The Annals of Statistics
2021, Vol. 49, No. 1, 531–556
https://doi.org/10.1214/20-AOS1967

CONCENTRATION OF KERNEL MATRICES WITH APPLICATION TO KERNEL SPECTRAL CLUSTERING

BY ARASH A. AMINI[1] AND ZAHRA S. RAZAEE[2]

[1]*Department of Statistics, University of California, Los Angeles, aaamini@ucla.edu*
[2]*Biostatistics and Bioinformatics Research Center, Cedars-Sinai Medical Center, zahra.razaee@cshs.org*

We study the concentration of random kernel matrices around their mean. We derive nonasymptotic exponential concentration inequalities for Lipschitz kernels assuming that the data points are independent draws from a class of multivariate distributions on $\mathbb{R}^d$, including the strongly log-concave distributions under affine transformations. A feature of our result is that the data points need not have identical distributions or zero mean, which is key in certain applications such as clustering. Our bound for the Lipschitz kernels is dimension-free and sharp up to constants. For comparison, we also derive the companion result for the Euclidean (inner product) kernel for a class of sub-Gaussian distributions. A notable difference between the two cases is that, in contrast to the Euclidean kernel, in the Lipschitz case, the concentration inequality does not depend on the mean of the underlying vectors. As an application of these inequalities, we derive a bound on the misclassification rate of a kernel spectral clustering (KSC) algorithm, under a perturbed nonparametric mixture model. We show an example where this bound establishes the high-dimensional consistency (as $d \to \infty$) of the KSC, when applied with a Gaussian kernel, to a noisy model of nested nonlinear manifolds.

1. Introduction. Kernel methods are quite widespread in statistics and machine learning, since many "linear" methods can be turned into nonlinear ones by replacing the Gram matrix with one based on a nonlinear kernel, the so-called kernel trick. The approach is often motivated as follows: One first maps the data $x \in \mathbb{R}^d$ to a point $\Phi(x)$ in a higher dimensional space H via a nonlinear feature map $\Phi : \mathbb{R}^d \to H$. In this new space, the data are better behaved (e.g., linearly separated in the case of classification), hence one can run a simple linear algorithm. Often this algorithm relies only on the inner products $\langle \Phi(x), \Phi(y) \rangle = K(x, y)$. Thus the transformation is effectively equivalent to replacing the usual inner product $\langle x, y \rangle$ with the kernelized version $K(x, y)$, keeping the computational cost of the algorithm roughly the same. This way of introducing nonlinearity without sacrificing efficiency, works well for many commonly used algorithms such as principal component analysis, ridge regression, support vector machines, k-means clustering, and so on [5, 23, 30].

To be concrete, let the data be the random sample $X_1, \dots, X_n \in \mathbb{R}^d$ drawn independently from unknown distributions $P_1, \dots, P_n$. Then the kernel trick replaces the Gram matrix $(\langle X_i, X_j \rangle) \in \mathbb{R}^{n \times n}$ with the random *kernel matrix* $K(X) := (K(X_i, X_j)) \in \mathbb{R}^{n \times n}$. Understanding the behavior of this random matrix, and especially how well it concentrates around its mean is key in evaluating the performance of the underlying kernel methods. This problem has been studied in the literature, but often in the asymptotic setting, including the classical asymptotics where d is fixed and $n \to \infty$ or in the (moderately) high-dimensional regime where $d, n \to \infty$ and $d/n \to \gamma \in (0, 1)$.

In this paper, we study finite-sample concentration of $K(X)$ around its mean in the ℓ_2 operator norm, that is, $\|K(X) - \mathbb{E}K(X)\|$. We will make no assumptions about the relative

Received September 2019; revised January 2020.
MSC2020 subject classifications. 62H20, 60E15, 62G99.
Key words and phrases. Concentration inequalities, kernel matrices, nonasymptotic bounds, kernel spectral clustering.

sizes of d and n; our results hold for any scalings of the pair (n, d). We also do not assume the kernel (function) to be positive semidefinite, using the term kernel broadly to refer to any symmetric real-valued function defined on $\mathbb{R}^d \times \mathbb{R}^d$. We consider the class of Lipschitz kernels and provide a concentration inequality when the data distributions $\{P_i\}$ correspond to certain classes of distributions, including the strongly log-concave distributions in $\mathbb{R}^d$. In particular, the result holds for general Gaussian distributions $P_i = N(\mu_i, \Sigma_i)$, $i = 1, \ldots, n$. For comparison, we also derive a concentration inequality for the usual Euclidean kernel, for certain classes of sub-Gaussian vectors. Our results highlight differences in dimension dependence between the concentration of Lipschitz kernels versus that of the Euclidean one. Another interesting observation is that, in contrast to the Euclidean case, the concentration inequality for Lipschitz kernels does not depend on the mean kernel $\mathbb{E}K(X)$.

A feature of our results is that the data, although independent, are not assumed to be identically distributed. This is important, for example, when studying clustering problems and implies that the mean kernel matrix $\mathbb{E}K(X)$ is nontrivial and can carry information about the underlying data distribution. Thus, one can study the behavior of a kernel method on the mean matrix $\mathbb{E}K(X)$ and then translate the results to a random sample, using the concentration equality.

We illustrate this approach by analyzing a kernel spectral clustering algorithm which is recently introduced in the context of network clustering. We adapt the algorithm to general kernel clustering, and provide bounds on its misclassification rate under a (nonparametric) mixture model that is perturbed by noise. Due to our concentration results, the bound we derive allows for anisotropic noise models as well as noise structures that vary with the signal. This, in turn, allows one to investigate an interesting trade-off between the noise and signal structure. There could be multiple ways of breaking the data into the signal and noise components. For example, consider $X_i = \mu_i + \varepsilon_i$ where μ_i is the signal component and $\varepsilon_i \sim N(0, \Sigma)$ the independent isotropic noise. An alternative decomposition is

$$X_i = \mu_i' + \varepsilon_i' \quad \text{for } \mu_i' = \mu_i + \Pi_{\mu_i}^{\perp}\varepsilon_i, \varepsilon_i' = \Pi_{\mu_i}\varepsilon_i,$$

where Π_{μ_i} is the operator projecting onto the span of $\{\mu_i\}$, and $\Pi_{\mu_i}^{\perp} = I_d - \Pi_{\mu_i}$ is its complementary projection operator. This latter decomposition has varying anisotropic noise $\varepsilon_i' \sim N(0, \Pi_{\mu_i}\Sigma\Pi_{\mu_i})$, but could allow for faster concentration of the kernel matrix (conditioned on $\{\mu_i\}$) when $\max_i \|\Pi_{\mu_i}\Sigma\Pi_{\mu_i}\|$ is smaller than $\|\Sigma\|$. We illustrate the application of our concentration bound by analyzing a nested sphere cluster model under isotropic and radial noise models, and show that the proposed kernel spectral clustering algorithm achieves high-dimensional consistency under both noise structures.

In addition to the trade-off in decomposition, the bound on the misclassification rate also shows an interesting trade-off between the approximation (by a block-constant matrix) and estimation errors. This trade-off is controlled by certain parameters of the mean kernel $\mathbb{E}K(X)$, denoted as γ^2 and $\overline{v}^2$ in Section 3, that characterize the between-cluster distance and the within-cluster variation. Both of these are further affected by the noise level σ and, in the case of the Gaussian kernel, by the kernel bandwidth.

1.1. *Related work.* Most of the prior work on the concentration of kernel matrices focuses on the asymptotic behavior. For fixed d, as $n \to \infty$, the eigenvalues of the normalized kernel matrix $K(X)/n$ converge to the eigenvalues of the associated integral operator if (and only if) the operator is Hilbert–Schmidt. This is shown in [17] which also provides rates of convergence and distributional limits.

More recently, the so-called high-dimensional asymptotic regime where $n, d \to \infty$ while d/n converges to a constant is considered. The study of kernel matrices in this regime was initiated by [10] where it was shown that for kernels with entries of the form $f(X_i^T X_j)$ and

$f(\|X_i - X_j\|)$, under a certain scaling of the distribution of $\{X_i\}$, the empirical kernel matrix asymptotically behaves similar to that obtained from a linear (i.e., Euclidean) kernel.

In particular, it was shown in [10] that the operator norm distance between the kernel matrix and its linearized version vanishes asymptotically; hence, for example, the corresponding spectral densities approach each other. The limiting spectral density (i.e., the limit of the empirical density of the eigenvalues) has been further studied for kernels with entries of the form $f(X_i^T X_j)$ and $f(\|X_i - X_j\|)$ in [8, 9, 12] under various (often relaxed) regularity assumptions on f and the distribution of $\{X_i\}$. In parallel work, [11] considers a signal-plus-noise model for X_i and shows that the kernel matrix, in this case, approaches a kernel matrix which is based on the signal component alone. Although the results are mostly asymptotic, they have similarities with our approach. We make a detailed comparison with [11] in Remark 1 and Section 3.4.

Early results on finite-sample concentration bounds for kernel matrices include [5, 6, 22] for individual eigenvalues or their partial sums. In [5, 6], the deviation of the eigenvalues of the empirical kernel matrices (or their partial sums) from their counterparts based on the associated integral operator are considered. In [22], nonasymptotic concentration bounds on the eigenvalues have been obtained for bounded kernels. In our notation, these bounds show that $|\lambda_i(K) - \mathbb{E}\lambda_i(K)|$ are small. In contrast, a consequence of our results is a control on $|\lambda_i(K) - \lambda_i(\mathbb{E}K)|$. In applications, getting a handle on $\lambda_i(\mathbb{E}K)$ is often much easier than $\mathbb{E}\lambda_i(K)$.

More recently, sharp nonasymptotic upper bounds on the operator norm of random kernel matrices were obtained in [16] for the case of polynomial and Gaussian kernels. These results focus on the case where X_i are *centered* sub-Gaussian vectors and provide direct bounds on the operator norm of the kernel matrix: $\|K\|$. In contrast, we focus on the case where X_i have a nonzero mean μ_i and $\mathbb{E}K$ has nontrivial information about these mean vectors, and we provide bounds on the deviation of K from $\mathbb{E}K$.

Much of the work on the analysis of spectral clustering focuses on the Laplacian-based approach. In a line of work, the convergence of the *adaptive* graph Laplacian to the corresponding Laplace–Beltrami operator is established [4, 13–15, 24]. For a *fixed* kernel, the convergence of the (empirical) graph Laplacian to the corresponding population-level integral operator is studied in [20, 27], and bounds on the deviation of the corresponding spectral projection operators are derived. More recently, a finite-sample analysis for fixed kernels is provided in [21] assuming an explicit mixture model for the data. Our work is close in spirit to [21] with notable differences. We consider an adjacency-based kernel spectral clustering, based on a recently proposed algorithm for network clustering, and provide direct bounds on its misclassification rate. Our bound requires no assumption on the signal structure, and the overall bound is simpler and in terms of explicit quantities related to the statistical properties of a mean kernel. We separate the contributions of the noise and signal (in contrast to [21]), which allows for a more refined analysis. In particular, we show how this could lead to high-dimensional consistency of the proposed kernel spectral clustering in some examples. Another recent work in the same spirit as ours is that of [29] where both a spectral method and a SDP relaxation are analyzed for clustering based on a kernel matrix. A mixture model with isotropic sub-Gaussian noise is considered in [29] and consistency results are obtained for both approaches, based on entrywise concentration bounds for the kernel matrix. We provide more detailed comparisons with the existing literature on kernel clustering in Section 3.4.

The rest of the paper is organized as follows: In Section 2, we derive the concentration inequalities for the Lipschitz and Euclidean kernels. Section 3 presents an application of these results in deriving misclassification bounds for kernel spectral clustering. In Section 3.5, we present simulation results corroborating the theory. We conclude by giving the proofs of the main results in Section 4, leaving some details to the Appendices in the Supplementary Material [3].

2. Concentration of kernel matrices. Throughout, $\{X_i, i = 1, \dots, n\}$ will be a collection of independent random vectors in $\mathbb{R}^d$. The sequence is not assumed i.i.d., that is, the distribution of X_i could in general depend on i. This, for example, is relevant to clustering applications. We will collect $\{X_i\}$ into the data matrix $X = (X_1, \dots, X_n) \in \mathbb{R}^{d\times n}$. We also use the notation $X = (X_1 \mid \cdots \mid X_n)$ to emphasize that X_i is the ith column of X. For a vector $x \in \mathbb{R}^n$, $\|x\| = \|x\|_2$ denotes the ℓ_2 norm. For a matrix $A \in \mathbb{R}^{n\times n}$, we use $\|A\|$ to denote the ℓ_2 operator norm, also known as the spectral norm.

We are interested in bounds on the deviation $\|K - \mathbb{E}K\|$, where $K = (K_{ij}) \in \mathbb{R}^{n\times n}$ is a kernel matrix. That is, $K_{ij} = K(X_i, X_j)$, where with some abuse of notation, we will use the same symbol K to denote both the kernel matrix and the kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Occasionally, we write $K(X)$ for the kernel matrix when we want to emphasize the dependence on X. Thus,

$$K(X) = \big(K(X_i, X_j)\big) \in \mathbb{R}^{n\times n}. \tag{1}$$

For a random vector X_i, we denote its covariance matrix as $\operatorname{cov}(X_i)$. We often work with Lipschitz functions. A function $f : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz with respect to (w.r.t.) metric δ on $\mathbb{R}^d$ if it has a finite Lipschitz seminorm:

$$\|f\|_{\mathrm{Lip}} := \sup_{x,y} \frac{|f(x) - f(y)|}{\delta(x, y)} < \infty.$$

It is called L-Lipschitz if $\|f\|_{\mathrm{Lip}} \le L$. If the metric is not specified, it is assumed to be the Euclidean metric, $\delta(x, y) := \|x - y\|$.

We consider the data model $X_i = \mu_i + \sqrt{\Sigma_i}W_i$, $i = 1, \dots, n$, where Σ_i is a *generalized square-root* of the positive semidefinite matrix Σ_i, in the sense that $\sqrt{\Sigma_i}\sqrt{\Sigma_i}^T = \Sigma_i$. Note that $\sqrt{\Sigma_i}$ need not be symmetric.

