Gaussian Sketching yields a J-L Lemma in RKHS

Samory Kpotufe
Statistics, Columbia University

Bharath K. Sriperumbudur
Statistics, Pennsylvania State University

Abstract

The main contribution of the paper is to show that Gaussian sketching of a kernel-Gram matrix $\boldsymbol{K}$ yields an operator whose counterpart in an RKHS $\mathcal{H}$, is a *random projection* operator—in the spirit of Johnson-Lindenstrauss (J-L) lemma. To be precise, given a random matrix Z with i.i.d. Gaussian entries, we show that a sketch $Z\boldsymbol{K}$ corresponds to a particular random operator in (infinite-dimensional) Hilbert space $\mathcal{H}$ that maps functions $f \in \mathcal{H}$ to a low-dimensional space $\mathbb{R}^d$, while preserving a weighted RKHS inner-product of the form $\langle f, g\rangle_\Sigma \doteq \langle f, \Sigma^3 g\rangle_\mathcal{H}$, where Σ is the *covariance* operator induced by the data distribution. In particular, under similar assumptions as in kernel PCA (KPCA), or kernel k-means (K-k-means), well-separated subsets of feature-space $\{K(\cdot, x) : x \in \mathcal{X}\}$ remain well-separated after such operation, which suggests similar benefits as in KPCA and/or K-k-means, albeit at the much cheaper cost of a random projection. In particular, our convergence rates suggest that, given a large dataset $\{X_i\}_{i=1}^N$ of size N, we can build the Gram matrix $\boldsymbol{K}$ on a much smaller subsample of size $n \ll N$, so that the sketch $Z\boldsymbol{K}$ is very cheap to obtain and subsequently apply as a projection operator on the original data $\{X_i\}_{i=1}^N$. We verify these insights empirically on synthetic data, and on real-world clustering applications.

1 Introduction

The Gram matrix $\boldsymbol{K}$, defined as $\boldsymbol{K}_{ij} = K(X_i, X_j)$ over a (sub) sample $\boldsymbol{X} \doteq \{X_i\}_{i=1}^n$, for a PSD kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, plays a central role in *kernel machines*, where learning tasks in a (reproducing kernel) Hilbert space $\mathcal{H}$ can be performed in sample space $\mathcal{X}$ via $\boldsymbol{K}$. Sketching of $\boldsymbol{K}$, i.e., multiplying by a random matrix (or matrices) $Z \in \mathbb{R}^{d\times n}$—as a form of rank reduction, is now ubiquitous in the design of computationally efficient approaches to kernel machines (Wang et al., 2019; Williams and Seeger, 2001; Yang et al., 2017). The simplest sketching approach consists of random subsampling of columns of $\boldsymbol{K}$, i.e., a data reduction, while the usual alternative of a Gaussian sketch (of the form $Z\boldsymbol{K}$, $Z_{i,j}{\sim}\mathcal{N}(0, 1)$) has less immediate interpretation. A main aim of this paper is to derive an operator-theoretic interpretation of Gaussian sketching, i.e., understand its effect in kernel space $\mathcal{H}$ on embedded data $K(x, \cdot)$. The analysis reveals interesting norm preservation properties of $Z\boldsymbol{K}$, in the spirit of the Johnson-Linderstauss (J-L) lemma, even when $\boldsymbol{K}$ is viewed as a smaller submatrix of an initial gram-matrix $\boldsymbol{K}_N$ on $N \gg n$ samples; these new insights imply an alternative use of Gaussian sketching in important applications such as kernel clustering or PCA, while yielding faster preprocessing than even vanilla Nyström.

Results Overview. It has been folklore in the community that Gaussian sketching corresponds to some form of *random projection*, although it remained unclear in which formal sense this is true. To draw the link to operators on $\mathcal{H}$, we consider linear operations of the form $Z\boldsymbol{K}f_{|\boldsymbol{X}} \in \mathbb{R}^d$, where $f_{|\boldsymbol{X}} \doteq (f(X_1), \ldots, f(X_n))^\top$ denotes the *sampled* version of $f \in \mathcal{H}$. We show that, $Z\boldsymbol{K}$, viewed in this sense as an operator, corresponds to a *random* operator Θ which maps (potentially infinite-dimensional) $\mathcal{H}$ to lower-dimensional $\mathbb{R}^d$, while preserving—in the spirit of the J-L lemma (Johnson and Lindenstrauss, 1984; Dasgupta and Gupta, 2003)—a weighted RKHS inner-product of the form $\langle f, g\rangle_\Sigma \doteq \langle f, \Sigma^3 g\rangle_\mathcal{H}$, Σ

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

being the *covariance* operator induced by the data-generating distribution (as defined in Section 2).

The corresponding random operator Θ *projects*—in the informal sense of dimension reduction—any $f \in \mathcal{H}$ onto d i.i.d Gaussian directions[1] $\{v_i\}_{i=1}^d$ in $\mathcal{H}$: formally, given i.i.d. Gaussians $\{v_i\}_{i=1}^d \sim \mathcal{N}_{\mathcal{H}}(0, \Sigma^3)$, Θ maps any $f \in \mathcal{H}$ to the vector $\frac{1}{\sqrt{d}}(\langle f, v_1\rangle_{\mathcal{H}}, \ldots, \langle f, v_d\rangle_{\mathcal{H}})^\top \in \mathbb{R}^d$. We refer the reader to Section 3 for details.

In Section 4 (see Theorem 1), we show the following correspondence between $Z\boldsymbol{K}$ and Θ: just as Θ preserves $\left\langle g, \Sigma^3 f\right\rangle_{\mathcal{H}}$, so does $Z\boldsymbol{K}$ (properly normalized), i.e., with high probability, we have $\forall f, g \in \mathcal{H}$,

$$\frac{1}{n^3 d}\left\langle (Z\boldsymbol{K})g_{|\boldsymbol{X}}, (Z\boldsymbol{K})f_{|\boldsymbol{X}}\right\rangle_2 \approx \left\langle g, \Sigma^3 f\right\rangle_{\mathcal{H}} \approx \left\langle \Theta g, \Theta f\right\rangle_2, \quad (1)$$

where $\langle\cdot,\cdot\rangle_2$ denotes the inner-product in $\mathbb{R}^d$. The result holds simultaneously $\forall f, g \in \mathcal{H}$, with an approximation rate of order $n^{-1/2} + d^{-1/2}$, for n, d greater than *effective dimension* terms (s_Σ or s_{Σ^3} of Theorem 1). In other words such approximation holds for both n and d small, whenever the *effective dimension* is small; our experiments suggest this is often the case.

Time complexity. The above result suggests a novel use of sketching where, given a larger dataset $\boldsymbol{X}_N = \{X_i\}_{i=1}^N$, we re-map all $X_i \in \boldsymbol{X}_N$ to $\mathbb{R}^d$ using n subsamples $\boldsymbol{X} \subset \boldsymbol{X}_N$, $n, d \ll N$, to form a *projection* operator $Z\boldsymbol{K}$. In other words, we re-map feature functions $K(\cdot, X_i), X_i \in \boldsymbol{X}_N$ to $\frac{1}{n^{3/2}\sqrt{d}}(Z\boldsymbol{K})K(\cdot, X_i)_{|\boldsymbol{X}}$, following the intuition that useful properties of kernel feature maps $K(\cdot, x)$ are preserved. We refer to such a mapping as Kernel JL (K-JL) for short. The time complexity is exactly $d \cdot n^2$ for forming $Z\boldsymbol{K}$, in addition to $N \cdot d \cdot n$ for the subsequent mapping of all N datapoints. The leading constant is 1 in all cases. In contrast, the cheapest Nyström approximation using n subsampled columns costs $O(d \cdot n^2)$ (for pseudo-inverse computation, where constants depend on desired precision) plus $N \cdot d \cdot n$ for mapping datapoints. K-JL avoids eigen-decompositions or matrix inversion steps, besides requiring smaller n for stability (see Section 5, for details including Nyström formulation).

