On the optimality of kernels for high-dimensional clustering

Leena Chennuru Vankadara
University of Tübingen, IMPRS-IS

Debarghya Ghoshdastidar
Technical University of Munich

Abstract

This paper studies the optimality of kernel methods in high-dimensional data clustering. Recent works have studied the large sample performance of kernel clustering in the high-dimensional regime, where Euclidean distance becomes less informative. However, it is unknown whether popular methods, such as kernel k-means, are optimal in this regime. We consider the problem of high-dimensional Gaussian clustering and show that, for a class of dot-product kernels, the sufficient conditions for partial recovery of clusters using the NP-hard kernel k-means objective matches the known information-theoretic limit up to a factor of $\sqrt{2}$ for large k. It also exactly matches the known upper bounds for the non-kernel setting. We also show that a semi-definite relaxation of the kernel k-means procedure matches upto constant factors, the spectral threshold, below which no polynomial time algorithm is known to succeed. This is the first work that provides such optimality guarantees for the kernel k-means as well as its convex relaxation. Our proofs demonstrate the utility of the less known polynomial concentration results for random variables with exponentially decaying tails in higher-order analysis of kernel methods.

1 Introduction

Kernel methods are one of the most empirically successful class of machine learning techniques. While being easy to implement, kernel methods are well known to improve empirical performance of algorithms and are also related to other successful machine learning principles such as Gaussian process and neural networks (Kanagawa et al., 2018; Jacot, Gabriel, and Hongler, 2018). At the heart of kernel based learning lies the *kernel trick* which implicitly maps the data to a high, possibly infinite, dimensional *reproducing kernel Hilbert space* (RKHS), and hence, induces non-linearity into classical linear learning models such as support vector machines, principle component analysis or k-means. Kernel methods are based on a solid theoretical foundation, which makes them conducive to theoretical analysis. There has been considerable theoretical research on kernel based supervised learning from a statistical perspective (Steinwart and Christmann, 2008; Mendelson and Neeman, 2010), and to some extent, in the context of semi-supervised learning (Wasserman and Lafferty, 2008; Mai and Couillet, 2018). Perhaps surprisingly, much less is known about the statistical performance of kernel methods beyond such settings, for instance, kernel based clustering.

A long-standing issue in the theoretical study of clustering, and also kernel based clustering, has been the lack of a universally accepted notion of *goodness* of clustering. A popular definition of good clustering is one that consistently or near-optimally partitions the data domain. Based on this perspective, there exist *approximation guarantees* for solving kernel based cost functions (S. Wang, Gittens, and Mahoney, 2019) and *consistency results* showing that the clustering asymptotically approaches a limiting clustering (Luxburg, Belkin, and Bousquet, 2008). In such analyses, the optimal cost function is inherently tied to the chosen kernel and hence can be arbitrarily far from the "ground truth." For instance, even an arbitrary clustering can be optimal (can achieve maximal clustering objective) for trivial kernels such as constant or identity kernels. Another approach to measure the performance of a clustering algorithm is by establishing recovery guarantees under distributional assumptions, sometimes known as *planted models*. Distributional assumptions, or specifically (sub)-Gaussian mixture model assumption, is often considered in the theory of clustering. While learning a mixture of Gaussians has always been an important research problem, Dasgupta (1999), for the time, presented a provable clustering algorithm to

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

learn a mixture of *high-dimensional* Gaussians. Theoretical research on learning high-dimensional Gaussians have ever since been highly significant, owing to the ubiquity of high-dimensional data in practice. Recent works in this direction provide *phase transitions* for both clustering and parameter estimation of a mixture of high-dimensional Gaussians (Banks et al., 2018; Ashtiani et al., 2018).

Couillet and Benaych-Georges (2016) initiated the theoretical study of kernel methods for high-dimensional Gaussian clustering, and in particular, presented the large sample behaviour of kernel spectral clustering in the regime where number of samples grow linearly with the data dimension. The statistical difficulty in this regime stems from the fact the Euclidean distance tends to be less informative in high dimensions and intra-cluster distances could be systematically larger than inter-cluster distances. Yan and Sarkar (2016) generalised the problem setup to sub-Gaussian mixtures and derived sufficient conditions for achieving zero clustering error using convex relaxations of the kernel k-means objective. In both works, the analysis is restricted to computationally efficient clustering algorithms and the optimality of kernel methods, in terms of comparing necessary and sufficient conditions for clustering, is not addressed.

In this paper, we study the phase transitions — sharp information-theoretic thresholds below which no algorithm can, provably, recover the true clustering better than chance — of the high-dimensional Gaussian clustering problem. Our setting is similar to Banks et al. (2018), where the number of samples is linear in the problem dimension. However, we focus on the case where one has access to only a kernel matrix. In other words, while the information-theoretic thresholds inherent to the Gaussian clustering problem are expected to remain unchanged in the kernel setting with a non-trivial kernel, we prove that one can nearly achieve such thresholds using popular kernel methods. The **main contributions** in this paper are the following: **(1)** We identify the smallest separation between the means of latent clusters such that the clusters are statistically distinguishable under a kernel k-means objective in the sense of partial recovery, that is, error smaller than random guessing. Our result **matches the phase transition** for high-dimensional Gaussian clustering without kernels (Banks et al., 2018). **(2)** We analyse a common **semi-definite relaxation** of the kernel k-means objective and present sufficient conditions for partial recovery that **match, up to constant factors, the known spectral threshold** — akin to the Kesten-Stigum threshold in the community detection under stochastic block model literature (Baik, Arous, and Péché, 2005; Paul, 2007).

Our main results obtained from the analysis of the two kernel-based clustering algorithms and the best known results for the same problem in a non-kernel setup are summarized in the table below. k is the number of clusters, and α is the ratio of sample size to the data dimension which remains asymptotically finite in our setting.

The lower and the upper bounds are on the minimum separation of the clusters required to achieve partial recovery. The first column contains the bounds for the information-theoretic threshold. The second column contains the bounds corresponding to the computational class of poly-time algorithms.