2.1. *Lipschitz kernels.* Our first result is for the case where the kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz, in the following sense:

$$\big|K(x_1, x_2) - K(y_1, y_2)\big| \le L\big(\|x_1 - y_1\| + \|x_2 - y_2\|\big). \tag{2}$$

This class includes any kernel function which is L-Lipschitz w.r.t. the ℓ_2 norm on $\mathbb{R}^{2d}$. It also includes the important class of *distance kernels* of the form (see Appendix A.1 in the Supplementary Material [3]):

$$K(x_1, x_2) = f\big(\|x_1 - x_2\|\big), \quad f : \mathbb{R} \to \mathbb{R} \text{ is } L\text{-Lipschitz}, \tag{3}$$

which in turn includes the important case of the Gaussian kernel where $f(t) \propto e^{-t^2/2\sigma^2}$. We also need the following definition.

DEFINITION 1. We say that a random vector $Z \in \mathbb{R}^d$ is strongly log-concave with curvature α^2 if it has a density $f(x) = e^{-U(x)}$ (w.r.t. the Lebesgue measure) such that $\nabla^2 U(x) \succeq \alpha^2 I_d$ for every $x \in \mathbb{R}^d$, that is, the Hessian of U exists and is uniformly bounded below.

We often work with the following class of multivariate distributions.

DEFINITION 2 (LC class). We say that random vector $X \in \mathbb{R}^d$ belongs to class $\mathrm{LC}(\mu, \Sigma, \omega)$ for some vector $\mu \in \mathbb{R}^d$, a $d \times d$ semidefinite matrix Σ and $\omega > 0$, if we can write $X = \mu + \sqrt{\Sigma}W$ where $W \in \mathbb{R}^d$ is a random vector whose jth coordinate, W_j, satisfies $\mathbb{E}W_j = 0$ and $\mathbb{E}W_j^2 = 1$ for all j. Moreover, either of the following conditions holds:

(a) $W_j = \phi_j(Z_j)$, for some function ϕ_j with $\|\phi_j\|_{\mathrm{Lip}} \le \omega$, for all j, and $\{Z_j\}$ is a collection of independent standard normal variables; or

(b) $\{W_j\}$ are independent and W_j has a density (w.r.t. the Lebesgue measure) uniformly bounded below by $1/\omega$; or

(c) W is strongly log-concave with curvature $\alpha^2 \ge 1/\omega^2$, and $\mathbb{E}WW^T = I_d$.

For part (b) of Definition 2, we say that a density f is uniformly bounded below, if $f(x) \ge 1/\omega > 0$ for all x in the support of the distribution. Part (b) thus includes the case where the marginals of X are uniformly distributed on bounded subsets of $\mathbb{R}$ and $\mathrm{cov}(X)^{-1/2}(X - \mathbb{E}X)$ has independent coordinates. Note that a multivariate Gaussian random vector is a special case of Definition 2 with $\omega = 1$. Our main result for the Lipschitz kernels is the following.

THEOREM 1. *Let $X_i \in \mathrm{LC}(\mu_i, \Sigma_i, \omega)$, $i = 1, \dots, n$, be a collection of independent random vectors, and let $K = K(X)$ be the kernel matrix in (1) with kernel function satisfying (2). Then, for some universal constant $c > 0$, with probability at least $1 - \exp(-ct^2)$,*

$$\|K - \mathbb{E}K\| \le 2L\omega\sigma_\infty(Cn + \sqrt{n}t), \tag{4}$$

where $\sigma_\infty^2 := \max_i \|\Sigma_i\|$ and $C = c^{-1/2}$. When all X_i s are multivariate Gaussians, one can take $c = 1/2$.

Although this result is stated for the LC classes of random vectors, it holds more broadly. In fact, we can even relax the independence assumption on $W_1, \dots, W_n$. Inspection of the proof shows that the result holds as long as $\vec{W} \in \mathbb{R}^{dn}$, which is obtained by stacking $\{W_i\}$ on top of each other, satisfies the so-called *concentration property*; see Definition 3 in Section 4.

Bound (4) is dimension-free. To see this, consider the case where $\Sigma_i = \sigma^2 I_d$ for all i. Then we have $\frac{1}{n}\|K - \mathbb{E}K\| = O(L\omega\sigma)$ with probability at least $1 - e^{-cn}$, for all d. The bound is also independent of $\{\mu_i\}$. The following proposition shows the bound is sharp.

PROPOSITION 1. *Let $X_i, i = 1, \dots, n$ be i.i.d. draws from a symmetric distribution with $\mathbb{P}(|X_i| > \sigma) = 1/2$, for example, the uniform distribution on $(-2\sigma, 2\sigma)$. Then, for any $\sigma > 0$, there is an L-Lipschitz kernel function on $\mathbb{R}$ such that, when $n \ge 8$, the corresponding kernel matrix $K = K(X)$ satisfies*

$$\mathbb{P}\big(\|K - \mathbb{E}K\| > L\sigma n/8\big) \ge 1 - e^{-n/8}. \tag{5}$$

The $1/2$ in assumption $\mathbb{P}(|X_i| > \sigma) = 1/2$, is for convenience. It can be replaced with any positive constant by modifying the constants in (5).

REMARK 1. As an intermediate step in proving Theorem 1, we obtain (cf. Proposition 4),

$$\frac{1}{n^2}\mathbb{E}\|K - \mathbb{E}K\|_F^2 \le \frac{4}{c}L^2\omega^2 \max_i \|\Sigma_i\|. \tag{6}$$

This is a significant strengthening of a result that follows from Theorem 1 in [11]: After a rescaling to match the two models, the result there implies

$$\frac{1}{n^2}\mathbb{E}\|K - \widetilde{K}\|_F^2 \le CL^2\big[\mathrm{tr}(\Sigma^2) + C_1\|\Sigma\|\big] \tag{7}$$

for the case where $\Sigma_i = \Sigma$ for all i, the kernel is of the form (3) and $\widetilde{K}$ is a modified kernel matrix where $f(\cdot)$ is replaced with $f(\cdot + \mathrm{tr}(\Sigma))$ off the diagonal and with $f(0)$ on the diagonal.

Our result is much sharper since the bound does not scale with d. It is also more general in some aspects, namely, that it applies to any Lipschitz kernel, not necessarily of the form (3), and we allow for heterogeneity in the covariance matrices of the data points. Our result is stated in terms of the mean matrix $\mathbb{E}K$ which is a more natural object. Moreover, we prove a full concentration result in Theorem 1 which goes beyond controlling the mean of the deviation as in (6) and (7). On the other hand, the result in [11] is more general in another direction: it applies to $X_i = \mu_i + \sqrt{\Sigma}W_i$ where W_i have independent coordinates with bounded fourth moments. (Note that Σ is the same for all data points in [11].) Since we seek exponential concentration, we need stronger control of the tail probabilities.

EXAMPLE 1 (Gaussian kernel and isotropic noise). Let us consider the implications of Theorem 1 for the Gaussian kernel, assuming that the underlying random vectors follow:

$$X_i = \mu_i + \frac{\sigma_i}{\sqrt{d}} w_i, \quad w_i \overset{\text{i.i.d.}}{\sim} N(0, I_d). \tag{8}$$

As will be discussed in Section 3, by allowing μ_i to vary over some latent clusters in the data, (8) provides a simple model for studying clustering problems. The scaling of the noise variances by $\sqrt{d}$ is so that the two terms μ_i and $(\sigma_i/\sqrt{d})w_i$ are balanced in size as $d \to \infty$. Without the scaling, since $\|w_i\|$ concentrates around $\sqrt{d}$, the noise $\sigma_i w_i$ will wash out the information in the signal μ_i (assuming $\|\mu_i\| = O(1)$ as $d \to \infty$).

Consider the Gaussian kernel on $(\mathbb{R}^d)^2$ with bandwidth parameter τ:

$$K(x, y) = \exp\left(-\frac{1}{2\tau^2}\|x - y\|^2\right) = f_\tau(\|x - y\|), \qquad f_\tau(t) := e^{-t^2/2\tau^2}. \tag{9}$$

This is a Lipschitz kernel with $L = \|f'_\tau\|_\infty = \sqrt{2}/(e\tau)$. The expected kernel matrix $\mathbb{E}K$ has the following entries (see Appendix B.4):

$$[\mathbb{E}K]_{ij} = \frac{1}{s_{ij}^d} \exp\left(-\frac{\|\mu_i - \mu_j\|^2}{2s_{ij}^2\tau^2}\right), \quad s_{ij}^2 = 1 + \frac{\sigma_i^2 + \sigma_j^2}{d\tau^2}, i \neq j.$$

Consider the special case where $\sigma_i = \sigma$ for all i, and let $s^2 = 1 + 2\sigma^2/(d\tau^2)$. Then the mean kernel matrix $\mathbb{E}K$ is itself a kernel matrix, based on a Gaussian kernel with updated bandwidth parameter τs, applied to mean vectors $\{\mu_i\}$, that is,

$$\widetilde{K}_\sigma(\mu_i, \mu_j) := [\mathbb{E}K]_{ij} = s^{-d} f_{\tau s}(\|\mu_i - \mu_j\|).$$

Note that the mean kernel matrix depends on the noise variance σ. Also, because of the scaling of the variance in (8), the prefactor s^{-d} stabilizes as $d \to \infty$, that is, $s^{-d} = (1 + 2\sigma^2/(d\tau^2))^{-d/2} \to e^{-\sigma^2/\tau^2}$ and the kernel function approaches the standard Gaussian kernel $f_{s\tau} \to f_1$. (Without the variance scaling, the prefactor would go to zero.)

Applying (4) with $\sigma_\infty = \sigma$, $\omega = 1$, $c = 1/2$, $L = \sqrt{2}/(e\tau)$ and replacing t with $\sqrt{2}t$,

$$\frac{1}{n}\|K - \mathbb{E}K\| \leq \frac{4}{e}\frac{\sigma}{\tau}\frac{1}{\sqrt{d}}\left(1 + \frac{t}{\sqrt{n}}\right), \quad \text{w.p.} \geq 1 - e^{-t^2}. \tag{10}$$

It is interesting to note that the deviation is controlled by the ratio σ/τ. For example, we could have started with the alternative model without the scaling of the standard deviation by $\sqrt{d}$, that is, model (8) with $\sigma_i/\sqrt{d}$ replaced with σ, but instead rescaled the bandwidth by changing τ to $\tau\sqrt{d}$. Then we would have the same exact concentration bound as in (10). This observation somewhat justifies the rule of thumb used in practice where one sets the bandwidth $\propto \sqrt{d}$ in the absence of additional information. According to the above discussion, this choice roughly corresponds to the belief that the per-coordinate standard deviation is $O(1)$ as $d \to \infty$.

Example 1 can be easily extended to the case of anisotropic noise, using the invariance of both the Gaussian kernel and the Gaussian distribution to unitary transformations. More generally, consider an extension of model (8) as follows:

$$X_i = \mu_i + \frac{1}{\sqrt{d}} w_i, \quad w_i \overset{\text{i.i.d.}}{\sim} N(0, \Sigma). \tag{11}$$

This is similar to the model in [11], assuming in addition the Gaussianity of the noise. Applying (4), replacing Σ with $\Sigma/\sqrt{d}$, we have for model (11),

$$\frac{1}{n}\|K - \mathbb{E}K\| \le 2\sqrt{2}L\sqrt{\frac{\|\Sigma\|}{d}}\left(1 + \frac{t}{\sqrt{n}}\right), \quad \text{w.p.} \ge 1 - e^{-t^2}. \tag{12}$$

In practice, it is often reasonable to assume $\|\Sigma\| = O(1)$. Then $\frac{1}{n}\|K - \mathbb{E}K\| = O_p(d^{-1/2})$ as $d \to \infty$, that is, we get consistency in estimating $\mathbb{E}(K/n)$ by K/n, as dimension d grows.

2.2. *Euclidean kernel.* We now consider the kernel function $K(x_1, x_2) = \langle x_1, x_2 \rangle$ which we refer to as the *Euclidean* or *inner product* kernel. The kernel matrix in this case is the Gram matrix of $\{X_i\}$:

$$K(X) = \big(\langle X_i, X_j \rangle\big) = X^T X. \tag{13}$$

Our main result for the Euclidean kernel is the following.

THEOREM 2. *Let $X_i = \mu_i + \sqrt{\Sigma_i} W_i$, where $\{W_i, i = 1, \dots, n\} \subset \mathbb{R}^d$ is a collection of independent centered random vectors, each with independent sub-Gaussian coordinates. Here, $\mu_i = \mathbb{E}[X_i] \in \mathbb{R}^d$ and each Σ_i is a $d \times d$ positive semidefinite matrix, with generalized square root $\sqrt{\Sigma_i}$. Let*

$$M = (\mu_1 \mid \cdots \mid \mu_n) \in \mathbb{R}^{d \times n}, \qquad \kappa = \max_{i,j} \|W_{ij}\|_{\psi_2},$$

$$\sigma_\infty^2 := \max_i \|\Sigma_i\|, \qquad \eta = d + \left(\frac{\|M\|}{\kappa \sigma_\infty}\right)^2.$$

For $K = K(X)$ as in (13) and for any $u \ge 0$, with probability at least $1 - 4n^{-c_1}\exp(-c_2 u^2)$,

$$\|K - \mathbb{E}K\| \le 2\kappa^2 \sigma_\infty^2 \eta \max\big(\delta^2, \delta\big) \quad \textit{where } \delta = \sqrt{\frac{n}{\eta}} + \frac{u}{\sqrt{\eta}}.$$

In particular, with probability at least $1 - 4n^{-c_1}$,

$$\begin{aligned} \|K - \mathbb{E}K\| &= O\big(\kappa^2 \sigma_\infty^2 (n + \sqrt{n\eta})\big) \\ &= O\big(\kappa^2 \sigma_\infty^2 (n + \sqrt{nd}) + \kappa \sigma_\infty \sqrt{n} \|M\|\big). \end{aligned} \tag{14}$$

A special case of this result, when X_i s are centered and isotropic ($\mu_i = 0$, $\Sigma_i = I_d$ and $\mathbb{E}W_{ij}^2 = 1$ for all i and j), appears in [25], Section 5.5. The normalized $n \times n$ kernel matrix $\frac{1}{n}X^T X$ is dual to the $d \times d$ matrix $\frac{1}{n}XX^T = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$ which is the main component of the sample covariance matrix of $\{X_i\}$. Thus, Theorem 2 is dual to the well-known concentration results for covariance matrices. However, a major difference with covariance estimation is that with Gram matrices, the data points need not have identical distributions.

