Kernelized Diffusion Maps

Loucas Pillaud-Vivien LPILLAUDVIVIEN@FLATIRONINSTITUTE.ORG
NYU Courant - Flatiron Institute

Francis Bach FRANCIS.BACH@INRIA.FR
INRIA Paris - DI ENS - PSL

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Spectral clustering (Ng et al., 2001) and diffusion maps (Coifman and Lafon, 2006) are celebrated dimensionality reduction algorithms built on eigen-elements related to the diffusive structure of the data. The core of these procedures is the approximation of a Laplacian through a graph kernel approach (Hein et al., 2007), however this local average construction is known to be cursed by the high-dimension d. In this article, we build a different estimator of the Laplacian's eigenvectors, via a reproducing kernel Hilbert space method, which adapts naturally to the regularity of the problem. We provide non-asymptotic statistical rates proving that the kernel estimator we build can circumvent the curse of dimensionality when the problem is well conditioned. Finally we discuss techniques (Nyström subsampling, Fourier features) that enable to reduce the computational cost of the estimator while not degrading its overall performance.

Keywords: spectral clustering, diffusion maps, graph Laplacians, reproducing kernel Hilbert Spaces, dimensionality reduction.

1. Introduction

One of the reasons of the success of learning with reproducing kernel Hilbert spaces (RKHS) is that they naturally select problem-adapted bases of test functions. Even more interestingly, leveraging the underlying regularity of the target function, RKHSs have the ability to circumvent the curse of dimensionality. This is exactly where all techniques resting on local averages fail: approximating a problem will always be cursed by the high-dimension d, because one will need $n^{-1/d}$ points to perform well. This main difference echoes in the nature of the kernels: *pointwise positive kernels* in the non-parametric estimation literature (Nadaraya, 1964) and *positive semi-definite (PSD) kernels* in modern kernel learning (Steinwart and Christmann, 2008; Schölkopf and Smola, 2002).

Solving a problem with PSD kernels that used to be tackled with local techniques is at the heart of this work. Indeed, we estimate the diffusion operator (or Laplacian) related to a measure μ through its principal eigen-elements. When cast into an *unsupervised* learning problem, this can be seen as a dimensionality reduction technique resting on the diffusive nature of the data. This is the core of the celebrated spectral clustering algorithm (Von Luxburg, 2007) and of diffusion maps (Coifman and Lafon, 2006) in the context of molecular dynamics. However, as introduced before, these algorithms are based on graph Laplacians; an intrinsically local construction that scales poorly with the dimension (Hein et al., 2007) and does not benefit from all the recent works on PSD kernels that tackle potential high-dimensional settings (Martinsson and Tropp, 2020; Meanti et al., 2020).

© 2023 L. Pillaud-Vivien & F. Bach.

Let us explain the fundamental difference between the approach of this work and that of graph Laplacians. When we want to estimate the diffusion operator (or its eigenvectors)

$$\mathcal{L} := -\Delta + \langle \nabla V, \nabla \cdot \rangle, \tag{1}$$

one of the difficult aspects is to approximate differential operators. While currently, people use local kernel smoothing techniques, our approach is different. It leverages the reproducing property of derivatives in RKHS and the self-adjointness of $\mathcal{L}$ to circumvent this difficulty: this strategy has shown fruitful results in numerical analysis for partial differential equations, where it is called *meshless methods* (Schaback and Wendland, 2006).

In another direction, it is interesting to note that Salinelli (1998) tried to show that considering the first eigenvectors of $\mathcal{L}$ was *the good way* of generalizing the principal components analysis procedure (Pearson, 1901; Hotelling, 1933) *in a non-linear fashion*. At this time, (i) neither the theory behind diffusions and weighted Sobolev spaces (ii) nor the theory of RKHS were mature. Hence, he clearly explained (i) that the theoretical framework of his analysis was limited but could be extended, and (ii) that at this point solving numerically the problem was impossible in high-dimension as it necessitates to discretize the Laplacian. Quite surprisingly, the literature on graph Laplacians seems to have overlooked Salinelli's seminal contribution. Our work can be considered as a natural continuation of his: pushing further the theoretical comprehension of this *non-linear principal component analysis* with modern tools and giving a way to solve it efficiently.

Note also that this work has been motivated by applications in *molecular dynamics* where diffusion maps is an important dimensional reduction technique to find reaction coordinates, i.e., the slow diffusion modes of the high-dimensional dynamics (Coifman et al., 2006). This article can also be seen as a natural extension of Pillaud-Vivien et al. (2020) whose aim was to estimate the first non-zero eigenvalue of $\mathcal{L}$ (this is saying, its spectral gap). Besides being more mature, the focus of this work is quite different: we focus here on the estimation of the whole spectrum of $\mathcal{L}$ and try to be more precise regarding its convergence properties. Finally remark that the procedure we are going to describe can be seen as a data estimation of the *Koopman generator* of the dynamics generated by $\mathcal{L}$ (Klus et al., 2020), that leverages crucially it self-adjointness property.

2. Diffusion operator

Consider a probability measure $d\mu$ on $\mathbb{R}^d$ which has a density with respect to the Lebesgue measure and can be written under the following form: $d\mu(x) = e^{-V(x)}dx$, where V is called the *potential function*. Consider $H^1(\mu)$ the subspace of functions of $L^2(\mu)$ (i.e., which are square integrable) that also have all their first order derivatives in L^2, that is, $H^1(\mu) = \{f \in L^2(\mu),\ \int_{\mathbb{R}^d} f^2 d\mu + \int_{\mathbb{R}^d} \|\nabla f\|^2 d\mu < \infty\}$, where ∇f is the gradient of f and $\|\cdot\|$ the standard Euclidean norm.

The aim of this work is to estimate the diffusion operator $\mathcal{L}$, associated with measure μ, given access to $x_1, \ldots, x_n$, i.i.d. samples distributed according to μ. It is defined by

$$\mathcal{L}\phi := -\Delta\phi + \nabla V \cdot \nabla\phi, \tag{2}$$

where ϕ is a smooth enough test function.

2.1. Langevin diffusion

Let us consider the overdamped Langevin diffusion in $\mathbb{R}^d$, that is the solution of the following stochastic differential equation:

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \tag{3}$$

where $(B_t)_{t\geqslant 0}$ is a d-dimensional Brownian motion. It is well-known (Bakry et al., 2014) that the law of $(X_t)_{t\geqslant 0}$ converges to the Gibbs measure $d\mu$ and that the Poincaré constant (see Remark 1 below) controls the rate of convergence to equilibrium in $L^2(\mu)$. Let us denote by $P_t(f)$ the Markovian semi-group associated with the Langevin diffusion $(X_t)_{t\geqslant 0}$. It is defined in the following way: $P_t(f)(x) = \mathbb{E}[f(X_t)|X_0 = x]$. This semi-group satisfies the dynamics

$$\frac{d}{dt}P_t(f) = -\mathcal{L}P_t(f),$$

where $\mathcal{L}\phi = -\Delta\phi + \nabla V \cdot \nabla\phi$ is a differential operator called the infinitesimal generator of the Langevin diffusion (3) (Δ denotes the standard Laplacian on $\mathbb{R}^d$). Note that by integration by parts, the semi-group $(P_t)_{t\geqslant 0}$ is reversible with respect to $d\mu$, that is: $\int f(\mathcal{L}g)\,d\mu = \int \nabla f \cdot \nabla g\,d\mu = \int (\mathcal{L}f)g\,d\mu$. This also shows that $\mathcal{L}$ is a symmetric positive definite operator on $H^1(\mu)$.

Remark 1 (Link with Poincaré constant) *Let us call π the orthogonal projector of $L^2(\mu)$ on constant functions: $\pi f : x \in \mathbb{R}^d \mapsto \int f d\mu$ and define $L_0^2(\mu) := \operatorname{Ker}\pi$. Under Assumption 0 (see below) the first non-zero eigenvalue of $\mathcal{L}$ is:*

$$\mathcal{P}^{-1} = \inf_{f\in(H^1(\mu)\cap L_0^2(\mu))\setminus\{0\}} \frac{\langle f, \mathcal{L}f\rangle_{L^2(\mu)}}{\|f\|^2_{L^2(\mu)}}, \tag{4}$$

where $\mathcal{P}$ is also known as the Poincaré constant of the distribution $d\mu$ (Cécile et al., 2000).

2.2. Some useful properties of the diffusion operator

Positive semi-definiteness. The first property that we saw is symmetry and positiveness of $\mathcal{L}$ in $H^1(\mu)$. It comes from the following integration by part identity:

$$\int f(\mathcal{L}g)\,d\mu = \int \nabla f \cdot \nabla g\,d\mu = \int (\mathcal{L}f)g\,d\mu, \tag{5}$$

showing that the quadratic form induced by $\mathcal{L}$ is also the Dirichlet energy

$$\langle \mathcal{L}f, f\rangle_{L^2(\mu)} = \int \|\nabla f\|^2 d\mu =: \mathcal{E}(f).$$

Link with Schrödinger operator. In the field of partial differential equations (PDEs) we say that an operator is of Schrödinger type if it is the sum of the Laplacian and a multiplicative operator, this comes from the fact that this is the type of operator that governs the dynamics of quantum systems (Helffer and Nier, 2005). Here, let us define the Schrödinger operator $\widetilde{\mathcal{L}} := -\Delta + \mathcal{V}$, where $\mathcal{V} := \frac{1}{2}\Delta V - \frac{1}{4}\|\nabla V\|^2$. We can show that $\widetilde{\mathcal{L}}$ and $\mathcal{L}$ are conjugate to each other: indeed, a rapid calculation shows that

$$\widetilde{\mathcal{L}} = e^{-V/2}\mathcal{L}\left[e^{V/2}\,\cdot\right].$$

As Schrödinger operators are well-studied, we can infer from this fact interesting properties on the spectrum of $\mathcal{L}$. Indeed,

$$(\lambda, u) \text{ eigen-elements of } \widetilde{\mathcal{L}} \ \Leftrightarrow\ (\lambda, e^{V/2}u) \text{ eigen-elements of } \mathcal{L},$$

and we also have the following equality for f smooth enough:

$$\frac{\langle \mathcal{L}f, f\rangle_{L^2(\mu)}}{\|f\|^2_{L^2(\mu)}} = \frac{\langle \widetilde{\mathcal{L}}f, f\rangle_{L^2(\mathbb{R}^d)}}{\|f\|^2_{L^2(\mathbb{R}^d)}}.$$

Spectrum of $\mathcal{L}$. The most important property that we can infer from this is the nature of the spectrum of $\mathcal{L}$. Indeed, it is well known (Reed and Simon, 2012) that if $\mathcal{V}$ is locally integrable, bounded from below and coercive ($\mathcal{V}(x) \longrightarrow +\infty$, when $\|x\| \to +\infty$), then the Schrödinger operator has a compact resolvent. In particular, we will assume throughout the article the following

Assumption 0 (Spectrum of $\mathcal{L}$) *Assume that* $\frac{1}{2}\Delta V(x) - \frac{1}{4}\|\nabla V\|^2 \longrightarrow +\infty$, *when* $\|x\| \to +\infty$.