	Information-theoretic limit	**Poly-time solvable**
Lower bounds	$\sqrt{\frac{2(k-1)\log(k-1)}{\alpha}}$	$\frac{k-1}{\sqrt{\alpha}}$
Upper bounds (non-kernel)	$2\sqrt{\frac{k\log k}{\alpha}}+2\log k$	$O(k-1\vee\frac{k-1}{\sqrt{\alpha}})$
Upper bounds (kernel)	$2\sqrt{\frac{k\log k}{\alpha}}+2\log k$	$O(k\vee\frac{k}{\sqrt{\alpha}})$

As noted in Couillet and Benaych-Georges (2016), one requires a second-order analysis since first-order approximation of the kernel function does not suffice for the analysis in the high-dimensional setting. To this end, our proofs show that recent polynomial concentration inequalities (Götze, Sambale, and Sinulis, 2019) can be useful for second-order analysis of kernel methods.

2 Background and Setting

Notation: We denote the set of natural numbers $\{1,2,\ldots,k\}$ by $[k]$. For any matrix A, $\|A\|_F$ refers to the Frobenius norm of the matrix. For any vector x, $\|x\|$ and $\|x\|_1$ refer to the Euclidean and l_1 norms of the vector. $\mathbb{I}$ denotes the identity matrix. For any $A\in\mathbb{R}^{m\times m}$, $\|A\|_{\infty\to 1}$ refers to the $\infty\to 1$ operator norm and defined as $\sup_{y,z\in\{\pm 1\}^m}(y^TAz)$. For any n real numbers $\{a_i\}_{i=1}^n$, $(a_1\vee a_2\ \ldots\vee a_n)$ refers to the maximum of the sequence:$\max_i a_i$. For any random variable x, $\mathbb{E}x$ denotes the expectation of x.

Setting: Our setting is akin to the one used in Banks et al. (2018), specifically due to the existence of a near-optimal phase transition for the information-theoretic threshold in the setting. We assume that the data is generated according to the following process. Let k be the number of clusters. Then, k points $\{\mu_1',\mu_2',...,\mu_k'\}\in\mathbb{R}^p$ are generated independently according to a normal distribution with mean 0 and co-

variance $\frac{k}{k-1}\mathbb{I}_p$. The k points are then centered by subtracting their sample mean from each entry and the resulting centered vectors are denoted by $\{\mu_1, \mu_2, ..., \mu_k\}$. Let $m = \alpha p$ for some $\alpha > 0$, a fixed parameter. Then for each $i \in [k]$, generate $\frac{m}{k}$ points from a normal distribution with mean $\sqrt{\frac{\rho}{p}}\mu_i$ and covariance matrix $\mathbb{I}$ for some fixed parameter $\rho > 0$. Observe that the parameter ρ represents the separation between the clusters and can be treated as the parameter indicating the "statistical ease" with respect to the clustering problem or alternatively as the signal-to-noise ratio in this setting (we have an identity covariance matrix). We are interested in studying the large sample behaviour of clustering approaches in the high-dimensional setting: $m, p \to \infty$ and $\frac{m}{p} = O(1)$.

We denote the resulting set of m points by $\{x_1, x_2, ..., x_m\}$. Let $\sigma : [m] \to [k]$ denote a balanced partition of m points into k clusters and let σ_* denote the true partition: $\sigma_*(i) = s$ if $\mathbb{E}x_i = \sqrt{\frac{\rho}{p}}\mu_s$. Let X^* denotes the ground truth clustering matrix defined as follows:

$$X^*_{i,j} = \begin{cases} 1 & \text{if } \sigma_*(i) = \sigma_*(j) \\ 0 & \text{otherwise.} \end{cases}$$

For any arbitrary partition σ, define the $k \times k$ overlap matrix $\beta(\sigma, \sigma_*)$, for each $s, t \in [k]$ as the fraction of all points assigned by σ to the s^{th} cluster **and** the fraction of all points assigned by σ_* to the t^{th} cluster,

$$\beta(\sigma, \sigma_*)_{s,t} = \frac{k|\sigma^{-1}(s) \cap \sigma_*^{-1}(t)|}{m}.$$

Then $\|\beta(\sigma, \sigma_*)\|_F^2$ is a measure of similarity of the partition, σ with the true partition, σ_*. Observe that if the partitions are completely uncorrelated, then β is the constant matrix of $1/k$ and $\|\beta(\sigma, \sigma_*)\|_F^2 = 1$. If the partitions are identical up to permutations over the labels, then β would be the permutation matrix and $\|\beta(\sigma, \sigma_*)\|_F^2 = k$.

Alternatively, for the sake of analytical tractability, we sometimes, use the quantity $err(\sigma, \sigma_*)$ to denote the fraction of points misclassified by σ.

$$err(\sigma, \sigma_*) = 1 - \frac{\max_{\pi} \text{Trace}(\pi\beta(\sigma, \sigma_*))}{k}.$$

where $\pi\beta$ refers to the matrix resulting from a permutation of β over the cluster labels and the maximum is over all possible such permutations.

Clustering with k-means: The clustering objective of the k-means procedure (Pollard, 1981) is given as follows:

$$\min_{\sigma:[m]\to[k]} \sum_{s=1}^{k} \sum_{i \in \sigma^{-1}(s)} \left\| x_i - \frac{k}{m} \sum_{\sigma(j)=s} x_j \right\|^2.$$

This is equivalent to the following optimization problem:

$$\max_{\sigma:[m]\to[k]} \sum_{s=1}^{k} \sum_{i,j \in \sigma^{-1}(s)} \langle x_i, x_j \rangle.$$

Kernel k-means: For any partition $\sigma : [m] \to [k]$, define

$$\mathcal{F}(\sigma) = \sum_{s=1}^{k} \sum_{i,j \in \sigma^{-1}(s)} k(x_i, x_j).$$