Performance. Now, whether K-JL preserves useful properties of feature mapping depends on how the inner-product $\left\langle g, \Sigma^3 f\right\rangle_{\mathcal{H}}$ relates to the natural inner-product $\langle g, f\rangle_{\mathcal{H}}$ of the RKHS $\mathcal{H}$. In the present work, we consider clustering and PCA applications, which require that properties such as separation (in $\mathcal{H}$ distance) between given subsets of *feature space* $\{K(\cdot, x) : x \in \mathcal{X}\} \subset \mathcal{H}$ are preserved. At first glance, there seems to be little hope, since in the worst-case over $\mathcal{H}$, there exist $f \in \mathcal{H}$ such that $\|f\|_{\mathcal{H}}^2 \doteq \langle f, f\rangle_{\mathcal{H}}$ is large but $\left\langle f, \Sigma^3 f\right\rangle_{\mathcal{H}}$ is close to 0 (e.g., eigenfunctions f of Σ with eigenvalues tending to 0).

Interestingly however, as we argue in Section 5, we can expect well-separated subsets of feature space $\{K(\cdot, x) : x \in \mathcal{X}\} \subset \mathcal{H}$ to remain well-separated after K-JL, under conditions favorable to kernel PCA (KPCA) (Blanchard et al., 2007; Mika et al., 1999; Schölkopf et al., 1998), or conditions favorable to kernel k-means (K-k-means) (Dhillon et al., 2004). Namely, if feature maps $K(\cdot, x)$ lie close to a low-dimensional subspace, or feature maps *cluster* well (in which case the means of clusters lie close to a low-dimensional subspace), then the worst-case distortions between the two inner-products happen outside of feature space $\{K(\cdot, x) : x \in \mathcal{X}\}$. This entails similar benefits as in KPCA and or K-k-means, albeit at the cheaper cost of a random projection. This intuition holds empirically, as we verify on a mix of synthetic data and real-world clustering applications.

Further Related Work. We note that *sketching* is of general interest outside the present context, motivated by the need for efficient approximations of general matrices appearing in numerical and data analysis (Woodruff et al., 2014; Andoni et al., 2016, 2018). Finally, we note that the benefits of Johnson-Linderstrauss type projections in Hilbert spaces were considered in Biau et al. (2008), however under the assumptions of a theoretical procedure which requires explicit Fourier coefficients (basis expansion) of Hilbert space elements $K(X_i, \cdot)$.

Paper Outline

Section 2 covers definitions and basic assumptions used throughout the paper. In Section 3 we develop some initial intuition about JL-type *random projections* in $\mathcal{H}$, followed by formal results in Section 4. Omitted proofs and supporting results are collected in an appendix.

2 Preliminaries

Let $\mathcal{X}$ denote a separable topological space on which a Borel probability measure ρ_X is defined. We assume that $\mathcal{H}$, consisting of functions $\mathcal{X} \to \mathbb{R}$, is a reproducing kernel Hilbert space (RKHS) with a continuous

[1] The notion of a Gaussian measure $\mathcal{N}_{\mathcal{H}}$ on $\mathcal{H}$ has to be suitably defined so as to ensure that random draws $v \sim \mathcal{N}_{\mathcal{H}}$ are indeed elements of $\mathcal{H}$.

and bounded reproducing kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ where $\sup_{x\in\mathcal{X}} K(x,x) =: \kappa < \infty$. For any $f \in \mathcal{H}$, the *outer-product* notation $f \otimes_{\mathcal{H}} f$ denotes the operator $g \mapsto \langle g, f\rangle_{\mathcal{H}} f$. We let $\Sigma : \mathcal{H} \to \mathcal{H}$ denote the uncentered covariance operator, which is defined as

$$\Sigma \doteq \int K(\cdot,x) \otimes_{\mathcal{H}} K(\cdot,x)\, d\rho_X(x),$$

in the sense of Bochner integration (Diestel and Uhl, 1977). Given data $\{X_i\}_{i=1}^n \overset{i.i.d.}{\sim} \rho_X$ where $n \geq 1$, the empirical counterpart of Σ is defined as

$$\Sigma_n \doteq \frac{1}{n}\sum_{i=1}^n K(\cdot,X_i) \otimes_{\mathcal{H}} K(\cdot,X_i).$$

Given two normed spaces $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ and $(\mathcal{G}, \|\cdot\|_{\mathcal{G}})$, let $A : \mathcal{F} \to \mathcal{G}$ and $B : \mathcal{F} \to \mathcal{F}$ be two linear operators. The operator norm of A is defined as $\|A\|_{\text{op}} \doteq \sup_{f\in\mathcal{F}} \frac{\|Af\|_{\mathcal{G}}}{\|f\|_{\mathcal{F}}}$.

The trace of a non-negative self-adjoint operator B, operating on a separable Hilbert space $\mathcal{F}$, is defined as $\text{tr}(B) \doteq \sum_\ell \langle Be_\ell, e_\ell\rangle_{\mathcal{F}}$, where $(e_\ell)_\ell$ is any orthonormal basis in $\mathcal{F}$. The Hilbert-Schmidt norm of B is then defined as $\|B\|_{\mathcal{L}_2(\mathcal{F})} \doteq \sqrt{\text{tr}(B^*B)}$.

A random element v of $\mathcal{H}$ is said to have **Gaussian** measure, denoted $\mathcal{N}_{\mathcal{H}}$, if for any $f \in \mathcal{H}$, $\langle v, f\rangle_{\mathcal{H}}$ is Gaussian. It is known that such a measure is well-defined, in the sense that $v \sim \mathcal{N}_{\mathcal{H}}$ has finite norm $\|v\|_{\mathcal{H}}$ w.p. 1, whenever its corresponding *covariance operator* $\mathcal{C} \doteq \mathbb{E}\, v \otimes_{\mathcal{H}} v - \mu \otimes_{\mathcal{H}} \mu$, $\mu = \mathbb{E}v$, is trace-class, i.e., has finite trace (see e.g., Bogachev, 1998). We can then parametrize the measure as $\mathcal{N}_{\mathcal{H}}(\mu, \mathcal{C})$.

3 Intuition on Random Projection in RKHS $\mathcal{H}$

As mentioned in Section 1, a key contribution of this paper is in showing that the random projection operator Θ is related to the Gaussian sketch of a kernel matrix. Before we present and prove a rigorous result in Section 4, in this section, we heuristically demonstrate the connection. In particular, we elucidate why Σ^3 shows up (rather than e.g. Σ, given that a priori $\boldsymbol{K}$ seems most naturally related to Σ), and why the origin of the peculiar normalization by $n^{3/2}\sqrt{d}$.

Given a set of N datapoints in $\mathbb{R}^D$, classical random projections in the style of Johsohn-Lindenstrauss (J-L) consists of projecting the datapoints onto d random directions which are sampled from a standard Gaussian distribution. The same idea can be intuitively carried forward to an RKHS, $\mathcal{H}$ by sampling functions from a Gaussian measure on $\mathcal{H}$—these functions act as directions along which a function in $\mathcal{H}$ can be projected. Now, consider the random directions $v_i \overset{i.i.d.}{\sim} \mathcal{N}_{\mathcal{H}}(0, \Sigma^3)$ and define the random projection of $f \in \mathcal{H}$ to $\mathbb{R}^d$ through the random operator $\Theta : \mathcal{H} \to \mathbb{R}^d$

$$f \mapsto \frac{1}{\sqrt{d}}(\langle v_1, f\rangle_{\mathcal{H}}, \ldots, \langle v_d, f\rangle_{\mathcal{H}})^\top. \tag{2}$$

It is important to note that random directions cannot be sampled from a Gaussian measure with identity covariance operator $I_{\mathcal{H}}$ (i.e., similar to the classical setting) as such a measure is not well-defined for infinite dimensional Hilbert spaces since $I_{\mathcal{H}}$ has infinite trace. The above normalization by $d^{-1/2}$ ensures that, with high-probability, $\langle \Theta g, \Theta f\rangle_2 \xrightarrow{d\to\infty} \langle g, \Sigma^3 f\rangle_{\mathcal{H}}$ (see Proposition 1).