An interesting feature of bound (14) is its dependence on the mean of the underlying vectors through $\|M\|$. Contrast this with the result of Theorem 1 where the bound is not affected by the mean of the random vectors X_i. Under the assumptions of Theorem 2, the mean kernel matrix is $\mathbb{E}K = \operatorname{diag}(\mathbb{E}\|\tilde{X}_i\|^2, i \in [n]) + M^T M$, where $\tilde{X}_i = X_i - \mu_i$ is the

centered version of X_i. The second term has operator norm $\|M^T M\| = \|M\|^2$, whereas the relevant term in (14) is of lower order in $\|M\|$. More precisely, $\frac{\|K-\mathbb{E}K\|}{\|\mathbb{E}K\|} \lesssim \frac{1}{\|M\|}$ as $\|M\| \to \infty$, confirming that (14) is indeed a concentration result.

EXAMPLE 2. Let us continue with model (8) of Example 1. The model corresponds to $\Sigma_i = \sigma_i^2 I_d/d$ and $\kappa \lesssim 1$ in Theorem 2. Assume that $\sigma_i \le \sigma$ for all i. It follows that $\sigma_\infty \le \sigma/\sqrt{d}$ and Theorem 2 gives

$$\frac{1}{n}\|K - \mathbb{E}K\| \lesssim \sigma^2\Big(\frac{1}{d} + \frac{1}{\sqrt{nd}}\Big) + \frac{\sigma}{\sqrt{nd}}\|M\|, \quad \text{w.p.} \ge 1 - 4n^{-c_1}.$$

Compared with (10), the deviation bound improves as d is increased. On the other hand, the bound directly depends on the mean matrix $M = \mathbb{E}X$, as opposed to (10).

The bound in (14) is sharp in general. To see this, first consider the term $\kappa^2\sigma_\infty^2(n + \sqrt{nd})$. Without loss of generality, assume $\sigma_\infty^2 = 1$. Consider the case $X_i \sim N(0, I_d)$, drawn i.i.d., and let y_k be the kth row of $(X_1 \mid \cdots \mid X_n)$. Then y_k, $k = 1, \dots, d$ are i.i.d. draws from $N(0, I_n)$. Hence, $\frac{1}{d}\|K - \mathbb{E}K\| = \|\frac{1}{d}\sum_k y_i y_i^T - I_n\|$ is the deviation of a sample covariance matrix from its expectation which is known to scale as $\sqrt{\frac{n}{d}} + \frac{n}{d}$; see, for example, [26], Theorem 4.7.1.

The last term in (14) is also unavoidable when $n \ge Cd$ for a sufficiently large constant C. To see this, let $X = \sigma_\infty^{-1} M + \sigma_\infty W \in \mathbb{R}^{d\times n}$ where X_i, M_i and W_i are the ith columns of X, M and W, respectively, and $W_i \sim N(0, I_d)$ drawn i.i.d. Letting $\sigma_\infty \to 0$, we have $\|K - \mathbb{E}K\| \to 2\|M^T W\|$. Note that $\frac{1}{n}WW^T$ is a sample covariance matrix, concentrated around I_d. By taking $n \ge Cd$ for a large constant C, we have $\frac{1}{n}WW^T \succeq \frac{1}{2}I_d$, with high probability. It follows that $2\|M^T W\| = 2\sqrt{\|M^T WW^T M\|} \ge 2(\frac{n}{2}\|M^T M\|)^{1/2} \ge \sqrt{2n}\|M\|$, which is proportional to bound (14) after replacing M with $\sigma_\infty^{-1}M$ and letting $\sigma_\infty \to 0$.

3. Kernel spectral clustering. We now consider how the concentration bounds of Section 2 can be used to derive performance bounds for the kernel spectral clustering.

3.1. *A kernel clustering algorithm.* Let $\mu \mapsto \Sigma(\mu)$ be a map from $\mathbb{R}^d$ to positive semidefinite matrices, and let $\sqrt{\Sigma(\mu)}$ denote its matrix square-root. We consider a nonparametric mixture model perturbed by noise, as follows:

$$X_i = \mu_i + \frac{\sigma}{\sqrt{d}}\sqrt{\Sigma(\mu_i)}w_i, \quad \mu_i \overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{R} \bar{\pi}_k P_k, \ w_i \overset{\text{i.i.d.}}{\sim} N(0, I_d), \tag{15}$$

for $i = 1, \dots, n$, where μ_i is the signal, w_i is the noise, and the two pieces are independent. Note that the distribution of X_i goes beyond a nonparametric mixture model unless $\mu \mapsto \Sigma(\mu)$ is constant. The reason for introducing the extra parameter σ is the convenience of setting $\Sigma(\mu_i) = I_d$ to study the case of isotropic noise. We think of $\Sigma(\mu)$ as a normalized covariance matrix (say $\|\Sigma(\mu)\| \le 1$) measuring anisotropy of the noise, and of σ as the overall noise level. The Gaussian assumption for w_i is for simplicity; the result holds for all the cases in Theorem 1. Here, $\{P_k\}$ are the distributions constituting the mixture components, and $\bar{\pi}_k \in [0, 1]$ are the class priors. In a typical case, components $\{P_k\}$ are supported on lower-dimensional submanifolds of $\mathbb{R}^d$, singular w.r.t. the Lebesgue measure and singular w.r.t. to each other; see, for example, Figure 1. Although, none of these assumptions are required for the result we present. Intuitively, the kernel clustering should perform well if we only observe $\{\mu_i\}$ and we would like to study the effect of adding noise to such ideal clustered data.

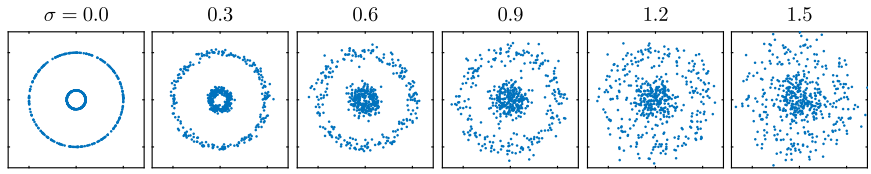

FIG. 1. *Example of the signal-plus-noise clustering model* (16) *with two signal component* P_1 *and* P_2, *each a uniform distribution on a circle in* $d = 2$ *dimensions, and* $\Sigma_0 = I_2$. *The plots correspond to different noise levels* σ.

Model (15) is sufficiently general to allow the noise structure to vary based on the signal. A special case is when $\Sigma(\mu) = \Sigma_0$ is constant, in which case the model is equivalent to

$$X_i = \mu_i + \frac{\sigma}{\sqrt{d}} w_i', \quad \mu_i \overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{R} \bar{\pi}_k P_k, \, w_i' \overset{\text{i.i.d.}}{\sim} N(0, \Sigma_0). \tag{16}$$

This special case is often encountered in the literature.

Given a kernel function, we can form the kernel matrix $K = K(X)$ as in (1). Throughout this section, unless otherwise stated, we condition on $\mu = (\mu_i)$, hence the expectations and probability statements are w.r.t. the randomness in $w = (w_i)$. Let $\widetilde{K}_\sigma(\mu) := \mathbb{E}[K(x_i, x_j)]$, which should be interpreted as $\widetilde{K}_\sigma(\mu) = \mathbb{E}[K(x_i, x_j) \mid \mu]$, by the convention just discussed. The mean kernel matrix $\widetilde{K}(\mu)$ has the following off-diagonal entries under model (15):

$$\big[\widetilde{K}_\sigma(\mu)\big]_{ij} = [\mathbb{E}K]_{ij} = \widetilde{K}_\sigma(\mu_i, \mu_j), \quad i \neq j, \tag{17}$$

where, with some abuse of the notation regarding $\widetilde{K}_\sigma$, we have defined:

$$\widetilde{K}_\sigma(u, v) := \mathbb{E}\left[K\left(u + \frac{\sigma}{\sqrt{d}} \sqrt{\Sigma(u)} w_1, v + \frac{\sigma}{\sqrt{d}} \sqrt{\Sigma(v)} w_2 \right) \right], \quad u \neq v. \tag{18}$$

Here, the expectation is w.r.t. the randomness in w_1 and w_2. Note that we are using $\widetilde{K}_\sigma$ to refer to both the mean kernel matrix and the corresponding kernel function. In the special case of constant noise covariance, $\Sigma(\mu) = \Sigma_0$, we simply have

$$\widetilde{K}_\sigma(u, v) := \mathbb{E}\left[K\left(u + \frac{\sigma}{\sqrt{d}} w_1', v + \frac{\sigma}{\sqrt{d}} w_2' \right) \right], \quad u \neq v, \tag{19}$$

where w_1' and w_2' are independent $N(0, \Sigma_0)$ variates. The properties of the new kernel matrix $\widetilde{K}_\sigma(\mu)$ plays a key role in our analysis.

We analyze the kernel-based spectral clustering (KSC) approach summarized in Algorithm 1 which is based on the recent SC-RRE algorithm of [31] for network clustering. An

Algorithm 1 A kernel spectral clustering (KSC) algorithm

Input: (a) Data points $x_1, \ldots, x_n \in \mathbb{R}$, (b) the number of clusters R and (c) the kernel function $(x, y) \mapsto K(x, y)$, not necessarily positive semidefinite.

Output: Cluster labels.

1: Form the normalized kernel matrix $A := (K(x_i, x_j)/n) \in \mathbb{R}^{n \times n}$.

2: Obtain $A^{(R)} = \hat{U}_1 \hat{\Lambda}_1 \hat{U}_1^T$, the R-truncated eigenvalue decomposition (EVD) of A. That is, if $A = \hat{U} \hat{\Lambda} \hat{U}^T$ is the full EVD of A, where $\hat{\Lambda} = \operatorname{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_n)$ with $|\hat{\lambda}_1| \geq \cdots \geq |\hat{\lambda}_n|$, then $\hat{\Lambda}_1 = \operatorname{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_R)$, and $\hat{U}_1 \in \mathbb{R}^{n \times R}$ collects the first R columns of $\hat{U}$.

3: Apply an isometry-invariant, constant-factor, k-means algorithm (with R clusters) on $\hat{U}_1 \hat{\Lambda}_1$ to recover the cluster labels.

advantage of this spectral algorithm is that we can provide theoretical guarantees that are explicitly expressed in terms of the original parameters of the model, avoiding eigenvalues in the statement of the bounds. The connection with network clustering is as follows: We can treat $K/n \in \mathbb{R}^{n\times n}$ as a similarity matrix, effectively defining a weighted network among n entities, and then use the adjacency-based spectral clustering described in [31].

Algorithm 1 proceeds by forming the R-truncated eigenvalue decomposition of the similarity matrix $A = K/n$, denoted as $A^{(R)} = \hat{U}_1\hat{\Lambda}_1\hat{U}_1^T$. One then performs a constant-factor approximate k-means algorithm on the rows of $\hat{U}_1\hat{\Lambda}_1$ to obtain the estimated cluster labels. The details of this step are as follows: For a set $\mathcal{Y} = \{y_1, \dots, y_R\} \subset \mathbb{R}^D$ and any point $x \in \mathbb{R}^D$, let $d(x, \mathcal{Y}) = \min_{y\in\mathcal{Y}} \|x - y\|$. The k-means problem, with R clusters, seeks to minimize $\sum_{i=1}^n d(X_i, \mathcal{Y})^2$ over R-element subsets $\mathcal{Y}$ of $\mathbb{R}^D$. This problem is in general NP-hard. However, it is possible to find κ-approximate solutions in polynomial-time, that is, $\widehat{\mathcal{Y}}$ such that $\sum_i d(X_i, \widehat{\mathcal{Y}})^2 \le \kappa \cdot \min_{\mathcal{Y}} \sum_i d(X_i, \mathcal{Y})^2$. Given, $\widehat{\mathcal{Y}}$, every point X_i is mapped to the closest element of $\widehat{\mathcal{Y}}$, producing cluster labels. We further assume that the algorithm for deriving the κ-approximate solution is isometry-invariant, that is, it only depends on the pairwise distances among $\{X_i\}$. Examples of such algorithms for deriving a $\kappa = 1 + \varepsilon$ approximation are the approach of [18] with time complexity $O(2^{\mathrm{poly}(R/\varepsilon)}nD)$ [1] and that of [7] with complexity $O(nDR + 2^{\mathrm{poly}(R/\varepsilon)}D^2\log^{D+2} n)$. Since we apply these algorithms with $\varepsilon = O(1)$ and $D = R$, assuming $R = O(1)$, both algorithms run in $O(n)$ time.

3.2. *Finite-sample bounds on misclassification error.* Let $z_i \in \{0,1\}^R$ be the label of data point i, determining the component of the mixture to which μ_i belongs. We use one-hot encoding for z_i, so that $z_{ik} = 1$ if and only if data point i belongs to cluster k, that is, $\mu_i \sim P_k$. Let $\mathcal{C}_k := \{i : z_{ik} = 1\}$ denote the indices of data points in the kth cluster, $n_k := |\mathcal{C}_k|$ and $\pi_k := n_k/n$, the size and the (empirical) proportion of the kth cluster, respectively.

For $k, \ell \in [R]$, let $\widehat{P}_{k,\ell}$ be the empirical measure on $\mathbb{R}^d \times \mathbb{R}^d$ given by

$$\widehat{P}_{k\ell} := \widehat{P}_{k\ell}(\mu) = \frac{1}{n_k n_\ell} \sum_{(i,j)\in[n]^2} z_{ik}z_{j\ell}\delta_{(\mu_i,\mu_j)} = \frac{1}{n_k n_\ell} \sum_{i\in\mathcal{C}_k, j\in\mathcal{C}_\ell} \delta_{(\mu_i,\mu_j)},$$

where $\delta_{(\mu_i,\mu_j)}$ is a point-mass measure at (μ_i, μ_j). In words, $\widehat{P}_{k\ell}$ is the empirical measure when the data consists of pairs (μ_i, μ_j), as i and j range over the kth and ℓth clusters, respectively. Consider the mean and variances of these empirical measures:

$$\Psi_{k\ell} := \mathbb{E}\big[\widetilde{K}_\sigma(X, Y)\big], \qquad v_{k\ell}^2 := \mathrm{var}\big(\widetilde{K}_\sigma(X, Y)\big), \quad (X, Y) \sim \widehat{P}_{k\ell}. \tag{20}$$

Let $\overline{v}^2$ be the average variance

$$\overline{v}^2 := \sum_{k,\ell\in[R]} \pi_k\pi_\ell v_{k\ell}^2, \tag{21}$$

and define the following minimum separations:

$$\gamma^2 := \min_{k\neq\ell} D_{k\ell}, \qquad \widetilde{\gamma}^2 := \min_{k\neq\ell} \pi_\ell D_{k\ell}, \quad D_{k\ell} := \sum_{r=1}^R \pi_r(\Psi_{kr} - \Psi_{\ell r})^2. \tag{22}$$

When the clusters are roughly balanced, we have $\pi_k \asymp 1/R$ for all $k \in [R]$, hence $\widetilde{\gamma}^2 \asymp \gamma^2/R$. If the number of clusters does not grow with n, then $\widetilde{\gamma}^2 \asymp \gamma^2$.

