Overcoming the curse of dimensionality with Laplacian regularization in semi-supervised learning

Vivien Cabannes, Loucas Pillaud-Vivien, Francis Bach, Alessandro Rudi

Overcoming the curse of dimensionality with Laplacian regularization in semi-supervised learning

Vivien Cabannes
ENS – INRIA – PSL
Paris, France
vivien.cabannes@gmail.com

Loucas Pillaud-Vivien
EPFL
Lausanne, Switzerland

Francis Bach
ENS – INRIA – PSL
Paris, France

Alessandro Rudi
ENS – INRIA – PSL
Paris, France

Abstract

As annotations of data can be scarce in large-scale practical problems, leveraging unlabelled examples is one of the most important aspects of machine learning. This is the aim of semi-supervised learning. To benefit from the access to unlabelled data, it is natural to diffuse smoothly knowledge of labelled data to unlabelled one. This induces to the use of Laplacian regularization. Yet, current implementations of Laplacian regularization suffer from several drawbacks, notably the well-known curse of dimensionality. In this paper, we provide a statistical analysis to overcome those issues, and unveil a large body of spectral filtering methods that exhibit desirable behaviors. They are implemented through (reproducing) kernel methods, for which we provide realistic computational guidelines in order to make our method usable with large amounts of data.

In the last decade, machine learning has been able to tackle amazingly complex tasks, which was mainly allowed by computational power to train large learning models on large annotated datasets. For instance, ImageNet is made of tens of millions of images, which have all been manually annotated by humans [16]. The greediness in data annotation of such a current learning paradigm is a major limitation. In particular, when annotation of data demands in-depth expertise, relying on techniques that require zillions of labelled data is not viable. This motivates several research streams to overcome the need for annotations, such as self-supervised learning for images or natural language processing [17]. Aiming for generality, semi-supervised learning is the most classical one, assuming access to a vast amount of input data, but among which only a scarce percentage is labelled. To leverage the presence of unlabelled data, most semi-supervised techniques assume a form of low-density separation hypothesis, as detailed in the recent review of van Engelen and Hoos [52], and illustrated by state-of the-art models [6, 54]. This hypothesis assumes that the function to learn from the data varies smoothly in highly populated regions of the input space, but might vary more strongly in scarcely populated areas, or that the decision frontiers between classes lie in regions with low-density. In such a setting, it is natural to enforce constraints on the variations of the function to learn. While semi-supervised learning is an important learning framework, it has not provided as much exciting realizations as one could have expected. This might be related to the fact that it is classically approached through graph-based Laplacian, a technique that does not scale well with the dimension of the input space [5].

Paper organization. In Section 1, we motivate Laplacian regularization, and recall drawbacks of naive implementations. These limitations are overcome in Section 2 where we expose a theoretically principled path to derive well-behaved algorithms. More precisely, we unveil a vast class of estimates

based on spectral filtering. We turn to implementation in Section 3 where we provide realistic guidelines to ensure scalability of the proposed algorithms. Statistical properties of our estimators are stated in Section 4.

Contributions. They are two folds. (*i*) Statistically, we explain that Laplacian regularization can be properly leveraged based on functional space considerations, and that those considerations can be turned into concrete implementations thanks to kernel methods. As a result, we provide consistent estimators that exhibit fast convergence rates under a low density separation hypothesis, and that, in particular, do not suffer from the curse of dimensionality. (*ii*) Computationally, we avoid dealing with large matrices of derivatives by providing a low-rank approximation that allows to deal with $n^\gamma \log(n) \times n^\gamma \log(n)$ matrices, with a parameter $\gamma \in (0, 1]$ depending on the regularity of the problem, instead of $n(d + 1) \times n(d + 1)$ matrices, thus cutting down to $\mathcal{O}(\log(n)^2 n^{1+2\gamma} d)$ the potential $\mathcal{O}(n^3 d^3)$ training cost.

Related work. Interplays between graph theory and machine learning were proven successful in the 2000s [47]. The seminal paper of Zhu et al. [62] introduced graph-Laplacian as a transductive method in the context of semi-supervised learning. A smoothing variant was proposed by [60], which is coherent with the fact that enforcing constraints on labelled points leads to spikes [1]. Interestingly, graph Laplacians do converge to diffusion operators linked with the weighted Laplace Beltrami operator [24, 21]. However, these local diffusion methods are known to suffer from the curse of dimensionality [5]. That is, local averaging methods are intuitive learning methods that have been used for more than half a century [20]. Yet, those methods do not scale well with the dimension of the input space [56]. This is related with the fact that to cover $[0, 1]^d$, we need ε^{-d} balls of radius ε. Interestingly, if the function to learn is m times differentiable with smooth partial derivatives, it is possible to leverage more information from function evaluations and overcome the curse of dimensionality when $m \gtrsim d$. This property is related to covering numbers (*a.k.a.* capacity) of Sobolev spaces [28] and is leveraged by (reproducing) kernel methods [49, 10]. The crux of this paper is apply this fact to Laplacian regularization techniques. Note that derivative with reproducing kernel methods in machine learning have already been considered in different settings by [61, 44, 19].

1 Laplacian regularization

In this section, we introduce the notations and concepts related to the semi-supervised learning regression problem, noting that most of our results extend to any convex loss beyond least-squares. We motivate and describe Laplacian regularization that will allow us to leverage the low-density separation hypothesis. We explain statistical drawbacks usually linked with Laplacian regularization, and discuss on how to circumvent them.

In the following, we denote by $\mathcal{X} = \mathbb{R}^d$ the input space, $\mathcal{Y} = \mathbb{R}$ the output space, and by $\rho \in \Delta_{\mathcal{X} \times \mathcal{Y}}$ the joint distribution on $\mathcal{X} \times \mathcal{Y}$. For simplicity, we assume that ρ has compact support. In the following, we denote by $\rho_{\mathcal{X}}$ the marginal of ρ over $\mathcal{X}$, and by $\rho|_x$ the conditional distribution of Y given $X = x$. As usual, for $p \in \mathbb{N}^*$, $L^p(\mathbb{R}^d)$ is the space of functions f such that f^p is integrable. Moreover, we define usual Sobolev spaces: for $s \in \mathbb{N}$, $W^{s,p}(\mathbb{R}^d)$ stands for the space of functions whose weak derivatives of order s-th are in $L^p(\mathbb{R}^d)$. When $p = 2$, they have a Hilbertian structure and we denote, $H^s(\mathbb{R}^d) = W^{s,2}(\mathbb{R}^d)$ these Hilbertian spaces. Ideally, we would like to retrieve the mapping $g^* : \mathcal{X} \to \mathcal{Y}$ defined as

$$g^* = \operatorname*{arg\,min}_{g \in L^2(\rho_{\mathcal{X}})} \mathbb{E}_{(X,Y) \sim \rho} \left[\|g(X) - Y\|^2 \right] = \operatorname*{arg\,min}_{g \in L^2(\rho_{\mathcal{X}})} \|g - g_\rho\|^2_{L^2(\rho_{\mathcal{X}})} = g_\rho, \qquad (1)$$

where $g_\rho : \mathcal{X} \to \mathcal{Y}$ is defined as $g_\rho(x) = \mathbb{E}[Y \mid X = x]$. In semi-supervised learning, we assume that we do not have access to ρ but we have access to n independent samples $(X_i)_{i \leq n} \sim \rho_{\mathcal{X}}^{\otimes n}$, among which we have n_ℓ labels $Y_i \sim \rho|_{X_i}$ for $i \leq n_\ell$, with n_ℓ potentially much smaller than n. In other terms, we have n_ℓ supervised pairs $(X_i, Y_i)_{i \leq n_\ell}$, and $n - n_\ell$ unsupervised samples $(X_i)_{n_\ell < i \leq n}$. While we restrict ourselves to real-valued regression for simplicity, our exposition indeed applies generically to partially supervised learning. In particular, it can be used off-the-shelve to complement the approaches of [7, 8] as we detailed in Appendix A.

