation, we also use a, b, a', b' to denote the mean of element values in the corresponding set in $\mathbf{v}_2$ and $\tilde{\mathbf{v}}_2$, respectively. Referring to the value δ^2 as the "energy," we aim to determine the maximum number of flippings m (i.e., missed clusterings) for any (randomly) given energy δ^2; this yields an upper bound for η.

Let $A_1,\ldots,A_{k_1}$ and $B_1,\ldots,B_{k_2}$ denote the zero mean fluctuations around a and b, respectively, and let $U_1,\ldots,U_{k_1-m_-}$ and $V_1,\ldots,V_{k_2-m_+}$ denote the zero mean fluctuations around a' and b',

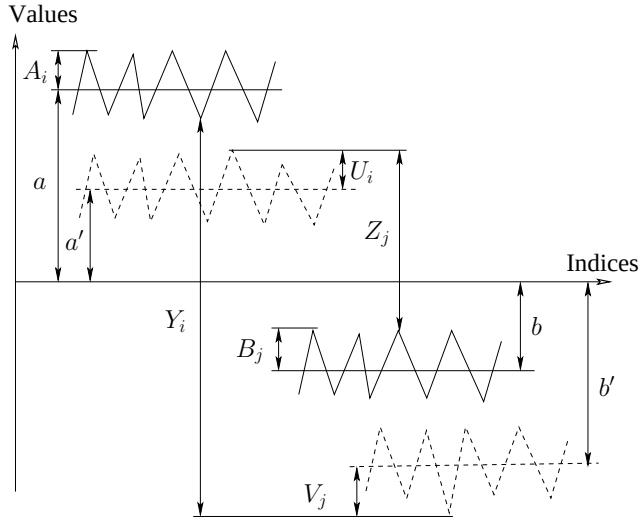

Figure 8: Notation in the Proof of Proposition 1.

maximize $\quad m = m_+ + m_-$

subject to:

$$\sum_{i=1}^{k_1}(a + A_i)^2 + \sum_{j=1}^{k_1}(b + B_j)^2 = 1 \tag{17}$$

$$\sum_{i=1}^{k_1-m_-}(a' + U_i)^2 + \sum_{j=1}^{k_2-m_+}(b' + V_j)^2 + \sum_{i=1}^{m_-}(a + A_i - Y_i)^2 + \sum_{j=1}^{m_+}(b + B_j + Z_j)^2 = 1 \tag{18}$$

$$\sum_{i=1}^{k_1-m_-}(a + A_i - a' - U_i)^2 + \sum_{j=1}^{k_2-m_+}(b + B_j - b' - V_j)^2 + \sum_{i=1}^{m_-}Y_i^2 + \sum_{j=1}^{m_+}Z_j^2 = \delta^2 \tag{19}$$

Figure 9: The optimization problem for mis-clustering rate.

respectively. Let variables $Y_1, \ldots, Y_{m_-} \geq 0$ denote the long-range (random) jumps from a to b' for elements in set $a \to b'$, and let $Z_1, \ldots, Z_{m_+} \geq 0$ for elements in set $b \to a'$; Apparently, we have $Z_j = a' + U_j - b - B_j$. We have the optimization problem described in Fig. 9 for the maximum number of mis-clusterings m. In the constraints, Eqs. (17) and (18) refer to the unit length of $\mathbf{v}_2$ and $\tilde{\mathbf{v}}_2$, respectively; and Eq. (19) refers to the total energy constraint.

Without loss of generality we consider the case described in Fig. 8 with $a, a' > 0, b, b' \leq 0$, and $(a' - b)^2 < (a - b')^2$. Due to symmetry, the same argument follows when we consider the other cases. According to assumption B3, we have $\mathrm{E}Y_i = a - b'$ and $\mathrm{E}Z_j = a' - b$, for all i and j. By working with expectations, we have

$$\begin{aligned}
\mathrm{E}Y_1^2 + ... + \mathrm{E}Y_{m_-}^2 &\geq \frac{1}{m_-}(\mathrm{E}Y_1 + ... + \mathrm{E}Y_{m_-})^2 = m_-(a - b')^2 \\
\mathrm{E}Z_1^2 + ... + \mathrm{E}Z_{m_+}^2 &\geq \frac{1}{m_+}(\mathrm{E}Z_1 + ... + \mathrm{E}Z_{m_Z Z+})^2 = m_+(a' - b)^2.
\end{aligned}$$

Summing and substituting into Eq. (19), we get

$$m_+(a' - b)^2 + m_-(a - b')^2 \leq \mathrm{E}\sum_{i=1}^{m_-}Y_i^2 + \mathrm{E}\sum_{j=1}^{m_+}Z_j^2 \leq \delta^2 \tag{20}$$

To maximize the objective function $m = m_- + m_-$ given (20) and given $(a' - b)^2 < (a - b')^2$, we should set $m_- = 0$ and $m = m_+$. This turns (20) into $m(a' - b)^2 \leq \delta^2$, under which m achieves its maximum when $b = 0$.