Now define $(u_i)_{i=1}^d \overset{i.i.d.}{\sim} \mathcal{N}_{\mathcal{H}}(0, \Sigma)$ so that $(v_i)_{i=1}^d$ can be written as $v_i = \Sigma u_i$ and

$$\begin{aligned}\langle v_i, f\rangle_{\mathcal{H}} = \langle \Sigma u_i, f\rangle_{\mathcal{H}} &= \int f(x) u_i(x)\, d\rho_X(x) \\ &= \langle u_i, f\rangle_{L^2(\mathcal{X},\rho_X)}.\end{aligned}$$

The above can be approximated empirically, using $\boldsymbol{X} \doteq \{X_i\}_{i=1}^n \overset{i.i.d.}{\sim} \rho_X$, as

$$\begin{aligned}\langle u_i, f\rangle_{L^2(\rho_X)} &\approx \frac{1}{n}\sum_{j=1}^n f(X_j) u_i(X_j) \\ &= \frac{1}{n}\langle S_{\boldsymbol{X}} u_i, S_{\boldsymbol{X}} f\rangle_2,\end{aligned} \tag{3}$$

where

$$S_{\boldsymbol{X}} : \mathcal{H} \to \mathbb{R}^n,\ f \mapsto (f(X_1), \ldots, f(X_n))^\top$$

is a *sampling operator* (Smale and Zhou, 2007) whose adjoint is given by

$$S_{\boldsymbol{X}}^* : \mathbb{R}^n \to \mathcal{H},\ \boldsymbol{\beta} \mapsto \sum_{i=1}^n \beta_i K(\cdot, X_i).$$

It follows from Proposition B.1 (in the appendix) that, conditioned on the sample $\boldsymbol{X}$, $S_{\boldsymbol{X}} u_i$ is distributed as $\mathcal{N}(0, M)$ where $M \in \mathbb{R}^{n\times n}$ is defined as

$$\begin{aligned}M_{jl} &\doteq \langle K(\cdot, X_j), \Sigma K(\cdot, X_l)\rangle_{\mathcal{H}} \\ &= \int_{\mathcal{X}} K(x, X_j) K(x, X_l)\, d\rho_X(x),\end{aligned} \tag{4}$$

with $X_j, X_l \in \boldsymbol{X}$. Based on $\boldsymbol{X}$, M can be further approximated as $\hat{M}$ where

$$\begin{aligned}\hat{M}_{jl} &\doteq \langle K(\cdot, X_j), \Sigma_n K(\cdot, X_l)\rangle_{\mathcal{H}} \\ &= \frac{1}{n}\sum_{i=1}^n K(X_i, X_j) K(X_i, X_l) = \frac{1}{n}(\boldsymbol{K}^2)_{jl},\end{aligned} \tag{5}$$

where $\boldsymbol{K}$ is the Gram matrix based on $\boldsymbol{X}$. To summarize, we have carried out the following sequence of approximations to $\langle v_i, f\rangle_{\mathcal{H}}$:

$$\langle v_i, f\rangle_{\mathcal{H}} = \langle u_i, f\rangle_{L^2(\mathcal{X},\rho_X)} \approx \frac{1}{n}\langle S_{\boldsymbol{X}} u_i, S_{\boldsymbol{X}} f\rangle_2$$

where $S_{\boldsymbol{X}} u_i \sim \mathcal{N}(0, M) \approx \mathcal{N}\left(0, \frac{1}{n}\boldsymbol{K}^2\right)$. This means an approximation to $\langle v_i, f\rangle_{\mathcal{H}}$ can be obtained by sampling, say $\hat{v}_i$ from $\mathcal{N}\left(0, \frac{1}{n}\boldsymbol{K}^2\right)$ and computing $\frac{1}{n}\langle \hat{v}_i, S_{\boldsymbol{X}} f\rangle_2$. Recalling the form of Θ (2), define

$$\hat{V} = \frac{1}{n\sqrt{d}}[\hat{v}_1, \ldots, \hat{v}_d] \in \mathbb{R}^{n\times d}.$$

The *approximate random projection operator* is then $\hat{V}^\top S_{\boldsymbol{X}} : \mathcal{H} \to \mathbb{R}^d$, where

$$f \mapsto \hat{V}^\top S_{\boldsymbol{X}} f = \frac{1}{n\sqrt{d}}(\langle \hat{v}_1, S_{\boldsymbol{X}} f\rangle_2, \ldots, \langle \hat{v}_d, S_{\boldsymbol{X}} f\rangle_2)^\top.$$

Note that $\hat{V}^\top = \frac{1}{n\sqrt{d}}[\hat{v}_1, \ldots, \hat{v}_d]^\top = \frac{1}{n\sqrt{nd}} Z\boldsymbol{K}$ with $Z \in \mathbb{R}^{d\times n}$ having i.i.d. $\mathcal{N}(0, 1)$ entries.

4 Main Results

In this section, we formalize the relation between Θ and $\hat{V}^\top S_{\boldsymbol{X}}$ by showing that, with high-probability, $\left\langle \hat{V}^\top g_{|\boldsymbol{X}}, \hat{V}^\top f_{|\boldsymbol{X}}\right\rangle_2 \approx \langle g, \Sigma^3 f\rangle_{\mathcal{H}} \approx \langle \Theta g, \Theta f\rangle_2$. This relation is established in Proposition1 (proved in Section A.1), and Theorem 1. In the sequel, we let $a \wedge b \doteq \min\{a, b\}$ and $a \vee b \doteq \max\{a, b\}$.

Proposition 1. *Define $s_\Sigma = \frac{\operatorname{tr}(\Sigma)}{\|\Sigma\|_{\mathrm{op}}}$. For any $\tau \geq 1$ and $d \geq (s_{\Sigma^3} \vee \tau)$, with probability at least $1 - e^{-\tau}$,*

$$\sup_{f,g\in\mathcal{H}} \frac{\left|\langle \Theta g, \Theta f\rangle_2 - \langle g, \Sigma^3 f\rangle_{\mathcal{H}}\right|}{\|f\|_{\mathcal{H}}\|g\|_{\mathcal{H}}} \leq \mathfrak{C}\|\Sigma\|_{\mathrm{op}}^3 \frac{\sqrt{s_{\Sigma^3}} + \sqrt{\tau}}{\sqrt{d}},$$

where $\mathfrak{C}$ is a universal constant independent of Σ, τ and d.

Theorem 1 (Convergence of inner products)**.** *Let $\tau \geq 1$. Define $s_\Sigma = \frac{\operatorname{tr}(\Sigma)}{\|\Sigma\|_{\mathrm{op}}}$. Suppose*

$$n \geq \frac{6272\kappa s_\Sigma^5 \tau}{\|\Sigma\|_{\mathrm{op}}}.$$

Then, with probability at least $1 - 5e^{-\tau}$ jointly over the choice of $\{\hat{v}_i\}_{i=1}^d$ and $\{X_i\}_{i=1}^n$:

$$\sup_{f,g\in\mathcal{H}} \frac{\left|\left\langle \hat{V}^\top S_{\boldsymbol{X}} g, \hat{V}^\top S_{\boldsymbol{X}} f\right\rangle_2 - \langle g, \Sigma^3 f\rangle_{\mathcal{H}}\right|}{\|f\|_{\mathcal{H}}\|g\|_{\mathcal{H}}}$$
$$\leq 3\mathfrak{C}\|\Sigma\|_{\mathrm{op}}^3 \frac{\sqrt{2s_{\Sigma^3}} + \sqrt{\tau}}{\sqrt{d}} + \frac{28\|\Sigma\|_{\mathrm{op}}^{5/2}\sqrt{2\kappa s_\Sigma \tau}}{\sqrt{n}},$$

where $\mathfrak{C}$ is a universal constant that does not depend on n, d, κ and Σ.