2

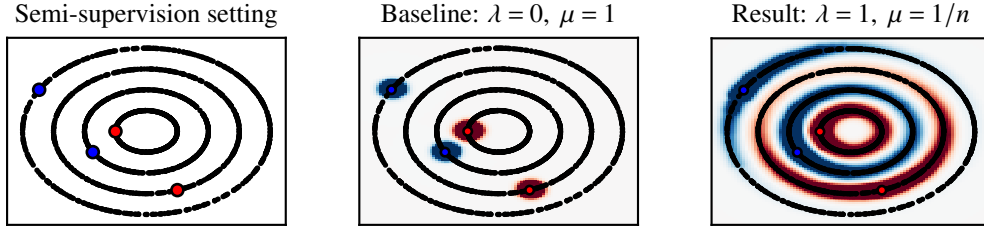| Semi-supervision setting | Baseline: $\lambda = 0$, $\mu = 1$ | Result: $\lambda = 1$, $\mu = 1/n$ |

Figure 1: Motivating example. (Left) We suppose given $n = 2000$ points in $\mathcal{X} = \mathbb{R}^2$, represented as black dots, spanning 4 concentric circles. Among those points are $n_\ell = 4$ labelled points, with labels being either 1 represented in red, and -1 represented in blue. In this setting, it is natural to assume that g^* should be constant on each circles, which can be encoded as $\|\nabla g^*\| = 0$ on $\operatorname{supp} \rho_{\mathcal{X}}$. (Middle) Kernel ridge regression estimate based on the labelled points with Gaussian kernel of bandwidth $\sigma = .2r$, r being the radius of the innermost circle. (Right) Laplacian regularization reconstruction. The reconstruction is based on approximate empirical risk minimization with $p = n$, which ensures a computational complexity of $O(p^2 n d)$, instead of $O(n^3 d^3)$ needed to recover the exact empirical risk minimizer (5).

1.1 Diffusion operator $\mathcal{L}$

In order to leverage unlabelled data, we will assume that g^* varies smoothly on highly populated regions of $\mathcal{X}$, and might vary highly on low density regions. For example, this is the case when data are clustered in well separated regions of space, and labels are constant on clusters. This is captured by the fact that the Dirichlet energy

$$\int_{\mathcal{X}} \|\nabla g^*(x)\|^2 \, \rho_{\mathcal{X}}(\mathrm{d}x) = \mathbb{E}_{X \sim \rho_{\mathcal{X}}} \left[\|\nabla g^*(X)\|^2 \right] =: \left\| \mathcal{L}^{1/2} g \right\|_{L^2(\rho_{\mathcal{X}})}^2, \tag{2}$$

is assumed to be small. Because the quadratic functional (2) will play a crucial role in our exposition, we define $\mathcal{L}$ as the self-adjoint operator on $L^2(\rho_{\mathcal{X}})$, extending the operator on $H^1(\rho_{\mathcal{X}})$ representing this functional. Under mild assumptions on $\rho_{\mathcal{X}}$, $\mathcal{L}^{-1}$ can be shown to be a compact operator, which we will assume in the following. In essence, we will assume that if we have a lot of unlabelled data and $\left\| \mathcal{L}^{1/2} g \right\|$ can be well approximated for any function g, then we do not need a lot of labelled data to estimate correctly g^*. To illustrate this, at one extreme, if we know that $\left\| \mathcal{L}^{1/2} g^* \right\| = 0$, then g^* is known to be constant on each connected component of $\rho_{\mathcal{X}}$ so that, along with the knowledge of $\rho_{\mathcal{X}}$, only a few labelled points would be sufficient to recover perfectly g^*. We illustrate those considerations on Figure 1.

1.2 Drawbacks of naive Laplacian regularization

Following the motivations presented precedently, it is natural to consider the regularized objective and solution defined, for $\lambda > 0$, as

$$\begin{aligned}
g_\lambda &= \underset{g \in H^1(\rho_{\mathcal{X}})}{\arg\min} \; \mathbb{E}_{(X,Y) \sim \rho} \left[\|g(X) - Y\|^2 \right] + \lambda \, \mathbb{E}_{X \sim \rho_{\mathcal{X}}} \left[\|\nabla g(X)\|_{\mathbb{R}^d}^2 \right] \\
&= \underset{g \in H^1(\rho_{\mathcal{X}})}{\arg\min} \; \|g - g_\rho\|_{L^2(\rho_{\mathcal{X}})}^2 + \lambda \left\| \mathcal{L}^{1/2} g \right\|_{L^2(\rho_{\mathcal{X}})}^2 = (I + \lambda \mathcal{L})^{-1} g_\rho.
\end{aligned} \tag{3}$$

This regularization has nice properties. In particular, for small λ, it can be seen as a first order approximation of the heat equation solution $e^{-\lambda \mathcal{L}} g_\rho$, which represents the temperature profile at time $t = \lambda$, instantiated with the initial profile g_ρ, and with $\rho_{\mathcal{X}}$ modelling the thermal conductivity. It also has interpretations in term of random walk and Langevin diffusion [40, 48]. In a word, g_λ is the diffusion of g_ρ with respect to the density $\rho_{\mathcal{X}}$, which relates to the idea of diffusing labelled data with respect to the intrinsic geometry of the data, which is the idea captured by [62].

However, from a learning perspective, Eq. (3) is linked with the prior that g^* belongs to $H^1(\rho_{\mathcal{X}})$, a prior that is not strong enough to overcome the curse of dimensionality as we saw in the related work section. Moreover, assuming we have enough unsupervised data to suppose known $\rho_{\mathcal{X}}$, and therefore $\mathcal{L}$, Eq. (3) leads to the naive empirical estimate $g_{\text{(naive)}} \in \arg\min_{g:\mathcal{X} \to \mathbb{R}} \sum_{i=1}^{n_\ell} \|g(X_i) - Y_i\|^2 + n_\ell \lambda \left\| \mathcal{L}^{1/2} g \right\|^2$. While the definition of $g_{\text{(naive)}}$ could seem like a great idea, in fact, such an estimate $g_{\text{(naive)}}$ is known to be mostly constant and spiking to interpolate the data (X_i, Y_i) as soon as $d > 2$

3

[37]. This is to be related with the capacity of the space associated with the pseudo-norm $\left\|\mathcal{L}^{1/2}g\right\|$ in L^2. This capacity, related to H^1, is too large for the Laplacian regularization term to constraint $g_{\text{(naive)}}$ in a meaningful way. In other terms, we need to regularize with stronger penalties.

1.3 Stronger regularization

In this subsection, we discuss techniques to overcome the issues encountered with $g_{\text{(naive)}}$. Those techniques are based on functional space constraints or on spectral filtering techniques.

Functional spaces. A solution to overcome the capacity issue of H^1 in L^2 is to constrain the estimate of g^* to belong to a smaller functional space. In the realm of graph Laplacian, [1] proposed to solve this problem by considering the r-Laplacian regularization reading $\Omega_r = \int_{\mathcal{X}} \|\nabla g(X)\|^r \rho(\mathrm{d}x)$, with $r > d$. In essence, this restricts g to live in $W^{1,r}(\rho_{\mathcal{X}})$ for $r > d$, and allows to avoid spikes associated with $g_{\text{(naive)}}$. However considering high power of the gradient is likely to introduce instability (think that d is the potentially really big dimension of the input space), and from a learning perspective, the capacity of $W^{1,r}$, which compares to the one of H^2, is still too big.

In this paper, we will rather keep the diffusion operator $\mathcal{L}$, and add a second penalty to reduce the space in which we look for the solution. With $\mathcal{G}$ an Hilbert space of functions, we could look for, with $\mu > 0$ a second regularization parameter

$$g_{\lambda,\mu} = \underset{g:\mathcal{G} \cap H^1(\rho_{\mathcal{X}})}{\arg\min} \|g - g_\rho\|^2_{L^2(\rho_{\mathcal{X}})} + \lambda \left\|\mathcal{L}^{1/2}g\right\|^2_{L^2(\rho_{\mathcal{X}})} + \lambda\mu \|g\|^2_{\mathcal{G}}. \tag{4}$$

This formulation restricts $g_{\lambda,\mu}$ to belong both to $H^1(\rho_{\mathcal{X}})$ (thanks to the term in λ) and $\mathcal{G}$ (thanks to the term in μ). In particular the resulting space $H^1(\rho_{\mathcal{X}}) \cap \mathcal{G}$ to which $g_{\lambda,\mu}$ belongs, has a smaller capacity in L^2 than the one of $\mathcal{G}$ in L^2. In practice, we do not have access to ρ and $\rho_{\mathcal{X}}$ but to $(X_i, Y_i)_{i \leq n_\ell}$ and $(X_i)_{i \leq n}$, and we might consider the empirical estimator defined through empirical risk minimization

$$g_{n_\ell, n} = \underset{g \in \mathcal{G}}{\arg\min}\, n_\ell^{-1} \sum_{i=1}^{n_\ell} \|g(X_i) - Y_i\|^2 + \lambda n^{-1} \sum_{i=1}^{n} \|\nabla g(X_i)\|^2 + \lambda\mu \|g\|^2_{\mathcal{G}}. \tag{5}$$