Then, by the use of the kernel trick, we can formulate the kernel k-means clustering objective as follows:

$$\max_{\sigma:[m]\to[k]} \mathcal{F}(\sigma) \tag{1}$$

where $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a kernel function. Minimizing this objective over all possible partitions is NP-hard (Garey, Johnson, and Witsenhausen, 1982; Aloise et al., 2009). Several convex relaxations of the k-means procedure exist in literature. A well known semi-definite program (SDP) relaxation of the kernel k-means (Peng and Wei, 2007) objective is given by:

$$\max_{X} \text{trace}(KX) \tag{2}$$
$$\text{s.t., } X \succeq 0, X \geq 0,\ X\mathbf{1} = \frac{m}{k}\mathbf{1},\ \text{diag}(X) = \mathbf{1},$$

where K refers to the kernel matrix for a given kernel function k: $K_{i,j} = k(x_i, x_j)$. This SDP can be solved in polynomial time. To obtain a partitioning $\hat{\sigma}$ of the data based on the optimal solution $\hat{X}$ of the SDP, a 7-approximate k-medians's procedure (Charikar et al., 2002) is applied on the rows of the matrix $\hat{X}$ in a similar fashion as Fei and Chen (2018). We denote the partition inferred by this procedure as $\hat{\sigma}$. The details of the k-median procedure can be found in Fei and Chen (2018, Algorithm 1).

Choice of kernel function: For the analysis, we consider the class of dot-product kernel functions — $k(x, y) = f(\langle x, y \rangle)$ where f is assumed to be twice continuously differentiable and $f'(0) > 0$. The well-known exponential kernel $k(x, y) = \exp(\langle x, y \rangle)$ belongs to this class of kernel functions.

3 Our Results

We denote the upper bound on the information-theoretic threshold in the non-kernel setting as $\rho^{upper}_{linear\,NP}$ and the best known lower bound as ρ^{lower}_{NP}. We denote the upper bound from the analysis of the NP-hard kernel k-means procedure as $\rho^{upper}_{kernel\,NP}$.

The central question we address in this section is the following: Does the kernel clustering procedure achieve information-theoretic optimality for high-dimensional clustering:

$$\rho^{upper}_{kernel\,NP} \stackrel{?}{=} \rho^{lower}_{NP}.$$

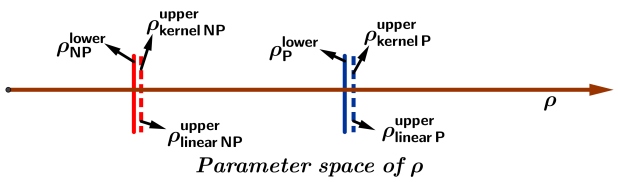

Figure 1: Our upper bounds are near optimal and exactly match those of the non-kernel setting.

The maximum likelihood estimator in the non-kernel setting is already known to achieve near optimality in an information-theoretic sense. Therefore, our principal objective can be rephrased, in essence, as: Is the the class of dot-product kernels more(or less) informative than the linear kernel:

$$\rho^{upper}_{kernel\,NP} \overset{?}{\leq} \rho^{upper}_{linear\,NP}.$$

As noted earlier, optimizing the kernel k-means objective is NP-hard. Therefore, for practical significance, it is also interesting to understand the information-theoretic optimality of kernels via kernelized, computationally efficient clustering algorithms. To this end, we analyze the kernel SDP given in (2). It has been observed in several clustering problems that the parameter space of the signal-to-noise ratio (SNR) ρ where polynomial time algorithms are known to succeed is, typically, strictly above the information-theoretic threshold. To evaluate if there is any information loss, due to the use of kernels in polynomial time clustering algorithms, we compare the SNR above which kernel SDP can provably recover the true clustering, $\rho^{upper}_{kernel\,P}$, with the known spectral threshold, ρ^{lower}_{P} below which no known poly-time algorithm is known to succeed. We also compare $\rho^{upper}_{kernel\,P}$ to the upper bound($\rho^{upper}_{linear\,P}$) derived from the analysis of a similar semidefinite relaxation of linear k-means:

$$\rho^{upper}_{kernel\,P} \overset{?}{=} \rho^{lower}_{P} \qquad \rho^{upper}_{kernel\,P} \overset{?}{=} \rho^{upper}_{linear\,P}.$$

We pictorially demonstrate all our results in Figure 1.

3.1 Optimality of kernel k-means

The following lower and upper bounds, ρ^{lower}_{NP} and $\rho^{upper}_{linear\,NP}$ respectively on the information-theoretic threshold appeared in Banks et al. (2018):

$$\rho^{upper}_{linear\,NP} = 2\sqrt{\frac{k\log k}{\alpha}} + 2\log k, \tag{3}$$

$$\rho^{lower}_{NP} = \begin{cases} \sqrt{1/\alpha} & k = 2 \\ \sqrt{\frac{2(k-1)\log(k-1)}{\alpha}} & k \geq 3\,. \end{cases} \tag{4}$$

We analyze the performance of the kernel k-means clustering algorithm and give the following upper bounds on the information-theoretic threshold:

Theorem 1 (**Optimality of kernel k-means**). *Let*

$$\rho^{upper}_{kernel\,NP} = 2\sqrt{\frac{k\log k}{\alpha}} + 2\log k. \tag{5}$$

If $\rho > \rho^{upper}_{kernel\,NP}$, then for large enough m, with high probability (w.h.p), it is possible to recover the true partition.

Our results show that there is no loss of information incurred due to the use of the kernel function in high-dimensional Gaussian clustering. This also matches the known information-theoretic lower bounds up to a factor of $\sqrt{2}$ when the number of clusters k is large (Banks et al., 2018).

Overview of the analysis: On a high level, the main line of argumentation of the proof is similar to the one in Banks et al. (2018). However, note that their analysis only holds for the linear k-means algorithm and extending the analysis to a second order expansion of the kernel k-means objective is considerably more complex and requires a different set of mathematical tools and techniques (see Section 4.1).

We consider the distribution of the objective of kernel k-means $\mathcal{F}(\sigma)$ as a function of the partition σ. We show that above the aforementioned threshold $\rho^{upper}_{kernel\,NP}$, with high probability, the distribution of $\mathcal{F}(\sigma_*)$ is disjoint with and higher than that of the distribution of $\max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2 \\ \leq 1+(k-1)\epsilon}} \mathcal{F}(\sigma)$, where $\epsilon > 0$ is an arbitrarily small constant. Let $\tilde{\sigma}$ denote the optimal solution to (1). Since, by definition, $\mathcal{F}(\tilde{\sigma}) \geq \mathcal{F}(\sigma_*)$, it follows that the support of the distribution of $\mathcal{F}(\tilde{\sigma})$ is disjoint with and higher than that of the distribution of $\max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2 \\ \leq 1+(k-1)\epsilon}} \mathcal{F}(\sigma)$.