Remark. (Main dependence on n and d) The leading constants above are in terms of $\|\Sigma\|_{\mathrm{op}} \leq \kappa \doteq \sup_x K(x, x)$, and are therefore expected to be small for common kernels such as Gaussian ($\kappa = 1$). Thus the main dependence on n and d in the rates are given by s_Σ and $s_{\Sigma^3} \leq s_\Sigma$, viewed as *effective dimension* terms. Thus, whenever s_Σ is small, both n and d can be chosen small while maintaining the guarantees of the above theorem. In our experiments of Section 5, $d \leq n \leq 100$ is often sufficient even for datasizes in excess of 40K points (see e.g. Figure 2).

Proof of Theorem 1. Define $A(f, g) \doteq \left\langle \hat{V}^\top S_{\boldsymbol{X}} g, \hat{V}^\top S_{\boldsymbol{X}} f\right\rangle_2 - \langle g, \Sigma^3 f\rangle_{\mathcal{H}}$. Since

$$\left\langle \hat{V}^\top S_{\boldsymbol{X}} g, \hat{V}^\top S_{\boldsymbol{X}} f\right\rangle_2 = \left\langle g, S_{\boldsymbol{X}}^* \hat{V}\hat{V}^\top S_{\boldsymbol{X}} f\right\rangle_{\mathcal{H}},$$

we have

$$\sup_{f,g\in\mathcal{H}} \frac{|A(f,g)|}{\|f\|_{\mathcal{H}}\|g\|_{\mathcal{H}}} = \sup_{f\in\mathcal{H}} \frac{\left\|\left(S_{\boldsymbol{X}}^* \hat{V}\hat{V}^\top S_{\boldsymbol{X}} - \Sigma^3\right) f\right\|_{\mathcal{H}}}{\|f\|_{\mathcal{H}}}$$
$$= \left\|S_{\boldsymbol{X}}^* \hat{V}\hat{V}^\top S_{\boldsymbol{X}} - \Sigma^3\right\|_{\mathrm{op}}.$$

In the following, we bound $\left\|S_{\boldsymbol{X}}^* \hat{V}\hat{V}^\top S_{\boldsymbol{X}} - \Sigma^3\right\|_{\mathrm{op}}$. To this end, consider

$$\begin{aligned} &S_{\boldsymbol{X}}^* \hat{V}\hat{V}^\top S_{\boldsymbol{X}} - \Sigma^3 \\ &= S_{\boldsymbol{X}}^* \left(\hat{V}\hat{V}^\top - \frac{1}{n^3}\boldsymbol{K}^2\right) S_{\boldsymbol{X}} + \frac{1}{n^3} S_{\boldsymbol{X}}^* \boldsymbol{K}^2 S_{\boldsymbol{X}} - \Sigma^3 \\ &\overset{(\dagger)}{=} S_{\boldsymbol{X}}^* \left(\hat{V}\hat{V}^\top - \frac{1}{n^3}\boldsymbol{K}^2\right) S_{\boldsymbol{X}} + \Sigma_n^3 - \Sigma^3, \end{aligned}$$

where in (†), we use the facts that $\boldsymbol{K} = S_{\boldsymbol{X}} S_{\boldsymbol{X}}^*$ and $\Sigma_n = \frac{1}{n} S_{\boldsymbol{X}}^* S_{\boldsymbol{X}}$. Therefore,

$$\begin{aligned} &\left\|S_{\boldsymbol{X}}^* \hat{V}\hat{V}^\top S_{\boldsymbol{X}} - \Sigma^3\right\|_{\mathrm{op}} \\ &\leq \left\|S_{\boldsymbol{X}}^* \left(\hat{V}\hat{V}^\top - \frac{1}{n^3}\boldsymbol{K}^2\right) S_{\boldsymbol{X}}\right\|_{\mathrm{op}} + \|\Sigma_n^3 - \Sigma^3\|_{\mathrm{op}} \\ &\overset{(\ddagger)}{\leq} \mathfrak{C}\left(\sqrt{\frac{\operatorname{tr}(\Sigma_n^3)\|\Sigma_n^3\|_{\mathrm{op}}}{d}} + \|\Sigma_n^3\|_{\mathrm{op}}\sqrt{\frac{\tau}{d}}\right) \\ &\qquad + \|\Sigma_n^3 - \Sigma^3\|_{\mathrm{op}}, \\ &\leq \mathfrak{C}\left(\sqrt{\frac{\operatorname{tr}(\Sigma_n^3)\|\Sigma_n^3 - \Sigma^3\|_{\mathrm{op}}}{d}} + \sqrt{\frac{\operatorname{tr}(\Sigma_n^3)\|\Sigma^3\|_{\mathrm{op}}}{d}}\right. \\ &\qquad \left. + \|\Sigma^3\|_{\mathrm{op}}\sqrt{\frac{\tau}{d}}\right) + \|\Sigma_n^3 - \Sigma^3\|_{\mathrm{op}}\left(1 + \mathfrak{C}\sqrt{\frac{\tau}{d}}\right), \end{aligned} \tag{6}$$

where (‡) follows from Lemma B.3, which holds with probability $1 - e^{-\tau}$ over the choice of $(\hat{v}_i)_{i=1}^d$ conditioned on $\boldsymbol{X}$ for any $\tau \geq 1$ and $d \geq$

$\left(\frac{\mathrm{tr}(\Sigma_n^3)}{\|\Sigma^3\|_{\mathrm{op}}-\|\Sigma^3-\Sigma_n^3\|_{\mathrm{op}}} \vee \tau\right) \geq \left(\frac{\mathrm{tr}(\Sigma_n^3)}{\|\Sigma_n^3\|_{\mathrm{op}}} \vee \tau\right)$. $\mathfrak{C}$ is a universal constant independent of $S_{\boldsymbol{X}}^* \boldsymbol{K}^2 S_{\boldsymbol{X}}$ and d.

We now bound $\mathrm{tr}(\Sigma_n^3)$ and $\|\Sigma_n^3 - \Sigma^3\|_{\mathrm{op}}$. Consider

$$\begin{aligned}\Sigma_n^3 - \Sigma^3 &= (\Sigma_n - \Sigma + \Sigma)^3 - \Sigma^3 \\ &= (\Sigma_n - \Sigma)^3 + (\Sigma_n - \Sigma)^2\Sigma + (\Sigma_n - \Sigma)\Sigma(\Sigma_n - \Sigma) \\ &\quad + (\Sigma_n - \Sigma)\Sigma^2 + \Sigma(\Sigma_n - \Sigma)^2 + \Sigma(\Sigma_n - \Sigma)\Sigma \\ &\qquad + \Sigma^2(\Sigma_n - \Sigma),\end{aligned}$$

which yields

$$\begin{aligned}\|\Sigma_n^3 - \Sigma^3\|_{\mathrm{op}} &\leq \|\Sigma_n - \Sigma\|_{\mathrm{op}}^3 + 3\|\Sigma_n - \Sigma\|_{\mathrm{op}}^2\|\Sigma\|_{\mathrm{op}} \\ &\quad + 3\|\Sigma_n - \Sigma\|_{\mathrm{op}}\|\Sigma\|_{\mathrm{op}}^2 \\ &\leq \|\Sigma_n - \Sigma\|_{\mathcal{L}_2(\mathcal{H})}^3 + 3\|\Sigma_n - \Sigma\|_{\mathcal{L}_2(\mathcal{H})}^2\|\Sigma\|_{\mathrm{op}} \\ &\quad + 3\|\Sigma_n - \Sigma\|_{\mathcal{L}_2(\mathcal{H})}\|\Sigma\|_{\mathrm{op}}^2\end{aligned}$$

and

$$\begin{aligned}\mathrm{tr}(\Sigma_n^3) &\leq \mathrm{tr}(\Sigma^3) + 3\|\Sigma\|_{\mathcal{L}_2(\mathcal{H})}\|\Sigma_n - \Sigma\|_{\mathcal{L}_2(\mathcal{H})}^2 \\ &\quad + \|\Sigma_n - \Sigma\|_{\mathcal{L}_2(\mathcal{H})}^3 + 3\|\Sigma\|_{\mathcal{L}_2(\mathcal{H})}^2\|\Sigma_n - \Sigma\|_{\mathcal{L}_2(\mathcal{H})}.\end{aligned}$$

It follows from Lemma B.2 that for any $\tau > 0$ and $n \geq \frac{32\kappa s_\Sigma \tau}{\|\Sigma\|_{\mathrm{op}}}$,

$$\|\Sigma_n^3 - \Sigma_n^3\|_{\mathrm{op}} \leq 28\|\Sigma\|_{\mathrm{op}}^2 \sqrt{\frac{2\kappa \mathrm{tr}(\Sigma)\tau}{n}} \qquad (7)$$

and

$$\mathrm{tr}(\Sigma_n^3) \leq \mathrm{tr}(\Sigma^3) + 28\|\Sigma\|_{\mathcal{L}_2(\mathcal{H})}^2 \sqrt{\frac{2\kappa \mathrm{tr}(\Sigma)\tau}{n}}, \qquad (8)$$

where each of the above inequalities hold with probability at least $1 - 2e^{-\tau}$ over $(X_1, \ldots, X_n)$. Using (7) and (8) in (6) yields the result, upon tying a few loose ends.