Assumption 0 implies that $\mathcal{L}$ has a compact resolvent. This also implies that $\mathcal{L}$ has a purely discrete spectrum and a complete set of eigenfunctions. Note that this assumption implies also a spectral gap for the diffusion operator $\mathcal{L}$ and hence that a Poincaré inequality holds. Throughout this work and even if not clearly stated, we will assume Assumption 0. For further discussions on the spectrum of $\mathcal{L}$, we refer to Bakry et al. (2014); Helffer and Nier (2005).

3. Approximation of the diffusion operator in the RKHS

Let $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ be an RKHS with positive definite kernel K. Let us suppose the following:

Assumption 1 (Universality) *$\mathcal{H}$ is dense in $H^1(\mu)$.*

Note that this is the case for most of the usual couples kernels/distribution: Gaussian and exponential kernels are universal if μ has compact support or subgaussian tails. Note that universality as depicted in Micchelli et al. (2006) is related to denseness with respect to the supremum norm (which ensures denseness in $L^2(\mu)$). Hence let us precise that for denseness in $H^1(\mu)$, one should also check universality of the RKHS built upon derivatives, that is, denseness in the $\mathcal{C}^1$ sense. As the expression of the diffusion operator in Eq. (2) involves derivatives of test functions, we will also need some regularity properties of the RKHS. Indeed, to represent ∇f in our RKHS we leverage crucially the partial derivative reproducing property of the kernel space. For this, we need:

Assumption 2 (Smoothness) *K is a positive definite kernel such that $K \in \mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}^d)$.*

For $i \in [\![1, d]\!]$, denote by $\partial_i = \partial_{x^i}$ the partial derivative operator with respect to the i-th component of x. It has been shown (Zhou, 2008) that under Assumption 2, we can define $\mathcal{H} \ni \partial_i K_x : y \to \partial_{x^i} K(x, y)$ and that a partial derivative reproducing property holds true: $\forall f \in \mathcal{H}$ and $\forall x \in \mathbb{R}^d$, $\partial_i f(x) = \langle \partial_i K_x, f\rangle_{\mathcal{H}}$. Hence, thanks to Assumption 2, ∇f is easily represented in the RKHS. We also need some boundedness properties of the kernel.

Assumption 3 (boundedness) *K is a kernel such that $\forall x \in \mathbb{R}^d$, $K(x,x) \leqslant \mathcal{K}$ and*[1] $\|\nabla K_x\|^2 \leqslant \mathcal{K}_d$, *where* $\|\nabla K_x\|^2 := \sum_{i=1}^d \langle \partial_i K_x, \partial_i K_x\rangle = \sum_{i=1}^d \frac{\partial^2 K}{\partial x^i \partial y^i}(x, x)$ *(see calculations below), x and y standing respectively for the first and the second variables of $(x, y) \mapsto K(x, y)$.*

1. The subscript d in $\mathcal{K}_d$ accounts for the fact that this quantity is expected to scale linearly with d (Gaussian kernel).

The equality in the expression of $\|\nabla K_x\|^2$ arises from the following computation: for all $x, y \in \mathbb{R}^d$, $\langle \partial_i K_x, \partial_i K_y \rangle = \partial_{x^i} (\partial_i K_y(x)) = \partial_{x^i} \partial_{y^i} K(x, y)$. Note that, for example, the Gaussian and exponential kernels satisfy Assumptions 1, 2, 3. Boundedness is stated here for the sake of clarity, however, up to logarithmic terms, the results of this paper would hold if we let $\|K_X\|^2, \|\nabla K_X\|^2$ be subgaussian random variables.

Example 1 (Gaussian kernel) *A prototypical example is the Gaussian kernel (or radial basis function), with bandwidth $\sigma > 0$, for which we can compute, for $i \neq j$,*

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right), \qquad \partial_{x^i}\partial_{y^j} K(x,y) = -\frac{(x_i - y_i)(x_j - y_j)}{\sigma^4} K(x,y),$$
$$\partial_{x^i} K(x,y) = -\frac{(x_i - y_i)}{\sigma^2} K(x,y), \qquad \partial_{x^i}\partial_{y^i} K(x,y) = \left(\frac{1}{\sigma^2} - \frac{(x_i - y_i)^2}{\sigma^4}\right) K(x,y).$$

3.1. Embedding the diffusion operator in the RKHS

Let us define the following operators from $\mathcal{H}$ to $\mathcal{H}$:

$$\Sigma = \mathbb{E}_\mu \left[K_X \otimes K_X\right], \qquad \mathsf{L} = \mathbb{E}_\mu \left[\nabla K_X \otimes_d \nabla K_X\right], \tag{6}$$

where $\otimes$ is the standard tensor product: $\forall f, g, h \in \mathcal{H}$, $(f \otimes g)(h) = \langle g, h \rangle_{\mathcal{H}} f$ and $\otimes_d$ is defined as follows: $\forall f, g \in \mathcal{H}^d$ and $h \in \mathcal{H}$, $(f \otimes_d g)(h) = \sum_{i=1}^d \langle g_i, h \rangle_{\mathcal{H}} f_i$. By the reproducing property of K, $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ injects canonically in $(L^2(\mu), \langle \cdot, \cdot \rangle_{L^2(\mu)})$ through an operator S, together with its adjoint S^* defined from $L^2(\mu)$ to $\mathcal{H}$ such that for all $x \in \mathbb{R}^d$:

$$\forall f \in \mathcal{H}, \ \ \mathsf{S}f(x) = \langle f, K_x \rangle_{\mathcal{H}} = f(x), \qquad \forall f \in L^2(\mu), \ \ \mathsf{S}^* f(x) = \mathbb{E}_\mu \left[K(x, X) f(X)\right].$$

Note that $\mathsf{S}^*\mathsf{S} = \Sigma$. With these definitions, and thanks to the symmetry property of $\mathcal{L}$ derived in Eq. (5), we can represent the diffusion operator $\mathcal{L}$ in the RKHS.

Proposition 2 (Embedding of $\mathcal{L}$ in the RKHS) *Suppose Assumptions 1,2 hold, then*

$$\mathsf{L} = S^* \mathcal{L} S, \tag{7}$$

where the equality stands for the equality between operators of $\mathcal{H}$.

Proof For $z \in \mathbb{R}^d$, $f \in \mathcal{H}$,

$$\begin{aligned}
\langle \mathsf{L} f, K_z \rangle_{\mathcal{H}} &= \int \nabla f(x) \cdot \nabla_x K(x,z) d\mu(x) \\
&= -\int \Delta f(x) K(x,z) d\mu(x) + \int \nabla f(x) \cdot \nabla V(x) K(x,z) d\mu(x) \\
&= \langle \mathcal{L}\mathsf{S} f, \mathsf{S} K_z \rangle_{L^2(\mu)} \\
&= \langle \mathsf{S}^* \mathcal{L} \mathsf{S} f, K_z \rangle_{\mathcal{H}},
\end{aligned}$$

hence the equality between operators. Note that to go from the first line to the second ones, we used the symmetry of $\mathcal{L}$. ∎

We want to construct an approximation of the eigen-elements of $\mathcal{L}$ with domain $H_0^1(\mu) := H^1(\mu) \cap L_0^2(\mu)$, where we recall that $L_0^2(\mu) = \mathrm{Ker}\,\pi$ stands for the space of square integrable functions without the constants. Similarly, let us denote $\mathcal{H}_0 = (\mathrm{Ker}\,\mathsf{L})^{\perp}$, the subspace of $\mathcal{H}$ without the constant functions. Note that this operator is invertible as a consequence of the spectral gap Assumption 0. In the following we will approximate the eigen-elements of $\mathcal{L}^{-1}$. First we give a representation of $\mathcal{L}^{-1}$ in the RKHS $\mathcal{H}$, then we construct an operator on $\mathcal{H}$ that has the same eigen-elements of $\mathcal{L}^{-1}$. Indeed, if we denote L^{-1} the inverse of L restricted on $(\mathrm{Ker}\mathsf{L})^{\perp}$, we have:

Proposition 3 (Representation of $\mathcal{L}^{-1}$) *Suppose Assumptions 0,1,2 hold, then*

$$\mathcal{L}^{-1} = \mathsf{S}\mathsf{L}^{-1}\mathsf{S}^*, \tag{8}$$

where the equality stands for the equality between operators whose domains are $H_0^1(\mu)$.

Thanks to Proposition 3, we have a representation of $\mathcal{L}^{-1}$ in the RKHS *through* the embedding S. But what we really would like is an operator on $\mathcal{H}$ that as the same eigen-elements as $\mathcal{L}^{-1}$. Such a representation allows for numerical computations: this is the purpose of the following proposition.

Theorem 4 (eigen-elements of $\mathcal{L}^{-1}$ as functions in the RKHS) *Decompose the inverse of the diffusion operator such that $\mathcal{L}^{-1} = \mathsf{S}\mathsf{L}^{-1}\mathsf{S}^* = \mathsf{S}\mathsf{L}^{-1/2}\mathsf{L}^{-1/2}\mathsf{S}^*$, then,*

(i) $\mathsf{S}\mathsf{L}^{-1/2}$ *is a bounded operator from $\mathcal{H}_0$ to $H_0^1(\mu)$.*

(ii) $\mathsf{L}^{-1/2}\Sigma\mathsf{L}^{-1/2}$ *is a self-adjoint compact operator on $\mathcal{H}_0$ with the same spectrum as $\mathcal{L}^{-1}$.*

(iii) If $\lambda \neq 0$ is an eigenvalue of $\mathsf{L}^{-1/2}\Sigma\mathsf{L}^{-1/2}$ *with eigenvector $u \in \mathcal{H}_0$, then λ is an eigenvalue of $\mathcal{L}^{-1}$ with eigenvector* $\mathsf{S}\mathsf{L}^{-1/2}u \in H_0^1(\mu)$.

This theorem will allow us to approximate the eigen-elements of $\mathcal{L}^{-1}$ with the ones of the operator $\mathsf{L}^{-1/2}\Sigma\mathsf{L}^{-1/2}$ (that is well-defined only on $\mathcal{H}_0$) with a finite set of samples. Its proof is the consequence of the representation of $\mathcal{L}^{-1}$ presented in the previous proposition and a technical lemma on Hilbert operators proven in Appendix (Lemma 10).

3.2. Definition of the estimator

Empirical operators. We define the empirical counterpart of L and Σ: they are defined by replacing expectation with respect to μ by expectations with respect to its empirical measure $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ where $x_1, \ldots, x_n$ are i.i.d. samples distributed according to $d\mu$.

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n K_{x_i} \otimes K_{x_i}, \quad \text{and} \quad \widehat{\mathsf{L}} = \frac{1}{n}\sum_{i=1}^n \nabla K_{x_i} \otimes_d \nabla K_{x_i}. \tag{9}$$

Hence, one could be tempted to define our estimator as $\widehat{\mathsf{L}}^{-1/2}\widehat{\Sigma}\widehat{\mathsf{L}}^{-1/2}$. However, this definition carries two main problems:

(i) If $f \in \mathrm{Ker}\,\widehat{\mathsf{L}}$, i.e., for all $i \leqslant n$, $\nabla f(X_i) = 0$, then $\|\widehat{\mathsf{L}}^{-1/2}\widehat{\Sigma}\widehat{\mathsf{L}}^{-1/2} f\| = +\infty$. This is an *overfitting-type issue*.