3.2 Optimality of kernel SDP

The following phase transition for spectral methods can be inferred from Paul (2007) and Baik, Arous, and Péché (2005) and appeared in Banks et al. (2018):

$$\rho^{lower}_{P} = \frac{k-1}{\sqrt{\alpha}}.$$

We give the following upper bound on the threshold below which no known computationally efficient polynomial clustering approaches (provably) achieve partial recovery.

Theorem 2 (**Optimality of kernel SDP**). *Let*

$$\rho^{upper}_{kernel\,P} = ck\left(1 \vee \frac{1}{\sqrt{\alpha}}\right). \tag{6}$$

for some fixed constant $c > 0$. If $\rho > \rho^{upper}_{kernel\,P}$, then for sufficiently large m, w.h.p, kernel SDP can recover the true partition.

Our results match the spectral threshold up to constant factors of approximation for large k. Our result also matches with the known upper bound ($\rho^{upper}_{linear\,P}$) for partial recovery via the linear k-means clustering procedure (Giraud and Verzelen, 2018) up to a factor of $\frac{(k-1)}{k}$. In agreement with the established conjecture, it is also evident from our results that the threshold at which a computationally efficient kernel clustering procedure can be guaranteed to succeed is strictly above the information-theoretic threshold. They differ by an order of $\sqrt{\frac{k}{\log k}}$.

Overview of the analysis: Denote $\kappa = f''(\tau + C_0\frac{\log p}{\sqrt{p}})$ We define the matrix $\tilde{K}$ that depends on the population parameters of the data distribution as follows:

$$\tilde{K}(i,j) = f(0)+$$

$$\begin{cases} \frac{f^{'}(0)\rho\langle\mu_i,\mu_j\rangle}{p^2} + \frac{\kappa\rho^2\langle\mu_i,\mu_j\rangle^2}{p^4} + \frac{\kappa}{p} & \text{if } i \neq j \\ \frac{f^{'}(0)(p^2+\rho\|\mu_i\|^2)}{p^2} + \frac{\kappa(p^2+\rho\|\mu_i\|^2)^2}{p^4} + \frac{\kappa}{p} & \text{otherwise.} \end{cases}$$

We show that the kernel matrix K concentrates around $\tilde{K}$ in the $\infty \to 1$ operator norm.

Let $\hat{X}$ denote the optimal solution to (2). Then, using Grothendieck's inequality (Grothendieck, 1956), we derive an upper bound on $\|\hat{X} - X^*\|_1$ in terms of $\|K - \tilde{K}\|_{\infty\to 1}$. Since $\hat{X}$ is not a partition matrix, we need a procedure that can infer a partition from $\hat{X}$. We use the 7-approximate k-median's procedure (Fei and Chen, 2018) on the rows of $\hat{X}$ to infer a partition $\hat{\sigma}$. Then Fei and Chen (2018) showed that the fraction of mis-classified vertices by the partition $\hat{\sigma}$ denoted by $err(\sigma_*, \hat{\sigma})$ can be upper bounded by a constant factor of $\frac{\left\|\hat{X}-X^*\right\|_1}{\|X^*\|_1}$. Thereby, we show that for $\rho > \rho^{upper}_{kernel\,P}$, the fraction of misclassified points $err(\hat{\sigma}, \sigma_*) < (1 - 1/k)$, which is the condition required for partial recovery.

SDPs, such as the one defined in (2), have been analyzed using the Grothendieck's inequality approach in community detection literature for stochastic block models (Guédon and Vershynin, 2016). However, the main technical challenges of our analysis lie in the choice of appropriate $\tilde{K}$ and showing that the matrix K concentrates around $\tilde{K}$ in the $\infty \to 1$ operator norm. Establishing the concentration results for $\|K - \tilde{K}\|_{\infty\to 1}$ is considerably harder compared to the analysis of similar quantities based on the adjacency matrix of a network generated from a stochastic block model. Unlike in the case of adjacency matrices, the entries of the kernel matrix encode dependencies between the data points and hence most classical concentration tools from random matrix theory fall short in the analysis of kernel matrices. Also the RKHS corresponding to the chosen class of kernel functions can be infinite dimensional and hence concentration inequalities that depend on the dimension of the feature space are also not applicable for analyzing functions of kernel entries.

To this end, we demonstrate that the polynomial concentration inequalities for exponentially decaying random variables in Götze, Sambale, and Sinulis (2019) can be used to analyze an entry-wise second order approximation of the kernel matrix. We make some further remarks about our proof, and possibilities for improving the result.

Remark 1: Finer upper bounds on $\langle K - \tilde{K}, X^* - \hat{X}\rangle$ can be obtained by applying an analysis similar to Fei and Chen (2018) to obtain better error rates. However, our bounds on ρ, essentially remain the same — which is the main emphasis of this paper.

Remark 2: The choice of $\tilde{K}$ can further be refined in the second order terms without changing the results of our analysis.

Remark 3: One could, alternatively, infer a partition by applying the k-means procedure on the rows of the eigenvectors of $\hat{X}$. Using Davis-Khan's theorem (Yu, T. Wang, and Samworth, 2014), one may similarly upper bound the fraction of misclassified nodes by a constant factor of $\frac{\|\hat{X}-X^*\|_1}{\|X^*\|_1}$. This approach gives a slightly worse approximation constant.

4 Proofs

4.1 Proof of Theorem 1

Overview of the technical steps: Let $\epsilon > 0$ be an arbitrarily small constant. The two main ingredients required to establish conditions of recovery are as follows:

- Upper tail bounds for $\max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2 \\ \leq 1+(k-1)\epsilon}} \mathcal{F}(\sigma)$.
- Lower tail estimates for the distribution of $\mathcal{F}(\sigma_*)$.