For example, we could consider $\mathcal{G}$ to be the Sobolev space $H^m(\mathrm{d}x)$. Note the difference between $\mathcal{G}$ linked with $\mathrm{d}x$, the Lebesgue measure, that is known, and $\mathcal{L}$ linked with $\rho_{\mathcal{X}}$, the marginal of ρ over $\mathcal{X}$, that is not known. In this setting, the regularization $\|\mathcal{L}^{1/2}g\|^2 + \mu\|g\|^2_{\mathcal{G}}$ reads $\int_{\mathcal{X}} \|Dg(x)\|^2 \rho_{\mathcal{X}}(\mathrm{d}x) + \mu \int_{\mathcal{X}} \sum_{\alpha=0}^{m} \|D^\alpha g(x)\|^2 \mathrm{d}x$. Because of the size of H^m in L^2, this allows for efficient approximation of $g_{\lambda,\mu}$ based on empirical risk minimization. In particular, if $n = +\infty$, we expect the minimizer (5) to converge toward $g_{\lambda,\mu}$ at rates in L^2 scaling similarly to $n_\ell^{-m/d}$ in n_ℓ. To complete the picture, depending on a prior on g_ρ, $g_{\lambda,\mu}$ might exhibit good convergence properties towards g_ρ as λ and μ go to zero. This contrasts with the problem encountered with $g_{\text{(naive)}}$. Those considerations are exactly what reproducing kernel Hilbert space will provide, additionally with a computationally friendly framework to perform the estimation. Note that quantity similar to $g_{\lambda,\mu}$ were considered in [61, 44].

Spectral filtering. Without looking for higher power-norm, [37] proposed to overcome the capacity issue by considering approximation of the operator $\mathcal{L}$ based on the graph-based technique provided by [4, 14] and to reduce the search of g_{n_ℓ} on the space spanned by the first few eigenvectors of the Laplacian. In particular, on Figure 1, g^* could be searched in the null space of $\mathcal{L}$, that is, among functions that are constant on each connected component of $\mathrm{supp}\,\rho_{\mathcal{X}}$. This technique exhibits two parts, the "unsupervised" estimation of $\mathcal{L}$ that will depend on the total number of data n, and the "supervised" search for g_ρ on the first few eigenvectors of $\mathcal{L}$ that will depend on the number of labels n_ℓ. While, at first sight, this technique seems to be completely different than Tikhonov regularization (4), it can be cast, along with gradient descent, into the same *spectral filtering* framework [30]. This point of view enables the use of a wide range of techniques offered by spectral manipulations on the diffusion operator $\mathcal{L}$.

This paper is motivated by the fact that current well-grounded semi-supervised learning techniques are implemented based on graph-based Laplacian, which is a local averaging method that does not leverage smartly functional capacity. In particular, as recalled earlier, graph-based Laplacian is known

to suffer from the curse of dimensionality, in the sense that the convergence of the empirical estimator $\widehat{\mathcal{L}}$ towards the $\mathcal{L}$ exhibits a rate of convergence of order $\mathcal{O}(n^{-1/d})$ with d the dimension of the input space $\mathcal{X}$ [24]. In this work, we will bypass this curse of dimensionality by looking for g in a smooth universal reproducing kernel Hilbert space, which will lead to efficient empirical estimates.

2 Spectral Filtering with Kernel Laplacian

In this section, we approach Laplacian regularization from a functional analysis perspective. We first introduce kernel methods and derivatives in reproducing kernel Hilbert space (RKHS). We then translate the considerations provided in Section 1.3 in the realm of kernel methods.

2.1 Kernel methods and derivatives evaluation maps

In this subsection, we introduce kernel methods (see [3, 46, 49] for more details Consider $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ a reproducting kernel Hilbert space, that is a Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$ such that the evaluation functionals $L_x : \mathcal{H} \to \mathbb{R}; g \to g(x)$ are continuous linear forms for any $x \in \mathcal{X}$. Such forms can be represented by $k_x \in \mathcal{H}$ such that, for any $g \in \mathcal{H}$, $L_x(g) = \langle k_x, g \rangle_{\mathcal{H}}$. A reproducing kernel Hilbert space can alternatively be defined from a symmetric positive semi-definite kernel $k : \mathcal{X} \to \mathcal{X} \to \mathbb{R}$, that is a function such that for any $n \in \mathbb{N}$ and $(x_i)_{i \leq n} \in \mathcal{X}^n$ the matrix $(k(x_i, x_j))_{i,j}$ is symmetric positive semi-definite, by building $(k_x)_{x \in \mathcal{X}}$ such that $k(x, x') = \langle k_x, k_{x'} \rangle_{\mathcal{H}}$. From a learning perspective, it is useful to use the evaluation maps to rewrite $\mathcal{H} = \{g_\theta : x \to \langle k_x, \theta \rangle_{\mathcal{H}} \mid \theta \in \mathcal{H}\}$. As such, kernel methods can be seen as "linear models" with features k_x, allowing to parameterize large spaces of functions [34]. In the following, we will differentiate θ seen as an element of $\mathcal{H}$ and g_θ seen as its embedding in L^2. To make this distinction formal, we define the embedding $S : (\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}}) \hookrightarrow (L^2(\rho_{\mathcal{X}}), \langle \cdot, \cdot \rangle_{L2}); \theta \to g_\theta$, as well as its adjoint $S^\star : L^2(\rho_{\mathcal{X}}) \to \mathcal{H}$.

Given a linear parametric model of functions $g_\theta(x) = \langle \theta, k_x \rangle_{\mathcal{H}}$, it is possible to compute derivatives of g_θ based on derivatives of the feature vector – think of $\mathcal{H} = \mathbb{R}^p$ and of $k_x = \varphi(x)$ as a feature vector with $\varphi : \mathbb{R}^d \to \mathbb{R}^p$. For $\alpha \in \mathbb{N}^d$, with $|\alpha| = \sum_{i \leq d} \alpha_i$, we have the following equality of partial derivatives, when k is $2|\alpha|$ times differentiable,

$$D^\alpha g_\theta(x) = \langle \theta, D^\alpha k_x \rangle, \qquad \text{where} \qquad D^\alpha = \frac{\partial^{|\alpha|}}{(\partial x_1)^{\alpha_1}(\partial x_2)^{\alpha_2} \cdots (\partial x_d)^{\alpha_d}}.$$

Here and $D^\alpha k_x$ has to be understood as the partial derivative of the mapping of $x \in \mathcal{X}$ to $k_x \in \mathcal{H}$, which can be shown to belong to $\mathcal{H}$ [61]. In the following, we assume that k is twice differentiable with continuous derivatives, and will make an extensive use of derivatives of the form $\partial_i k_x = \partial k_x / \partial x_i$ for $i \leq d$ and $x \in \mathcal{X}$. Note that, as well as we can describe completely the Hilbertian geometry of the space $\mathrm{Span}\{k_x \mid x \in \mathcal{X}\}$ through $k(x, x') = \langle k_x, k'_x \rangle$, for $x, x' \in \mathcal{X}$, we can describe the Hilbertian geometry of $\mathrm{Span}\{k_x \mid x \in \mathcal{X}\} + \mathrm{Span}\{\partial_i k_x \mid x \in \mathcal{X}\}$, through

$$\partial_{1,i} k(x, x') = \langle \partial_i k_x, k_{x'} \rangle_{\mathcal{H}}, \qquad \text{and} \qquad \partial_{1,i}\partial_{2,j} k(x, x') = \langle \partial_i k_x, \partial_j k_{x'} \rangle_{\mathcal{H}},$$

where $\partial_{1,i}$ denotes the partial derivative with respect to the i-th coordinates of the first variable. This echoes to so-called "representer theorems".

Example 1 (Gaussian kernel). *A classical kernel is the Gaussian kernel, also known as radial basis function, defined for $\sigma > 0$ as the following k, and satisfying, for $i \neq j$, the following equalities,*

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

$$\partial_{1,i} k(x, y) = -\frac{(x_i - y_i)}{\sigma^2} k(x, y), \qquad \partial_{1,i}\partial_{2,i} k(x, y) = \left(\frac{1}{\sigma^2} - \frac{(x_i - y_i)^2}{\sigma^4}\right) k(x, y),$$

where x_i designs the i-th coordinates of the vector $x \in \mathcal{X} = \mathbb{R}^d$.
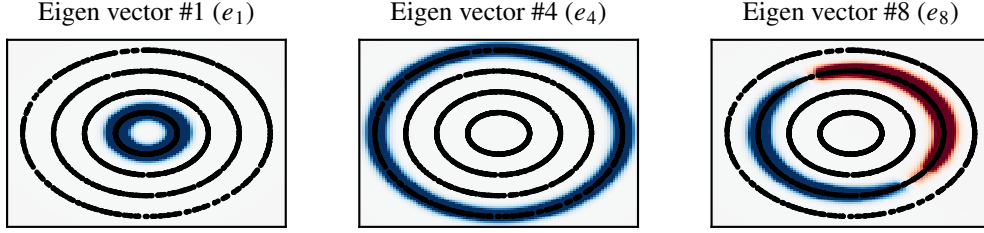
5

Figure 2: Few of the first generalized eigenvectors of $(\hat{\Sigma}; \hat{L} + \mu I)$ (with $\mu = 1/n$). The first four eigenvectors correspond to constant functions on each circle, as shown with e_1 and e_4. The few eigenvectors after correspond to second harmonics localized on a single circle as shown with e_8.