(ii) Another problem is related to the fact that finding the eigen-elements of $\mathsf{L}^{-1/2}\Sigma\mathsf{L}^{-1/2}$ is equivalent to solving the generalized eigenvalue problem: $\Sigma f = \sigma \mathsf{L} f$. Such systems are known to be numerically unstable as mentioned by Crawford (1976). This would be especially the case when replacing the operators by their empirical counterpart. This is a *stability issue*.

Regularization. These two concerns recall the pitfall of overfitting for regression tasks. Hence, as for kernel ridge regression, a natural idea is to regularize with some parameter λ. This leads to the following definition of our estimator and its empirical counterpart:

Definition 5 (Definition of the estimator) *Under Assumptions 0,1,2,3, we define the two estimators of the inverse diffusion operator $\mathcal{L}^{-1}$:*

$$\textit{Biased estimator:} \qquad (\mathsf{L} + \lambda I)^{-1/2} \Sigma (\mathsf{L} + \lambda I)^{-1/2} \tag{10}$$

$$\textit{Empirical estimator:} \qquad (\widehat{\mathsf{L}} + \lambda I)^{-1/2} \widehat{\Sigma} (\widehat{\mathsf{L}} + \lambda I)^{-1/2}. \tag{11}$$

In the following, to shorten notations, let us define $\mathsf{L}_\lambda = \mathsf{L} + \lambda I$ and $\widehat{\mathsf{L}}_\lambda = \widehat{\mathsf{L}} + \lambda I$. Obviously, the main drawback of this regularization is that it induces a bias in our estimation: more precisely the acute reader will recognize that the bigger the λ the closer the problem is to kernel-PCA (Mika et al., 1999). In other words, the scale of λ controls the magnitude of the diffusive information we want to retrieve from the data (this point of view can be further studied but we leave this for future work at this point).

When analyzing the performances of our empirical estimator, we will draw a particular attention to the comparison with the standard algorithm that computes the eigen-elements of the operator: diffusion maps (Coifman and Lafon, 2006; Hein et al., 2007). We emphasize that the RKHS method we present allows to benefit from the numerous positive aspects of RKHS methods (Schölkopf and Smola, 2002): both *on the statistical side* regarding the dependency on the dimension, the adaptivity to the regularity of the target (Caponnetto and De Vito, 2007), and on *on the computational side* benefiting from the techniques developed in the literature like column subsampling or the use of random features (Martinsson and Tropp, 2020).

3.3. What quantities are we interested in approximating?

Requirements of the problem. The natural and general goal of the present work is to give an approximation of the diffusion operator based on i.i.d. samples. However, there are in fact more precise practical objects that the reader may want to have an approximation of:

- **The whole operator.** Either its representation in $\mathcal{H}$ either in $H^1(\mu)$. This can lead, as recalled in Subsection 2.2, to the estimation of Schrödinger operators. This can also be used to regularize a semi-supervised problem with the Dirichlet energy of the unlabeled data to leverage its structure (Cabannes et al., 2021, 2022).
- **The semigroup.** In fact, as $\mathcal{L}$ is the infinitesimal generator of the dynamics, we can be interested in the convergence to the associated semigroups $e^{t\mathcal{L}}$ (Klus et al., 2020).
- **Eigenvectors.** As one of the main applications of this estimator could be the computation of a low-dimensional embedding of the data through the eigenvectors of $\mathcal{L}$, we are directly interested in the approximation of the eigenvectors. Either eigenvector per eigenvector, either finite dimensional subspaces spanned by few of them. Note that we are mostly interested in the small eigenvalues of $\mathcal{L}$, corresponding to the large eigenvalues of $\mathcal{L}^{-1}$, because they are those governing the behaviour of the dynamics (Lelièvre, 2013).
- **Eigenvalues.** As it has already been done in previous work for the top eigenvalue (Pillaud-Vivien et al., 2020), one would like to approximate a set of eigenvalues. Another application is the construction of the diffusion distance used for clustering (Coifman and Lafon, 2006).

Previous results. In previous works, e.g., Hein et al. (2007) and Coifman and Lafon (2006) proved the convergence of the estimated operator. However, note that the convergence theorems are given *pointwise*, for *bounded domains* and have a *bad dependency in the dimension* as $n^{-1/d}$. We will try to overpass these three limiting results. Please note that the operator norm convergence to the diffusion operator *implies the convergence of all the quantities mentioned earlier:* **(i) semigroup** at finite time, thanks to the inequality: $\|e^B - e^A\| \leq \|B - A\| e^{\max\{\|A\|,\|B\|\}}$, **(ii) eigenvectors and eigenvalues,** directly by perturbation theory arguments. Importantly, refined bounds are discussed if one want to approximate k-dimensional subspaces, similarly to Zwald and Blanchard (2005).

4. Statistical analysis of the estimator

As said earlier, to shorten the notations, let us define for an operator A, the operator $A_\lambda := A + \lambda I$. We will split the problem in two: a bias term and a variance term

$$\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\widehat{\Sigma}\widehat{\mathsf{L}}_\lambda^{-1/2} - \mathsf{L}^{-1/2}\Sigma\mathsf{L}^{-1/2}\right\| \leqslant \underbrace{\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\widehat{\Sigma}\widehat{\mathsf{L}}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2}\right\|}_{\text{variance}} + \underbrace{\left\|\mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2} - \mathsf{L}^{-1/2}\Sigma\mathsf{L}^{-1/2}\right\|}_{\text{bias}}$$

The variance term corresponds to the statistical error coming from the fact that we have only access to a finite set of n samples of the distribution μ. The bias comes from the introduction of a regularization of the operator L scaled by λ. We first derive bounds for the variance term.

4.1. Variance analysis

Proposition 6 (Analysis of the statistical error) *Suppose Assumptions 0,1,2,3, hold true. For any* $\delta \in (0, 1/3)$, $0 < \lambda \leqslant \|\mathsf{L}\|$ *and any integer* $n \geqslant 15\frac{\mathcal{K}_d}{\lambda}\log\frac{4\,\mathrm{Tr}\mathsf{L}}{\lambda\delta}$, *with probability at least* $1 - 2\delta$,

$$\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\widehat{\Sigma}\widehat{\mathsf{L}}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2}\right\| \leqslant \frac{8\mathcal{K}}{\lambda\sqrt{n}}\log(2/\delta) + \mathrm{o}\left(\frac{1}{\lambda\sqrt{n}}\right). \tag{12}$$

Note that the analysis behind the proof of Proposition 6 is not completely new: in Pillaud-Vivien et al. (2020), the convergence of the largest eigenvalue was studied using similar tools, the main difference being that in Eq. (12), the bound is in operator norm. Note also that for the sake of clarity, we only emphasized the inequality in the regime where $\lambda\sqrt{n}$ is large but an explicit non-asymptotic bound is given in Lemmas 14 of the Appendix. Finally we emphasize that: (i) the bound is dimension-free, (ii) the bound is **in operator norm** which is a *strong* bound for the operator convergence as it implies many others: eigenvalue and eigenvector convergence by perturbation theory results, bound on the associated semi-group, pointwise convergence or other forms of weak convergence for operators in infinite dimension; (iii) we could have derived refined bounds based on additional assumption on the *capacity* of the RKHS (i.e. through the eigenvalue decay of L), but, for the sake of clarity, we decided to present less general but more interpretable bounds.

4.2. Bias analysis

The bias analysis is harder although all objects are now deterministic. We know that $\mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2}$ is a compact operator (Σ is compact and $\mathsf{L}_\lambda^{-1/2}$ bounded) so that its spectrum is discrete and is formed by isolated points except from 0. On the same manner (Reed and Simon, 2012, Theorem XIII.67) the inverse of the diffusion operator $\mathcal{L}^{-1}$ is compact so that we can talk of the approximation of the k-th eigen-element of $\mathcal{L}^{-1}$ by the one of $\mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2}$ as λ goes to 0 (or eigenspaces if the eigenvalues are not isolated).

Consistency of the estimator. First, if we are only interested in consistency of the estimator and not on rates of convergence we have the following consistency result:

Proposition 7 (Convergence of the bias) *Under Assumptions 0,1,2,3, we have the following convergence in operator norm:*

$$\left\|\mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2} - \mathsf{L}^{-1/2}\Sigma\mathsf{L}^{-1/2}\right\| \underset{\lambda\to 0}{\longrightarrow} 0. \tag{13}$$

This results crucially relies on the fact that the operator $\mathsf{L}^{-1/2}\Sigma\mathsf{L}^{-1/2}$ is compact (shown in Theorem 4), combined with some algebraic manipulations.

Fast rates for smooth eigenfunctions. Without more *a priori knowledge* on the distribution μ (and the RKHS), it is hard to derive universal rates of convergence of the bias with respect to the regularization parameter λ. In fact, even deriving quantitative perturbation results solely on the *first eigenvalue*, $|\|\mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2}\| - \|\mathsf{L}^{-1/2}\Sigma\mathsf{L}^{-1/2}\||$, which corresponds to the Poincaré constant of the distribution, is known to be a difficult problem (Cécile et al., 2000). This is out of the scope of this paper. However, eigenfunctions of these elliptic operators are known to be smooth under standard assumptions on the distribution (typically, smoothness of the Gibbs potential and fast decay of μ tails (Bogachev et al., 2022)). Hence, similarly to what is done in non-parametric regression, we can exploit this and quantify the difficulty of the problem (Caponnetto and De Vito, 2007) by understanding how smooth (w.r.t. the RKHS) the target function is. This is what is often referred to as a *source condition* in this literature (Dieuleuveut, 2017). Here, for most of the applications (Coifman and Lafon, 2006), we want to approximate the p-eigen-elements corresponding to the largest eigenvalues of $\mathcal{L}^{-1}$ (smallest eigenvalues of $\mathcal{L}$) for some $p \in \mathbb{N}^*$. Let us make here a natural assumption on their smoothness.

Assumption 4 (Regularity of the problem) *The p first eigenvectors of $\mathcal{L}$ belongs to $\mathcal{H}$.*

An prototypical example of when it happens for any $p \in \mathbb{N}^*$ is if we consider a distribution with compact support Ω, density $\mu = e^{-V} \in \mathcal{C}^\infty(\Omega)$, and the Gaussian kernel. Note that for the eigenvectors to belong to the RKHS, they need to be at least $d/2$-differentiable: hence this assumption reveals an *a priori* hypothesis on the target. However, eigenvectors of $\mathcal{L}$ are known to be infinitely smooth by elliptic theory of PDEs (Evans, 2022) so that this a priori relatively natural.

Let us denote $\Pi^p : \mathcal{H} \to \mathcal{H}$ the spectral projector onto the span of the p largest eigenvectors of $\mathsf{L}^{-1/2}\Sigma\mathsf{L}^{-1/2}$. Technically speaking, Assumption 4 implies that for all $v \in \operatorname{span}\Pi^p$,

$$\|\mathsf{L}^{-1/2}v\|_{\mathcal{H}} < +\infty,$$

or equivalently in terms of operators: $\|\Pi^p\mathsf{L}^{-1/2}\Pi^p\| < \infty$. Indeed, thanks to Theorem 4-(iii), for such a $v \in \mathcal{H}$, $S\mathsf{L}^{-1/2}v \in \mathcal{H}$, so that, by isometry this means that $\|(SS^*)^{-1/2}S\mathsf{L}^{-1/2}v\|_{L^2} < +\infty$, which is equivalent to the conditions above.