To obtain these bounds, for any fixed σ, we first apply the Taylor's theorem with mean value form of the reminder to obtain a 2nd order polynomial approximation of each of the kernel entry and obtain a tight lower bound $\mathcal{F}_l(\sigma)$ and an upper bound $\mathcal{F}_u(\sigma)$ on $\mathcal{F}(\sigma)$.

For any fixed σ such that $\|\beta(\sigma,\sigma_*)\|_F^2 \leq 1 + (k-1)\epsilon$, we compute $\mathcal{F}_u(\sigma)$ which is a 4th order polynomial of normally distributed random variables and carefully upper bound all the terms of this polynomial using various known concentration results in literature. By an

union bound over all such partitions, we obtain upper tail bounds for $\max_{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2 \le 1+(k-1)\epsilon} \mathcal{F}_u(\sigma)$. Similarly, we compute $\mathcal{F}_l(\sigma_*)$ which is a 4th order polynomial of normally distributed random variables and obtain lower bounds for all the involved terms. Therefore, we obtain:

$$\mathcal{F}(\sigma_*) \ge \mathcal{F}_l(\sigma_*) > \omega_l$$
$$\max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2 \\ \le 1+(k-1)\epsilon}} \mathcal{F}(\sigma) \le \max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2 \\ \le 1+(k-1)\epsilon}} \mathcal{F}_u(\sigma) \le \omega_u.$$

By comparing ω_u and ω_l, we obtain the conditions on ρ under which $\omega_l \ge \omega_u$.

Notation: We use the following notation for some recurring terms for improved readability.

For any $i, j \in [m]$, set $\tau = \frac{\mathbb{E}\|x_i\|^2}{p} = 1 + O(1/p)$. For any σ, we use the following notation:

$$Q_{1\sigma} = \tfrac{k}{m} \sum_{s\in[k]} \sum_{i,j\in\sigma^{-1}(s)} \tfrac{\langle x_i, x_j\rangle}{p};$$

$$Q_{2\sigma} = \tfrac{k}{m} \sum_{s\in[k]} \sum_{i,j\in\sigma^{-1}(s)} \tfrac{\langle x_i, x_j\rangle^2}{p^2};$$

$$Q_3 = \tfrac{k(f'(\tau)-f'(0))}{m} \sum_{i\in[m]} \left(\tfrac{\|x_i\|^2}{p} - \tau\right);$$

$$Q_4 = \tfrac{k}{2m} \sum_{i\in[m]} \left(\tfrac{\|x_i\|^2}{p} - \tau\right)^2; \quad Q_5 = \sum_{i\in[m]} \tfrac{k\tau\|x_i\|^2}{mp};$$

$$\gamma_1 = f''(C_0 \tfrac{\log p}{\sqrt{p}}); \ \gamma_2 = f''(\tau + C_0 \tfrac{\log p}{\sqrt{p}})$$

$$\gamma_3 = f''(-C_0 \tfrac{\log p}{\sqrt{p}}); \ \gamma_4 = f''(\tau - C_0 \tfrac{\log p}{\sqrt{p}})$$

for some constant $C_0 > 0$.

All the lemmas we state below hold with high probability $(1 - \Omega(\frac{1}{p}))$ and the proofs of all the lemmas are provided in the supplementary.

Outline of the proof: Recall that for any partition $\sigma : [m] \to [k]$, $\mathcal{F}(\sigma) = \frac{k}{m} \sum_{s\in[k]} \sum_{i,j\in\sigma^{-1}(s)} k(x_i, x_j)$.

Lemma 1 (**Upper and lower bounds for inner products**).

$$\max_{i,j} \frac{|\langle x_i, x_j\rangle|}{p} = \tau \mathbf{1}_{i=j} + O\left(\frac{\log p}{\sqrt{p}}\right), \text{ and}$$
$$\min_{i,j} \frac{\langle x_i, x_j\rangle}{p} = \tau \mathbf{1}_{i=j} + \Omega\left(-\frac{\log p}{\sqrt{p}}\right).$$

By a second order Taylor expansion of each $k(x_i, x_j)$ where $i \ne j$ around 0 and expanding each $k(x_i, x_i)$ around τ, and using Lemma 1, for any σ, we can write $\mathcal{F}(\sigma) \le \mathcal{F}_u(\sigma) =$

$$f'(0)Q_{1\sigma} + \gamma_1 Q_{2\sigma} + Q_3 + (\gamma_2 - \gamma_1)Q_4 - \gamma_1 Q_5 - k\tau f'(\tau) + kf(\tau) + (m-k)f(0) + \frac{k\gamma_1\tau^2}{2}.$$

and for any σ, $\mathcal{F}(\sigma) \ge \mathcal{F}_l(\sigma) =$

$$f'(0)Q_{1\sigma} + \gamma_3 Q_{2\sigma} + Q_3 + (\gamma_4 - \gamma_3)Q_4 - \gamma_3 Q_5 - k\tau f'(\tau) + kf(\tau) + (m-k)f(0) + \frac{k\gamma_3\tau^2}{2}.$$

Upper bounds for $\max_{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2 \le 1+(k-1)\epsilon} \mathcal{F}(\sigma)$**:** We derive upper bounds for all the terms that constitute $\mathcal{F}_u(\sigma)$, which simultaneously hold for all σ such that $\|\beta(\sigma, \sigma_*)\|_F^2 \le 1 + (k-1)\epsilon$.