2.2 Tikhonov, spectral filtering and dimensionality reduction

Given the kernel k, its associated RKHS $\mathcal{H}$ and S the embedding of $\mathcal{H}$ in L^2, we rewrite Eq. (4) under its "parameterized" version

$$g_{\lambda,\mu} = S \arg\min_{\theta \in \mathcal{H}} \left\{ \|S\theta - g_\rho\|^2_{L^2(\rho_\mathcal{X})} + \lambda \left\|\mathcal{L}^{1/2}S\theta\right\|^2_{L^2(\rho_\mathcal{X})} + \lambda\mu \|\theta\|^2_{\mathcal{H}} \right\}. \tag{4}$$

Do not hesitate to refer to Table 1 to keep track of notations. In the following, we will use that $\left\|\mathcal{L}^{1/2}S\theta\right\|^2_{L^2(\rho_\mathcal{X})} + \mu \|\theta\|^2_{\mathcal{H}} = \left\|(S^\star\mathcal{L}S + \mu I)^{1/2}\theta\right\|^2_{\mathcal{H}}$. This equality explains why we consider $\mu\lambda$ instead of μ in the last term. In the RKHS setting, the study of Eq. (4) unveils the three operators Σ, L, and I on $\mathcal{H}$, (indeed $g_{\lambda,\mu} = S \arg\min_{\theta \in \mathcal{H}} \{\theta^\star(\Sigma + \lambda L + \lambda\mu)\theta - 2\theta^\star S^\star g_\rho\}$) where I is the identity, and, as we detail in Appendix C,

$$\Sigma = S^\star S = \mathbb{E}_{X \sim \rho_\mathcal{X}}\left[k_X \otimes k_X\right], \qquad \text{and} \qquad L = S^\star\mathcal{L}S = \mathbb{E}_{X \sim \rho_\mathcal{X}}\left[\sum_{i=1}^{d} \partial_j k_X \otimes \partial_j k_X\right]. \tag{6}$$

Regularization and spectral filtering have been well-studied in the inverse-problem literature. In particular, the regularization Eq. (4) is known to be linked with the generalized singular value decomposition of $[\Sigma; L + \mu I]$ (see, *e.g.*, [18]), which is linked to the generalized eigenvalue decomposition of $(\Sigma, L + \mu I)$ [22]. We derive the following characterization of Eq. (4), whose proof is reported in Appendix D.

Proposition 1. *Let* $(\lambda_{i,\mu})_{i \in \mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$, $(\theta_{i,\mu})_{i \in \mathbb{N}} \in \mathcal{H}^{\mathbb{N}}$ *be the generalized eigenvalue decomposition of the pair* $(\Sigma, L + \mu I)$*, that is* $(\theta_{i,\mu})$ *generating* $\mathcal{H}$ *and such that for any* $i, j \in \mathbb{N}$, $\Sigma\theta_{i,\mu} = \lambda_{i,\mu}(L + \mu I)\theta_{i,\mu}$, *and* $\langle\theta_{i,\mu}, (L + \mu I)\theta_{j,\mu}\rangle = \mathbf{1}_{i=j}$. *Eq. (4) can be rewritten as*

$$g_{\lambda,\mu} = \left(\sum_{i \in \mathbb{N}} \psi(\lambda_{i,\mu})S\theta_{i,\mu} \otimes S\theta_{i,\mu}\right) g_\rho = \sum_{i \in \mathbb{N}} \psi(\lambda_{i,\mu}) \langle S^\star g_\rho, \theta_{i,\mu}\rangle S\theta_{i,\mu}, \tag{7}$$

with $\psi : \mathbb{R}_+ \to \mathbb{R}; x \to (x + \lambda)^{-1}$. *Eq. (7) should be seen as a specific instance of spectral filtering based on a filter function* $\psi : \mathbb{R}_+ \to \mathbb{R}$.

Interestingly, the generalized eigenvalue decomposition of the pair $(\Sigma, L + \mu I)$ was already considered by Pillaud-Vivien [40] to estimate the first eigenvalue of the Laplacian. Moreover, Pillaud-Vivien [41] suggests to leverage this decomposition for dimensionality reduction based on the first eigenvectors of the Laplacian. As well as Eq. (4) contrasts with graph-based semi-supervised learning techniques, this dimensionality reduction technique contrasts with methods based on graph Laplacian provided by [4, 14]. Remarkably, the semi-supervised learning algorithm that consists in using the unsupervised data to perform dimensionality reduction based on the Laplacian eigenvalue decomposition, before solving a small linear regression problem on the small resulting space, can be seen as a specific instance of spectral filtering, based on regularization by thresholding/cutting-off eigenvalue, which corresponds to $\psi : x \to x^{-1}\mathbf{1}_{x>\lambda}$ for a given threshold $\lambda > 0$ in Eq. (7).

3 Implementation

In this section, we discuss on how to practically implement estimates for Eq. (7) based on empirical data $(X_i, Y_i)_{i \le n_\ell}$ and $(X_i)_{n_\ell < i \le n}$. We first review how we can approximate the integral operators of

Eq. (6) based on data. We then discuss on how to implement our methods practically on a computer. We end this section by considering approximations that allow to cut down high computational costs associated with kernel methods involving derivatives.

Algorithm 1: Empirical estimates based on spectral filtering.

Data: $(X_i, Y_i)_{i \leq n_\ell}$, $(X_i)_{n_\ell < i \leq n}$, a kernel k, a filter ψ, a regularizer μ
Result: $\hat{g}_p$ through $c \in \mathbb{R}^p$ defining $\hat{g}_p(x) = \sum_{i=1}^n c_i k(x, X_i) = k_x^\star T_a c$
Compute $S_n T_a = (k(X_i, X_j))_{i \leq n, j \leq p} \in \mathbb{R}^{n \times p}$ in $\mathcal{O}(pn)$
Compute $Z_n T_a = (\partial_{1,j} k(X_l, X_i))_{(j \leq d, l \leq n), i \leq p} \in \mathbb{R}^{nd \times p}$ in $\mathcal{O}(pnd)$
Build $T_a^\star \hat{\Sigma} T_a = n^{-1} (S_n T_a)^\top (S_n T_a)$ in $\mathcal{O}(p^2 n)$
Build $T_a^\star \hat{L} T_a = n^{-1} (Z_n T_a)^\top (Z_n T_a)$ in $\mathcal{O}(p^2 nd)^a$
Build $T_a^\star T_a = (k(X_i, X_j))_{i,j \leq p} \in \mathbb{R}^{p \times p}$ in $\mathcal{O}(1)$ as a partial copy of $S_n T_a$
Get $(\lambda_{i,\mu}, u_{i,\mu})_{i \leq n}$ the generalized eigenelements of $(T_a^\star \hat{\Sigma} T_a, T_a^\star (\hat{L} + \mu I) T_a)$ in $\mathcal{O}(p^3)$
Get $b = T_a^\star \hat{\theta} = (n_\ell^{-1} \sum_{i=1}^{n_\ell} Y_i k(X_i, X_j))_{j \leq p} \in \mathbb{R}^p$ in $\mathcal{O}(pn_\ell)$
Return $c = \sum_{i=1}^n \psi(\lambda_i) u_i u_i^\top b \in \mathbb{R}^p$ in $\mathcal{O}(p^3)$.

---

^{a}Building this matrix can be avoided by using the generalized singular value decomposition rather than the generalized eigenvector decomposition. Implemented with Lapack, such a procedure will also requires $O(p^2 nd)$ floating point operations, but with a smaller constant in the big O [22].