Define $\Delta \doteq 28\|\Sigma\|_{\mathcal{L}_2(\mathcal{H})}^2 \sqrt{\frac{2\kappa \mathrm{tr}(\Sigma)\tau}{n}}$ and $\Delta' \doteq \frac{\|\Sigma\|_{\mathrm{op}}^2}{\|\Sigma\|_{\mathcal{L}_2(\mathcal{H})}^2}\Delta$. As aforementioned, (6) holds if $d \geq \left(\frac{\mathrm{tr}(\Sigma_n^3)}{\|\Sigma^3\|_{\mathrm{op}}-\|\Sigma_n^3-\Sigma^3\|_{\mathrm{op}}} \vee \tau\right)$ which is the case whenever $d \geq \left(\frac{\mathrm{tr}(\Sigma^3)+\Delta}{\|\Sigma^3\|_{\mathrm{op}}-\Delta'} \vee \tau\right)$. Under the assumed conditions on n, it follows that $\|\Sigma\|_{\mathrm{op}}^3 \geq 2\Delta'$ and $\mathrm{tr}(\Sigma^3) \leq \Delta$, which yields that $d \geq \left(\frac{\mathrm{tr}(\Sigma^3)+\Delta}{\|\Sigma^3\|_{\mathrm{op}}-\Delta'} \vee \tau\right)$ is true whenever $d \geq (4s_{\Sigma^3} \vee \tau)$. □

Theorem 1 shows that the approximate random projection operator $\hat{V}^\top S_{\boldsymbol{X}}$ preserves the inner product $\langle g, \Sigma^3 f\rangle_{\mathcal{H}}$ uniformly over all $f, g \in \mathcal{H}$ at an approximation rate of $n^{-1/2} + d^{-1/2}$.

The following result (proved in Section **??**) provides a different angle by which Θ relates to $\hat{V}^\top S_{\boldsymbol{X}}$, by showing that, for all $\boldsymbol{\alpha} \in \mathbb{R}^d$ and $f \in \mathcal{H}$, $\langle \boldsymbol{\alpha}, \hat{V}^\top S_{\boldsymbol{X}} f\rangle_2$ converges in probability to $\langle \boldsymbol{\alpha}, \Theta f\rangle_2$ at the rate of $d^{-1/2}$, provided n is large enough for $\|\Sigma_n\|_{\mathrm{op}} \lesssim \|\Sigma\|_{\mathrm{op}}$. Recall that two operators $A, B : \mathcal{H} \to \mathbb{R}^d$ are equal (in a *weak sense*) if $\forall f \in \mathcal{H}$, $\forall \boldsymbol{\alpha} \in \mathbb{R}^d$, we have $\langle \boldsymbol{\alpha}, Af\rangle_2 = \langle \boldsymbol{\alpha}, Bf\rangle_2$.

Theorem 2 (Convergence of random projection operators). *Define $s_\Sigma = \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|_{\mathrm{op}}}$. For any $\boldsymbol{\alpha} \in \mathbb{R}^d$, $f \in \mathcal{H}$, $\tau > 0$ and*

$$n \geq \kappa s_\Sigma \tau \left(\frac{32}{\|\Sigma\|_{\mathrm{op}}} \vee 1\right),$$

with probability at least $1 - 4e^{-\tau}$ jointly over the choice of $\{\hat{v}_i\}_{i=1}^d$, $\{v_i\}_{i=1}^d$ and $\{X_i\}_{i=1}^n$:

$$\begin{aligned}&\left|\langle \boldsymbol{\alpha}, \hat{V}^\top S_{\boldsymbol{X}} f\rangle_2 - \langle \boldsymbol{\alpha}, \Theta f\rangle_2\right| \\ &\qquad \leq \frac{16\sqrt{2\tau}\|\boldsymbol{\alpha}\|_2\|f\|_{\mathcal{H}}\left(\|\Sigma\|_{\mathrm{op}}^{3/2} \vee \|\Sigma\|_{\mathrm{op}}^{5/4}\right)}{\sqrt{d}}. \qquad (9)\end{aligned}$$

Proof. Note that $\langle \boldsymbol{\alpha}, \hat{V}^\top S_{\boldsymbol{X}} f\rangle_2 = \langle \hat{V}\boldsymbol{\alpha}, S_{\boldsymbol{X}} f\rangle_2 = \frac{1}{n\sqrt{d}}\left\langle \sum_{i=1}^d \alpha_i \hat{v}_i, S_{\boldsymbol{X}} f\right\rangle_2$. Since $(\hat{v}_i) \overset{i.i.d.}{\sim} \mathcal{N}(0, \frac{1}{n}\boldsymbol{K}^2)$, conditioned on $\boldsymbol{X}$ it follows that

$$\begin{aligned}&\frac{1}{n\sqrt{d}}\left\langle \sum_{i=1}^d \alpha_i \hat{v}_i, S_{\boldsymbol{X}} f\right\rangle_2 \\ &\qquad \sim \mathcal{N}\left(0, \tfrac{1}{n^3 d}\|\boldsymbol{\alpha}\|_2^2 \langle S_{\boldsymbol{X}} f, \boldsymbol{K}^2 S_{\boldsymbol{X}} f\rangle_2\right) \\ &\qquad = \mathcal{N}\left(0, \tfrac{1}{d}\|\boldsymbol{\alpha}\|_2^2 \langle f, \Sigma_n^3 f\rangle_{\mathcal{H}}\right),\end{aligned}$$

where we used $\boldsymbol{K} = S_{\boldsymbol{X}} S_{\boldsymbol{X}}^*$ and $n\Sigma_n = S_{\boldsymbol{X}}^* S_{\boldsymbol{X}}$. On the other hand, $\langle \boldsymbol{\alpha}, \Theta f\rangle_2 = \frac{1}{\sqrt{d}}\sum_{i=1}^d \alpha_i \langle v_i, f\rangle_{\mathcal{H}} \sim \mathcal{N}\left(0, \frac{1}{d}\|\boldsymbol{\alpha}\|_2^2 \langle f, \Sigma^3 f\rangle_{\mathcal{H}}\right)$, which follows from the fact that $(v_i)_{i=1}^d \overset{i.i.d.}{\sim} \mathcal{N}_{\mathcal{H}}(0, \Sigma^3)$ which in turn implies $\langle v_i, f\rangle_{\mathcal{H}} \sim \mathcal{N}(0, \langle f, \Sigma^3 f\rangle_{\mathcal{H}})$. Therefore

$$\langle \boldsymbol{\alpha}, \hat{V}^\top S_{\boldsymbol{X}} f - \Theta f\rangle_2 \sim \mathcal{N}\left(0, \frac{\|\boldsymbol{\alpha}\|_2^2}{d}\langle f, (\Sigma^3 + \Sigma_n^3) f\rangle_{\mathcal{H}}\right)$$