Let $\{\widehat{z}_i\}$ be the labels outputted by Algorithm 1 and let $\overline{\mathrm{Mis}}$ be the corresponding average misclassification rate relative to the true labels. That is, $\overline{\mathrm{Mis}} = \min_\sigma \frac{1}{n}1\{\sigma(\widehat{z}_i) \neq z_i\}$ where the minimum is take over all permutations $\sigma : [R] \to [R]$. (Here, we treat both $\widehat{z}_i$ and z_i as elements of $[R]$.) We are now ready to state our result on the performance of kernel spectral clustering.

THEOREM 3. *Assume that the data points* $\{X_i, i = 1, \dots, d\} \subset \mathbb{R}^d$ *follow the nonparametric noisy mixture model* (15). *Consider the kernel spectral clustering Algorithm* 1 *with an* L*-Lipschitz kernel function as in* (2). *Let* $\overline{v}^2$ *and* γ^2 *be defined, based on* $\widetilde{K}_\sigma$, *as given in* (18). *Fix* $t \geq 0$, *and let*

$$F(\gamma^2, \overline{v}^2) := \frac{16R}{\gamma^2}\left[\frac{4L^2\sigma^2}{d}\left(1 + \frac{t}{\sqrt{n}}\right)^2 \max_i \|\Sigma(\mu_i)\| + \overline{v}^2\right] \tag{23}$$

and $C_1 := 4(1+\kappa)^2$ *where* κ *is the approximation factor of the* k*-means algorithm. Assume that* $F(\widetilde{\gamma}^2, \overline{v}^2) \leq C_1^{-1}$. *Then, with probability at least* $1 - \exp(-t^2)$, *the average misclassification rate of Algorithm* 1 *satisfies*

$$\overline{\text{Mis}} \leq C_1 F(\gamma^2, \overline{v}^2). \tag{24}$$

A similar result can be stated for the Euclidean kernel of Section 2.2. Consider the special case where $\Sigma(\mu) = \Sigma_0$ for all μ. The quantity $\overline{v}^2/\gamma^2$ in (23) is a measure of the hardness of the noiseless clustering problem, which we refer to as the approximation error. The first term in bound (23) is the contribution due to noise, the so-called estimation error. Both quantities depend on the noise level σ as well as the noise structure $\Sigma(\mu_i)$, through $\widetilde{K}_\sigma$ in (19). Thus, a more precise statement is that $\overline{v}^2/\gamma^2$ measures the hardness of the noiseless problem *at the appropriate level determined by the noise level* σ, *and noise structure* $\Sigma(\mu_i)$. This dependence on noise can become negligible in the high-dimensional setting where $d \to \infty$; see Section 3.3.

The geometry of the signal directly affects the approximation error. In a classical parametric mixture model, $P_k = \delta_{\mu_k^*}$, that is, point masses at $\{\mu_1^*, \dots, \mu_R^*\}$, in which case $\overline{v}^2 = 0$, that is, the approximation error vanishes. Another example with $\overline{v}^2 = 0$, is the case of separating a point mass at the origin, $P_1 = \delta_0$, from the sphere, $P_2 = \text{Unif}(S^{d-1})$. A more elaborate example is the nested sphere model discussed below.

In addition, both the approximation and estimation errors depend on the choice of the kernel function $K(\cdot, \cdot)$: the estimation error through the Lipschitz constant L and approximation error clearly as the definitions of $\overline{v}^2$ and γ^2 show. When the kernel class has a tuning parameter, one might be able to trade-off the contributions of these terms as the following example shows.

EXAMPLE 3 (Spectral clustering with Gaussian kernel). Consider the case of constant isotropic noise $\Sigma_0 = I_d$ and the Gaussian kernel (9) with bandwidth τ. As discussed in Example 1, the Lipschitz constant is $L \lesssim 1/\tau$. Thus the misclassification bound (23) in this case reduces to

$$\overline{\text{Mis}} \lesssim \frac{R}{\gamma^2}\left[\frac{\sigma^2}{\tau^2}\frac{1}{d}\left(1 + \frac{t}{\sqrt{n}}\right)^2 + \overline{v}^2\right] \tag{25}$$

which holds with probability $\geq 1 - e^{-t^2}$. Roughly speaking, assuming $R = O(1)$, the estimation error is $\lesssim \sigma^2/(\gamma^2\tau^2 d)$ and the approximation error $\lesssim \overline{v}^2/\gamma^2$. The estimation error is $O(d^{-1})$, as $d \to \infty$, assuming that γ^2 stays away from 0, which is the case as discussed in Section 3.3.

As argued in Example 1, $\widetilde{K}_\sigma$ is again a Gaussian kernel, with modified bandwidth:

$$\widetilde{K}_\sigma(\mu_i, \mu_j) = s^{-d} f_{\tau s}(\|\mu_i - \mu_j\|), \qquad s^2 = 1 + \frac{2\sigma^2}{d\tau^2}. \tag{26}$$

Since $\overline{v}^2$ and γ^2 are defined based on $\widetilde{K}_\sigma$, both the approximation and estimation errors depend on the normalized bandwidth τ/σ. In addition, the approximation error also depends on

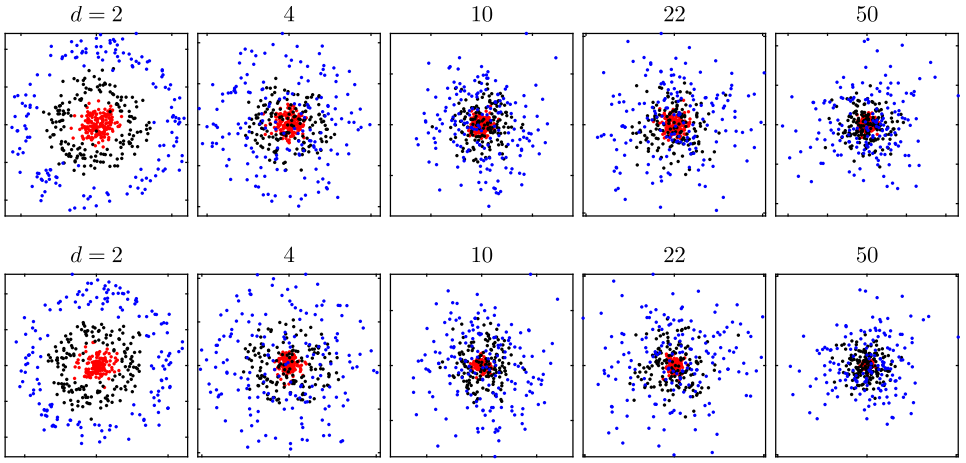

FIG. 2. *Plots of the first two coordinates of X_i for the "nested spheres" example, with radii $r_i = 1, 5, 10$, noise level $\sigma = 1.5$ and variable d. The top and bottom rows corresponds to the isotropic versus radial noise models, respectively. The plots look qualitatively the same in both cases. As can be seen, for large d, it is very hard to distinguish the clusters from a low-dimensional projection.* (*The scale of the plots varies with d.*)

the bandwidth-normalized pairwise distances of the signal component, that is, $\|\mu_i - \mu_j\|/\tau$, for $i, j \in [n]$. It is interesting to note that the dependence of the approximation error on the noise level σ vanishes as $d \to \infty$. In Example 5 below, we provide explicit limit expressions for $\overline{v}^2$ and γ^2.

EXAMPLE 4 (Nested spheres with radial noise). Assume that the signal mixture components $\{P_k\}$ are uniform distributions on nested spheres in $\mathbb{R}^d$ of various radii: $r_1, \ldots, r_R$. Assume that to each μ_i, drawn from the mixture, we add a Gaussian noise in the direction perpendicular to the sphere, that is,

$$X_i = \mu_i + \frac{\sigma}{\sqrt{d}} \frac{\mu_i}{\|\mu_i\|} \xi_i, \quad \xi_i \overset{\text{i.i.d.}}{\sim} N(0, 1). \tag{27}$$

Figure 2 illustrates an example of this noise setup in comparison with the isotropic case. The radial noise structure falls under model (15) with $\Sigma(\mu) := \mu\mu^T/\|\mu\|^2$, i.e., the rank-one projection onto the span of μ. Since $\max_i \|\Sigma(\mu_i)\| = 1$, the misclassification bound obtained from (23) is similar to (25) in the isotropic case, with $1/\tau^2$ replaced with L^2. Thus, the dominant term in the estimation error is $\lesssim (RL^2\sigma^2)/(\gamma^2 d)$ which is $O(1/d)$ as $d \to \infty$, assuming that γ^2 stays bounded away from 0. (This is the case as discussed in Section 3.3.) Note that the behavior of the estimation error is the same as that of the isotropic case. Let us also compute the mean kernel function, assuming as the base, the usual Gaussian kernel (9). We have

$$\begin{aligned}\widetilde{K}_\sigma(u, v) &:= \mathbb{E}\left[K\left(u + \frac{\sigma}{\sqrt{d}}\widetilde{u}\xi_1, v + \frac{\sigma}{\sqrt{d}}\widetilde{v}\xi_2\right)\right] \\ &= \mathbb{E}\exp\left(-\frac{\|u - v + (\sigma/\sqrt{d})w\|^2}{2\tau^2}\right),\end{aligned}$$

where $\widetilde{u} = u/\|u\|$, $\widetilde{v} = v/\|v\|$, and $w = \widetilde{u}\xi_1 - \widetilde{v}\xi_2 \sim N(0, \widetilde{u}\widetilde{u}^T + \widetilde{v}\widetilde{v}^T)$. One can show that

$$\widetilde{K}_\sigma(u, v) = \frac{1}{s_1 s_2} \exp\left\{-\frac{1}{2\tau^2}\left[\frac{\lambda_1}{2s_1^2}\big(\|u\| - \text{sign}(\alpha)\|v\|\big)^2\right.\right.$$

$$+\frac{\lambda_2}{2s_2^2}\big(\|u\|+\operatorname{sign}(\alpha)\|v\|\big)^2\Big]\Big\}, \tag{28}$$

$$s_i^2 = 1+\frac{\sigma^2\lambda_i}{\tau^2 d}, \quad i=1,2,$$

$$\lambda_1 = 1+|\alpha|, \qquad \lambda_2 = 1-|\alpha|, \quad \alpha = \frac{\langle u, v\rangle}{\|u\|\|v\|}$$

assuming that $\alpha \neq 0$, and $u \neq v$. See Appendix B.1 for details. It is interesting to note that this mean kernel mostly depends on the norms of u and v. The dependence on α, the angle between u and v, is quite weak (through s_i^2 and $\operatorname{sign}(\alpha)$) and mostly goes away as $d \to \infty$. In the next section, we argue that the approximation error $\overline{v}^2/\gamma^2$ based on this kernel also goes to zero as $n, d \to \infty$.

3.3. *Population-level parameters*. The quantities $v_{k\ell}^2$ and $\Psi_{k\ell}^2$ that underlie $\overline{v}^2$ and γ^2, and control the approximation error in Theorem 3, are defined based on the empirical measures $\widehat{P}_{k\ell}$. But it is also possible to state them directly in terms of the underlying population-level components $\{P_k\}$ and the related integrals. The main idea is that $\widehat{P}_{k\ell}$, in general, has a well-defined limit:

$$\widehat{P}_{k\ell} \to P_k \otimes P_\ell \quad \text{as } n \to \infty, \text{ w.h.p.}, \tag{29}$$

where the convergence can be interpreted in various senses (e.g., weak convergence of probability measures, or convergence in L^p Wasserstein distances). The notation $P_k \otimes P_\ell$ represents a product measure, that is, if $(X, Y) \sim P_k \otimes P_\ell$, then X and Y are independent variables with marginal distributions P_k and P_ℓ. The convergence in (29) holds even when $k = \ell$ (cf. Proposition 2 below). Let

$$\Psi_{k\ell}^* := \int \widetilde{K}_\sigma(\mu, \mu')\, dP_k(\mu)\, dP_\ell(\mu'), \qquad \big(v_{k\ell}^*\big)^2 := \operatorname{var}\big(\widetilde{K}_\sigma(X, Y)\big),$$

where $(X, Y) \sim P_k \otimes P_\ell$. Similarly, let $D_{k\ell}^*$, γ_*^2 and $\overline{v}_*^2$ be the population-level versions of $D_{k\ell}$, γ^2 and $\overline{v}^2$ obtained by replacing $\Psi_{k\ell}$ and $v_{k\ell}^2$ with their starred versions in the corresponding definitions. The above discussion suggests that for large n, $\Psi_{k\ell} \approx \Psi_{k\ell}^*$ and $v_{k\ell}^2 \approx (v_{k\ell}^*)^2$ and similarly for the other related quantities. The following result formalizes these ideas.

PROPOSITION 2. *Assume that $\widetilde{K}_\sigma$ has constant diagonal and is uniformly bounded on the union of the supports of P_k, $k \in [R]$, so that $|\widetilde{K}_\sigma(\mu_i, \mu_j)| \le b$ a.s. for all $i, j \in [n]$ and some $b > 0$. Then, with probability at least $1 - 4R^2\exp(-t^2)$, for all $k, \ell \in [R]$,*

$$\big|\Psi_{k\ell} - \Psi_{k\ell}^*\big| \le \frac{3bt}{\sqrt{n_k \wedge n_\ell}} =: \delta_{k\ell}, \qquad \big|v_{k\ell}^2 - \big(v_{k\ell}^*\big)^2\big| \le \frac{9b^2t}{\sqrt{n_k \wedge n_\ell}}.$$

Letting $\pi_{\min} = \min_k \pi_k$, on the same event, we have

$$\gamma^2 \ge \gamma_*^2 - \frac{24b^2t}{\sqrt{\pi_{\min}}}\frac{1}{\sqrt{n}}, \qquad \overline{v}^2 \le \overline{v}_*^2 + \frac{9b^2t}{\sqrt{\pi_{\min}}}\frac{1}{\sqrt{n}}. \tag{30}$$

Note that the bounds in (30) are dimension-free: Assume that $\pi_{\min}$ is bounded below. Then, as long as n is sufficiently large, both γ_*^2 and $\overline{v}_*^2$ are good approximations for their empirical versions, irrespective of how large d is. When γ_*^2 is bounded below, we can replace γ^2 and $\overline{v}^2$ in the misclassification bound in Theorem 3 and only pay a price of $O(n^{-1/2})$.