Lemma 2 (**Upper bounds for** $Q_{1\sigma}$**,** Q_5**).**

$$\max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2 \\ \le 1+(k-1)\epsilon}} Q_{1\sigma} \le k + \alpha\rho\epsilon + 2(1+\epsilon)\alpha\log k + 2\sqrt{(1+\epsilon)(k+2\alpha\rho\epsilon)\alpha\log k} + O(\sqrt{\log p/p}).$$

$$\max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2 \\ \le 1+(k-1)\epsilon}} -\gamma_1 Q_5 \le - \frac{k\gamma_1\tau}{mp}\left(mp + p\alpha\rho - 2\sqrt{(mp + 2p\alpha\rho)\log p}\right).$$

Proof (sketch). The terms $Q_{1\sigma}$ and Q_5 can be expressed as sums of independent non-central chi-squared random variables and applying the known upper tail bounds for such sums, followed by a union bound over all $\sigma : \|\beta(\sigma, \sigma_*)\|_F^2 \le 1 + (k-1)\epsilon$, we have the results from Lemma 2. □

Lemma 3 (**Upper bound for** $Q_{2\sigma}$**).**

$$\max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2 \\ \le 1+(k-1)\epsilon}} \gamma_1 Q_{2\sigma} \le \gamma_1\left(1 + \frac{1}{k} + O\left(\frac{1}{p}\right)\right) + C_2\gamma_1 O\left(\sqrt{\frac{\alpha}{p}} \vee \alpha\sqrt{\frac{\alpha}{p}} \vee \sqrt{\frac{1}{\alpha p}}\right)$$

for some constant $C_2 > 0$.

Proof (sketch). Controlling the typical behavior of the term $Q_{2\sigma}$ is the most demanding part of the proof. We use the concentration results established for polynomials of sub-Gaussian random variables (see the supplementary for a definition) in Götze, Sambale, and Sinulis (2019) to establish the result in Lemma 3. □

Lemma 4 (**Upper bound for** Q_4**).**

$$\max_{\substack{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2 \\ \le 1+(k-1)\epsilon}} (\gamma_2 - \gamma_1)Q_4 \le C_4 k(\gamma_2 - \gamma_1)(\log p)^2/2p.$$

for some constant $C_4 > 0$.

Proof (sketch). The term Q_4 is small relative to the other terms and hence a crude upper bound based on the inequality: For any two vectors a, b, $\sum_{i=1}^{n} a_i \cdot b_i \leq \sup_{i\in[n]} |b_i| \cdot \sum_{i=1}^{n} |a_i|$, followed by an application of Lemma 1 suffices to establish the behavior of this term. □

Lower bounds for $\mathcal{F}(\sigma_*)$. Similarly, we derive lower bounds for all terms that arise in $\mathcal{F}_l(\sigma_*)$.

Lemma 5 (Lower bound for $Q_{1\sigma_*}$ and Q_5).

$$\begin{aligned} Q_{1\sigma_*} &> k + \alpha\rho - O(\sqrt{\log p/p}), \ \textit{and} \\ -\gamma_3 Q_5 &> -\frac{k\gamma_3\tau}{mp}\left(mp + p\alpha\rho + 2\log p\right) \\ &\quad - \frac{k\gamma_3\tau}{mp}\left(2\sqrt{(mp + 2p\alpha\rho)\log p}\right). \end{aligned}$$

Proof (sketch). From upper tail estimates for sums of non-central chi-squared random variables, we establish the result of Lemma 5. □

Lemma 6 (Lower bound for $Q_{2\sigma_*}$).

$$\begin{aligned} \gamma_3 Q_{2\sigma_*} &> \gamma_3\left(1 + \frac{1}{k} + O\left(\frac{1}{p}\right)\right) \\ &\quad - C_2\gamma_3\left(\sqrt{\frac{\log p}{p^2}} \vee \alpha\sqrt{\frac{\log p}{p^2}}\right). \end{aligned}$$

Proof (sketch). $Q_{2\sigma_*}$, as discussed earlier, is a 4^{th} order polynomial of sub-Gaussian random variables (see the supplementary for a definition). Therefore from lower tail estimates for polynomials of sub-Gaussian random variables in Götze, Sambale, and Sinulis (2019), we establish the result in Lemma 6. □

Since Q_4 is a smaller term, the following lower bound suffices to control its behavior:

$$Q_4 > 0. \tag{7}$$

Using the mean-value theorem, we can write $\gamma_1 - \gamma_2 = f'''(\xi)2C_0\log p/\sqrt{p}$, where $\xi \in (-C_0\log p/\sqrt{p}, C_0\log p/\sqrt{p})$. By assumption, f is twice continuously differentiable on the compact interval $[-C_0, C_0]$ and thereby $f'''(\xi)$ is bounded. Hence $\gamma_1 - \gamma_2 \to 0$ as $p \to \infty$. Similarly, $\gamma_3 - \gamma_4 \to 0$ as $p \to \infty$. From Lemmas 2 to 6 and Equation (7), we obtain that for $\rho > 2\sqrt{k\log k/\alpha} + 2\log k$, for large enough p, with high probability, $\max_{\sigma} \mathcal{F}(\sigma) \geq \mathcal{F}(\sigma_*) \geq \max_{\sigma:\|\beta(\sigma,\sigma_*)\|_F^2 \leq 1+(k-1)\epsilon} \mathcal{F}(\sigma)$.

4.2 Proof of Theorem 2

Overview of the technical steps:

- We define a matrix $\tilde{K}$ which relies on the model parameters of the data distribution.
- We upper bound the l_1 norm of the difference between the ground truth clustering matrix and the optimal solution of the SDP $\|\hat{X} - X^*\|_1$ by a constant factor of the inner product between $K - \tilde{K}$ and $X^* - \hat{X}$, that is, $\langle K - \tilde{K}, X^* - \hat{X}\rangle$.
- We use the Grothendieck's inequality to upper bound $\langle K - \tilde{K}, X^* - \hat{X}\rangle$ by a constant factor of $\|K - \tilde{K}\|_{\infty\to 1}$.
- We establish the upper tail estimates of the deviation of the kernel matrix K from $\tilde{K}$ in the $\infty \to 1$ norm.
- Thereby, we have an upper bound on $\|\hat{X} - X^*\|_1$ which translates to an upper bound on $err(\hat{\sigma}, \sigma_*)$. By setting $err(\hat{\sigma}, \sigma_*) < 1 - 1/k$, we derive the desired conditions on ρ.