conditioned on $\boldsymbol{X}$. For $Y \sim \mathcal{N}(0, \sigma^2)$, the Gaussian concentration inequality yields that for any $\tau > 0$, with probability at least $1 - 2e^{-\tau}$, $|Y| \leq \sqrt{2\sigma^2\tau}$. Hence it follows that for any $\tau > 0$, with probability at least $1 - 2e^{-\tau}$ jointly over $\{v_i\}_{i=1}^d$, $\{\hat{v}_i\}_{i=1}^d$ and conditioned on $\boldsymbol{X}$, we obtain

$$\begin{aligned}\left|\langle \boldsymbol{\alpha}, \hat{V}^\top S_{\boldsymbol{X}} f - \Theta f\rangle_2\right| &\leq \|\boldsymbol{\alpha}\|_2 \sqrt{\frac{2\tau \langle f, (\Sigma^3 + \Sigma_n^3) f\rangle_{\mathcal{H}}}{d}} \\ \leq \|\boldsymbol{\alpha}\|_2 \|f\|_{\mathcal{H}} \sqrt{2\tau} &\left(\frac{\sqrt{2}\|\Sigma\|_{\mathrm{op}}^{3/2}}{\sqrt{d}} + \frac{\|\Sigma^3 - \Sigma_n^3\|_{\mathrm{op}}^{1/2}}{\sqrt{d}}\right),\end{aligned} \qquad (10)$$

which follows from the fact that $\langle f, (\Sigma^3 + \Sigma_n^3) f\rangle_{\mathcal{H}} = 2\langle f, \Sigma^3 f\rangle_{\mathcal{H}} + \langle f, (\Sigma_n^3 - \Sigma^3) f\rangle_{\mathcal{H}} \leq 2\|\Sigma\|_{\mathrm{op}}^3 \|f\|_{\mathcal{H}}^2 +$

$\|f\|^2\|\Sigma_n^3 - \Sigma^3\|_{\mathrm{op}}$. Therefore, by unconditioning w.r.t. $\boldsymbol{X}$, (10) holds with probability at least $1 - 2e^{-\tau}$ jointly over the choice of $\{\hat{v}_i\}_{i=1}^d$, $\{v_i\}_{i=1}^d$ and $\{X_i\}_{i=1}^n$. The result therefore follows by invoking (7) to bound $\|\Sigma^3 - \Sigma_n^3\|_{\mathrm{op}}$ in (10). □

In Theorem 2, we require $d \to \infty$ to achieve $\langle \boldsymbol{\alpha}, \hat{V}^\top S_{\boldsymbol{X}} f\rangle_2 \to \langle \boldsymbol{\alpha}, \Theta f\rangle_2$ in probability for all $\boldsymbol{\alpha} \in \mathbb{R}^d$ and $f \in \mathcal{H}$, for sufficiently large n. Instead in the following result, we keep d fixed, and show the convergence in distribution of $\hat{V}^\top S_{\boldsymbol{X}} f$ to Θf as $n \to \infty$ for all $f \in \mathcal{H}$ as $n \to \infty$.

Theorem 3. *For all $f \in \mathcal{H}$ we have that*

$$\hat{V}^\top S_{\boldsymbol{X}} f \xrightarrow{dist} \Theta f, \text{ as } n \to \infty.$$

5 K-JL relations to KPCA and K-k-means

In the next two subsections we argue, through simple corollaries to Theorem 1 and experimental evaluation, that K-JL preserves geometric and clustering aspects of kernel PCA (KPCA) (Mika et al., 1999; Schölkopf et al., 1998) and kernel k-means (K-k-means) (Dhillon et al., 2004), at the cheaper costs of random projection, whenever favorable conditions for KPCA, resp. K-k-means hold in practice.

Recall that, given a large dataset $\boldsymbol{X}_N$ of size N, K-JL consists of remapping each $X_i \in \boldsymbol{X}_N$ as $\hat{V}^\top S_{\boldsymbol{X}} K(\cdot, X_i)$, where $\boldsymbol{X}$ is the size n subsample of $\boldsymbol{X}_N$ used to compute $\hat{V}^\top \doteq \frac{1}{n\sqrt{nd}} Z\boldsymbol{K}$.

5.1 Preserving Low-dimensional Separation

In this section, we develop the intuition that, Kernel JL has similar advantages as Kernel PCA (KPCA) under situations favorable to KPCA. In particular, KPCA works under the assumption that the data in feature space $\{K(\cdot, x) : x \in \mathcal{X}\}$ lies close to a lower-dimensional eigenspace of the covariance operator Σ (Blanchard et al., 2007; Mika et al., 1999; Schölkopf et al., 1998). We formalize this assumption below.

Assumption 1 (KPCA). *Let $\Sigma = \sum_i \lambda_i (f_i \otimes_{\mathcal{H}} f_i)$ denote a spectral decomposition of Σ (with non-increasing eigenvalues $\{\lambda_i\}$, and assume $\mathbb{E}_{X\sim\rho_X} K(\cdot, X) = 0$. For any $k \in \mathbb{N}$, let P_k denote the projection operator onto span$\{f_i : i \in [k]\}$. There exists $k \in \mathbb{N}$, and $0 < \epsilon, \eta < 1$ such that*

$$\rho_X \left\{x : \|P_k K(\cdot, x)\|_{\mathcal{H}}^2 \geq (1-\epsilon)\|K(\cdot, x)\|_{\mathcal{H}}^2\right\} \geq 1 - \eta.$$

We start with some theoretical intuition using the following formal example.

Example 1 (Well-separated subsets of feature space)**.** *Let Assumption 1 hold for some $0 < \epsilon, \eta < 1$. Let $\tau > 1$. Let $\boldsymbol{X}_N$ denote an i.i.d. sample of size N from ρ_X (not necessarily independent from $\boldsymbol{X}$, since the results of Theorem 1 hold uniformly over $\mathcal{H}$).*

The following holds with probability at least $1 - e^{-\tau} - N\eta$, over any subsets $\mathcal{F}, \mathcal{G}$ of $\{K(\cdot, x) : x \in \boldsymbol{X}_N\}$ satisfying $\min_{f\in\mathcal{F},\, g\in\mathcal{G}} \|f - g\|_{\mathcal{H}}^2 = \Delta$, for some separation $\Delta = \Delta(\mathcal{F}, \mathcal{G}) > 0$. We have for some C_1, C_2, both functions of (K, ρ_X), that for $n \wedge d > C_1$:

$$\inf_{f\in\mathcal{F},\, g\in\mathcal{G}} \left\|\hat{V}^\top S_{\boldsymbol{X}}\,(f - g)\right\|^2 \geq \lambda_k^3 \cdot (\Delta - 2\epsilon\cdot\kappa) - C_2\sqrt{\tfrac{\tau}{n\wedge d}}. \quad (11)$$

On the other hand, independent of Assumption 1, we have with probability at least $1 - e^{-\tau}$ that for all $f, g \in \mathcal{H}$ we have the upper-bound

$$\left\|\hat{V}^\top S_{\boldsymbol{X}}\,(f - g)\right\|^2 \leq \lambda_1^3 \cdot \|f - g\|_{\mathcal{H}}^2 + C_2\sqrt{\frac{\tau}{n\wedge d}}. \quad (12)$$

The above is obtained by noticing that, for any $h \in \mathcal{H}$

$$\begin{aligned}\left\langle h, \Sigma^3 h\right\rangle_{\mathcal{H}} &= \sum_{i=1}^{\infty} \lambda_i^3 \left\langle h, f_i\right\rangle^2 \geq \lambda_k^3 \cdot \sum_{i=1}^{k} \left\langle h, f_i\right\rangle^2 \\ &= \lambda_k^3 \cdot \|P_k h\|_{\mathcal{H}}^2\,, \text{ and similarly} \\ \left\langle h, \Sigma^3 h\right\rangle_{\mathcal{H}} &\leq \lambda_1^3 \cdot \sum_{i=1}^{\infty} \left\langle h, f_i\right\rangle^2 = \lambda_1^3 \cdot \|h\|_{\mathcal{H}}^2\,.\end{aligned}$$

Now, under Assumption 1 and Theorem 1, take $h \doteq f - g$ to obtain the statements of (11) and (12). For (11), notice further that, under Assumption 1, we have with probability at least $1 - N\eta$ that

$$\begin{aligned}\|P_k(f-g)\|_{\mathcal{H}}^2 &= \|f-g\|_{\mathcal{H}}^2 - \|P_k^\perp(f-g)\|_{\mathcal{H}}^2 \\ &\geq \|f-g\|_{\mathcal{H}}^2 - 2\left(\|P_k^\perp f\|_{\mathcal{H}}^2 + \|P_k^\perp g\|_{\mathcal{H}}^2\right) \\ &\geq \|f-g\|_{\mathcal{H}}^2 - 2\epsilon\cdot\kappa. \quad \square\end{aligned}$$

From the above, if the two subsets $\mathcal{F}, \mathcal{G}$ are well-separated in feature space under KPCA, they remain well-separated after K-JL, provided the *condition number* λ_1/λ_k is not too large: distances are rescaled below by λ_k^3, but rescaled above by λ_1^3. In the favorable case where $\lambda_1/\lambda_k \approx 1$, we see from (11) and (12) that K-JL should achieve similar separation properties as KPCA, provided Δ is large w.r.t. to interpoint distances in $\mathcal{F}$ and $\mathcal{G}$. This intuition is formalized in the following example where we consider a scale-free notion of separation.