Proposition 8 (Fast rates under source condition) *Under Assumptions 0,1,2,3,4, we have the following bound in operator norm:*

$$\left\|\Pi^p\left(\mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2} - \mathsf{L}^{-1/2}\Sigma\mathsf{L}^{-1/2}\right)\Pi^p\right\| \leq 2\lambda\mathcal{P}\|\Pi^p\mathsf{L}^{-1/2}\Pi^p\|. \tag{14}$$

This proposition means that, under the source condition, the bias in the first p eigenvectors depends linearly on λ. Furthermore, this is remarkable that the two crucial regularity assumptions appear in this bound: (i) the measure of complexity of the measure μ, through its Poincaré constant $\mathcal{P}$, (ii) the smooth *a priori* on the target eigenvectors we want to approximate. Here we decided to showcase, for the sake of clarity, the case where the target belongs to the RKHS, but remark that refined bounds could be easily adapted from this results under more precise (and technical) source assumptions.

4.3. Consistency and convergence rates under source assumption

To summarize the results and the discussion of the two previous sections, let us state here the overall consistency of the estimator as well a final bound on the empirical estimator with respect to the data.

Theorem 9 (Consistency and convergence rates) *Under Assumptions 0,1,2,3, for any $\delta \in (0, 1/2)$ and any integer $n \geqslant \mathsf{K} \log \frac{1}{\delta}$, with K depending on $\mathcal{K}, \mathcal{K}_d$ with probability at least $1 - 2\delta$, take a sequence of regularizers such that $\lambda_n \to 0$ and $\lambda_n \sqrt{n} \to +\infty$, then our estimator is consistent*

$$\widehat{\mathsf{L}}_{\lambda_n}^{-1/2} \widehat{\Sigma} \widehat{\mathsf{L}}_{\lambda_n}^{-1/2} \underset{n\to\infty}{\longrightarrow} \mathsf{L}^{-1/2} \Sigma \mathsf{L}^{-1/2}, \tag{15}$$

where the convergence holds in operator norm. Furthermore, assume 4, then, if $\lambda = \mathcal{K} n^{-1/4}$,

$$\left\| \Pi^p \left(\widehat{\mathsf{L}}_{\lambda_n}^{-1/2} \widehat{\Sigma} \widehat{\mathsf{L}}_{\lambda_n}^{-1/2} - \mathsf{L}^{-1/2} \Sigma \mathsf{L}^{-1/2} \right) \Pi^p \right\| \leq \frac{8 + 2\mathcal{K}\, \mathcal{P} \|\Pi^p \mathsf{L}^{-1/2} \Pi^p\|}{n^{1/4}} + \mathrm{o}\left(n^{-1/4} \right), \tag{16}$$

where, $\forall p \in \mathbb{N}^*$, Π^p is the spectral projector over the largest p eigenvectors of $\mathsf{L}^{-1/2} \Sigma \mathsf{L}^{-1/2}$.

The theorem quantifies the statistical performance of the built estimator: we emphasize that, under the smoothness Assumption 4, the rate of convergence to any estimated eigenfunction of $\mathcal{L}$ *does not depend on the dimension*. This contrasts with the $n^{-1/d}$ rates of graph Laplacian/diffusion maps. This difference echoes the more general and intrinsic difference between local approximation techniques and kernel methods, which adapt to the underlying regularity of the problem.

5. Numerical construction of the estimator

Beyond their statistical performance, kernel methods also enjoy good numerical strategies to reduce their computational cost while keeping their overall precision (Martinsson and Tropp, 2020, Section 19). We discuss informally how to apply them in our context.

We note that if the reader wants to apply concretely the developed method, and go deeper on this Laplacian's eigenvector representation, Vivien Cabannes has developed in Cabannes (2023) a very nice perspective around this *spectral embedding*, as well as a package available on its GitHub page: `https://github.com/VivienCabannes/`, in the kernel laplacian repository ("klap" package available via `$ pip install klap`).

Computing the estimator: naive approach. To compute the estimator $\widehat{\mathsf{L}}_{\lambda}^{-1/2} \widehat{\Sigma} \widehat{\mathsf{L}}_{\lambda}^{-1/2}$, one needs to be able to represent the operators

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} K_{x_i} \otimes K_{x_i}, \qquad \widehat{\mathsf{L}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \partial_j K_{x_i} \otimes \partial_j K_{x_i},$$

whose expressions are recalled here for the sake of clarity. In fact, it suffices to represent them on $\mathrm{Span}\{K_{x_i}\}_{i\leq n} + \mathrm{Span}\{\partial_j K_{x_i}\}_{i\leq n,\, j\leq d}$. Once such matrices $(\Sigma, L) \in \mathbb{R}^{(n+nd)\times(n+nd)s}$ are built, an efficient way to compute the operator is by solving the **generalized eigenvalue problem**: i.e., find all $(\psi_k, \mu_k) \in \mathbb{R}^{n+nd} \times \mathbb{R}$, for $k \in [\![1, n+nd]\!]$ such that

$$\Sigma\psi_k = \mu_k L_\lambda \psi_k, \tag{17}$$

then we can write thanks the eigenvalue decomposition $\widehat{\mathsf{L}}_\lambda^{-1/2}\widehat{\Sigma}\widehat{\mathsf{L}}_\lambda^{-1/2} = \sum_{k=1}^{n+nd} \mu_k f_k \otimes f_k$, with

$$f_k = \sum_{i=1}^{n} \psi_k[i] K_{x_i} + \sum_{i=1}^{n}\sum_{j=1}^{d} \psi_k[n+ij]\partial_j K_{x_i}.$$

Obviously, one of the bottleneck is to build the large matrices Σ and L that have approximately n^2d^2 coefficients, and then solve their related generalized eigenvalue problem. Hence, this procedure becomes intractable if n or d is too large. Fortunately, there is a pass forward: to implement the kernel method we need only an approximation of the kernels matrices. We propose below two well-developed method used to reduce the computations.

Nyström approximation/column subsampling. The idea of this method is to build low-rank approximations of Σ, L by selecting only $p \in \mathbb{N}^*$ columns among them. Note that, in favorable cases, p can be chosen as low as $\log(n)$ without hurting the statistical performances (Rudi et al., 2015). Let us choose only the p columns that refer to the elements $(K_{x_i})_{i\leq p}$. The algorithm below, presented for other purposes in Cabannes et al. (2021), returns the eigenvectors we want to approximate:

Algorithm 1 Compute the eigenvectors by Nyström method

Data: $(x_i)_{i\leq n}$, a kernel k, and a regularizer λ
Compute $S_p = (k(x_i, x_l))_{i\leq n, l\leq p} \in \mathbb{R}^{n\times p}$
Compute $D_p = (\partial_{1,j} k(x_i, x_l))_{(i\leq n, j\leq d), l\leq p} \in \mathbb{R}^{nd\times p}$
Build $\Sigma_p = S_p^\top S_p \in \mathbb{R}^{p\times p}$ and $\mathsf{L}_p = D_p^\top D_p \in \mathbb{R}^{p\times p}$
Get $(\psi_k, \mu_k)_{k\leq p}$ the generalized eigen-elements of $(\Sigma_p, \mathsf{L}_p + \lambda I_p)$

At the end of Algorithm 1, for $k \leq p$ the k-th approximated eigenvalue is μ_k, and its associated eigenfunction writes $f_k(x) = \sum_{i=1}^p \Psi_k[i] K(x_i, x)$, where $\Psi_k = (\mathsf{L}_p + \lambda I_p)^{-1/2}\psi_k$. In terms of numerical complexity, the main costs are due to building L_p in $\mathcal{O}(p^2 nd)$ and finding the generalized eigen-elements in $\mathcal{O}(p^3)$. In the course of the algorithm we used the notation $\partial_{1,j} k(x_i, x_l)$ to stress that the derivative should apply to the first variable.

Random features. Random features (Rahimi and Recht, 2008) is another way to circumvent the problem by building explicitly features that approximate any translation invariant kernel $k(x, y) = k(x-y)$. More precisely, let $p \in \mathbb{N}^*$ be the number of random features, $(w_l)_{l\leq p}$ be random variables of $\mathbb{R}^d$ independently and identically distributed according to $\mathbb{P}(dw) = \int_{\mathbb{R}^d} \mathrm{e}^{-\mathrm{i}w^\top\delta} K(\delta) d\delta\, dw$ and $(b_l)_{l\leq p}$ be independently and identically distributed according to the uniform law on $[0, 2\pi]$, then the feature vector $\phi_p(x) = \sqrt{\frac{2}{p}}\left(\cos(w_1^\top x + b_1), \ldots, \cos(w_M^\top x + b_M)\right)^\top \in \mathbb{R}^M$ satisfies $K(x, x') \approx \langle \phi_p(x), \phi_p(x')\rangle_2$. Therefore, random features allow to approximate Σ and L by $p \times p$ matrices.

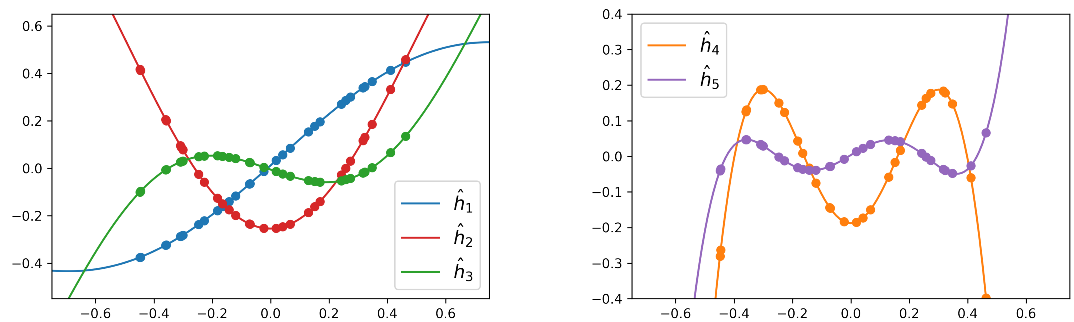

Figure 1: Estimation of the first five eigenfunctions of the Ornstein Uhlenbeck operator: the Hermite polynomials.

Algorithm 2 Compute the eigenvectors with random features

Data: $(x_i)_{i\leq n}$ and a regularizer λ
Compute $S_p = (\cos(w_l^\top x_i + b_l))_{i\leq n, l\leq p} \in \mathbb{R}^{n\times p}$
Compute $D_p^j = (-w_l[j]\sin(w_l^\top x_i + b_l))_{i\leq n, l\leq p} \in \mathbb{R}^{n\times p}$, for all $j \leq d$
Build $D_p = \sum_{j=1}^d D_p^j \in \mathbb{R}^{n\times p}$
Build $\Sigma_p = S_p^\top S_p \in \mathbb{R}^{p\times p}$ and $\mathsf{L}_p = D_p^\top D_p \in \mathbb{R}^{p\times p}$
Get $(\psi_k, \mu_k)_{k\leq p}$ the generalized eigen-elements of $(\Sigma_p, \mathsf{L}_p + \lambda I_p)$

At the end of Algorithm 2, for $k \leq p$ the k-th approximated eigenvalue is μ_k, and its associated eigenfunction writes $f_k(x) = \langle \Psi_k, \phi_p(x)\rangle$, where $\Psi_k = (\mathsf{L}_p + \lambda I_p)^{-1/2}\psi_k$. Similarly as before, the main costs are due to building L_p in $\mathcal{O}(p^2 nd)$ and finding the generalized eigen-elements in $\mathcal{O}(p^3)$.