3.1 Integral operators approximation

The classical empirical risk minimization in Eq. (5) can be understood as the plugging of the approximate distributions $\hat{\rho} = n_\ell^{-1} \sum_{i=1}^{n_\ell} \delta_{X_i} \otimes \delta_{Y_i}$ and $\hat{\rho}_\mathcal{X} = n^{-1} \sum_{i=1}^n \delta_{X_i}$ instead of ρ and $\rho_\mathcal{X}$ in Eq. (4). It can also be understood as the same replacement when dealing with integral operators, leading to the three following important quantities to rewrite Eq. (7),

$$\hat{\Sigma} := n^{-1} \sum_{i=1}^n k_{X_i} \otimes k_{X_i}, \quad \hat{L} := n^{-1} \sum_{i=1}^n \sum_{j=1}^d \partial_j k_{X_i} \otimes \partial_j k_{X_i}, \quad \hat{\theta} := \widehat{S^\star g_\rho} := n_\ell^{-1} \sum_{i=1}^{n_\ell} Y_i k_{X_i}. \quad (8)$$

It should be noted that while considering n in the definition of $\hat{\Sigma}$ is natural from the spectral filtering perspective, to make it formally equivalent with the empirical risk minimization (5), it should be replaced by n_ℓ. Eq. (8) allows to rewrite Eq. (7) without relying on the knowledge of ρ, by considering $(\hat{\lambda}_{i,\mu}, \hat{\theta}_{i,\mu})$ the generalized eigenvalue decomposition of $(\hat{\Sigma}, \hat{L})$ and considering

$$\hat{g} = \sum_{i \in \mathbb{N}} \psi(\hat{\lambda}_{i,\mu}) \left\langle \widehat{S^\star g_\rho}, \hat{\theta}_{i,\mu} \right\rangle S \hat{\theta}_{i,\mu}, \quad (9)$$

We present the first eigenvectors (after plunging them in L^2 through S) of the generalized eigenvalue decomposition of $(\hat{\Sigma}, \hat{L} + \mu I)$ on Figure 2. The first eigenvectors allow to recover the null space of $\mathcal{L}$. This explains clearly the behavior on the right of Figure 1.

3.2 Matrix representation and approximation of operators

Currently, we are dealing with operators $(\hat{\Sigma}, \hat{L})$ and vectors (*e.g.*, $\hat{\theta}$) in the Hilbert space $\mathcal{H}$. It is natural to wonder on how to represent this on a computer. The answer is the object of representer theorems (see Theorem 1 of [61]), and consists in noticing that all the objects introduced are actually defined in, or operate on, $\mathcal{H}_n + \mathcal{H}_{n,\partial} \subset \mathcal{H}$, with $\mathcal{H}_n = \text{Span}\{k_{X_i} \,|\, i \leq n\}$ and $\mathcal{H}_{n,\partial} = \text{Span}\{\partial_j k_{X_i} \,|\, i \leq n, j \leq d\}$. This subspace of $\mathcal{H}$ is of dimension at most $n(d+1)$ and if $T : \mathbb{R}^p \to \mathcal{H}_n + \mathcal{H}_{n,\partial}$ (with $p \leq n(d+1)$) parameterizes $\mathcal{H}_n + \mathcal{H}_{n,\partial}$, our problem can be cast in $\mathbb{R}^p$ by considering the $p \times p$ matrices $T^\star \hat{\Sigma} T$ and $T^\star (\hat{L} + \mu I) T$ instead of the operators $\hat{\Sigma}$ and $\hat{L} + \mu I$. The canonical representation consists in taking $p = n(d+1)$ and considering for $c \in \mathbb{R}^{n(d+1)}$, the mapping $T_c c = \sum_{i=1}^n c_{i0} k_{X_i} + \sum_{j=1}^d c_{ij} \partial_j k_{X_i}$ [61, 44].

This exact implementation implies dealing and finding the generalized eigen value decomposition of $p \times p$ matrices with $p = n(d+1)$, which leads to computational costs in $\mathcal{O}(n^3 d^3)$, which can be prohibitive. Two solutions are known to cut down prohibitive computational costs of kernel methods.

Both methods consist in looking for a space that can be parameterized by $\mathbb{R}^p$ for a small p and that approximates well the space $\mathcal{H}_n + \mathcal{H}_{n,\partial} \subset \mathcal{H}$. The first solution is provided by random features [42]. It consists in approximating $\mathcal{H}$ with a space of small dimension $p \in \mathbb{N}$, linked with an explicit representation $\varphi : \mathcal{X} \to \mathbb{R}^p$ that approximate $k(x, x') \simeq k_\varphi(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathbb{R}^p}$. In theory, it substitutes the kernel k by k_φ. In practice, all computations can be done with the explicit feature φ.

Approximate solution. The second solution, which we are going to use in this work, consists in approximating $\mathcal{H}_n + \mathcal{H}_{n,\partial}$ by $\mathcal{H}_p = \text{Span} \{k_{X_i}\}_{i \leq p}$ for $p \leq n$. This method echoes the celebrated Nyström method [55], as well as the Rayleigh–Ritz method for Sturm–Liouville problems. In essence, [45] shows that, when considering subsampling based on leverage score, $p = n^\gamma \log(n)$, with $\gamma \in (0, 1]$ linked to the "size" of the RKHS and the regularity of the solution, is a good enough approximation, in the sense that it only downgrades the sample complexity by a constant factor. In theory, we know that the space $\mathcal{H}_p$ will converge to $\mathcal{H} = \text{Closure Span} \{k_x\}_{x \in \text{supp} \rho_{\mathcal{X}}}$ as p goes to infinity. In practice, it means considering the approximation mapping $T_a : \mathbb{R}^p \to \mathcal{H}; c \to \sum_{i=1}^p c_i k_{X_i}$, and dealing with the $p \times p$ matrices $T_a^\star \Sigma T_a$ and $T_a^\star L T_a$. It should be noted that the computation of $T_a^\star L T_a$ requires to multiply a $p \times nd$ matrix by its transpose. Overall, training this method can be done with $\mathcal{O}(p^2 nd)$ basic operations, and inference with this method can be done in $\mathcal{O}(p)$. The saving cost of this approximate method is huge: without compromising the precision of our estimator, we went from $O(n^3 d^3)$ run time complexities to $O(\log(n)^2 n^{1+2\gamma} d)$ computations, with γ possibly very small. Similarly, the memory cost went from $O(n^2 d^2)$ down to $O(nd + n^{2\gamma})$.[1]

4   Statistical analysis

In this section, we are interested in quantifying the risk of the learnt mapping $\hat{g}$. We study it through the generalization bound, which consists in obtaining a bound on the averaged excess risk $\mathbb{E}_{\text{data}} \|\hat{g} - g_\rho\|_{L^2}^2$. In particular, we want to answer the following points.

1. How, and under which assumptions, Laplacian regularization boost learning?
2. How the excess of risk relates to the number of labelled and unlabelled data?

In terms of priors, we want to leverage a low-density separation hypothesis. In particular, we can suppose that when diffusing g_ρ with $e^{-t\mathcal{L}}$ we stay close to g_ρ, or that g_ρ is supported on a finite dimensional space of functions on which $\|\mathcal{L}^{1/2} g\|$ (which measures the variation of g) is small. Both those assumptions can be made formal by assuming the g_ρ is supported by the first eigenvectors of the diffusion operator $\mathcal{L}$.

Assumption 1 (Source condition)**.** *g_ρ is supported on a finite dimensional space that is left stable by the diffusion operator $\mathcal{L}$. In other terms, if $(e_i) \in (L^2)^{\mathbb{N}}$ are the eigenvectors of $\mathcal{L}$, there exists $r \in \mathbb{N}$, such that $g_\rho \in \text{Span} \{e_i\}_{i \leq r}$.*

We will also assume that the diffusion operator $\mathcal{L}$ can be well approximated by the RKHS associated with k. In practice, under mild assumptions, *c.f.* Appendix C, the eigenvectors of the Laplacian are known to be regular, in particular to belong to H^m for $m \in \mathbb{N}$ bigger than d. As such, many classical kernels would allow to verify the following assumption.

Assumption 2 (Approximation condition)**.** *The eigenvectors (e_i) of $\mathcal{L}$ belongs to the RKHS $\mathcal{H}$.*

We add one technical assumptions regarding the eigenvalue decay of the operator Σ compared to the operator L, with $\preceq$ denoting the Löwner order (*i.e.*, for A and B symmetric, $A \preceq B$ if $B - A$ is positive semi-definite).

Assumption 3 (Eigenvalue decay)**.** *There exists $a \in [0, 1]$ and $c > 0$ such that $L \preceq c\Sigma^\alpha$.*

Note that, in our setting, L is compact and bounded and Assumption 3 is always satisfied with $a = 0$. For translation-invariant kernel, such as Gaussian or Laplace kernels, based on considerations linking eigenvalue decay of operators with functional space capacities [49], under mild assumptions, we can take $a > 1 - 2/d$. We discuss all assumptions in more details in Appendix C.