COROLLARY 1. *Consider the setup of Theorem 3 and further assume that γ_*^2 is bounded below, as $d \to \infty$. Then, for any $t \geq 0$, there is a constant $C_2 = C_2(\pi_{\min}, b, t)$, such that for $n \geq C_2 \gamma_*^{-2}$, with probability at least $1 - 5R^2 \exp(-t^2)$, the average misclassification rate of Algorithm 1 satisfies*

$$\overline{\text{Mis}} \leq 2C_1 F(\gamma_*^2, \overline{v}_*^2) + \frac{C_3(t)}{\sqrt{n}}, \tag{31}$$

where $C_3(t) = 18b^2 t/(\sqrt{\pi_{\min}}\gamma_*^2)$, assuming that $F(\widetilde{\gamma}_*^2, \overline{v}_*^2) + \frac{C_3}{\sqrt{n}} \leq C_1^{-1}$.

The boundedness assumption in Proposition 2 holds if either $\widetilde{K}_\sigma$ is uniformly bounded on $\mathbb{R}^d$ (as in the case of the Gaussian kernel), or $\{P_k\}$ are supported on some bounded manifolds and $\widetilde{K}_\sigma$ is continuous. The second assumption is quite reasonable since it assumes the "true" signal μ_i to be bounded whereas the noisy observation x_i can still have an unbounded distribution.

In some cases, one might be able to explicitly compute γ_*^2 and $\overline{v}_*^2$ as the next examples show.

EXAMPLE 5 (Nested spheres with isotropic noise). Consider the case where $\{P_k\}$ are uniform distributions on nested spheres in $\mathbb{R}^d$ of various radii: $r_1, \ldots, r_R$. Recalling the definition of s in (26), let

$$\widetilde{r}_k = \frac{r_k}{\tau s}, \qquad \widetilde{u}_k := s^{-d/2} e^{-\widetilde{r}_k^2/2} \quad \text{and} \quad u_k := e^{-(r_k^2+\sigma^2)/2\tau^2}$$

for $k \in [R]$. Let θ and θ' be independent variables distributed uniformly on the unit sphere S^{d-1}, and set $\psi_d(u) = \mathbb{E}\exp(u\langle\theta, \theta'\rangle)$. Then it is not hard to see that

$$\Psi_{k\ell}^* = \mathbb{E}\big[\widetilde{K}_\sigma(r_k\theta, r_\ell\theta')\big] = \widetilde{u}_k\widetilde{u}_\ell\psi_d(\widetilde{r}_k\widetilde{r}_\ell),$$

$$(v_{k\ell}^*)^2 = \operatorname{var}\big[\widetilde{K}_\sigma(r_k\theta, r_\ell\theta')\big] = \widetilde{u}_k^2\widetilde{u}_\ell^2\big[\psi_d(2\widetilde{r}_k\widetilde{r}_\ell) - \psi_d(\widetilde{r}_k\widetilde{r}_\ell)\big].$$

Although, ψ_d can be written as a Beta integral, let us consider the case of large d (high-dimensional data) which simplifies the expressions. As $d \to \infty$, both $\widetilde{r}_k$ and $\widetilde{u}_k$ stabilize since $s \to 1$ and $s^{-d/2} \to e^{-\sigma^2/2\tau^2}$ (see Example 1). It follows that $\widetilde{r}_k \to r_k/\tau$ and $\widetilde{u}_k \to u_k$. One can also show that $\psi_d(u) \approx \exp(u^2/4d)$ for $u \ll d$ (see Appendix B.2). Then $\Psi_{k\ell} \to u_k u_\ell$ and $v_{k\ell}^2 \to 0$ as $d \to \infty$, assuming that the bandwidth τ and the radii $\{r_k\}$ remain fixed.

The population-level approximation error is bounded (up to constants) by

$$\frac{\overline{v}_*^2}{\gamma_*^2} = O\left(\frac{C_1(u)}{C_2(u)}\frac{(r_k r_\ell)^2}{\tau^4 d}\right) = O\left(\frac{1}{d}\right) \quad \text{as } d \to \infty, \tag{32}$$

which is vanishing as d gets large. Here,

$$C_1(u) = \max_k u_k^4, \qquad C_2(u) = \left(\sum_t \pi_t u_t^2\right)\min_{k\neq\ell}(u_k - u_\ell)^2.$$

To simplify the numerator, we have used $\psi(u)/\psi(2u) \approx 1 - e^{-3u^2/4d} \approx 3u^2/4d$ as $d \to \infty$. Note that the prefactor in (32) makes intuitive sense: The bound is controlled by the closest sphere to the origin (having largest u_k, hence largest variance) in the numerator and the two closest spheres in the denominator.

Let us now consider the population-level estimation error. As discussed in Example 3, the estimation error is bounded up to constants by

$$\frac{1}{\gamma_*^2}\frac{\sigma^2}{\tau^2}\frac{1}{d} \asymp \frac{1}{C_2(u)}\frac{\sigma^2}{\tau^2}\frac{1}{d}.$$

Increasing τ^2 decreases the effect of noise by reducing σ^2/τ^2, but increases $1/\gamma_*^2 \asymp 1/C_2(u)$ by making $\{u_k\}$ closer, since all u_k approach 1 as $\tau \to \infty$. This also increases the approximation error (32) in general. Thus the bandwidth to noise level τ/σ plays a subtle role in balancing the effect of the two terms. Since both the estimation and approximation errors go down as $O(d^{-1})$, KSC is consistent at an overall rate of $O(d^{-1} + n^{-1/2})$, as implied by (31).

EXAMPLE 6 (Nested spheres with radial noise). Consider again the nested spheres as the signal model, but this time with (anisotropic) radial noise model of Example 4. We can proceed as in Example 5 in estimating parameters γ_*^2 and $\overline{v}_*^2$. The only difference is that we need to use the appropriate kernel mean matrix $\widetilde{K}_\sigma$, given by (28) in this case. Let $u_k = e^{-r_k^2/2\tau^2}$. Then one can show that (cf. Appendix B.3) as $d \to \infty$,

$$\Psi_{k\ell}^* = \mathbb{E}\big[\widetilde{K}_\sigma(r_k\theta, r_\ell\theta')\big] \to u_k u_\ell, \qquad (v_{k\ell}^*)^2 = \operatorname{var}\big[\widetilde{K}_\sigma(r_k\theta, r_\ell\theta')\big] \asymp \frac{u_k^2 u_\ell^2}{\tau^4 d}.$$

These estimates are similar to those obtained in Example 5, hence the same bound (32) holds for $\overline{v}_*^2/\gamma_*^2$ in this case; that is, the population-level approximation error goes down as $O(d^{-1})$, similar to the case of the isotropic noise. Since the estimation error also goes down as $O(d^{-1})$ in this case (cf. Example 4), KSC is consistent at an overall rate of $O(d^{-1}+n^{-1/2})$, as implied by (31).

Let us summarize our analysis for the nested spheres example with the isotropic and radial noise models. Assume that σ, τ (the kernel bandwidth), $\pi_{\min}$ and the radii of the spheres remain constant. For both noise models, the bound on the approximation error vanishes at a rate $O(d^{-1} + n^{-1/2})$, while the bound on the estimation error vanishes at a rate $O(d^{-1})$, for sufficiently large n. Irrespective of which noise structure one assumes (i.e., radial or isotropic), the KSC is consistent at a rate $O(d^{-1} + n^{-1/2})$ for the nested sphere signal. This conclusion is corroborated by simulations in Section 3.5.

3.4. *Comparison with existing literature.* The work of [29] considers a model of the form (15) with $P_k = \delta_{\mu_k^*}$, that is, point masses at $\{\mu_1^*, \dots, \mu_R^*\}$, $\sigma = 1$ and $\Sigma(\mu_k^*) = \sigma_k^2 I_d$ for $k = 1, \dots, R$. They consider clustering based on a kernel of the form $K(x, y) = f(\|x - y\|^2)$ where f is both bounded and Lipschitz. They analyze a Laplacian-based spectral clustering algorithm, using a row and column normalized version of the kernel matrix K, and obtain bounds on its misclassification rate, involving the eigenvalues of a block-constant version of K. When all the noise variances, and pairwise distances among $\{\mu_k^*\}$, are equal, the eigenvalue bound simplifies to give a consistency rate of $O(\log d/d)$.

When $d \geq 2$, the class of multivariate Lipschitz kernels allowed by Theorem 3 is much larger than that of the bounded distance-based kernels considered in [29]. For example, a kernel that takes two input images and processes them through a ReLU neural network, with operator-norm bounded weight matrices, falls within the Lipschitz class we consider. We note, however, that the particular form considered in [29] is not necessarily a Lipschitz kernel unless $\sup_t |tf'(t^2)| < \infty$, hence could fall outside our class. Our result also allows for more general noise and signal structures. In particular, the signal $P_k = \delta_{\mu_k^*}$ considered in [29] corresponds to the classical parametric mixtures. This model gives $\overline{v}^2 = 0$ in our result, leading to zero approximation error, hence a $O(1/d)$ convergence rate from Theorem 3, improving the rate of [29] by a $\log d$ factor for Lipschitz kernels. This holds even if the entire covariance matrix of the noise changes at every data point, as long as $\max_i \|\Sigma_i\| \lesssim 1$. It is also worth noting that, in contrast to bounds based on eigenvalues which are often hard to interpret, our bound is directly in terms of interpretable quantities $\overline{v}^2$ and γ^2. On the other hand, [29] allows for the existence of outliers which we do not consider. They also obtain a strong consistency

(exact recovery) result for a semidefinite programming variant of kernel clustering, which falls outside the scope of this paper.

The work of [21] considers a finite nonparametric mixture model on a *compact* space. Their model is equivalent to (16) with the noise component set to zero ($\sigma = 0$), that is, assuming $X_i \sim \sum_k \bar{\pi}_k P_k$ with P_k compactly supported. In contrast, we assume $X_i \sim \sum_k \bar{\pi}_k P_k * N(0, \sigma^2 \Sigma_0/d)$, in the special case of constant covariance noise. Here, $*$ denotes convolution. In fact, we can allow for the convolution with any member of the LC class defined earlier, including strongly log-concave densities. This allows us to model mixture components with infinite support on $\mathbb{R}^d$, a more realistic setup not covered in [21]. In addition, compactness together with the continuity of the kernel function, assumed in [21], implies a bounded kernel while we allow for unbounded Lipschitz kernels. Moreover, in contrast to [21], we do not require the kernel function to be positive semidefinite. The main focus of [21] is to establish a geometric property for the embedding of the data points obtained from a Laplacian-based kernel representation.

Under suitable conditions, [21] establishes what they call an (α, θ)-orthogonal cone structure (OCS) for that embedding [21], Theorem 2. This means that a $1-\alpha$ fraction of the points from each mixture component lie within a cone of angle α centered at one of the coordinate axes. They also show that under further assumptions on α and θ, a randomized k-means algorithm applied to an embedding, with an (α, θ)-OCS structure, leads to a misclassification rate at most α [21], Proposition 1. The implicit nature of the multiple conditions on α and θ in these two results, however, makes it difficult to parse out an explicit misclassification rate. Moreover, α at best is a constant and cannot go to zero to establish consistency. In contrast, we provide an explicit misclassification bound in terms of easily computable quantities and derive explicit rates of convergence as d and n diverge.

It is worth noting that our results apply to model (15) which in its general form (with variable covariance structure) goes beyond even a finite nonparametric mixture model for $\{X_i\}$. As far as we know, the general case of model (15) has not been analyzed for clustering before. The special case in model (16) is the same as the signal plus noise model of [11] with covariance matrix Σ in that paper replaced with $\sigma \Sigma_0$. However, in contrast to (16), [11] does not consider any structure for the signal and the problem there is only to establish the closeness of the kernel matrix based on the pure signal and that based on the contaminated signal.

Finally, our work is based on the technical machinery developed in [31] for the analysis of network spectral clustering. In particular, we leveraged the approach of [31] in deriving eigenvalue-free bounds on misclassification rate. The results of [31], however, are not directly applicable to kernel clustering, since the (symmetric) deviation matrix $A - \mathbb{E}A$ there, is assumed to have independent entries on and above the diagonal. In contrast, the deviation $K - \mathbb{E}K$ for a kernel matrix does not have independent entries on and above the diagonal. Deriving a concentration bound for such a matrix was the main focus of this paper, allowing us to provide the main missing ingredient of the analysis.

3.5. *Simulations*. We now provide some simulations to corroborate the theory we developed for the kernel spectral clustering. We use the "nested spheres" example that we analyzed in Sections 3.2 and 3.3. We compare the performance of the kernel spectral clustering described in Algorithm 1 with the Lloyd's algorithm (with `kmeans++` initialization) applied directly to the data points.

For the kernel function, we consider the Gaussian kernel with bandwidth set as $\tau^2 = \alpha(1 + \sigma^2)$, for $\alpha = 1, 2$. This scaling of τ^2 in terms of σ^2 is motivated by the concentration bounds, where the estimation error is controlled by σ^2/τ^2. Constant 1 is added to avoid degeneracy when $\sigma \to 0$.

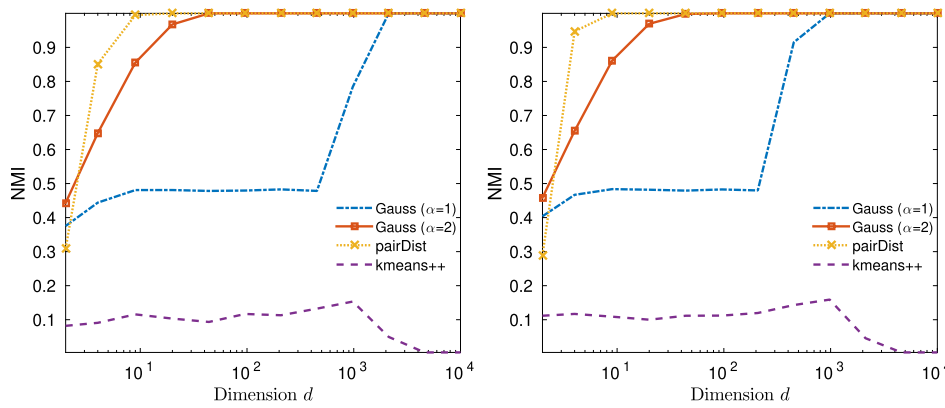

FIG. 3. *Plots of NMI versus dimension for kernel spectral clustering Algorithm 1, under the noisy "nested spheres" model with radii $r_i = 1, 5, 10$ (three clusters). Left and right plots correspond to isotropic versus radial noise, respectively. Here, $n = 500$, $\sigma = 1.5$ and $\tau^2 = \alpha(1 + \sigma^2)$.*

In addition to the Gaussian kernel, we also use the simple *pairwise distance* (`pairDist`) kernel $K(x, y) = \|x - y\|$. Since this kernel is 1-Lipschitz, all the theory developed in the paper applies in this case, with appropriate modifications to the mean kernel $\widetilde{K}_\sigma$. In particular, one can argue as in Examples 4 and 6 that for the radial noise model, this kernel is also consistent as $n, d \to \infty$. Note that although a more appropriate choice would be $(x, y) \mapsto -\|x - y\|$ to make the kernel a similarity measure, the sign is irrelevant in spectral clustering.