Hermite polynomials. To conclude this numerical section, and illustrate the results, we exhibit a prototypical example where the eigenfunctions of $\mathcal{L}$ are known; and we estimate them. Indeed, take $\mu(x) = e^{-x^2/2}$, the one dimensional Gaussian. Then $\mathcal{L}f = f'' - xf'$ is the *Ornstein–Uhlenbeck* operator and it is known that its eigenfunctions are the Hermite polynomials (Bakry et al., 2014). We estimate the first five Hermite polynomials with our method, thanks to Algorithm 1, with $n = p = 30$, regularization parameter $\lambda = 0.1$ and the Gaussian kernel. The result is displayed in Figure 1. The data points are displayed with dots and the built eigenfunctions are plotted with plain lines. Note that the approximation is only valid on $H^1(e^{-x^2/2})$, hence the estimated eigenfunctions behave poorly outside of the dataset: a striking example of this fact is the behavior of $\hat{h}_1$, that is a linear function on the dataset interval but diverges from it rapidly when there are no data.

6. Conclusion and further thoughts

Comparison to graph Laplacians. In this work, we proved that we could estimate the eigen-elements of the diffusion operator. This construction relies on PSD kernel methods, whereas previous rely on local averaging techniques. This leads to efficient estimations in high dimension in comparison to graph Laplacians (Hein et al., 2007). More precisely, under smoothness of the targeted eigenvectors, we showed that the statistical rates of the RKHS estimator does not depend on the dimension. We finally discuss computationally efficient ways to construct these eigen-elements resting on Nyström approximation (Williams and Seeger, 2000) or Fourier Feautures (Rahimi and Recht, 2008). Yet, this article focuses on the mathematical foundation of the estimator and its statistical performances: a precise computational and experimental comparison with graph Laplacians, diffusion maps and spectral clustering, as they are currently used, is left for a future work.

The kernel choice. Another discussion that we only sketched is the choice of the kernel. In fact, when it comes to specific applications, the art of kernel engineering is a central question. For this problem, as emphasized by the source condition, the RKHS should be chosen to approximate well the eigenfunctions of the diffusion operator in $H^1(\mu)$. Generally speaking, the interplay between $H^1(\mu)$ and $\mathcal{H}$ is a fundamental question at the core of RKHS approximation theory, and understanding this link should enable dimensionless approach in most of the cases.

Markov chains. A recent literature in applied probability aims at estimating the spectral gaps of Markov chain given the first n iterates of it (Hsu et al., 2015). The first eigenvalue of the estimator seems to do exactly the same, and understanding the difference between our algorithm and theirs is something worth of exploration. This will require to adapt a bit our algorithm and change our i.i.d. assumption on the samples to a Markovian one.

Acknowledgments

We thank a bunch of people and funding agency.

References

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014.

Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model, 2020.

Vladimir I. Bogachev, Nicolai V. Krylov, Michael Röckner, and Stanislav V. Shaposhnikov. *Fokker–Planck–Kolmogorov Equations*, volume 207. American Mathematical Society, 2022.

Vivien Cabannes. Going deeper with spectral embeddings. *arXiv preprint arXiv:2306.00742*, 2023.

Vivien Cabannes, Loucas Pillaud-Vivien, Francis Bach, and Alessandro Rudi. Overcoming the curse of dimensionality with laplacian regularization in semi-supervised learning. *Advances in Neural Information Processing Systems*, 34:30439–30451, 2021.

Vivien Cabannes, Alberto Bietti, and Randall Balestriero. On minimal variations for unsupervised representation learning. *arXiv preprint arXiv:2211.03782*, 2022.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Ané Cécile, Blachère Sébastien, Fougères Pierre, Gentil Ivan, Malrieu Florent, Roberto Cyril, and Scheffer Grégory. *Sur les Inégalités de Sobolev Logarithmiques*. Panoramas et Synthèses 10, Société Mathématique de France (SMF), 2000.

Ronald Coifman, Nadler Boaz, Stéphane Lafon, and Ioannis Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(12):113–127, 2006.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1), 2006.

Charles R Crawford. A stable generalized eigenvalue problem. *SIAM Journal on Numerical Analysis*, 13(6):854–860, 1976.

Aymeric Dieuleuveut. *Stochastic Approximation in Hilbert Spaces*. PhD thesis, ENS - INRIA, 2017.

Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Society, 2022.

Matthias Hein, Jean-Yves Audibert, and Ulrike von Luxburg. Graph laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, pages 1325–1368, 2007.

Bernard Helffer and Francis Nier. Hypoelliptic estimates and spectral theory for Fokker-Planck operators and Witten Laplacians. *Lecture Notes in Mathematics*, 1862, 2005.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417, 1933.

Daniel Hsu, Aryeh Kontorovich, and Csaba Szepesvári. Mixing time estimation in reversible markov chains from a single sample path. In *Advances in Neural Information Processing Systems*, pages 1459–1467, 2015.

Stefan Klus, Feliks Nüske, Sebastian Peitz, Jan-Hendrik Niemann, Cecilia Clementi, and Christof Schütte. Data-driven approximation of the koopman generator: Model reduction, system identification, and control. *Physica D: Nonlinear Phenomena*, 406:132416, 2020.

Tony Lelièvre. Two mathematical tools to analyze metastable stochastic processes. In *Numerical Mathematics and Advanced Applications 2011*, pages 791–810, Berlin, Heidelberg, 2013. Springer.

Per-Gunnar Martinsson and Joel A. Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, 2020.

Giacomo Meanti, Luigi Carratino, Lorenzo Rosasco, and Alessandro Rudi. Kernel methods through the roof: handling billions of points efficiently. *Advances in Neural Information Processing Systems*, 2020.

Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

Sebastian Mika, Bernhard Schölkopf, Alex J. Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. Kernel PCA and de-noising in feature spaces. In *Advances in Neural Information Processing Systems*, pages 536–542, 1999.

Elizbar A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964.

Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

Loucas Pillaud-Vivien. *Learning with Reproducing Kernel Hilbert Spaces: Stochastic Gradient Descent and Laplacian Estimation*. PhD thesis, Université Paris sciences et lettres, 2020.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems*, 31, 2018.

Loucas Pillaud-Vivien, Francis Bach, Tony Lelièvre, Alessandro Rudi, and Gabriel Stoltz. Statistical estimation of the Poincaré constant and application to sampling multimodal distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 2753–2763, 2020.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics: Functional Analysis*, volume IV. Elsevier, 2012.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3218–3228, 2017.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pages 1657–1665, 2015.

Ernesto Salinelli. Nonlinear principal components i. absolutely continuous random variables with positive bounded densities. *Annals of Statistics*, 26(2):596–616, 1998.

Robert Schaback and Holger Wendland. Kernel techniques: from machine learning to meshless methods. *Acta Numerica*, 15:543, 2006.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

Minsker Stanislav. On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics and Probability Letters*, 127:111–119, 2017.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

Joel A. Tropp. User-friendly tools for random matrices: an introduction. *NIPS Tutorials*, 2012.

Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of sgd for least-squares in the interpolation regime. *Advances in Neural Information Processing Systems*, 34:21581–21591, 2021.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.

Vadim Vladimirovich Yurinsky. *Gaussian and Related Approximations for Distributions of Sums*, pages 163–216. Springer Berlin Heidelberg, 1995.

Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220(1):456–463, 2008.

Laurent Zwald and Gilles Blanchard. On the convergence of eigenspaces in kernel principal component analysis. *Advances in Neural Information Processing Systems*, 18, 2005.

Appendix

Note that is all the appendix, to avoid cumbersome notations we will make no distinctions between $\mathcal{H}_0$ and $\mathcal{H}$ and $H_0^1(\mu)$ and $H^1(\mu)$ unless it is strictly necessary.

Appendix A. Proof on the embedding of the diffusion operator and its inverse

A.1. Proof of Proposition 3

We want to show that on $H_0^1(\mu)$, $\mathcal{L}^{-1} = \mathsf{S}\mathsf{L}^{-1}\mathsf{S}^*$.

Proof [of Proposition 3] Let $g \in \operatorname{Ran} \mathsf{S}$, there exists $f \in \mathcal{H}$ such that $g = \mathsf{S}f$. Let us calculate:

$$\mathsf{S}\mathsf{L}^{-1}\mathsf{S}^*\mathcal{L}g = \mathsf{S}\mathsf{L}^{-1}\mathsf{S}^*\mathcal{L}\mathsf{S}f = \mathsf{S}\mathsf{L}^{-1}\mathsf{L}f = \mathsf{S}f = g.$$

Moreover, as $\mathcal{L}$ is invertible on $\operatorname{Ran} \mathsf{S} \cap H^1(\mu) \cap L_0^2(\mu)$, the left and right inverse are the same. Hence, $\mathcal{L}^{-1}$ and $\mathsf{S}\mathsf{L}^{-1}\mathsf{S}^*$ are equal on $\operatorname{Ran} \mathsf{S}$. Furthermore we can notice that $\mathcal{L}^{-1}$ and $\mathsf{S}\mathsf{L}^{-1}\mathsf{S}^*$ are bounded on $L^2(\mu)$. Indeed,

$$\begin{aligned}\mathcal{P} = \sup_{f \in (\operatorname{Ker} \mathsf{L})^\perp} \frac{\langle f, \mathsf{S}^*\mathsf{S}f\rangle_{\mathcal{H}}}{\langle f, \mathsf{L}f\rangle_{\mathcal{H}}} \geqslant \sup_{f \in (\operatorname{Ker} \mathsf{L})^\perp} \frac{\langle \mathsf{L}^{-1/2}f, \mathsf{S}^*\mathsf{S}\mathsf{L}^{-1/2}f\rangle_{\mathcal{H}}}{\langle \mathsf{L}^{-1/2}f, \mathsf{L}\mathsf{L}^{-1/2}f\rangle_{\mathcal{H}}} &= \sup_{f \in (\operatorname{Ker} \mathsf{L})^\perp} \frac{\langle f, \mathsf{L}^{-1/2}\mathsf{S}^*\mathsf{S}\mathsf{L}^{-1/2}f\rangle_{\mathcal{H}}}{\|f\|_{\mathcal{H}}^2} \\ &= \|\mathsf{L}^{-1/2}\mathsf{S}^*\mathsf{S}\mathsf{L}^{-1/2}\|_{\mathcal{H}} \\ &= \|\mathsf{S}\mathsf{L}^{-1}\mathsf{S}^*\|_{L^2(\mu)}.\end{aligned}$$

As $\mathcal{L}^{-1}$ and $\mathsf{S}\mathsf{L}^{-1}\mathsf{S}^*$ are equal and continuous on $\operatorname{Ran} \mathsf{S}$, they are also equal on its closure. ■

A.2. Proof of Theorem 4 through a technical result on operators between Hilbert spaces

The lemma below gives the proof of Theorem 4 considering $A = S\mathsf{L}^{-1/2}$, $\mathcal{H}_1 = \mathcal{H}_0$ and $\mathcal{H}_2 = H_0^1(\mu)$.