Notation: Denote $\kappa = f''(\tau + \frac{C_0\log p}{\sqrt{p}})$. For ease of notation, we define the $m \times m$ matrices $R^{(1)}$ and $R^{(2)}$ as follows:

$$R^{(1)}_{i,j} = \begin{cases} \frac{f'(0)\langle x_i, x_j\rangle}{p} - \frac{f'(0)\rho\langle \mu_i, \mu_j\rangle}{p^2} & \text{if } i \neq j, \\ \frac{f'(0)\|x_i\|^2}{p} - \frac{f'(0)p^2 + f'(0)\rho\|\mu_i\|^2}{p^2} & \text{otherwise.} \end{cases}$$

$$R^{(2)}_{i,j} = \begin{cases} \frac{\langle x_i, x_j\rangle^2}{p^2} - \frac{\rho^2\langle \mu_i, \mu_j\rangle^2}{p^4} - \frac{1}{p} & \text{if } i \neq j, \\ \frac{\|x_i\|^4}{p^2} - \frac{(p^2 + \rho\|\mu_i\|^2)^2}{p^4} - \frac{1}{p} & \text{otherwise.} \end{cases}$$

All the lemmas hold with high probability: with probability $1 - \Omega(1/p)$ and the proofs of the lemmas are provided in the supplementary.

Outline of the proof: We begin by establishing the following upper bound on $\|X^* - \hat{X}\|_1$.

Lemma 7 (Upper bound on $\|X^* - \hat{X}\|_1$).

$$\|X^* - \hat{X}\|_1 \leq \frac{2\langle \tilde{K}, X^* - \hat{X}\rangle}{\frac{\rho}{p}\left(\frac{k}{k-1} + O(\sqrt{\log p/p}) + \kappa\rho O(\frac{1}{p})\right)}.$$

Observe that, by definition, $\langle K, \hat{X}\rangle \geq \langle K, X^*\rangle \implies \langle K, \hat{X} - X^*\rangle \geq 0$, and therefore,

$$\begin{aligned} \langle K, X^* - \hat{X}\rangle &\leq \langle K - \tilde{K}, X^* - \hat{X}\rangle \\ &\leq 2 \sup_{\substack{X\succeq 0 \\ diag(X)\leq 1}} |\langle K - \tilde{K}, X\rangle|. \end{aligned}$$

Using Grothendieck's inequality (Grothendieck, 1956), we arrive at the following (see the appendix for a statement of Grothendieck's inequality):

$$2 \sup_{\substack{X \succeq 0 \\ diag(X) \leq 1}} |\langle K - \tilde{K}, X \rangle| \leq K_G \|K - \tilde{K}\|_{\infty \to 1}.$$

where $K_G \approx 1.783$ is the Grothendieck's constant. For any pair of fixed vectors $z, y \in \{\pm 1\}^m$, by a 2nd order Taylor's expansion of each $K_{i,j}$ around 0, and applying the result from Lemma 1, we can see that:

$$y^T(K - \tilde{K})z \leq y^T(R^{(1)} + \kappa R^{(2)})z. \tag{8}$$

Lemma 8 (**Upper bounds for $R^{(1)}$**).

$$\sup_{z,y \in \{\pm\}^n} y^T R^{(1)} z \leq C_1 \alpha \left(\sqrt{mp} \vee m\right) \tag{9}$$

for some constant $C_1 > 0$.

Proof (sketch): Linear combinations of entries of the matrix $R^{(1)}$ can be re-written as sums of independent sub-exponential random variables (see the supplementary for a definition). By an application of Bernstein's inequality for each fixed $\{z, y \in \pm 1\}^m$, followed by an union bound over all possible z, y we establish the result. □

Lemma 9 (**Upper bounds for $R^{(2)}$**).

$$\sup_{\{z,y \in \pm 1\}^m} \kappa y^T R^{(2)} z \leq$$
$$\frac{C_2' \kappa}{p^2}(\rho m(m-1) + m + (mp\sqrt{m} \vee m^2\sqrt{m} \vee p^2\sqrt{m})).$$

for some constant $C_2' > 0$.

Proof. In order to bound the linear combinations of entries of $R^{(2)}$, for each fixed $\{z, y \in \pm 1\}^m$ we apply the concentration results for polynomials of independent sub-Gaussian random variables. (Götze, Sambale, and Sinulis, 2019). In order to bound the maximum of the second order terms over all $\{z, y \in \pm 1\}^m$, we use the union bound. □

From Lemmas 7, 8 and 9, we have that: $\|\hat{X} - X^*\|_1 \leq$

$$\frac{2C'}{\phi p}\left((m^2/\sqrt{\alpha} \vee m^2) + \frac{\kappa}{p}(\rho m^2 + O(m))\right)$$
$$+ \frac{2C'\kappa}{\phi p^2}\left(mp\sqrt{m} \vee m^2\sqrt{m} \vee p^2\sqrt{m}\right), \tag{10}$$

where $\phi = \frac{\rho}{p}\left(\frac{k}{k-1} + O(\sqrt{\log p/p}) + \kappa\rho O(\frac{1}{p})\right)$, for some constant, $C' > 0$.

Let $\hat{\sigma}$ be the partition generated by applying the η-approximate k-median's procedure on $\hat{X}$.

Proposition 1 (**Fraction of misclassified nodes** (Fei and Chen, 2018)). *The fraction of mis-classified points corresponding to the partition $\hat{\sigma}$:*

$$err(\hat{\sigma}, \sigma_*) \leq 2(1 + 2\eta)\frac{\|\hat{X} - X^*\|_1}{\|X^*\|_1}$$

Observe that $\|X^*\|_1 = \frac{m^2}{k}$. Applying the result from Proposition 1, we have that for large enough p, if $\rho \gtrsim k(\frac{1}{\sqrt{\alpha}} \vee 1)$, $err(\hat{\sigma}, \sigma_*) < 1 - \frac{1}{k}$.

5 Discussion

In this paper, we study the large sample behaviour of the kernel k-means algorithm for high-dimensional clustering. The principal focus lies in investigating the information-theoretic optimality of the kernel k-means procedure. Recent works have demonstrated that the linear k-means algorithm is near optimal in this sense. Therefore another aspect of our work resides in understanding the informativeness of specific kernels for high-dimensional clustering in relation to the linear kernel. A thorough understanding of these aspects is fundamental to the use of kernels in any unsupervised high-dimensional learning problem.