Figure 3 shows the results. The plots show the normalized mutual information (NMI) versus dimension d, for a fixed value of $\sigma = 1.5$ and a sample size of $n = 500$. The "nested spheres" signal with radii $r_i = 1, 5, 10$ (three clusters) is considered along with both the isotropic and radial noise models. The plots show the normalized mutual information (NMI) obtained by the KSC algorithm (relative to the true labels) as the dimension varies from $d = 2$ to $d = 10^4$. The NMI is a similarity measure between two cluster assignments, more aggressive than the average accuracy. A random clustering against the truth produces NMI ≈ 0, while a prefect match gives NMI $= 1$. The plots are obtained by averaging over 12 independent replicates.

The right and left panels in Figure 3 correspond to the radial and isotropic noise model, respectively. The plots show that, with either noise structure, the KSC Algorithm 1 is consistent for the pairwise distance as well as the Gaussian kernel, for both values of α, eventually, as d grows. These results are as predicted by the theory. Note that for the Gaussian kernel with $\alpha = 2$, consistency in the isotropic case is achieved at a "slightly higher dimension d," consistent with the intuition that the isotropic model corresponds to the radial case with spheres "slightly closer." The tuition is based on translating the isotropic model to the radial model by projecting the noise onto the sphere. However, the linear projection is imperfect in putting the transverse noise component exactly on the sphere, hence causing the spheres to appear closer relative to the purely radial noise.

4. Proofs of the main results. Let us start by giving high-level ideas of the proofs. For Theorem 1, we first show that $\|K - \mathbb{E}K\|$ is a Lipschitz function of X. The distributions in class LC have the property that any Lipschitz function of X concentrates around its mean. This allows us to show that $\|K - \mathbb{E}K\|$ is concentrated near $\mathbb{E}\|K - \mathbb{E}K\|$. We bound this latter expectation by $\mathbb{E}\|K - \mathbb{E}K\|_F$ which in turn is bounded by controlling $\operatorname{var}(K(X_i, X_j))$ for all pairs (i, j), again using the Lipschitz concentration property.

For Theorem 2, we first derive a tail bound for $|z^T(K - \mathbb{E}K)z|$, given a fixed $z \in S^{n-1}$. This bound requires an extension of the Hanson–Wright inequality to noncentered variables,

which is presented and proved in Appendix A.3. Equipped with the tail bound, we use a discretization argument to obtain uniform control over S^{n-1} and complete the proof.

For Theorem 3, we first approximate the normalized kernel matrix $A = K/n$, in operator norm, by a block-constant matrix, denoted as K^*_σ/n. Next, we argue that the eigenvalue-truncated version of A, namely $A^{(R)}$, is close to K^*_σ/n in Frobenius norm. Finally, we use k-means perturbation results to show that the misclassification error is bounded, up to constants, by $\|A^{(R)} - K^*_\sigma/n\|_F^2/\gamma^2$ where γ^2/n is related to the minimum center separation among the rows of K^*_σ/n. Combining these bounds gives the desired inequality (24).

In the rest of this section, we give details of the proofs, starting with some preliminary concentration results.

4.1. *Preliminaries.* Let us start with the following definition (borrowed from [2] with modifications).

DEFINITION 3. A random vector $Z \in \mathbb{R}^d$ satisfies the *concentration property* with constant $\kappa > 0$ if for any Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, with respect to the ℓ_2 norm, we have

$$\mathbb{P}\big(f(Z) - \mathbb{E}f(Z) > t\|f\|_{\mathrm{Lip}}\big) \le \exp\big(-\kappa t^2\big) \quad \forall t > 0. \tag{33}$$

Note that it is enough to have (3) for 1-Lipschitz functions (i.e., $\|f\|_{\mathrm{Lip}} = 1$) which then implies the general case by rescaling. The following result is well known [19]; see also [26], Theorem 5.2.2.

THEOREM 4. *A standard Gaussian random vector $Z \sim N(0, I_d)$ satisfies the concentration property with constant $\kappa = 1/2$.*

A similar result holds for a strongly log-concave random vector [26], Theorem 5.2.15.

THEOREM 5. *A strongly log-concave random vector $Z \in \mathbb{R}^d$ with curvature $\alpha^2 > 0$ satisfies the concentration property with constant $\kappa = C\alpha^2$ for some universal constant $C > 0$.*

This result can be easily extended to a collection of independent strongly log-concave random vectors.

COROLLARY 2. *Let $Z_1, \dots, Z_n \in \mathbb{R}^d$ be independent strongly log-concave random vectors with curvatures $\alpha_i^2 > 0$. Then $\vec{Z} \in \mathbb{R}^{nd}$ obtained by concatenating $Z_1, \dots, Z_n$ is strongly log-concave with curvature $\alpha^2 := \min_i \alpha_i^2$. In particular, $\vec{Z}$ satisfies the concentration property with constant $\kappa = C\alpha^2$.*

PROOF. It is enough to note that $\vec{Z}$ has density $f(z) = \prod_i e^{-U_i(z_i)} = e^{-U(z)}$ where $U(z) := \sum_i U_i(z_i)$ whose Hessian is block-diagonal with diagonal blocks $\nabla^2 U_i(z_i) \succeq \alpha_i^2 I_d$. □

We write $S^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ for the sphere in $\mathbb{R}^n$. We frequently use the following vector and matrix notation: For $X_1, \dots, X_n \in \mathbb{R}^d$, we write $X = [X_1 \mid \cdots \mid X_n]$ for the $d \times n$ matrix with columns X_i, and let

$$X \mapsto \vec{X} : \mathbb{R}^{d\times n} \to \mathbb{R}^{dn} \tag{34}$$

be the operator that maps a matrix X to a vector $\vec{X}$ by concatenating its columns.

4.2. *Proof of Theorem 1.* The key is the following lemma due to M. Rudelson which is proved in Appendix A.2.

LEMMA 1 (Rudelson). *Assume that $K(X)$ is as in (1) and the kernel function is L-Lipschitz as in (2). Then,*

(a) $\|K(X) - K(X')\|_F^2 \le 4nL^2\|X - X'\|_F^2$ *for any* $X, X' \in \mathbb{R}^{d\times n}$, *and*
(b) *for any* $a \in \mathbb{R}$, $X \mapsto \|K(X) - a\|$ *is* $2\sqrt{n}L$*-Lipschitz w.r.t. the Frobenius norm.*

Part (a) of Lemma 1 can be interpreted as showing that the matrix-valued map $X \mapsto K(X) : \mathbb{R}^{d\times n} \to \mathbb{R}^{n\times n}$ is $(2\sqrt{n}L)$-Lipschitz, assuming that both spaces are equipped with the Frobenius norm. As a consequence of Lemma 1, we get the following concentration inequality.

PROPOSITION 3. *Let $X_i = \mu_i + \sqrt{\Sigma_i}W_i \in \mathbb{R}^d$, $i = 1, \dots, n$ be random vectors and set $W = [W_1 \mid \cdots \mid W_n] \in \mathbb{R}^{d\times n}$. Assume that the random vector $\vec{W} \in \mathbb{R}^{dn}$ satisfies the concentration property (33) with constant $\kappa = c/\omega^2 > 0$. Let $K(X)$ be as defined in (1) with a kernel function satisfying (2). Then, $V := \|K - \mathbb{E}K\|$ is sub-Gaussian, and*

$$\mathbb{P}(V - \mathbb{E}V \ge 2\sqrt{n}L\sigma_\infty\omega t) \le \exp(-ct^2), \quad t \ge 0,$$

where $\sigma_\infty^2 := \max_i \|\Sigma_i\|$.

An equivalent (up to constant) statement of this result is

$$\big\|\|K - \mathbb{E}K\|\big\|_{\psi_2} \lesssim \sqrt{n}L\sigma_\infty\omega,$$

where $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm.

PROOF. Set $S_i = \sqrt{\Sigma_i}$ and let $S = \operatorname{diag}(S_1, \dots, S_n)$ be the $dn \times dn$ block diagonal matrix with diagonal blocks $\{S_i\}$. Also, let $X, W, \mu \in \mathbb{R}^{d\times n}$ be the matrices with columns $\{X_i\}$, $\{W_i\}$ and $\{\mu_i\}$, respectively. Using vector notation (34), we have $\vec{X} = \vec{\mu} + S\vec{Z}$. With some abuse of notation, we write $K(\vec{X})$ to denote $K(X)$ as defined in (1). Note that, $\|\vec{W}\| = \|W\|_F$, that is, the ℓ_2 norm of vector $\vec{W}$ is the same as the Frobenius norm of matrix W.

For any $a \in \mathbb{R}$, we claim that $\vec{W} \mapsto F(\vec{W}) := \|K(\vec{\mu} + S\vec{W}) - a\|$ is $(2\sqrt{n}L\sigma_\infty)$-Lipschitz w.r.t. the ℓ_2 norm on $\mathbb{R}^{dn}$. Indeed,

$$\begin{aligned}\big|F(\vec{W}) - F(\vec{W}')\big| &\le 2\sqrt{n}L\big\|S\vec{W} - S\vec{W}'\big\| \quad \text{(By Lemma 1(b))}\\ &\le 2\sqrt{n}L\|S\|\big\|\vec{W} - \vec{W}'\big\|\end{aligned}$$

and $\|S\| = \max_i \|S_i\| = \sigma_\infty$, since $\|S_i\|^2 = \|\Sigma_i\|$. The result now follows from (33) after replacing t with ωt. □

Next, we bound the expectation of $\|K - \mathbb{E}K\|$. Here, we pass to the Frobenius norm, giving us an upper bound on the expectation.

PROPOSITION 4. *Assume that $\{X_i\}_{i=1}^n$ satisfy the assumption of Proposition 3, and let $K = K(X)$ be as defined in (1) and satisfies (2). Then, with $C = 2/\sqrt{c}$,*

$$\mathbb{E}\|K - \mathbb{E}K\|_F \le CnL\omega\sigma_\infty.$$

PROOF. By Lemma 2 below,

$$\mathbb{E}\|K-\mathbb{E}K\|_F^2 = \sum_{i,j=1}^{n} \operatorname{var}\big(K(X_i,X_j)\big)$$
$$\le n^2 \max_{i,j} \operatorname{var}\big(K(X_i,X_j)\big) \le C^2 n^2 L^2 \omega^2 \sigma_\infty^2.$$

Noting that $\mathbb{E}\|K-\mathbb{E}K\|_F \le (\mathbb{E}\|K-\mathbb{E}K\|_F^2)^{1/2}$ completes the proof. □

LEMMA 2. *Assume that $X_i = \mu_i + \sqrt{\Sigma_i} W_i \in \mathbb{R}^d$ are independent for $i=1,2$, and $\vec{W} = (W_1, W_2) \in \mathbb{R}^{2d}$ satisfies the concentration property (33) with $\kappa = c/\omega^2 > 0$. Then, with $C^2 = 4/c$,*

$$\operatorname{var}\big(K(X_1,X_2)\big) \le C^2 L^2 \omega^2 \max\big\{\|\Sigma_1\|, \|\Sigma_2\|\big\},$$
$$\operatorname{var}\big(K(X_1,X_1)\big) \le C^2 L^2 \omega^2 \|\Sigma_1\|.$$

PROOF. For $x, y \in \mathbb{R}^d$, let $\vec{z} = (x,y)$ and define $\widetilde{K} : \mathbb{R}^{2d} \to \mathbb{R}$ by $\widetilde{K}(\vec{z}) := K(x,y)$. Note that $\widetilde{K}$ is $\sqrt{2}L$-Lipschitz w.r.t. to the ℓ_2 norm on $\mathbb{R}^{2d}$, that is, $|\widetilde{K}(\vec{z}) - \widetilde{K}(\vec{y})| \le \sqrt{2}L\|\vec{z}-\vec{y}\|$ for any $\vec{z}, \vec{y} \in \mathbb{R}^{2d}$. Let $\vec{\mu} = (\mu_1, \mu_2) \in \mathbb{R}^{2d}$, $\vec{W} = (W_1, W_2) \in \mathbb{R}^{2d}$ and $\mathbf{\Sigma} = \operatorname{diag}(\Sigma_1, \Sigma_2) \in \mathbb{R}^{2d\times 2d}$. We have $K(X_1,X_2) = \widetilde{K}(\vec{\mu} + \mathbf{\Sigma}^{1/2}\vec{W})$. We note that

$$\|\vec{W} \mapsto \widetilde{K}(\vec{\mu} + \mathbf{\Sigma}^{1/2}\vec{W})\|_{\mathrm{Lip}} \le \sqrt{2}L\|\mathbf{\Sigma}^{1/2}\| = \sqrt{2}L\sigma_\infty^{(12)}, \tag{35}$$

where $\sigma_\infty^{(12)} := \|\mathbf{\Sigma}^{1/2}\| = \max\{\|\Sigma_1^{1/2}\|, \|\Sigma_2^{1/2}\|\}$. From the concentration property, it follows that

$$\mathbb{P}\big(|K(X_1,X_2) - \mathbb{E}K(X_1,X_2)| > t\sqrt{2}L\sigma_\infty^{(12)}\omega\big) \le 2\exp(-ct^2) \quad \forall t > 0.$$

Letting $\Delta = K(X_1,X_2) - \mathbb{E}K(X_1,X_2)$ and $\alpha = \sqrt{2}L\sigma_\infty^{(12)}\omega$, we have

$$\mathbb{E}\Delta^2 = \int_0^\infty 2t\mathbb{P}\big(|\Delta| > t\big)\,dt = 2\alpha^2 \int_0^\infty t\mathbb{P}\big(|\Delta| > \alpha t\big)\,dt \le 4\alpha^2 \int_0^\infty te^{-ct^2}\,dt = \frac{2}{c}\alpha^2,$$

which gives the desired result for $\operatorname{var}(K(X_1,X_2))$ with $C^2 = 4/c$.