To study the consistency of our algorithms, we can reuse the extensive literature on kernel ridge regression [10, 30]. This literature body provides an extensive picture on convergence rates relying

---
[1]Our code is available online at `https://github.com/VivienCabannes/partial_labelling`.
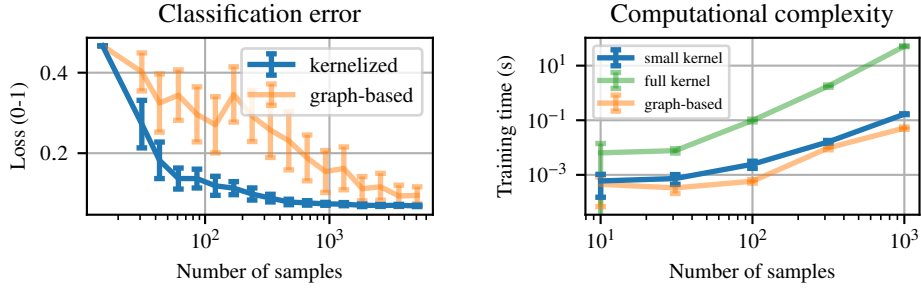
Figure 3: (Left) Comparison between our kernelized Laplacian method (Tikhonov regularization version with $\lambda = 1$, $\mu_n = n^{-1}$, $p = 50$) and graph-based Laplacian based on the same Gaussian kernel with bandwidth $\sigma_n = n^{-\frac{1}{d+4}} \log(n)$ as suggested by graph-based theoretical results [24]. We report classification error as a function of the number of samples n. The error is averaged over 50 trials, with errorbars representing standard deviations. We fixed the ratio n_ℓ / n to one tenth, and generated the data according to two Gaussians in dimension $d = 10$ with unit variance and whose centers are at distance $\delta = 3$ of each other (similar to the setting of [11, 29]). Our method discovers the structure of the data much faster than graph-based Laplacian (to get a 20% error we need 40 points, while graph-based need 700). (Right) Time to perform training with graph-based Laplacian in orange, with Algorithm 1 in blue (with the specification of the left figure), and with the naive representation in $\mathbb{R}^{n(d+1)}$ of the empirical minimizer Eq. (5) in green. When dealing with 1000 points, our algorithm, as well as graph-based Laplacian, can be computed in about one tenth of a second on a 2 GHz processor, while the naive kernel implementation requires 10 seconds. We show in Appendix B that this cut in costs is not associated with a loss in performance.

on various filters and assumptions of capacity, *a.k.a* effective dimension, and source conditions. Our setting is slightly different and showcases two specificities: (*i*) the eigenelements $(\lambda_{i,\mu}, \theta_{i,\mu})$ are dependent of μ; (*ii*) the low-rank approximation in Algorithm 1 is specific to settings with derivatives. We end our exposition with the following convergence result, proven in Appendix E. Note that the dependency of p in n can be improved based on subsampling techniques that leverage expressiveness of the different (k_{X_i}) [45]. Moreover, universal consistency results could also be provided when the RKHS is dense in H^1, as well as convergence rates for other filters and laxer assumptions which we discuss in Appendix E (in particular, the source condition can be relaxed by considering the biggest $q \in (0, 1]$ such that $g \in \operatorname{im} \mathcal{L}^q$).

Theorem 1 (Convergence rates). *Under Assumptions 1, 2 and 3, for $n_\ell, n \in \mathbb{N}$, when considering the spectral filtering Algorithm 1 with $\psi_\lambda : x \to (x + \lambda)^{-1}$, there exists a constant C independent of n, n_ℓ, λ, μ and p such that the estimate $\hat{g}_p$ defined in Algorithm 1 verifies*

$$\mathbb{E}_{\mathcal{D}_n}\left[\|\hat{g}_p - g_\rho\|_{L^2}^2\right] \leq C\left(\lambda^2 + \lambda\mu + \frac{\sigma_\ell^2 n_\ell^{-1} + n_\ell^{-2} + n^{-1}}{\lambda\mu} + \frac{\log(p)}{p} + \lambda\frac{\log(p)^a}{p^a}\right), \quad (10)$$

with σ_ℓ^2 is a variance parameter that relates to the variance of the variable $Y(I + \lambda\mathcal{L})^{-1}\delta_X$, inheriting its randomness from $(X, Y) \sim \rho$. In particular, when the ratio $r = n_\ell/n$ is fixed, with the regularization scheme $\lambda_n = \lambda_0 n^{-1/4}$, $\mu_n = \mu_0 n^{-1/4}$, for any $\lambda_0 > 0$ and $\mu_0 > 0$, and the subsampling scheme $p_n = p_0 n^s \log(n)$ for any $p_0 > 0$ and with $s = \max(1/2, 1/4a)$, there exists a constant C' independent of n and n_ℓ such that the excess of risk verifies

$$\mathbb{E}_{\mathcal{D}_n}\left[\|\hat{g}_p - g_\rho\|_{L^2}^2\right] \leq C'(n^{-1/2} + \sigma_\ell^2 n_\ell^{-1/2}). \quad (11)$$

Theorem 1 answers the two questions asked at the beginning of this section. In particular, it characterizes the dependency of the need for labelled data to a variance parameter linked with the diffusion of observations (X_i, Y_i) based on the density $\rho_{\mathcal{X}}$ through the operator $\mathcal{L}$. Intuitively, if the index set $I \subset \{1, 2, \cdots, n\}$ of data $(X_i)_{i \in I}$ we labelled does not change the profile of the diffusion solution $\hat{g}$, then we do not need that much labelled data – as this is the case on Figure 1.

Finally, Theorem 1 is remarkable in that it exhibits no dependency to the dimension of $\mathcal{X}$ in the power of n and n_ℓ. This contrasts with graph-based Laplacian methods that do not scale well with the input space dimensionality [5, 24]. Indeed, Figure 3 shows the superiority of our method over graph-based Laplacian in dimension $d = 10$, with a mixture of Gaussians. We provide details as well as additional experiments in Appendix B.

9

5 Conclusion

Diffusing information or enforcing regularity through penalties on derivatives are natural ideas to tackle many machine learning problems. Those ideas can be captured informally with graph-based techniques and finite element differences, or captured more formally with the diffusion operator we introduced in this work. This formalization allowed us to shed lights on Laplacian regularization techniques based on statistical learning considerations. In order to make our method usable in practice, we provided strong computational guidelines to cut down prohibitive cost associated with a naive implementation of our methods. In particular, we were able to develop computationally efficient semi-supervised techniques that do not suffer from the curse of dimensionality.

This work paves the way to many extensions beyond semi-supervised learning. For example, in Appendix A, we describe its usefulness to the partial supervised learning problem, where minimizing the Dirichlet energy provide a learning principle, in order to bypass the restrictive non-ambiguity assumption usually made in this setup [15, 31, 7, 8]. Moreover, in the context of active learning, retaking the strategy of Karzand and Nowak [27], this energy provides a computationally-effective, theoretically-grounded, data-dependent score to select the next point to query. As such, follow-ups would be of interest to see how this introductory theoretical paper makes its way into the world of concrete applications.

Acknowledgments and Disclosure of Funding

References

[1] Ahmed El Alaoui, Xiang Cheng, Aaditya Ramdas, Martin Wainwright, and Michael Jordan. Asymptotic behavior of ℓ_p-based Laplacian regularization in semi-supervised learning. In *Conference on Learning Theory*, 2016.

[2] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *Conference on Computer Vision and Pattern Recognition*, 2016.

[3] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 1950.

[4] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 2003.

[5] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. Label propagation and quadratic criterion. In *Semi-Supervised Learning*. MIT Press, 2006.

[6] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Neural Information Processing Systems*, 2019.

[7] Vivien Cabannes, Alessandro Rudi, and Francis Bach. Structured prediction with partial labelling through the infimum loss. In *International Conference on Machine Learning*, 2020.

[8] Vivien Cabannes, Alessandro Rudi, and Francis Bach. Disambiguation of weak supervision with exponential convergence rates. In *International Conference on Machine Learning*, 2021.

[9] Vivien Cabannes, Alessandro Rudi, and Francis Bach. Fast rates in structured prediction. In *Conference on Learning Theory*, 2021.

[10] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 2006.

[11] Vittorio Castelli and Thomas Cover. On the exponential value of labeled samples. *Pattern Recognition Letters*, 1995.

[12] Lin Chen and Sheng Xu. Deep neural tangent kernel and Laplace kernel have the same RKHS. In *International Conference on Learning Representations*, 2021.

[13] Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A general framework for consistent structured prediction with implicit loss embeddings. *Journal of Machine Learning Research*, 2020.

[14] Ronald Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 2006.