Lemma 10 (Link between $\mathsf{A}^*\mathsf{A}$ and $\mathsf{A}\mathsf{A}^*$ in the compact case.) *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ two Hilbert spaces. Let A be an operator from $\mathcal{H}_1$ to $\mathcal{H}_2$ such that $\mathsf{A}^*\mathsf{A}$ is a self-adjoint compact operator on $\mathcal{H}_1$. Then,*

(i) *A is a bounded operator from $\mathcal{H}_1$ to $\mathcal{H}_2$.*

(ii) *$\mathsf{A}\mathsf{A}^*$ is a self-adjoint compact operator on $\mathcal{H}_2$ with the same spectrum as $\mathsf{A}\mathsf{A}^*$.*

(iii) *If $\lambda \neq 0$ is an eigenvalue of $\mathsf{A}^*\mathsf{A}$ with eigenvector $u \in \mathcal{H}_1$, then λ is an eigenvalue of $\mathsf{A}\mathsf{A}^*$ with eigenvector $\mathsf{A}u \in \mathcal{H}_2$.*

Proof First let us notice that A is necessarily bounded. Indeed, let $u \in \mathcal{H}_1$,

$$\|\mathsf{A}u\|_{\mathcal{H}_2}^2 = \langle \mathsf{A}u, \mathsf{A}u\rangle_{\mathcal{H}_2} = \langle \mathsf{A}^*\mathsf{A}u, u\rangle_{\mathcal{H}_1} \leqslant \|\mathsf{A}^*\mathsf{A}u\|_{\mathcal{H}_1} \|u\|_{\mathcal{H}_1} \leqslant \|\mathsf{A}^*\mathsf{A}\| \|u\|_{\mathcal{H}_1}^2 .$$

Hence, $\|\mathsf{A}\| \leqslant \sqrt{\|\mathsf{A}^*\mathsf{A}\|}$.

Second, as $\mathsf{A}^*\mathsf{A}$ is self-adjoint and compact on $\mathcal{H}_1$, there exists $(\psi_i)_{i\in\mathbb{N}}$ an orthonormal basis $\mathcal{H}_1$ and a sequence of reals $(\lambda_i)_{i\in\mathbb{N}}$ such that:

$$\mathsf{A}^*\mathsf{A} = \sum_{i\geqslant 0} \lambda_i \psi_i \otimes \psi_i,$$

where the infinite sum stands for the strong convergence of operators. Now, by composing on the left side by A^* and on the right side by A^*, we get:

$$(\mathsf{A}\mathsf{A}^*)^2 = \mathsf{A}\mathsf{A}^*\mathsf{A}\mathsf{A}^* = \sum_{i\geqslant 0} \lambda_i (\mathsf{A}\psi_i) \otimes (\mathsf{A}\psi_i) = \sum_{i\geqslant 0} \lambda_i^2 (\mathsf{A}\frac{\psi_i}{\sqrt{\lambda_i}}) \otimes (\mathsf{A}\frac{\psi_i}{\sqrt{\lambda_i}}).$$

Hence, $\mathsf{A}\mathsf{A}^* = \sum_{i\geqslant 0} \lambda_i (\mathsf{A}\frac{\psi_i}{\sqrt{\lambda_i}}) \otimes (\mathsf{A}\frac{\psi_i}{\sqrt{\lambda_i}})$ and is a compact operator. We can of course check that $\left(\mathsf{A}\frac{\psi_i}{\sqrt{\lambda_i}}\right)_{i\in\mathbb{N}}$ is an orthonormal basis of $\mathcal{H}_2$: $\left\langle \mathsf{A}\frac{\psi_i}{\sqrt{\lambda_i}}, \mathsf{A}\frac{\psi_j}{\sqrt{\lambda_j}} \right\rangle = (\lambda_i\lambda_j)^{-1/2} \langle \psi_i, \mathsf{A}^*\mathsf{A}\psi_j \rangle = \sqrt{\frac{\lambda_j}{\lambda_i}} \langle \psi_i, \psi_j \rangle = \delta_{ij}$. ■

Appendix B. Bound on the variance term

We begin first to recall usual concentration inequalities that will help handling the variance term.

B.1. Concentration inequalities

We first begin by recalling some concentration inequalities for sums of random vectors and operators.

Proposition 11 (Bernstein's inequality for sums of random vectors) *Let $z_1, \dots, z_n$ be a sequence of independent identically and distributed random elements of a separable Hilbert space $\mathcal{H}$. Assume that $\mathbb{E}\|z_1\| < +\infty$ and note $\mu = \mathbb{E}z_1$. Let $\sigma, L \geqslant 0$ such that,*

$$\forall p \geqslant 2, \qquad \mathbb{E}\left\|z_1 - \mu\right\|_{\mathcal{H}}^p \leqslant \frac{1}{2} p! \sigma^2 L^{p-2}.$$

Then, for any $\delta \in (0, 1]$,

$$\left\| \frac{1}{n} \sum_{i=1}^n z_i - \mu \right\|_{\mathcal{H}} \leqslant \frac{2L\log(2/\delta)}{n} + \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}}, \tag{18}$$

with probability at least $1 - \delta$.

Proof This is a restatement of Theorem 3.3.4 of Yurinsky (1995). ■

Proposition 12 (Bernstein's inequality for sums of random operators) *Let $\mathcal{H}$ be a separable Hilbert space and let $X_1, \ldots, X_n$ be a sequence of independent and identically distributed self-adjoint random operators on $\mathcal{H}$. Assume that $\mathbb{E}(X_i) = 0$ and that there exist $T > 0$ and S a positive trace-class operator such that $\|X_i\| \leqslant T$ almost surely and $\mathbb{E}X_i^2 \preccurlyeq S$ for any $i \in \{1, \ldots, n\}$. Then, for any $\delta \in (0, 1]$, the following inequality holds:*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \right\| \leqslant \frac{2T\beta}{3n} + \sqrt{\frac{2\|S\|\beta}{n}}, \tag{19}$$

with probability at least $1 - \delta$ and where $\beta = \log \frac{2\mathrm{Tr}S}{\|S\|\delta}$.

Proof The theorem is a restatement of Theorem 7.3.1 of Tropp (2012) generalized to the separable Hilbert space case by means of the technique in Section 4 of Stanislav (2017). ■

B.2. Operator bounds

Lemma 13 *Under Assumptions 2 and 3, Σ, and L are trace-class operators.*

Proof We only prove the result for L, the proof for Σ being similar. Consider an orthonormal basis $(\phi_i)_{i\in\mathbb{N}}$ of $\mathcal{H}$. Then, as L is a positive self adjoint operator,

$$\begin{aligned} \operatorname{tr} \mathsf{L} &= \sum_{i=1}^{\infty} \langle \mathsf{L}\phi_i, \phi_i \rangle = \sum_{i=1}^{\infty} \mathbb{E}_\mu \left[\sum_{j=1}^{d} \langle \partial_j K_x, \phi_i \rangle^2 \right] = \mathbb{E}_\mu \left[\sum_{i=1}^{\infty} \sum_{j=1}^{d} \langle \partial_j K_x, \phi_i \rangle^2 \right] \\ &= \mathbb{E}_\mu \left[\sum_{j=1}^{d} \|\partial_j K_x\|^2 \right] \leqslant \mathcal{K}_d. \end{aligned}$$

Hence, L is a trace-class operator. ■

B.3. Proof of Proposition 6: the variance bound

We recall here the expression of the variance of our estimator we want to control:

$$\begin{aligned} \text{Variance} &:= \left\| \widehat{\mathsf{L}}_\lambda^{-1/2} \widehat{\Sigma} \widehat{\mathsf{L}}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2} \Sigma \mathsf{L}_\lambda^{-1/2} \right\| \\ &= \left\| \widehat{\mathsf{L}}_\lambda^{-1/2} \widehat{\Sigma} \widehat{\mathsf{L}}_\lambda^{-1/2} - \widehat{\mathsf{L}}_\lambda^{-1/2} \Sigma \widehat{\mathsf{L}}_\lambda^{-1/2} \right\| + \left\| \widehat{\mathsf{L}}_\lambda^{-1/2} \Sigma \widehat{\mathsf{L}}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2} \Sigma \mathsf{L}_\lambda^{-1/2} \right\| \\ &\leqslant \left\| \widehat{\mathsf{L}}_\lambda^{-1/2} (\widehat{\Sigma} - \Sigma) \widehat{\mathsf{L}}_\lambda^{-1/2} \right\| + \left\| \widehat{\mathsf{L}}_\lambda^{-1/2} \Sigma \widehat{\mathsf{L}}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2} \Sigma \widehat{\mathsf{L}}_\lambda^{-1/2} \right\| + \left\| \mathsf{L}_\lambda^{-1/2} \Sigma \widehat{\mathsf{L}}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2} \Sigma \mathsf{L}_\lambda^{-1/2} \right\| \\ &\leqslant \underbrace{\left\| \widehat{\mathsf{L}}_\lambda^{-1/2} (\widehat{\Sigma} - \Sigma) \widehat{\mathsf{L}}_\lambda^{-1/2} \right\|}_{\text{Lemma 15}} + \underbrace{\left\| (\widehat{\mathsf{L}}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2}) \Sigma \widehat{\mathsf{L}}_\lambda^{-1/2} \right\|}_{\text{Lemma 16}} + \underbrace{\left\| \mathsf{L}_\lambda^{-1/2} \Sigma (\widehat{\mathsf{L}}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2}) \right\|}_{\text{Lemma 16 as well}}. \end{aligned}$$

The following quantities are useful for the estimates in this section:

$$\mathcal{N}_\infty(\lambda) = \sup_{x \in \operatorname{supp}(\mu)} \left\| \mathsf{L}_\lambda^{-1/2} K_x \right\|_{\mathcal{H}}^2, \text{ and } \mathcal{F}_\infty(\lambda) = \sup_{x \in \operatorname{supp}(\mu)} \left\| \mathsf{L}_\lambda^{-1/2} \nabla K_x \right\|_{\mathcal{H}}^2.$$

Note that under Assumption 3, $\mathcal{N}_\infty(\lambda) \leqslant \frac{\mathcal{K}}{\lambda}$ and $\mathcal{F}_\infty(\lambda) \leqslant \frac{\mathcal{K}_d}{\lambda}$. Note also that under refined assumptions on the spectrum of L, we could have a better dependence of the latter bounds with respect to λ. We first state the overall result before showing all the auxiliary lemmas below:

Lemma 14 *For any $0 < \lambda < \|\mathsf{L}\|$, $n \geqslant 15\mathcal{F}_\infty(\lambda) \log \frac{4\,\mathrm{Tr}\mathsf{L}}{\lambda\delta}$ and any $\delta \in (0, 1/2]$,*

$$\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\widehat{\Sigma}\widehat{\mathsf{L}}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2}\right\| \leqslant \frac{4\mathcal{N}_\infty(\lambda)\log\frac{2\,\mathcal{P}\mathrm{Tr}\Sigma}{\lambda\delta}}{3n} + \left[\frac{2\,\mathcal{P}\,\mathcal{N}_\infty(\lambda)\log\frac{4\,\mathcal{P}\mathrm{Tr}\Sigma}{\lambda\delta}}{n}\right]^{1/2}$$
$$+ 4\left(\mathcal{P}\,\|\Sigma\|\right)^{1/2}\left(\frac{4\mathcal{F}_\infty(\lambda)\log\frac{4\,\mathrm{Tr}\mathsf{L}}{\lambda\delta}}{3n} + \sqrt{\frac{2\,\mathcal{F}_\infty(\lambda)\log\frac{4\,\mathrm{Tr}\mathsf{L}}{\lambda\delta}}{n}}\right),$$

with probability at least $1 - 2\delta$.