For the second assertion, let $J := \left[\begin{smallmatrix} I_d \\ I_d \end{smallmatrix}\right]$ and note that $K(X_1,X_1) = \widetilde{K}(J\mu_1 + J\Sigma_1^{1/2}W_1)$. We also have $\|W_1 \mapsto \widetilde{K}(J\mu_1 + J\Sigma_1^{1/2}W_1)\|_{\mathrm{Lip}} \le \sqrt{2}L\|\Sigma_1^{1/2}\|$. The rest of the argument follows as in the case of $K(X_1,X_1)$. □

Combining Propositions 3 and 4 and noting that $\mathbb{E}V \le \mathbb{E}\|K-\mathbb{E}K\|_F$ establishes the result for any collection of $\{W_i\}$ for which the concentration property holds for $\vec{W}$ with constant c/ω^2. It remains to verify that each case in Definition 2 has this property.

Verifying the three cases in the LC class. We first deduce the result for part (b) from (a). Fix i and j and let $f : \mathbb{R} \to \mathbb{R}$ denote the density of W_{ij} w.r.t. the Lebesgue measure, S the support of the distribution, and F the corresponding CDF, that is, $F(t) = \int_{-\infty}^t f(x)dx$. Pick $x \in S$ and note that x does not belong to flat parts of F. Then, by assumption $f(x) \ge 1/\omega$. Let $\nu = F(x)$ so that $x = F^{-1}(\nu)$. By the inverse function theorem, $Q := F^{-1}$ is differentiable at ν and we have $Q'(\nu) = 1/f(x) \le \omega$. Thus, Q is ω-Lipschitz on S. The range of Q restricted to S is $[0,1]$.

Let Φ be the CDF of the standard normal distribution which is $(1/\sqrt{2\pi})$-Lipschitz. If $Z_{ij} \sim N(0,1)$, then $U_{ij} := \Phi(Z_{ij})$ are uniformly distributed on $[0,1]$ and $Q(U_{ij})$ has the same distribution as W_{ij}. In other words, we can redefine $W_{ij} = \phi_{ij}(Z_{ij})$ for $\phi_{ij} = Q \circ \Phi$.

We have $\|\phi_{ij}\|_{\mathrm{Lip}} \le \|Q\|_{\mathrm{Lip}}\|\Phi\|_{\mathrm{Lip}} \le \omega/\sqrt{2\pi}$, and the problem is reduced to part (a), up to constants.

For part (a), we have $W_i = (W_{ij})$ with $W_{ij} = \phi_{ij}(Z_{ij})$ where $Z_{ij} \sim N(0,1)$ are independent across $i = 1,\dots,n$ and $j = 1,\dots,d$. We define $\vec{W}$ and $\vec{Z}$ based on the $d \times n$ matrices W and Z as in (34) and compactly write $\vec{W} = \phi(\vec{Z})$. Let $f : \mathbb{R}^{dn} \to \mathbb{R}$ be a 1-Lipschitz function and define $g(\vec{Z}) := f(\phi(\vec{Z})) = f(\vec{W})$. Then

$$\begin{aligned}\|g(\vec{Z}) - g(\vec{Z}')\|^2 &\le \sum_{ij} \bigl(\phi_{ij}(Z_{ij}) - \phi_{ij}(Z'_{ij})\bigr)^2 \\ &\le \sum_{ij} \|\phi_{ij}\|_{\mathrm{Lip}}^2 (Z'_{ij} - Z_{ij})^2 \le \omega^2 \|\vec{Z}' - \vec{Z}\|^2,\end{aligned}$$

for any vectors $\vec{Z}, \vec{Z}' \in \mathbb{R}^{dn}$. It follows that g is ω-Lipschitz, hence by the concentration of Gaussian measure (Theorem 4), we have

$$\mathbb{P}\bigl(g(\vec{Z}) - \mathbb{E}g(\vec{Z}) \ge \omega t\bigr) \le \exp(-t^2/2).$$

Since $g(\vec{Z}) = f(\vec{W})$, we have the concentration property for $\vec{W}$ with constant $1/(2\omega^2)$.

For part (c), since each W_i has a strongly log-concave density with curvature $\alpha_i^2 \ge 1/\omega^2$, it follows from Corollary 2 that $\vec{W}$ is strongly log-concave with curvature $1/\omega^2$. Then, by Theorem 5, $\vec{W}$ satisfies the desired concentration property with constant C/ω^2.

4.3. *Proof of Proposition 1.* Let us define $\phi : \mathbb{R} \to \mathbb{R}$ by setting $\phi(x)$ equal to $-\sqrt{L\sigma}$, $x\sqrt{L/\sigma}$ and $\sqrt{L\sigma}$ on $[-\infty, -\sigma]$, $[-\sigma, \sigma]$ and $[\sigma, \infty)$. Let $K(x,y) := \phi(x)\phi(y)$. We note that ϕ is $\sqrt{L\sigma}$-bounded and $\sqrt{L/\sigma}$-Lipschitz, hence $K(\cdot,\cdot)$ is L-Lipschitz. Let $u_i = \phi(X_i)$ and $u = (u_i) \in \mathbb{R}^n$. We have $\mathbb{E}K(X) = \alpha I_n$ where $\alpha = \mathbb{E}[\phi(X_1)]^2 \le L\sigma$, and $K(X) = uu^T$.

Let $Z_i = 1\{|X_i| > \sigma\}$. When $Z_i = 1$, $u_i = \pm\sqrt{L\sigma}$, hence $u_i^2 Z_i = L\sigma Z_i$. Assuming $\|u\|^2 \ge \alpha$, we have $\|K - \mathbb{E}K\| = \|u\|^2 - \alpha \ge \sum_i u_i^2 Z_i - \alpha \ge L\sigma(\sum_i Z_i - 1)$. Since $\sum_i Z_i \sim \mathrm{Bin}(n, \frac{1}{2})$, by the Hoeffding's inequality, $\mathbb{P}(\sum_i Z_i \le n/4) \le \exp(-n/8)$. On the complement of this event, $\sum_i Z_i - 1 \ge n/8$ when $n \ge 8$, completing the proof.

4.4. *Proof of Theorem 2.* We can write $K = X^T X$ where $X = (X_1 \mid \cdots \mid X_n) \in \mathbb{R}^{d\times n}$ has $\{X_i\}$ as its columns. Let us fix $z \in S^{n-1}$ and consider

$$Y_z := z^T(K - \mathbb{E}K)z = \|Xz\|^2 - \mathbb{E}\|Xz\|^2. \tag{36}$$

Let $\tilde{X}_i = X_i - \mu_i$ be the centered version of X_i, and let $\tilde{X} \in \mathbb{R}^{d\times n}$ be the matrix with columns $\{\tilde{X}_i\}$. Setting $\mu_z = Mz = \sum_i z_i \mu_i$, we have $Xz = \mu_z + \tilde{X}z$, hence

$$Y_z = \|\tilde{X}z\|^2 - \mathbb{E}\|\tilde{X}z\|^2 + 2\langle \mu_z, \tilde{X}z\rangle$$

using the fact that $\tilde{X}z$ is zero-mean.

LEMMA 3. *For any $z \in S^{n-1}$, Y_z in (36) based on $X_i = \mu_i + \sqrt{\Sigma_i}W_i$ is subexponential and*

$$\mathbb{P}\bigl(|Y_z| \ge \kappa^2\sigma_\infty^2 t\bigr) \le 4\exp\biggl[-c\min\biggl(\frac{t^2}{d + \kappa^{-2}\sigma_\infty^{-2}\|M\|^2}, t\biggr)\biggr]. \tag{37}$$

Recalling $\eta = d + \kappa^{-2}\sigma_\infty^{-2}\|M\|^2$, and changing t to ηt, (37) can be written as

$$\mathbb{P}\bigl(|Y_z| \ge \kappa^2\sigma_\infty^2 \eta t\bigr) \le 4\exp\bigl[-c\eta\min(t^2, t)\bigr].$$

Letting $\delta = (\sqrt{Cn} + u)/\sqrt{\eta}$ and setting $t = \max(\delta^2, \delta)$, we obtain

$$\mathbb{P}\big(|Y_z| \geq \kappa^2 \sigma_\infty^2 \eta \max(\delta^2, \delta)\big) \leq 4\exp(-c\eta\delta^2) \leq 4\exp\big[-c(Cn + u^2)\big].$$

We can now use a discretization argument. Let $\mathcal{N}$ be a $\frac{1}{4}$-net of S^{n-1}, so that $|\mathcal{N}| \leq 9^n$. We have $\|K - \mathbb{E}K\| = \sup_{z \in S^{n-1}} |Y_z| \leq 2 \max_{z \in \mathcal{N}} |Y_z|$; see, for example, [26], Exercise 4.4.3. Letting $\varepsilon = 2\kappa^2 \sigma_\infty^2 \eta \max(\delta^2, \delta)$, we have

$$\begin{aligned}\mathbb{P}\big(\|K - \mathbb{E}K\| \geq \varepsilon\big) &\leq \mathbb{P}\Big(\max_{z \in \mathcal{N}} |Y_z| \geq \varepsilon/2\Big) \\ &\leq 4 \cdot 9^n \exp\big[-c(Cn + u^2)\big] \leq 4\exp\big[-c(C_1 n + u^2)\big],\end{aligned}$$

where $C_1 = C - \log 9/c$ which can be made positive by take $C > \log 9/c$.

PROOF OF LEMMA 3. Without loss of generality, assume $\Sigma_i \succ 0$ for all i. Define Y_z as in (36) based on $X_i = \mu_i + \sqrt{\Sigma_i} W_i$. Using vector notation (34), we have $\vec{X} = \vec{\mu} + \sqrt{\Sigma}\vec{W}$ where $\Sigma = \operatorname{diag}(\Sigma_1, \ldots, \Sigma_n)$ is the $nd \times nd$ block diagonal matrix with diagonal blocks Σ_i. We have $z^T K z = \|\sum_i z_i X_i\|^2 = \|Xz\|^2$. Let $\Gamma_z = z^T \otimes I_d \in \mathbb{R}^{d \times nd}$ where $\otimes$ is the Kronecker matrix product. Then

$$\Gamma_z \vec{X} = (z^T \otimes I_d)\vec{X} = \begin{bmatrix} z_1 I_d & z_2 I_d & \cdots & z_n I_d \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = Xz. \tag{38}$$

It follows that

$$Xz = \Gamma_z \vec{\mu} + \Gamma_z \sqrt{\Sigma}\vec{W} = \Gamma_z \sqrt{\Sigma}(\Sigma^{-1/2}\vec{\mu} + \vec{W}).$$

Letting $\vec{\xi} := \Sigma^{-1/2}\vec{\mu} + \vec{W}$, we have

$$\|Xz\|^2 = \|\Gamma_z \sqrt{\Sigma}\vec{\xi}\|^2 = \vec{\xi}^T A_z \vec{\xi},$$

where $A_z := \sqrt{\Sigma}^T \Gamma_z^T \Gamma_z \sqrt{\Sigma}$. Hence, $Y_z := z^T(K - \mathbb{E}K)z = \vec{\xi}^T A_z \vec{\xi} - \mathbb{E}(\vec{\xi}^T A_z \vec{\xi})$ and we can apply the extension of Hanson–Wright inequality, Theorem 6 in Appendix A.3 (with $d = 1$ and n replaced with nd), to obtain

$$\mathbb{P}\big(|Y_z| \geq \kappa^2 t\big) \leq 4\exp\bigg[-c\min\bigg(\frac{t^2}{\|A_z\|_F^2 + \kappa^{-2}\|MA_z\|_F^2}, \frac{t}{\|A_z\|}\bigg)\bigg],$$

where $M = (\Sigma^{-1/2}\vec{\mu})^T \in \mathbb{R}^{1 \times nd}$. We obtain $MA_z = \vec{\mu}^T \Gamma_z^T \Gamma_z \sqrt{\Sigma}$. Using the inequality $\|AB\|_F \leq \|A\|\|B\|_F$ $(*)$ which holds for any two matrices A and B, we have

$$\|MA_z\|_F^2 = \big\|\sqrt{\Sigma}^T \Gamma_z^T \Gamma_z \vec{\mu}\big\|_2^2 \leq \|\sqrt{\Sigma}\|^2 \|\Gamma_z\|^2 \|\Gamma_z \vec{\mu}\|_2^2 \leq \sigma_\infty^2 \|\Gamma_z \vec{\mu}\|_2^2$$

since $\|\Gamma_z\| = \|z\|_2 \|I_d\| = 1$ and $\|\sqrt{\Sigma}\|^2 = \|\Sigma\| = \max_i \|\Sigma_i\| = \sigma_\infty^2$ where the last equality is by definition. Also, by identity (38), $\Gamma_z \vec{\mu} = Mz$. Hence, $\sup_{z \in S^{d-1}} \|\Gamma_z \vec{\mu}\| = \|M\|$. Putting the pieces together, $\|MA_z\|_F^2 \leq \sigma_\infty^2 \|M\|^2$.

Now, consider the operator norm of A_z, for which we have

$$\|A_z\| \leq \|\sqrt{\Sigma}\|^2 \|\Gamma_z\|^2 = \sigma_\infty^2.$$

Finally, for the Frobenious norm of A_z,

$$\|A_z\|_F \leq \|\sqrt{\Sigma}\|^2 \|\Gamma_z\| \|\Gamma_z\|_F = \sigma_\infty^2 \sqrt{d}$$

by repeated application of matrix inequality $(*)$ and $\|\Gamma_z\|_F^2 = d\|z\|_2^2 = d$. We obtain

$$\mathbb{P}\big(|Y_z| \geq \kappa^2 t\big) \leq 4\exp\bigg[-c\min\bigg(\frac{t^2}{\sigma_\infty^4 d + \kappa^{-2}\sigma_\infty^2 \|M\|^2}, \frac{t}{\sigma_\infty^2}\bigg)\bigg].$$

Changing t to $t\sigma_\infty^2$ gives the desired result. $\square$

4.5. *Proof of Theorem 3.* Consider a block-constant approximation of $\widetilde{K}(\mu)$, denoted as $K_\sigma^* \in \mathbb{R}^{n\times n}$, and defined as follows:

$$[K_\sigma^*]_{ij} = \Psi_{k\ell} \quad \text{whenever } (i,j) \in \mathcal{C}_k \times \mathcal{C}_\ell, \tag{39}$$

where $\{\Psi_{k\ell}\}$ are the empirical averages defined in (20). Let $Z \in \{0,1\}^{n\times K}$ be the membership matrix with rows z_i^T. It is not hard to see that $\frac{1}{n}K_\sigma^* = Z(\Psi/n)Z^T$ which resembles the mean matrix of a stochastic block model on the natural sparse scaling (see equation (4) in [31]).