Let σ_Z^2 denote the variance of Z_i, which may depend on m or n. The following arguments allow us to assume that $a'^2 = b'^2 + \mathrm{E}V_1^2 - \sigma_Z^2 - \mathrm{E}B_1^2$. The results $b = 0$ and $m = m_+$ simplify Eq. (18) into

$$\sum_{i=1}^{k_1}(a' + U_i)^2 + \sum_{j=1}^{k_2-m}(b' + V_j)^2 + \sum_{j=1}^{m}(B_j + Z_j)^2 = 1. \tag{21}$$

Taking an expectation on both sides and using the assumptions that the U_i are identically distributed with zero mean, as are the V_j, we get

$$k_1 a'^2 + k_1\mathrm{E}U_1^2 + (k_2 - m)b'^2 + (k_2 - m)\mathrm{E}V_1^2 + m(a'^2 + \sigma_Z^2) = 1, \tag{22}$$

which implies that $\sigma_Z^2 \leq \frac{1}{m}$. Hence we can assume that $\sigma_Z^2 = \frac{\beta}{m}$ for some $0 \leq \beta \leq 1$. Setting $b'^2 = a'^2 - \mathrm{E}V_1^2 + \sigma_Z^2 + \mathrm{E}B_1^2$ in Eq. (22) effectively disables this constraint for m, and Eq. (22) becomes:

$$k_1 a'^2 + k_1\mathrm{E}U_1^2 + k_2 b'^2 + k_2\mathrm{E}V_1^2 = 1.$$

Substituting $b'^2 = a'^2 - \mathrm{E}V_1^2 + \sigma_Z^2 + \mathrm{E}B_1^2$ into this equation, we obtain

$$k_1 a'^2 + k_1\mathrm{E}U_1^2 + k_2(a'^2 - \mathrm{E}V_1^2 + \sigma_Z^2) + k_2\mathrm{E}V_1^2 = na'^2 + k_2\sigma_Z^2 + k_1\mathrm{E}U_1^2 = 1,$$

from which we obtain

$$na'^2 = 1 - k_1\mathrm{E}U_1^2 - \frac{k_2\beta}{m} - k_2\mathrm{E}B_1^2. \tag{23}$$

Using $b = 0$ and $m = m_+$ also simplifies Eq. (19), which becomes:

$$\sum_{i=1}^{k_1}(a + A_i - a' - U_i)^2 + \sum_{j=1}^{k_2-m}(B_j - b' - V_j)^2 + \sum_{j=1}^{m}(a' - B_j + U_j)^2 = \delta^2.$$

Taking an expectation on both sides yields

$$k_1((a - a')^2 + \mathrm{E}A_1^2) + k_1\mathrm{E}U_1^2 + k_2\mathrm{E}B_1^2 + (k_2 - m)(b'^2 + \mathrm{E}V_1^2) + m(a'^2 + \mathrm{E}U_1^2) = \delta^2,$$

implying

$$k_1\mathrm{E}U_1^2 + k_2\mathrm{E}B_1^2 \leq \delta^2 - \beta - ma'^2 \tag{24}$$

$$(n - m)a'^2 \leq \frac{(n - m)(\delta^2 - \beta)}{m}. \tag{25}$$

Substituting (24) into (23) and combining with (25), we get

$$1 - \delta^2 + \beta - \frac{k_2}{m}\beta \leq (n - m)a'^2 \leq \frac{(n - m)(\delta^2 - \beta)}{m}$$

Rearranging terms we have

$$\begin{aligned} m &\leq n\delta^2 - n\beta + k_2\beta \leq n\delta^2 \\ \eta = \frac{m}{n} &\leq \delta^2 = \|\tilde{\mathbf{v}}_2 - \mathbf{v}_2\|^2. \end{aligned}$$

When $\|\tilde{\mathbf{v}}_2 - \mathbf{v}_2\|^2$ is small, and if we further assume that all components of $\tilde{\mathbf{v}}_2 - \mathbf{v}_2$ are independent, then $\|\tilde{\mathbf{v}}_2 - \mathbf{v}_2\|^2$ is highly concentrated around its mean asymptotically in the number of data points. In this case we obtain $\eta \leq (1 + o_p(1))\mathrm{E}\|\tilde{\mathbf{v}}_2 - \mathbf{v}_2\|^2$.