B.3.1. BOUND ON THE FIRST TERM

Lemma 15 *For any $\lambda > 0$, and any $\delta \in (0, 1]$,*

$$\left\|\mathsf{L}_\lambda^{-1/2}(\widehat{\Sigma} - \Sigma)\mathsf{L}_\lambda^{-1/2}\right\| \leqslant \frac{4\mathcal{N}_\infty(\lambda)\log\frac{2\,\mathcal{P}\mathrm{Tr}\Sigma}{\lambda\delta}}{3n} + \left[\frac{2\,\mathcal{P}\,\mathcal{N}_\infty(\lambda)\log\frac{4\,\mathcal{P}\mathrm{Tr}\Sigma}{\lambda\delta}}{n}\right]^{1/2},$$

with probability at least $1 - \delta$.

Proof [of Lemma 15] We apply some concentration inequality to the operator $\mathsf{L}_\lambda^{-1/2}\widehat{\Sigma}\mathsf{L}_\lambda^{-1/2}$ whose mean is exactly $\mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2}$. The calculation is the following:

$$\begin{aligned}\left\|\mathsf{L}_\lambda^{-1/2}(\widehat{\Sigma} - \Sigma)\mathsf{L}_\lambda^{-1/2}\right\| &= \left\|\mathsf{L}_\lambda^{-1/2}\widehat{\Sigma}\mathsf{L}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2}\right\| \\ &= \left\|\frac{1}{n}\sum_{i=1}^n\left[(\mathsf{L}_\lambda^{-1/2}K_{x_i}) \otimes (\mathsf{L}_\lambda^{-1/2}K_{x_i}) - \mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2}\right]\right\|.\end{aligned}$$

We use Proposition 12. To do this, we bound for $i \in [\![1, n]\!]$:

$$\begin{aligned}\left\|(\mathsf{L}_\lambda^{-1/2}K_{x_i}) \otimes (\mathsf{L}_\lambda^{-1/2}K_{x_i}) - \mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2}\right\| &\leqslant \left\|\mathsf{L}_\lambda^{-1/2}K_{x_i}\right\|_{\mathcal{H}}^2 + \left\|\mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2}\right\| \\ &\leqslant 2\mathcal{N}_\infty(\lambda),\end{aligned}$$

and, for the second order moment,

$$\begin{aligned}&\mathbb{E}\left((\mathsf{L}_\lambda^{-1/2}K_{x_i}) \otimes (\mathsf{L}_\lambda^{-1/2}K_{x_i}) - \mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2}\right)^2 \\ &\quad= \mathbb{E}\left[\left\|\mathsf{L}_\lambda^{-1/2}K_{x_i}\right\|_{\mathcal{H}}^2(\mathsf{L}_\lambda^{-1/2}K_{x_i}) \otimes (\mathsf{L}_\lambda^{-1/2}K_{x_i})\right] - \mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1}\Sigma\mathsf{L}_\lambda^{-1/2} \\ &\quad\preccurlyeq \mathcal{N}_\infty(\lambda)\mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2}.\end{aligned}$$

We conclude the proof by some estimation of the constant $\beta = \log \frac{2\,\mathrm{Tr}(\Sigma \mathsf{L}_\lambda^{-1})}{\left\|\mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2}\right\|\delta}$. To do this, we remark that, thanks to Proposition 16 of Pillaud-Vivien (2020),

$$\left\|\mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2}\right\| = (\mathcal{P}_\mu^\lambda)^{-1} \geqslant \mathcal{P}^{-1},$$

and using $\mathrm{Tr}\Sigma\mathsf{L}_\lambda^{-1} \leqslant \lambda^{-1}\mathrm{Tr}\Sigma$, it holds $\beta \leqslant \log \frac{2\,\mathcal{P}\mathrm{Tr}\Sigma}{\lambda\delta}$. Therefore,

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n}\left[(\mathsf{L}_\lambda^{-1/2}K_{x_i})\otimes(\mathsf{L}_\lambda^{-1/2}K_{x_i}) - \mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2}\right]\right\| \\ \leqslant \frac{4\mathcal{N}_\infty(\lambda)\log\frac{2\,\mathcal{P}\mathrm{Tr}\Sigma}{\lambda\delta}}{3n} + \left[\frac{2\,\mathcal{P}_\mu^\lambda\,\mathcal{N}_\infty(\lambda)\log\frac{2\,\mathcal{P}\mathrm{Tr}\Sigma}{\lambda\delta}}{n}\right]^{1/2}.\end{aligned}$$

This concludes the proof of Lemma 15. ■

B.3.2. BOUND ON THE SECOND TERM

Here, we want to bound the term $\left\|(\widehat{\mathsf{L}}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2})\Sigma\widehat{\mathsf{L}}_\lambda^{-1/2}\right\|$. Let us work on it a little bit more.

$$\begin{aligned}\left\|(\widehat{\mathsf{L}}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2})\Sigma\widehat{\mathsf{L}}_\lambda^{-1/2}\right\| &= \left\|\widehat{\mathsf{L}}_\lambda^{1/2}(\widehat{\mathsf{L}}_\lambda^{-1} - \mathsf{L}_\lambda^{-1})\mathsf{L}_\lambda^{1/2}\Sigma\widehat{\mathsf{L}}_\lambda^{-1/2}\right\| \\ &= \left\|\widehat{\mathsf{L}}_\lambda^{-1/2}(\widehat{\mathsf{L}}_\lambda - \mathsf{L}_\lambda)\mathsf{L}_\lambda^{-1/2}\Sigma\widehat{\mathsf{L}}_\lambda^{-1/2}\right\| \\ &= \left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\mathsf{L}_\lambda^{-1/2}(\widehat{\mathsf{L}}_\lambda - \mathsf{L}_\lambda)\mathsf{L}_\lambda^{-1/2}\Sigma^{1/2}\Sigma^{1/2}\mathsf{L}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\widehat{\mathsf{L}}_\lambda^{-1/2}\right\| \\ &\leq \left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\right\|\left\|\mathsf{L}_\lambda^{-1/2}(\widehat{\mathsf{L}}_\lambda - \mathsf{L}_\lambda)\mathsf{L}_\lambda^{-1/2}\right\|\left\|\Sigma^{1/2}\right\|\left\|\Sigma^{1/2}\mathsf{L}_\lambda^{-1/2}\right\|\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\right\| \\ &\leq (\mathcal{P}\left\|\Sigma\right\|)^{1/2}\underbrace{\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\right\|^2}_{\text{Lemma 18}}\underbrace{\left\|\mathsf{L}_\lambda^{-1/2}(\widehat{\mathsf{L}}_\lambda - \mathsf{L}_\lambda)\mathsf{L}_\lambda^{-1/2}\right\|}_{\text{Lemma 17}}.\end{aligned}$$

Hence, we can formulate the principal result of this subsection:

Lemma 16 *For any $0 < \lambda < \|\mathsf{L}\|$, $\delta \in (0,1)$, and $n \geqslant 15\mathcal{F}_\infty(\lambda)\log\frac{4\,\mathrm{Tr}\mathsf{L}}{\lambda\delta}$, it holds with probability at least $1-\delta$:*

$$\left\|(\widehat{\mathsf{L}}_\lambda^{-1/2} - \mathsf{L}_\lambda^{-1/2})\Sigma\widehat{\mathsf{L}}_\lambda^{-1/2}\right\| \leq 2\,(\mathcal{P}\left\|\Sigma\right\|)^{1/2}\left(\frac{4\mathcal{F}_\infty(\lambda)\log\frac{4\,\mathrm{Tr}\mathsf{L}}{\lambda\delta}}{3n} + \sqrt{\frac{2\,\mathcal{F}_\infty(\lambda)\log\frac{4\,\mathrm{Tr}\mathsf{L}}{\lambda\delta}}{n}}\right),$$

with probability at least $1-\delta$.

And now we prove the auxiliary lemmas.

Lemma 17 *For any $0 < \lambda < \|\mathsf{L}\|$ and any $\delta \in (0,1]$,*

$$\left\|\mathsf{L}_\lambda^{-1/2}(\widehat{\mathsf{L}} - \mathsf{L})\mathsf{L}_\lambda^{-1/2}\right\| \leqslant \frac{4\mathcal{F}_\infty(\lambda)\log\frac{4\,\mathrm{Tr}\mathsf{L}}{\lambda\delta}}{3n} + \sqrt{\frac{2\,\mathcal{F}_\infty(\lambda)\log\frac{4\,\mathrm{Tr}\mathsf{L}}{\lambda\delta}}{n}},$$

with probability at least $1-\delta$.

Proof [Proof of Lemma 17]

As in the proof of Lemma 15, we want to apply some concentration inequality to the operator $\mathsf{L}_\lambda^{-1/2}\widehat{\Delta}\mathsf{L}_\lambda^{-1/2}$, whose mean is exactly $\mathsf{L}_\lambda^{-1/2}\Delta\mathsf{L}_\lambda^{-1/2}$. The proof is almost the same as Lemma 15. We start by writing

$$\begin{aligned}\left\|\mathsf{L}_\lambda^{-1/2}(\widehat{\mathsf{L}}-\mathsf{L})\mathsf{L}_\lambda^{-1/2}\right\| &= \left\|\mathsf{L}_\lambda^{-1/2}\widehat{\mathsf{L}}\mathsf{L}_\lambda^{-1/2}-\mathsf{L}_\lambda^{-1/2}\mathsf{L}\mathsf{L}_\lambda^{-1/2}\right\| \\ &= \left\|\frac{1}{n}\sum_{i=1}^n\left[(\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i})\otimes(\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i})-\mathsf{L}_\lambda^{-1/2}\Delta\mathsf{L}_\lambda^{-1/2}\right]\right\|.\end{aligned}$$

In order to use Proposition 12, we bound for $i\in[\![1,n]\!]$,

$$\begin{aligned}\left\|(\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i})\otimes(\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i})-\mathsf{L}_\lambda^{-1/2}\mathsf{L}\mathsf{L}_\lambda^{-1/2}\right\| &\leqslant \left\|\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i}\right\|_{\mathcal{H}}^2+\left\|\mathsf{L}_\lambda^{-1/2}\mathsf{L}\mathsf{L}_\lambda^{-1/2}\right\| \\ &\leqslant 2\mathcal{F}_\infty(\lambda),\end{aligned}$$

and, for the second order moment,

$$\begin{aligned}&\mathbb{E}\left[\left((\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i})\otimes(\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i})-\mathsf{L}_\lambda^{-1/2}\mathsf{L}\mathsf{L}_\lambda^{-1/2}\right)^2\right] \\ &= \mathbb{E}\left[\left\|\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i}\right\|_{\mathcal{H}}^2(\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i})\otimes(\mathsf{L}_\lambda^{-1/2}\nabla K_{x_i})\right]-\mathsf{L}_\lambda^{-1/2}\mathsf{L}\mathsf{L}_\lambda^{-1}\mathsf{L}\mathsf{L}_\lambda^{-1/2} \\ &\preccurlyeq \mathcal{F}_\infty(\lambda)\mathsf{L}_\lambda^{-1/2}\mathsf{L}\mathsf{L}_\lambda^{-1/2}.\end{aligned}$$