[15] Timothée Cour, Benjamin Sapp, and Ben Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 2011.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, 2009.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[18] Alan Edelman and Yuyang Wang. The GSVD: where are the ellipses?, matrix trigonometry, and more. *Journal on Matrix Analysis and Applications*, 2020.

[19] David Eriksson, Kun Dong, Eric Hans Lee, David Bindel, and Andrew Gordon Wilson. Scaling Gaussian process regression with derivatives. In *Neural Information Processing Systems*, 2018.

[20] Evelyn Fix and Joseph Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. Technical report, SAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[21] Nicolás García Trillos, Moritz Gerlach, Matthias Hein, and Dejan Slepčev. Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace–Beltrami operator. *Foundations of Computational Mathematics*, 2019.

[22] Gene Golub and Charles Van Loan. *Matrix computations (4th editions)*. Johns Hopkins University Press, 2013.

[23] Yves Grandvalet. Logistic regression for partial labels. In *9th Information Processing and Management of Uncertainty*, 2002.

[24] Matthias Hein, Jean-Yves Audibert, and Ulrike von Luxburg. Graph Laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 2007.

[25] Eyke Hüllermeier, Johannes Fürnkranz, Weiwei Cheng, and Klaus Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 2008.

[26] Rong Jin and Zoubin Ghahramani. Learning with multiple labels. In *Neural Information Processing Systems*, 2002.

[27] Mina Karzand and Robert Nowak. Maximin active learning in overparameterized model classes. *IEEE Journal on Selected Areas in Information Theory*, 2020.

[28] Andrey Kolmogorov and Vladimir Tikhomirov. ε-entropy and ε-capacity of sets in functional spaces. *Uspekhi Matematicheskikh Nauk*, 1959.

[29] Marc Lelarge and Léo Miolane. Asymptotic Bayes risk for Gaussian mixture in a semi-supervised setting. In *International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, 2019.

[30] Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Applied and Computational Harmonic Analysis*, 2020.

[31] Li-Ping Liu and Thomas Dietterich. Learnability of the superset label learning problem. In *International Conference on Machine Learning*, 2014.

[32] Xinghua Lou and Fred Hamprecht. Structured learning from partial annotations. In *International Conference on Machine Learning*, 2012.

[33] Jie Luo and Francesco Orabona. Learning from candidate labeling sets. In *Neural Information Processing Systems*, 2010.

[34] Charles Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 2006.

[35] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *International Conference on Computer Vision*, 2019.

[36] Stanislav Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 2017.

[37] Boaz Nadler, Nathan Srebro, and Xueyuan Zhou. Statistical analysis of semi-supervised learning: The limit of infinite unlabelled data. In *Neural Information Processing Systems*, 2009.

[38] Nam Nguyen and Rich Caruana. Classification with partial labels. In *14th International Conference on Knowledge Discovery and Data Mining*, 2008.

[39] George Papandreou, Liang-Chieh Chen, Kevin Murphy, and Alan Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *International Conference on Computer Vision*, 2015.

[40] Loucas Pillaud-Vivien. Statistical estimation of the poincaré constant and application to sampling multimodal distributions. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[41] Loucas Pillaud-Vivien. *Learning with Reproducing Kernel Hilbert Spaces: Stochastic Gradient Descent and Laplacian Estimation*. Phd thesis, Ecole normale supérieure, 2020.

[42] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Neural Information Processing Systems*, 2007.

[43] Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 2007.

[44] Lorenzo Rosasco, Silvia Villa, Sofia Mosci, Matteo Santoro, and Alessandro verri. Nonparametric sparsity and regularization. *Journal of Machine Learning Research*, 2013.

[45] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Neural Information Processing Systems*, 2015.

[46] Bernhard Scholkopf and Alex Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[47] Alexander Smola and Risi Kondor. Kernels and regularization on graphs. In *Conference on Computational Learning Theory*, 2003.

[48] Boumediene Hamzi Stefan Klus, Feliks Nüske. Kernel-based approximation of the koopman generator and schrödinger operator. *Entropy*, 2020.

[49] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

[50] James Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 1958.

[51] Alan Turing. Computing machinery and intelligence. *Mind*, 1950.

[52] Jesper van Engelen and Holger Hoos. A survey on semi-supervised learning. *Machine Learning Journal*, 2020.

[53] Jakob Verbeek and William Triggs. Scene Segmentation with CRFs Learned from Partially Labeled Images. In *Neural Information Processing Systems*, 2008.

[54] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *International Joint Conference on Artificial Intelligence*, 2019.

[55] Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Neural Information Processing Systems*, 2000.

[56] Yuhong Yang. Minimax nonparametric classification. I. Rates of convergence. *IEEE Transactions on Information Theory*, 1999.

[57] Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon. Large-scale multi-label learning with missing labels. In *International Conference on Machine Learning*, 2014.

[58] Vadim Vladimirovich Yurinskii. On an infinite-dimensional version of S. N. Bernstein's inequalities. *Theory of Probability and Its Applications*, 1970.

[59] Mingyuan Zhang and Shivani Agarwal. Bayes consistency vs. $\mathcal{H}$-consistency: The interplay between surrogate loss functions and the scoring function class. In *Neural Information Processing Systems*, 2020.

[60] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Neural Information Processing Systems*, 2003.

[61] Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 2008.

[62] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *International Conference of Machine Learning*, 2003.

Ethical considerations

This work aims at advancing our understanding of weakly supervised learning. Weakly supervised learning enrolls in the quest of an automated artificial intelligence, free from the need of human supervision. Automation, which is at the basis of computer science [51], is known to increase productivity at a reduced human labor cost, and is associated with several political/societal issues. In term of concrete applications, reducing the need for annotations is especially useful when humans reproduce biases when annotating data, or when the lack of output annotation restricts the outreach of a method (*e.g.*, learning to translate languages by collecting input/output pairs based on books already translated by humans can hardly be applied to languages with few written resources).

A Extensions: Least-square surrogate and partially supervised learning

In this section, we first show how our work can be extended to generic semi-supervised learning problem, beyond real-valued regression. This first extension is based on the least-square surrogate introduced by Ciliberto et al. [13] for structured prediction problems. We later show how our work can be extended to generic partially-supervised learning. This second extension is based on the work of Cabannes et al. [7].

A.1 Structured prediction and least-square surrogate

Until now, we have considered the least-square problem with $Y \in \mathbb{R}$. Indeed, our work can be extended easily to a wide class of learning problem. Consider $\mathcal{Y}$ an output space, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a loss function, and keep $\mathcal{X} \subset \mathbb{R}^d$ and $\rho \in \Delta_{\mathcal{X} \times \mathcal{Y}}$. Suppose that we want to retrieve

$$f^* = \underset{f : \mathcal{X} \to \mathcal{Y}}{\arg\min} \mathcal{R}(f), \qquad \text{with} \qquad \mathcal{R}(f) = \mathbb{E}_{(X,Y) \sim \rho} \left[\ell(f(X), Y) \right]. \tag{12}$$

Ciliberto et al. [13] showed that as soon as ℓ can be decomposed through two mappings $\varphi : \mathcal{Y} \to \mathcal{H}_{\mathcal{Y}}$ and $\psi : \mathcal{Y} \to \mathcal{H}_{\mathcal{Y}}$ with $\mathcal{H}_{\mathcal{Y}}$ a Hilbert space as $\ell(y, z) = \langle \varphi(y), \psi(z) \rangle_{\mathcal{H}_{\mathcal{Y}}}$, it is possible to leverage the least-square regression by considering the surrogate problem

$$g^* \in \underset{g : \mathcal{X} \to \mathcal{H}_{\mathcal{Y}}}{\arg\min} \mathbb{E}_{(X,Y) \sim \rho} \left[\| g(X) - \varphi(Y) \|_{\mathcal{H}_{\mathcal{Y}}}^2 \right]. \tag{13}$$

This surrogate problem relates to the original one through the decoding d that relates a surrogate estimate $g : \mathcal{X} \to \mathcal{H}_{\mathcal{Y}}$ to an estimate of the original problem $f : \mathcal{X} \to \mathcal{Y}$ as $f = d(g)$ defined through, for $x \in \operatorname{supp} \rho_{\mathcal{X}}$,

$$f(x) = \underset{z \in \mathcal{Y}}{\arg\min} \langle \psi(z), g(x) \rangle_{\mathcal{H}_{\mathcal{Y}}}. \tag{14}$$

In the real-valued regression case, presented precedently, our estimates for g_n can all be written as $g_n(x) = \sum_{i=1}^{n_\ell} \beta_i(x) Y_i$, where $\beta_i(x)$ is a function of the $(X_i)_{i \leq n}$, involving the kernel k and its derivatives. Those estimates can be cast to vector-valued regression by considering coordinates-wise regression[2], which leads to $g_n(x) = \sum_{i=1}^{n_\ell} \beta_i(x) \varphi(Y_i)$, and to the original estimates, for any $x \in \operatorname{supp} \rho_{\mathcal{X}}$,

$$f_n(x) \in \underset{z \in \mathcal{Z}}{\arg\min} \sum_{i=1}^{n_\ell} \beta_i(x) \ell(z, Y_i). \tag{15}$$

The behavior of f_n being independent of the decomposition (φ, ψ) of ℓ was referred to as the loss trick. In particular, Ciliberto et al. [13] showed that convergence rates derived between $\| g_n - g^* \|_{L^2}$ does not change if we consider $g : \mathcal{X} \to \mathbb{R}$ or $g : \mathcal{X} \to \mathcal{H}_{\mathcal{Y}}$ and that those rates can be cast directly as convergence rates between $\mathcal{R}(f_n)$ and $\mathcal{R}(f^*)$ with $f_n = d(g_n)$ defined by Eq.(14). Moreover, when $\mathcal{Y}$ is a discrete output space, it is possible to get much better generalization bound on $\mathcal{R}(f_n) - \mathcal{R}^*$ by introducing geometrical considerations regarding g^* and decision frontier between classes [9].