Proof of Lemma 2 For the perturbation on the Laplacian matrix, we have

$$dL = I - (D + \Delta)^{-1}(K + dK) - I + D^{-1}K \tag{26}$$

Because the perturbation dK is small comparing to K, so is Δ comparing to D, and ΔD^{-1} is small. Using Taylor expansion for function $G(X) = (I + X)^{-1}$ around $X = 0_{n\times n}$, we have

$$(I + \Delta D^{-1})^{-1} = I - \Delta D^{-1} + O((\Delta D^{-1})^2).$$

Substituting it into Eq. (26), we get

$$\begin{aligned} dL &= -[I - \Delta D^{-1} + O((\Delta D)^{-2})]D^{-1}(K + dK) + D^{-1}K \\ &= -D^{-1}(K + dK) + \Delta D^{-2}(K + dK) - O((\Delta D)^{-2})D^{-1}(K + dK) + D^{-1}K \\ &= (1 + o(1))\,\Delta D^{-2}K - D^{-1}dK. \end{aligned}$$

Proof of Lemma 3 Let $S_{ij} := ||\mathbf{x}_i - \mathbf{x}_j + \epsilon_i - \epsilon_j||^2$. For $i = j$, the result holds trivially. For $i \neq j$ and given X, $\left(S_{ij}/2\sigma^2\right)$ follows a non-central chi-square distribution with parameter (d, λ_{ij}), where $\lambda_{ij} = \left(||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma^2\right)$. The mean is $d + \lambda_{ij}$, the variance is $2(d + 2\lambda_{ij})$, and the moment generating function is $M_{ij}(t) = \left[\exp\left(\frac{\lambda_{ij}t}{1-2t}\right)/(1-2t)^{d/2}\right]$. $\tilde{K}_{ij}$ is an exponential function of the non-central chi-square random variable $\left(S_{ij}/2\sigma^2\right)$, so the first two moments of $\tilde{K}_{ij}$ can be computed using the moment generating function $M_{ij}(t)$, which gives the results in Eq.(15) in the main paper.

Proof of Lemma 4 Let $S_{ij} := ||\mathbf{x}_i - \mathbf{x}_j + \epsilon_i - \epsilon_j||^2$. For $i = j$, the result holds trivially. For $i \neq j$, S_{ij} is the sum of d (the dimension of X) independent variables, and approximately follows a Gaussian distribution for large d. We only need to work out its mean and variance. Let $\lambda_{ij} := ||\mathbf{x}_i - \mathbf{x}_j||^2$. Given input data X, we have

$$\begin{aligned}
\mathrm{E}S_{ij} &= \mathrm{E}||\mathbf{x}_i - \mathbf{x}_j + \epsilon_i - \epsilon_j||^2 = \sum_{p=1}^{d}\left[(X_i^{(p)} - X_j^{(p)})^2 + 2\sigma^2\right] = \lambda_{ij} + 2d\sigma^2 \\
\mathrm{E}S_{ij}^2 &= \sum_{p=1}^{d}\sum_{q=1}^{d}\mathrm{E}\left[X_i^{(p)} - X_j^{(p)} + \epsilon_i^{(p)} - \epsilon_j^{(p)}\right]^2 \cdot \left[X_i^{(q)} - X_j^{(q)} + \epsilon_i^{(q)} - \epsilon_j^{(q)}\right]^2 \\
&= \lambda_{ij}^4 + 4(d^2 - d)\sigma^4 + d(2\mu^4 + 6\sigma^4) + 8\sigma^2\lambda_{ij}^2 + 4d\sigma^2\lambda_{ij}^2 \\
\mathrm{Var}(S_{ij}) &= \mathrm{E}S_{ij}^2 - (\mathrm{E}S_{ij})^2 = 2d\mu^4 + 2d\sigma^4 + 8\sigma^2\lambda_{ij}^2.
\end{aligned}$$

So using the moment-generating function, we obtain the following asymptotic results [16]:

$$\begin{aligned}
\mathrm{E}\left(\tilde{K}_{ij}\right) &= \mathrm{E}\left(\exp\left(-\frac{S_{ij}}{2\sigma_k^2}\right)\right) = M_{ij}\left(-\frac{1}{2\sigma_k^2}\right) \\
\mathrm{E}\left(\tilde{K}_{ij}^2\right) &= \mathrm{E}\left(\exp\left(-\frac{S_{ij}}{\sigma_k^2}\right)\right) = M_{ij}\left(-\frac{1}{\sigma_k^2}\right),
\end{aligned}$$

which completes the proof.