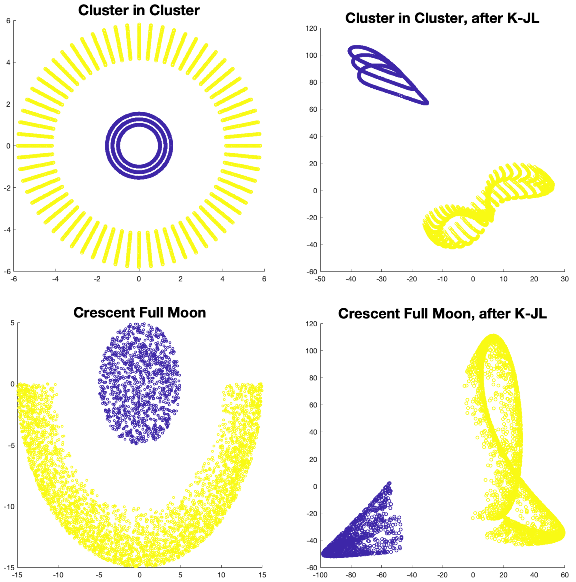

Figure 1: The data *Cluster in Cluster* and *Crescent Full Moon* each have 5000 points, and are shown before and after K-JL projection. K-JL behaves as a random version of KPCA in how it separates clusters.

Simulations. Next, we verify the above insights empirically. In particular, an empirical fact about KPCA, justifying its popularity, is that it can reveal separable subsets $\mathcal{F}, \mathcal{G}$ (in feature space) of data $\boldsymbol{X}_N$ that were not separable in original space $\mathcal{X}$. Per the above insights, this should also be the case with K-JL. In Figure 1 we show projection results, where, given $N = 5000$ points, we use a subsampling size $n = 100$ and projection dimension $d = 2$ to verify the intuition that K-JL (centralized) is able to separate subsets of data on typical examples (e.g., cluster in cluster) where KPCA is known to work well (Mika et al., 1999; Schölkopf et al., 1998).

5.2 Preserving Clustering Properties

In this section we argue that if the data is clusterable in feature space—an assumption underlying K-k-means, and uses of KPCA in clustering—then it remains clusterable after K-JL.

To develop intuition, we formalize *clusterability* in terms of the distribution ρ_X being given as a mixture of distributions with sufficiently separated means. We adapt traditional arguments given in the work on clustering mixtures of Gaussians (Dasgupta, 1999; Kannan et al., 2005; Sanjeev and Kannan, 2001) to the square norm $\left\langle f, \Sigma^3 f\right\rangle_{\mathcal{H}}$. In particular, these works develop the intuition that if the k cluster means are sufficiently separated, they then lie close to a k-dimensional subspace close to the top k-eigenspace of the data covariance. Such intuition holds in general Hilbert space, and in the sequel we illustrate this in the case of 2 clusters, while similar arguments extend to multiple clusters.

Example 2 (Clusterability of ρ_X)**.** *The following holds with probability at least $1 - e^{-\tau}$, $\tau > 1$.*

Let $\rho_X = \pi_1 \rho_{X,1} + \pi_2 \rho_{X,2}$, $0 < \pi_1, \pi_2 < 1$, $\pi_1 + \pi_2 = 1$; let μ_1, μ_2, Σ_1, Σ_2 are respectively the means and covariance operators of $\rho_{X,1}, \rho_{X,2}$, i.e., for $i = 1, 2$, $\mu_i = \mathbb{E}_{\rho_{X,i}} K(\cdot, X)$ and $\Sigma_i = \mathbb{E}_{\rho_{X,i}} K(\cdot, X) \otimes_{\mathcal{H}} K(\cdot, X) - \mu_i \otimes_{\mathcal{H}} \mu_i$.

Suppose the maximum eigenvalues of Σ_1, Σ_2 are upper-bounded by σ. We have for some C_1, C_2, both functions of (K, ρ_X), that for $n \wedge d > C_1$:

$$\left\| \hat{V}^\top S_{\boldsymbol{X}} \left(\mu_1 - \mu_2\right) \right\|^2 \geq \lambda_1^3 \left(\|\mu_1 - \mu_2\|_{\mathcal{H}}^2 - \frac{1}{\pi_1 \pi_2} \sigma \right) - C_2 \sqrt{\frac{\tau}{n \wedge d}}. \quad (13)$$

In other words, separation between cluster means are maintained. On the other hand, as a consequence of (12)*, inter-cluster distances are maintained (at the same scale λ_1^3).*

The above is a consequence of the following decomposition. Let $\gamma = \pi_1 / \pi_2$, so that $\mu_2 = \gamma \mu_1$:

$$\begin{aligned} \Sigma &\doteq \mathbb{E}_{\rho_X} K(\cdot, X) \otimes_{\mathcal{H}} K(\cdot, X) = \pi_1 \left(\Sigma_1 + \mu_1 \otimes_{\mathcal{H}} \mu_1\right) \\ &\qquad + \pi_2 \left(\Sigma_2 + \mu_2 \otimes_{\mathcal{H}} \mu_2\right) \\ &= \pi_1 \Sigma_1 + \pi_2 \Sigma_2 + \left(\pi_1 + \pi_2 \gamma^2\right) \mu_1 \otimes_{\mathcal{H}} \mu_1 \\ &= \pi_1 \Sigma_1 + \pi_2 \Sigma_2 + \gamma \mu_1 \otimes_{\mathcal{H}} \mu_1. \end{aligned}$$

It follows from the above that

$$\begin{aligned} \lambda_1 &\doteq \langle f_1, \Sigma f_1 \rangle_{\mathcal{H}} \leq \sigma + \gamma \left\langle f_1, (\mu_1 \otimes_{\mathcal{H}} \mu_1) f_1 \right\rangle_{\mathcal{H}} \\ &= \sigma + \gamma \left\langle f_1, \mu_1 \right\rangle_{\mathcal{H}}^2, \text{ and} \qquad (14) \\ \lambda_1 &\geq \frac{1}{\|\mu_1\|_{\mathcal{H}}^2} \left\langle \mu_1, \Sigma \mu_1 \right\rangle_{\mathcal{H}} \geq \gamma \frac{1}{\|\mu_1\|_{\mathcal{H}}^2} \left\langle \mu_1, \mu_1 \right\rangle_{\mathcal{H}}^2 \\ &= \gamma \|\mu_1\|_{\mathcal{H}}^2. \qquad (15) \end{aligned}$$

Combining (14) and (15), it follows that $\langle f_1, \mu_1 \rangle_{\mathcal{H}}^2 \geq \|\mu_1\|_{\mathcal{H}}^2 - \sigma/\gamma$. Noticing that $\mu_1 - \mu_2 = (1 + \gamma)\mu_1$, we therefore obtain

$$\begin{aligned} &\left\langle (\mu_1 - \mu_2), \Sigma^3 (\mu_1 - \mu_2) \right\rangle_{\mathcal{H}} = (1 + \gamma)^2 \left\langle \mu_1, \Sigma^3 \mu_1 \right\rangle_{\mathcal{H}} \\ &\geq (1 + \gamma)^2 \lambda_1^3 \left\langle f_1, \mu_1 \right\rangle_{\mathcal{H}}^2 \\ &\geq \lambda_1^3 \left(\|\mu_1 - \mu_2\|_{\mathcal{H}}^2 - \frac{(1 + \gamma)^2}{\gamma} \cdot \sigma \right) \\ &= \lambda_1^3 \left(\|\mu_1 - \mu_2\|_{\mathcal{H}}^2 - \frac{1}{\pi_1 \pi_2} \sigma \right). \end{aligned}$$

Equation (13) is then obtained by combining this last inequality with Theorem 1. □

Table 1: Data description

UCI Datasets	*Size N*	*Dimension*	*Num. of clusters*
Avila Bible	20867 (bible pages)	10	12 (scribes)
IoT	40000 (traffic traces)	115	5 (devices)
Bank Notes	1372 (images)	4	2 (forged or not)

Table 2: Clustering Results: Preprocessing time / Rand Index **(best 2 in bold)**.