We conclude by some estimation of $\beta=\log\frac{2\,\mathrm{Tr}(\mathsf{L}\mathsf{L}_\lambda^{-1})}{\|\mathsf{L}_\lambda^{-1}\mathsf{L}\|\delta}$. Since $\mathrm{Tr}(\mathsf{L}\mathsf{L}_\lambda^{-1})\leqslant\lambda^{-1}\mathrm{Tr}\mathsf{L}$ and for $\lambda\leqslant\|\mathsf{L}\|$, $\left\|\mathsf{L}_\lambda^{-1}\mathsf{L}\right\|\geqslant 1/2$, it follows that $\beta\leqslant\log\frac{4\,\mathrm{Tr}\mathsf{L}}{\lambda\delta}$. The conclusion then follows from (19). ■

Lemma 18 (Bounding operators) *For any $\lambda>0$, $\delta\in(0,1)$, and $n\geqslant 15\mathcal{F}_\infty(\lambda)\log\frac{4\,\mathrm{Tr}\mathsf{L}}{\lambda\delta}$, it holds with probability at least $1-\delta$:*

$$\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\right\|^2\leqslant 2,$$

The proof of this result relies on the following lemma (see proof by Rudi and Rosasco (2017, Proposition 8)).

Lemma 19 *Let $\mathcal{H}$ be a separable Hilbert space, A and B two bounded self-adjoint positive linear operators on $\mathcal{H}$ and $\lambda>0$. Then*

$$\left\|(A+\lambda I)^{-1/2}(B+\lambda I)^{1/2}\right\|\leqslant(1-\beta)^{-1/2},$$

with $\beta=\lambda_{\max}\left((B+\lambda I)^{-1/2}(B-A)(B+\lambda I)^{-1/2}\right)<1$, where $\lambda_{\max}(O)$ is the largest eigenvalue of the self-adjoint operator O.

We can now write the proof of Lemma 18.
Proof [Proof of Lemma 18] Thanks to Lemma 19, we see that

$$\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\right\|^2 \leqslant \left(1-\lambda_{\max}\left(\mathsf{L}_\lambda^{-1/2}(\widehat{\mathsf{L}}-\mathsf{L})\mathsf{L}_\lambda^{-1/2}\right)\right)^{-1},$$

and as $\left\|\mathsf{L}_\lambda^{-1/2}(\widehat{\mathsf{L}}-\mathsf{L})\mathsf{L}_\lambda^{-1/2}\right\| < 1$, we have:

$$\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\right\|^2 \leqslant \left(1-\left\|\mathsf{L}_\lambda^{-1/2}(\widehat{\mathsf{L}}-\mathsf{L})\mathsf{L}_\lambda^{-1/2}\right\|\right)^{-1}.$$

We can then apply the bound of Lemma 17 to obtain that, if λ is such that $\frac{4\mathcal{F}_\infty(\lambda)\log\frac{4\,\mathrm{Tr}\mathsf{L}}{\lambda\delta}}{3n} + \sqrt{\frac{2\,\mathcal{F}_\infty(\lambda)\log\frac{4\,\mathrm{Tr}\mathsf{L}}{\lambda\delta}}{n}} \leqslant \frac{1}{2}$, then $\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\right\|^2 \leqslant 2$ with probability $1-\delta$. The condition on λ is satisfied when $n \geqslant 15\mathcal{F}_\infty(\lambda)\log\frac{4\,\mathrm{Tr}\mathsf{L}}{\lambda\delta}$. ∎

B.3.3. BOUND ON THE THIRD AND LAST TERM

Here, we want to bound the term $\left\|\mathsf{L}_\lambda^{-1/2}\Sigma(\widehat{\mathsf{L}}_\lambda^{-1/2}-\mathsf{L}_\lambda^{-1/2})\right\|$. Let us apply the same tricks as previously.

$$\begin{aligned}\left\|\mathsf{L}_\lambda^{-1/2}\Sigma(\widehat{\mathsf{L}}_\lambda^{-1/2}-\mathsf{L}_\lambda^{-1/2})\right\| &= \left\|\mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2}(\mathsf{L}-\widehat{\mathsf{L}})\widehat{\mathsf{L}}_\lambda^{-1/2}\right\| \\ &\leq (\mathcal{P}\,\|\Sigma\|)^{1/2}\left\|\widehat{\mathsf{L}}_\lambda^{-1/2}\mathsf{L}_\lambda^{1/2}\right\|^2\left\|\mathsf{L}_\lambda^{-1/2}(\widehat{\mathsf{L}}_\lambda-\mathsf{L}_\lambda)\mathsf{L}_\lambda^{-1/2}\right\|.\end{aligned}$$

Hence, the same bound as Lemma 16 applies!

Appendix C. The bias term

C.1. Proof of the consistency: Proposition 7

To prove Proposition 7, we first need a general result on operator norm convergence.

Lemma 20 *Let $\mathcal{H}$ be a Hilbert space and suppose that $(A_n)_{n\geqslant 0}$ is a family of bounded operators such that $\forall n \in \mathbb{N}$, $\|A_n\| \leqslant 1$ and $\forall f \in \mathcal{H}$, $A_n f \xrightarrow{n\to\infty} Af$. Suppose also that B is a compact operator. Then, in operator norm,*

$$A_n B A_n^* \xrightarrow{n\to\infty} ABA^*.$$

Proof Let $\varepsilon > 0$. As B is compact, it can be approximated by a finite rank operator $B_{n_\varepsilon} = \sum_{i=1}^{n_\varepsilon} b_i\langle f_i,\cdot\rangle g_i$, where $(f_i)_i$ and $(g_i)_i$ are orthonormal bases, and $(b_i)_i$ is a sequence of nonnegative numbers with limit zero (singular values of the operator). More precisely, n_ε is chosen so that

$$\|B-B_{n_\varepsilon}\| \leqslant \frac{\varepsilon}{2}.$$

Moreover, ε being fixed, $A_n B_{n_\varepsilon} A_n^* = \sum_{i=1}^{n_\varepsilon} b_i\langle A_n f_i,\cdot\rangle A_n g_i \underset{n\infty}{\longrightarrow} \sum_{i=1}^{n_\varepsilon} b_i\langle Af_i,\cdot\rangle Ag_i = AB_{n_\varepsilon}A^*$ in operator norm, so that, for $n \geqslant N_\varepsilon$, with $N_\varepsilon \geqslant n_\varepsilon$ sufficiently large, $\|A_nB_{n_\varepsilon}A_n^* - AB_{n_\varepsilon}A^*\| \leqslant \frac{\varepsilon}{2}$. Finally, as $\|A\| \leqslant 1$, it holds, for $n \geqslant N_\varepsilon$

$$\begin{aligned}\|A_nB_{n_\varepsilon}A_n^* - ABA^*\| &\leqslant \|A_nB_{n_\varepsilon}A_n^* - AB_{n_\varepsilon}A^*\| + \|A(B_{n_\varepsilon}-B)A^*\| \\ &\leqslant \|A_nB_{n_\varepsilon}A_n^* - AB_{n_\varepsilon}A^*\| + \|B_{n_\varepsilon}-B\| \leqslant \varepsilon.\end{aligned}$$

This proves the convergence in operator norm of $A_n B A_n^*$ to ABA^* when n goes to infinity. ■

We can now prove Proposition 7.

Proof [of Proposition 7] Denoting by $B = \mathsf{L}^{-1/2}\Sigma\mathsf{L}^{-1/2}$ and by $A_\lambda = \mathsf{L}_\lambda^{-1/2}\mathsf{L}^{1/2}$ both defined on $\mathcal{H}_0$, we have $\mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2} = A_\lambda B A_\lambda^*$ with B compact and $\|A_\lambda\| \leqslant 1$. Furthermore, let $(\phi_i)_{i\in\mathbb{N}}$ be an orthonormal family of eigenvectors of the compact operator L associated to eigenvalues $(\nu_i)_{i\in\mathbb{N}}$. Then we can write, for any $f \in \mathcal{H}_0$,

$$A_\lambda f = \mathsf{L}_\lambda^{-1/2}\mathsf{L}^{1/2} f = \sum_{i=0}^{\infty} \sqrt{\frac{\nu_i}{\lambda+\nu_i}} \langle f, \phi_i \rangle_{\mathcal{H}}\, \phi_i \underset{\lambda\to 0}{\longrightarrow} f.$$

Hence by applying Lemma 20, we have the convergence in operator norm of $\mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2}$ to $\mathsf{L}^{-1/2}\Sigma\mathsf{L}^{-1/2}$. ■

C.2. Fast rates under source condition: Proposition 8

Proof [of Proposition 8] To show Proposition 8, we simply bound the bias term according to the following inequalities.

$$\begin{aligned}
\left\|\Pi^p \left(\mathsf{L}_\lambda^{-1/2}\Sigma\mathsf{L}_\lambda^{-1/2} - \mathsf{L}^{-1/2}\Sigma\mathsf{L}^{-1/2}\right)\Pi^p\right\| &\leq \left\|\Pi^p \left(\mathsf{L}_\lambda^{-1/2} - \mathsf{L}^{-1/2}\right)\Sigma\mathsf{L}_\lambda^{-1/2}\Pi^p\right\| \\
&\quad + \left\|\Pi^p\mathsf{L}^{-1/2}\Sigma\left(\mathsf{L}_\lambda^{-1/2} - \mathsf{L}^{-1/2}\right)\Pi^p\right\| \\
&\leq \left\|\Pi^p\mathsf{L}_\lambda^{-1/2}\mathsf{L}^{-1/2}\Sigma\mathsf{L}^{-1/2}\Pi^p\right\| + \left\|\Pi^p\mathsf{L}^{-1/2}\Sigma\mathsf{L}^{-1/2}\mathsf{L}_\lambda^{-1/2}\Pi^p\right\| \\
&\leq \left\|\Pi^p\mathsf{L}^{-1/2}\mathsf{L}^{-1/2}\Sigma\mathsf{L}^{-1/2}\Pi^p\right\| + \left\|\Pi^p\mathsf{L}^{-1/2}\Sigma\mathsf{L}^{-1/2}\mathsf{L}^{-1/2}\Pi^p\right\| \\
&\leq 2\mathcal{P}\left\|\Pi^p\mathsf{L}^{-1/2}\Pi^p\right\|,
\end{aligned}$$

which finally proves Proposition 8. ■

Of course, under refined a priori on how smooth are the eigenvectors of $\mathcal{L}$, i.e., on control like $\left\|\Pi^p\mathsf{L}^{-\theta}\Pi^p\right\|$, for $\theta \in [0, 1]$ that generalize the *source conditions* we used, we could get finer-grained rates (Pillaud-Vivien et al., 2018; Berthier et al., 2020; Varre et al., 2021).