We also study the large sample behaviour of a popular semi-definite relaxation of the kernel k-means objective. We emphasize on optimality and informativeness of kernels in computationally efficient algorithms for high-dimensional clustering. A widely believed conjecture in clustering literature, with support from well founded theoretical evidence, is that computationally efficient algorithms are sub-optimal in an information-theoretic sense. Therefore, in this paper, we consider the SDP to be information-theoretically optimal if its optimal in the class of computationally efficient algorithms. The best known result for this class arises from the well known spectral threshold in this setting.

We show that both the algorithms are near optimal in their computational class and as a consequence also demonstrate that their is no loss of information incurred by the use of the class of dot-product kernels over the linear kernel. By virtue of our proofs, we also demonstrate that the recent polynomial concentration inequalities for random variables with exponentially decaying tails can aid in the analysis of higher order kernel approximations.

Furthermore, this line of analysis can be extended to other empirically popular kernels. In particular, since the squared distance is known to be less informative in high dimensions, it would be interesting to investigate the informativeness of the popular Gaussian kernel which relies on the square distances.

Acknowledgements

This work was supported by the Baden-Württemberg Stiftung through the BW Eliteprogramm for Postdocs. We thank Prof.Dr.Ulrike von Luxburg for insightful discussions during the course of this work. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Leena C Vankadara.

References

Aloise, Daniel, Amit Deshpande, Pierre Hansen, and Preyas Popat (2009). "NP-hardness of Euclidean sum-of-squares clustering". In: *Machine learning* 75.2, pp. 245–248.

Ashtiani, Hassan, Shai Ben-David, Nicholas Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan (2018). "Nearly tight sample complexity bounds for learning mixtures of Gaussians via sample compression schemes". In: *Advances in Neural Information Processing Systems*, pp. 3412–3421.

Baik, Jinho, Gérard Ben Arous, and Sandrine Péché (2005). "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices". In: *The Annals of Probability* 33.5, pp. 1643–1697.

Banks, Jess, Cristopher Moore, Roman Vershynin, Nicolas Verzelen, and Jiaming Xu (2018). "Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization". In: *IEEE Transactions on Information Theory* 64.7, pp. 4872–4894.

Charikar, Moses, Sudipto Guha, Éva Tardos, and David B Shmoys (2002). "A constant-factor approximation algorithm for the k-median problem". In: *Journal of Computer and System Sciences* 65.1, pp. 129–149.

Couillet, Romain and Florent Benaych-Georges (2016). "Kernel spectral clustering of large dimensional data". In: *Electronic Journal of Statistics* 10.1, pp. 1393–1454.

Dasgupta, Sanjoy (1999). "Learning mixtures of Gaussians". In: *40th Annual Symposium on Foundations of Computer Science*. IEEE, pp. 634–644.

Fei, Yingjie and Yudong Chen (2018). "Exponential error rates of SDP for block models: Beyond Grothendieck's inequality". In: *IEEE Transactions on Information Theory* 65.1, pp. 551–571.

Garey, MR, D Johnson, and Hans Witsenhausen (1982). "The complexity of the generalized Lloyd-max problem (corresp.)" In: *IEEE Transactions on Information Theory* 28.2, pp. 255–256.

Giraud, Christophe and Nicolas Verzelen (2018). "Partial recovery bounds for clustering with the relaxed K-means". In: *CoRR* abs/1807.07547. arXiv: 1807.07547. URL: http://arxiv.org/abs/1807.07547.

Götze, Friedrich, Holger Sambale, and Arthur Sinulis (2019). "Concentration inequalities for polynomials in α-sub-exponential random variables". In: *arXiv preprint arXiv:1903.05964*.

Grothendieck, Alexander (1956). *Résumé of the metrological theory of topological tensor products*. Soc. of Matemática of São Paulo.

Guédon, Olivier and Roman Vershynin (2016). "Community detection in sparse networks via Grothendieck's inequality". In: *Probability Theory and Related Fields* 165.3-4, pp. 1025–1049.

Jacot, Arthur, Franck Gabriel, and Clement Hongler (2018). "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in neural information processing systems*, pp. 8571–8580.

Kanagawa, Motonobu, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur (2018). "Gaussian processes and kernel methods: A review on connections and equivalences". In: *arXiv preprint arXiv:1807.02582*.

Luxburg, Ulrike von, Mikhail Belkin, and Olivier Bousquet (2008). "Consistency of spectral clustering". In: *The Annals of Statistics*, pp. 555–586.

Mai, Xiaoyi and Romain Couillet (2018). "A random matrix analysis and improvement of semi-supervised learning for large dimensional data". In: *Journal of Machine Learning Research* 19.1, pp. 3074–3100.

Mendelson, Shahar and Joseph Neeman (2010). "Regularization in kernel learning". In: *The Annals of Statistics* 38.1, pp. 526–565.

Paul, Debashis (2007). "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model". In: *Statistica Sinica*, pp. 1617–1642.

Peng, Jiming and Yu Wei (2007). "Approximating k-means-type clustering via semidefinite programming". In: *SIAM Journal on Optimization* 18.1, pp. 186–205.

Pollard, David (1981). "Strong consistency of k-means clustering". In: *The Annals of Statistics* 9.1, pp. 135–140.

Steinwart, Ingo and Andreas Christmann (2008). *Support vector machines*. Springer Science & Business Media.

Wang, Shusen, Alex Gittens, and Michael Mahoney (2019). "Scalable Kernel k-Means Clustering with Nyström Approximation: Relative-Error Bounds". In: *Journal of Machine Learning Research* 20, pp. 1–49.

Wasserman, Larry and John D Lafferty (2008). "Statistical analysis of semi-supervised regression". In: *Advances in Neural Information Processing Systems*, pp. 801–808.

Yan, Bowei and Purnamrita Sarkar (2016). "On robustness of kernel clustering". In: *Advances in Neural Information Processing Systems*, pp. 3098–3106.

Yu, Yi, Tengyao Wang, and Richard J Samworth (2014). "A useful variant of the Davis–Kahan theorem for statisticians". In: *Biometrika* 102.2, pp. 315–323.