UCI Datasets	k-means	K-k-means	KPCA	K-JL
Avila Bible	NA / .683 ±.026	.112s / **.728** ±.003	.103s / **.725** ±.002	.086s / .718 ±.002
IoT	NA / .548 ±.086	.537s / .745 ±.039	.500s / **.759** ±.024	.447s / **.749** ±.006
Bank Notes	NA / .507 ±.0001	.020s / **.529** ±.067	.014s / .526 ±.041	.007s / **.527** ±.031

Experiments. We run clustering experiments on UCI datasets, $\boldsymbol{X}_N$ of sizes N, described in Table 1. We compare K-JL (i.e. k-means after centralized K-JL) against k-means clustering after KPCA (centralized), and K-k-means. For KPCA we use a fast implementation where eigen-decomposition is done on the centralized gram matrix $\bar{K}$ of a subsample of size n to approximate the top d eigenfunctions of the centralized gram-matrix $\bar{K}_N$ on N samples; that is, if $\boldsymbol{\alpha} \in \mathbb{R}^n$ is an eigenvector of $\bar{K}$, then $x \in \boldsymbol{X}_N$ is mapped to $\sum_{i\in[n]} \alpha_i K(x, x_i)$, $x_i \in \boldsymbol{X}$.

Nyström embedding. For K-k-means we use a fast Nyström embedding $\tilde{\mathbf{K}}_{\mathbf{N}}^{\mathbf{1/2}}$, where $\tilde{\mathbf{K}}_{\mathbf{N}}$ approximates the gram matrix $\boldsymbol{K}_N$ on N samples, using a rank d pseudo-inverse $\boldsymbol{K}_{(d)}^{\dagger}$ of the gram-matrix $\boldsymbol{K}$ on n subsamples (Calandriello and Rosasco, 2018; Wang et al., 2019; Williams and Seeger, 2001). That is, we use $\tilde{\boldsymbol{K}}_N = \boldsymbol{K}_{(N,n)}\boldsymbol{K}_{(d)}^{\dagger}\boldsymbol{K}_{(N,n)}^{\top}$, where $\boldsymbol{K}_{(N,n)}$ denotes the gram-matrix between $\boldsymbol{X}_N$ and $\boldsymbol{X}$. In all our implementations, $\boldsymbol{X}$ are n random subsamples of $\boldsymbol{X}_N$. We use a Gaussian kernel $K(x, x') \doteq \exp\{-\|x - x'\|^2/\sigma^2\}$, where σ is chosen as the 25th percentile of interpoint distances.

Relative performance. The results of Table 2 validate our intuition that K-JL achieves similar clustering as K-k-means and KPCA, in faster preprocessing time (for the mapping of $\boldsymbol{X}_N$, as implemented in Matlab without further optimization of matrix multiplications). For all methods, we set $d = 10k$, where k is the number of clusters, and $n = \max\{200, N/100\}$. All experiments are repeated 30 times, and mean and std of Rand Index (RI) are reported. Interestingly, K-JL also appears most stable in terms of RI: the higher instability of the other two methods is likely due to the fast-eigensolvers used in Matlab.

Effects of d and n. In Figure 2, we vary n, d on the IOT dataset, where to reduce running time we now set $N = 10000$ and use 10 repetitions (rather than 30 as above) per values of d and n. Average RI are reported. The main take-home is that the methods are most sensitive to the choice of n. We again observe that Kernel-JL appears overall most stable.

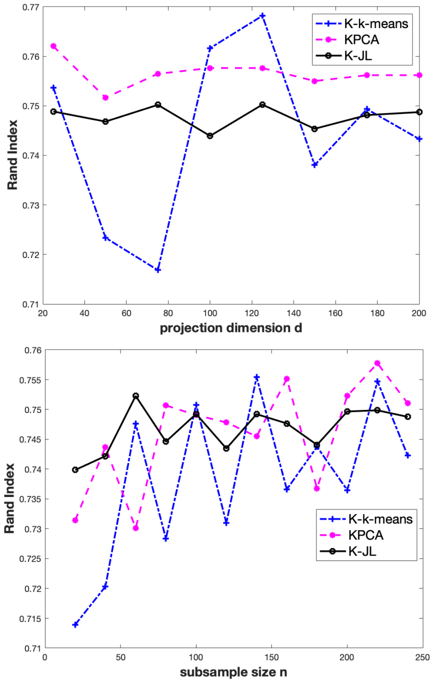

Figure 2: Effects of n and d, using $N = 10000$ samples from the IoT dataset. Left, we fix $n = 200$ and vary d. Right, we fix $d = 10$ and vary n. The choice of subsample size n seems most crucial.

Acknowledgments

Samory Kpotufe and B. K. Sriperumbudur thank the National Science Foundation for support under, respectively, grant id NSF-CPS-1739809, and NSF-DMS-1713011.

References

A. Andoni, J. Chen, R. Krauthgamer, B. Qin, D. P. Woodruff, and Q. Zhang. On sketching quadratic forms. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 311–319. ACM, 2016.

A. Andoni, R. Krauthgamer, and I. Razenshteyn. Sketching and embedding are equivalent for norms. *SIAM Journal on Computing*, 47(3):890–916, 2018.

G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory*, 54(2):781–790, 2008.

G. Blanchard, O. Bousquet, and L. Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2):259–294, 2007.

V. I. Bogachev. *Gaussian Measures*. American Mathematical Society, 1998.

D. Calandriello and L. Rosasco. Statistical and computational trade-offs in kernel k-means. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9357–9367. Curran Associates, Inc., 2018.

S. Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE, 1999.

S. Dasgupta and A. Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

I. Dhillon, Y. Guan, and B. Kulis. Kernel *k*-means, spectral clustering and normalized cuts. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2004.

J. Diestel and J. J. Uhl. *Vector Measures*. American Mathematical Society, Providence, USA, 1977.

W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz maps into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *International Conference on Computational Learning Theory*, pages 444–457. Springer, 2005.

V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23:110–133, 2017.

S. Mika, B. Schölkopf, A. J. Smola, K.-R. Müller, M. Scholz, and G. Rätsch. Kernel PCA and denoising in feature spaces. In *Advances in neural information processing systems*, pages 536–542, 1999.

S. I. Resnick. *A Probability Path*. Birkhäuser Basel, 2014.

A. Sanjeev and R. Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257. ACM, 2001.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172, 2007.

S. Wang, A. Gittens, and M. W. Mahoney. Scalable kernel k-means clustering with Nyström approximation: Relative-error bounds. *Journal of Machine Learning Research*, 20(12):1–49, 2019.

C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press, 2001.

D. P. Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Y. Yang, M. Pilanci, and M. J. Wainwright. Randomized sketches for kernels: Fast and optimal non-parametric regression. *Annals of Statistics*, 45 (3):991–1023, 2017.

V. Yurinsky. *Sums and Gaussian Vectors*. Springer, 1995.