The first step of the proof is to to show that the empirical (normalized) kernel matrix $K(X)/n$ is close of K_σ^*/n. Let us write

$$\sqrt{a} := \frac{1}{n}\|K(X) - K_\sigma^*\|, \qquad \sqrt{\omega} := \frac{1}{n}\|K(X) - \widetilde{K}_\sigma(\mu)\|,$$

and $\sqrt{b} := \frac{1}{n}\|\widetilde{K}_\sigma(\mu) - K_\sigma^*\|$. Using the definition of $v_{k\ell}$ in (20),

$$b \le \frac{1}{n^2}\|\widetilde{K}_\sigma(\mu) - K_\sigma^*\|_F^2 = \frac{1}{n^2}\sum_{k,\ell}\sum_{i,j} z_{ik}z_{j\ell}\big(\widetilde{K}_\sigma(\mu_i,\mu_j) - [K_\sigma^*]_{ij}\big)^2$$

$$= \frac{1}{n^2}\sum_{k,\ell} n_k n_\ell v_{k\ell}^2 = \overline{v}^2.$$

To control ω, note that $\widetilde{K}_\sigma(\mu) = \mathbb{E}[K(X)]$ and apply Theorem 1 with $\Sigma_i = \sigma^2\Sigma(\mu_i)/d$, $c = 1/2$, $C = \sqrt{2}$ and t replaced with $\sqrt{2}t$, to get with probability $\ge 1 - e^{-t^2}$,

$$n^2\omega = \|K(X) - \mathbb{E}K(X)\|^2 \le 4L^2\sigma_\infty^2(\sqrt{2}n + \sqrt{2n}t)^2$$

$$\le \frac{8L^2\sigma^2}{d}\max_i\|\Sigma(\mu_i)\|(n+\sqrt{n}t)^2.$$

By triangle inequality, $a \le 2(\omega + b)$. Thus, recalling the definition of $F(\gamma^2,\overline{v}^2)$,

$$a \le \frac{\gamma^2}{8R}F(\gamma^2,\overline{v}^2). \tag{40}$$

Let $A := K(X)/n$ and $A^{(R)}$ be obtained by truncating the EVD of A to its R largest eigenvalues in absolute value. The second step is to control the deviation of $A^{(R)}$ from the block-constant matrix K_σ^*/n. Lemma 6 in [31] gives

$$\|A^{(R)} - (K_\sigma^*/n)\|_F^2 \le 8R\|A - (K_\sigma^*/n)\|^2 = 8Ra =: \varepsilon^2. \tag{41}$$

The third and final step is to apply perturbation results for the k-means step of the algorithm. We note that K_σ^*/n is a k-means matrix with R centers, meaning that it has (at most) R distinct rows. Let us refer to these distinct vectors as $q_1,\dots,q_R \in \mathbb{R}^n$. Let δ_k be the minimum ℓ_2 distance of q_k from q_j, $j \ne r$. Then $n\delta_k^2 = \min_{\ell:\ell\ne k} D_{k\ell}$ where $D_{k\ell}$ is as defined in (22). Now, Corollary 1 in [31] implies that if $\varepsilon^2/(n\pi_k\delta_k^2) = \varepsilon^2/(n_k\delta_k^2) < [4(1+\kappa)^2]^{-1} = C_1^{-1}$, we have

$$\overline{\text{Mis}} \le C_1\frac{\varepsilon^2}{\min_k(n\delta_k^2)} = C_1\frac{8Ra}{\gamma^2} \le C_1F(\gamma^2,\overline{v}^2)$$

using the definition of γ^2 in (22) and inequality (40). Since, by definition, $\widetilde{\gamma}^2 = \min_k(n\pi_k\delta_k^2)$, the required condition holds if $8Ra/\widetilde{\gamma}^2 = \varepsilon^2/\widetilde{\gamma}^2 \le C_1^{-1}$. A further sufficient condition, in view of (40), is

$$F(\widetilde{\gamma}^2,\overline{v}^2) = \frac{\gamma^2F(\gamma^2,\overline{v}^2)}{\widetilde{\gamma}^2} \le C_1^{-1}.$$

This completes the proof for the case where one runs the k-means algorithm on the rows of $A^{(R)}$. Since the pairwise distance among the rows of $\hat{U}_1\hat{\Lambda}_1$ is the same as that of $A^{(R)}$, and the k-means algorithm is assumed isometry-invariant, the same result holds for $\hat{U}_1\hat{\Lambda}_1$. The proof is complete.

4.6. *Proof of Proposition 2.* Let $Y_1, \dots, Y_n$ be an independent sequence of variables and consider the U-statistic $U = \binom{n}{2}^{-1} \sum_{i<j} h(Y_i, Y_j)$ for some symmetric b-bounded function h. Then one has the following consequence of bounded difference inequality [28], Example 2.23:

$$\mathbb{P}\big(|U - \mathbb{E}U| > t\sqrt{8b^2/n}\big) \le 2e^{-t^2}.$$

Applying this result with $Y_i = \mu_i$ for $i \in \mathcal{C}_k$ and $h = \widetilde{K}_\sigma$, with probability at least $1 - 2e^{-t^2}$,

$$|\Psi_{kk} - \Psi^*_{kk}| \le t\frac{n_k - 1}{n_k}\sqrt{8b^2/n_k} \le t\sqrt{8b^2/n_k}.$$

Now assume that $Y_1, \dots, Y_n, Z_1, \dots, Z_m$ are independent and let

$$V = (nm)^{-1} \sum_{i,j} h(Y_i, Z_j).$$

Then, by a similar bounded difference argument,

$$\mathbb{P}\big(|V - \mathbb{E}V| > t\sqrt{8b^2/\min\{m,n\}}\big) \le 2e^{-t^2}.$$

For $k \neq \ell$, applying this result with $Y_i = \mu_i$, $i \in \mathcal{C}_k$ and $Z_j = \mu_j$, $j \in \mathcal{C}_\ell$ gives the desired result. For the variance, we have $v^2_{k\ell} = \mathbb{E}\widetilde{K}^2_\sigma(X, Y) - \Psi^2_{k\ell}$ where $(X, Y) \sim \widehat{P}_{k\ell}$. The first term is controlled similarly with b replaced with b^2, since $\widetilde{K}^2_\sigma$ is b^2-bounded. For the second term, assume that $|\Psi_{k\ell} - \Psi^*_{k\ell}| \le \delta_{k\ell}$. Then, $|\Psi^2_{k\ell} - (\Psi^*_{k\ell})^2| \le 2b\delta_{k\ell}$. Thus, under the event that the bounds hold, we have

$$|v^2_{k\ell} - (v^*_{k\ell})^2| \le \frac{t\sqrt{8}b^2}{\sqrt{n_k \wedge n_\ell}} + (2b)\frac{t\sqrt{8}b}{\sqrt{n_k \wedge n_\ell}}.$$

By a similar argument, $|D_{k\ell} - D^*_{k\ell}| \le 8b\delta_{k\ell}$. Applying union bound over $2R^2$ pairs, required for controlling $\Psi_{k\ell}$ and $v^2_{k\ell}$, completes the proof.

Acknowledgments. We thank Mark Rudelson for helpful comments, in particular, for the idea behind Lemma 1. We also thank the anonymous reviewer whose comment led to the strengthening of the results.

SUPPLEMENTARY MATERIAL

Supplement: Technical lemmas (DOI: 10.1214/20-AOS1967SUPP; .pdf). This supplement collects some technical results used in the paper.

REFERENCES

[1] ACKERMANN, M. R., BLÖMER, J. and SOHLER, C. (2010). Clustering for metric and nonmetric distance measures. *ACM Trans. Algorithms* **6** Art. 59, 26. MR2760422 https://doi.org/10.1145/1824777.1824779

[2] ADAMCZAK, R. (2015). A note on the Hanson–Wright inequality for random vectors with dependencies. *Electron. Commun. Probab.* **20** no. 72, 13. MR3407216 https://doi.org/10.1214/ECP.v20-3829

[3] AMINI, A. A. and RAZAEE, Z. S. (2021). Supplement to "Concentration of kernel matrices with application to kernel spectral clustering." https://doi.org/10.1214/20-AOS1967SUPP

[4] BELKIN, M. and NIYOGI, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput*. **15** 1373–1396.

[5] BLANCHARD, G., BOUSQUET, O. and ZWALD, L. (2007). Statistical properties of kernel principal component analysis. *Mach. Learn*. **66** 259–294.

[6] BRAUN, M. L. (2006). Accurate error bounds for the eigenvalues of the kernel matrix. *J. Mach. Learn. Res*. **7** 2303–2328. MR2274441 https://doi.org/10.1080/14685240600860923

[7] CHEN, K. (2009). On coresets for k-median and k-means clustering in metric and Euclidean spaces and their applications. *SIAM J. Comput*. **39** 923–947. MR2538844 https://doi.org/10.1137/070699007

[8] CHENG, X. and SINGER, A. (2013). The spectrum of random inner-product kernel matrices. *Random Matrices Theory Appl*. **2** 1350010, 47. MR3149440 https://doi.org/10.1142/S201032631350010X

[9] DO, Y. and VU, V. (2013). The spectrum of random kernel matrices: Universality results for rough and varying kernels. *Random Matrices Theory Appl*. **2** 1350005, 29. MR3109422 https://doi.org/10.1142/S2010326313500056

[10] EL KAROUI, N. (2010). The spectrum of kernel random matrices. *Ann. Statist*. **38** 1–50. MR2589315 https://doi.org/10.1214/08-AOS648

[11] EL KAROUI, N. (2010). On information plus noise kernel random matrices. *Ann. Statist*. **38** 3191–3216. MR2722468 https://doi.org/10.1214/10-AOS801

[12] FAN, Z. and MONTANARI, A. (2019). The spectral norm of random inner-product kernel matrices. *Probab. Theory Related Fields* **173** 27–85. MR3916104 https://doi.org/10.1007/s00440-018-0830-4

[13] GINÉ, E. and KOLTCHINSKII, V. (2006). Empirical graph Laplacian approximation of Laplace–Beltrami operators: Large sample results. In *High Dimensional Probability*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **51** 238–259. IMS, Beachwood, OH. MR2387773 https://doi.org/10.1214/074921706000000888

[14] HEIN, M. (2006). Uniform convergence of adaptive graph-based regularization. In *Learning Theory*. *Lecture Notes in Computer Science* **4005** 50–64. Springer, Berlin. MR2277918 https://doi.org/10.1007/11776420_7

[15] HEIN, M., AUDIBERT, J.-Y. and VON LUXBURG, U. (2005). From graphs to manifolds—Weak and strong pointwise consistency of graph Laplacians. In *Learning Theory*. *Lecture Notes in Computer Science* **3559** 470–485. Springer, Berlin. MR2203281 https://doi.org/10.1007/11503415_32

[16] KASIVISWANATHAN, S. P. and RUDELSON, M. (2015). Spectral norm of random kernel matrices with applications to privacy. In *Approximation*, *Randomization*, *and Combinatorial Optimization*. *Algorithms and Techniques*. *LIPIcs*. *Leibniz Int. Proc. Inform*. **40** 898–914. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3442004

[17] KOLTCHINSKII, V. and GINÉ, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli* **6** 113–167. MR1781185 https://doi.org/10.2307/3318636

[18] KUMAR, A., SABHARWAL, Y. and SEN, S. (2004). A simple linear time $(1+\varepsilon)$-approximation algorithm for k-means clustering in any dimensions. In *Annual Symposium on Foundations of Computer Science* **45** 454–462. IEEE Computer Society Press.

[19] LEDOUX, M. (2001). *The Concentration of Measure Phenomenon*. *Mathematical Surveys and Monographs* **89**. Amer. Math. Soc., Providence, RI. MR1849347

[20] ROSASCO, L., BELKIN, M. and DE VITO, E. (2010). On learning with integral operators. *J. Mach. Learn. Res*. **11** 905–934. MR2600634

[21] SCHIEBINGER, G., WAINWRIGHT, M. J. and YU, B. (2015). The geometry of kernelized spectral clustering. *Ann. Statist*. **43** 819–846. MR3325711 https://doi.org/10.1214/14-AOS1283

[22] SHAWE-TAYLOR, J., CRISTIANINI, N. and KANDOLA, J. S. (2002). On the concentration of spectral properties. In *Advances in Neural Information Processing Systems* 511–517.

[23] SHAWE-TAYLOR, J., WILLIAMS, C. K. I., CRISTIANINI, N. and KANDOLA, J. (2005). On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Trans. Inf. Theory* **51** 2510–2522. MR2246374 https://doi.org/10.1109/TIT.2005.850052

[24] SINGER, A. (2006). From graph to manifold Laplacian: The convergence rate. *Appl. Comput. Harmon. Anal*. **21** 128–134. MR2238670 https://doi.org/10.1016/j.acha.2006.03.004

[25] VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*: *Theory and Practice* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

[26] VERSHYNIN, R. (2018). *High-Dimensional Probability*: *An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge Univ. Press, Cambridge. MR3837109 https://doi.org/10.1017/9781108231596

[27] VON LUXBURG, U., BELKIN, M. and BOUSQUET, O. (2008). Consistency of spectral clustering. *Ann. Statist*. **36** 555–586. MR2396807 https://doi.org/10.1214/009053607000000640

[28] WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics*: *A Non-asymptotic Viewpoint*. *Cambridge Series in Statistical and Probabilistic Mathematics* **48**. Cambridge Univ. Press, Cambridge. MR3967104 https://doi.org/10.1017/9781108627771

[29] YAN, B. and SARKAR, P. (2016). On robustness of kernel clustering. In *Advances in Neural Information Processing Systems* 3098–3106.

[30] YANG, Y., PILANCI, M. and WAINWRIGHT, M. J. (2017). Randomized sketches for kernels: Fast and optimal nonparametric regression. *Ann*. *Statist*. **45** 991–1023. MR3662446 https://doi.org/10.1214/16-AOS1472

[31] ZHOU, Z. and AMINI, A. A. (2019). Analysis of spectral clustering algorithms for community detection: The general bipartite setting. *J*. *Mach*. *Learn*. *Res*. **20** Paper No. 47, 47. MR3948087