---

[2]To parameterize functions g from $\mathcal{X}$ to $\mathcal{H}_{\mathcal{Y}}$, we can parameterized independently each coordinates $\langle g, e_i \rangle_{\mathcal{H}_{\mathcal{Y}}}$, for (e_i) a basis of $\mathcal{H}_{\mathcal{Y}}$, by the space $\mathcal{G}$ – note that it is possible to generalize real-valued kernel to parameterize coordinates in a joint fashion [10]. The coordinate-wise parameterization corresponds to the tensorization $\mathcal{H}' = \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}$ and to the parametric space $\mathcal{G}' = \{ x \to \Theta k_x \,|\, \Theta \in \mathcal{H}' \}$ of functions from $\mathcal{X}$ to $\mathcal{H}_{\mathcal{Y}}$. $\mathcal{G}'$ naturally inherits of the Hilbertian structure of $\mathcal{H}'$, itself inherited from the structure of $\mathcal{H}$ and $\mathcal{H}_{\mathcal{Y}}$.

Example 2 (Binary classification). *This framework aims at generalizing well known surrogate considerations in the case of the binary classification. Binary classification corresponds to $\mathcal{Y} = \{-1, 1\}$, ℓ the $0 - 1$ loss. In this setting, $\mathcal{H}_{\mathcal{Y}} = \mathbb{R}$, $\varphi : \mathcal{Y} \to \mathbb{R}; y \to y$, and $\psi = -\varphi$. This definition verifies $\ell(y, z) = .5 - .5\varphi(y)^\top \varphi(z) \simeq \varphi(y)^\top \psi(z)$. This corresponds to the usual least-square surrogate, which is $\mathcal{R}_S(g) = \mathbb{E}[\|g(X) - Y\|^2]$, $g(x) = \mathbb{E}[Y \mid X = x]$ and $f = \text{sign } g$.*

Beyond least-squares. Considering a least-square surrogate assumes that retrieving g^* (13) is the way to solve the original problem (12) and that the low-density separation hypothesis can be expressed as Assumption 1 being verified by g^*. We would like to point out that the low-density separation could be expressed under a much weaker form, which is that there exists g such that $f^* = d(g)$ (14) and g verifies Assumption 1. In particular, the cluster assumption [43] could be understood as assuming that $g = \varphi(f^*)$, the trivial embedding of f^* in $\mathcal{H}_{\mathcal{Y}}$, is constant on clusters, with means that g belongs to the kernel of the Laplacian operator $\mathcal{L}$. Yet, $g^* : x \to \mathbb{E}[\varphi(Y)|X = x]$, which depends on the labelling noise, could be really non-smooth, even under the cluster assumption. Those considerations are related to an open problem in machine learning, which is that we do not know what is the best statistical way (and the best surrogate problem) to solve the fully supervised binary classification problem [see *e.g.* 59]. However, many points introduced in the work could be retaken with other surrogate, could it be SVM (which leads to $g^* = \varphi(f^*)$, with g^* minimizing the Hinge loss), softmax regression (used in deep learning) or others.

A.2 Partially supervised learning

Partial supervision is a popular instance of weak supervision, which generalizes semi-supervised learning. It has been known under the name of partial labelling [15], superset learning [31], as well as learning with partial label [23], with partial annotation [32], with candidate labeling set [33] or with multiple label [26]. It encompasses many problems such as "classification with partial labels" [38, 15], "multilabelling with missing labels" [57], "ranking with partial ordering" [25], "regression with censored data" [50], "segmentation with pixel annotation" [53, 39], as well as instances of "action retrieval", especially on instructional videos [2, 35].

It consists, for a given input x, in not observing its label $y \in \mathcal{Y}$, but observing a set of potential labels $s \in 2^{\mathcal{Y}}$ that contains the labels ($y \in s$). Typically, if $\mathcal{Y}$ is the space $\mathfrak{S}_m$ of orderings between m items (*e.g.* movies on a streaming website), for a given input x (*e.g.* some feature vectors characterizing a user) s might be specified by a partial ordering that the true label y should satisfy (*e.g.* the user prefers romantic movies over action movies).

In this setting, it is natural to create consensus between the different sets giving information on $(y|x)$, which has been formalized mathematically by the infimum loss $(z, s) \in \mathcal{Y} \times 2^{\mathcal{Y}} \to \inf_{y \in s} \ell(z, y) \in \mathbb{R}$ for $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a specified loss on the underlying fully supervised learning problem. This leads, for $\tau \in \Delta_{\mathcal{X} \times 2^{\mathcal{Y}}}$ encoding the distribution generating samples (X, S), to the formulation $f^* \in \mathcal{F} = \arg\min_{f:\mathcal{X} \to \mathcal{Y}} \mathbb{E}_{(X,S) \sim \tau} [\inf_{Y \in S} \ell(f(X), Y)]$. To study this problem, a non-ambiguity assumption is usually made [15, 33, 31, 7, 8]. This is a very strong assumption to ensure that $\mathcal{F}$ is, in essence, a singleton. Highly adequate to this setting, the Laplacian regularization allows to relax this assumption, assuming that $\mathcal{F}$ can be big, but that we can discriminate between function in $\mathcal{F}$ by looking for the smoothest one in the sense defined by the Laplacian penalty. Moreover, the loss trick (15) allows to endow, in a off-the-shelf fashion, the recent work of Cabannes et al. [7, 8] on the partial supervised learning problem with our considerations on Laplacian regularization.

B Experiments

B.1 Low-rank approximation

Cutting computation cost thanks to low-rank approximation, as we did by going from the naive exact empirical risk minimizer $\hat{g}$ Eq. (5) to the smart implementation $\hat{g}_p$ Algorithm 1, is associated with a trade-off between computational versus statistical performance. This trade-off can be studied theoretically thanks to Theorem 1, which shows that under mild assumptions, considering $p = n^{1/2} \log(n)$ does not lead to any loss in performance, in the sense that the convergence rates in n, the number of samples, are only changed by a constant factor. We show on Figure 4 that in the setting of Figure 3, our low-rank approximation is not associated with a loss in performance. Actually low-rank
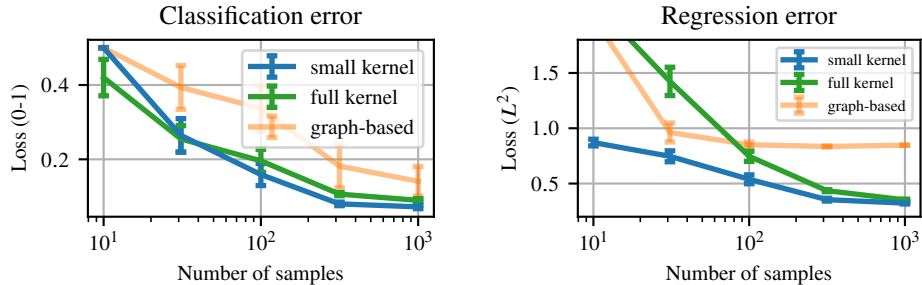
Figure 4: Cut in computation cost are not associated with a loss in performance. The estimate $\hat{g}_p$ Algorithm 1 (in blue), based on low-rank approximation that cut computation cuts performs as well as the exact computation of $\hat{g}$ Eq. (5). (Left) Classification error in the setting of Figure 3. (Right) Regression error in the same setting. The fact that the error of the graph-based method stalls around one, is due to the amplitude of the estimate being very small, which is coherent with behaviors described in [37].
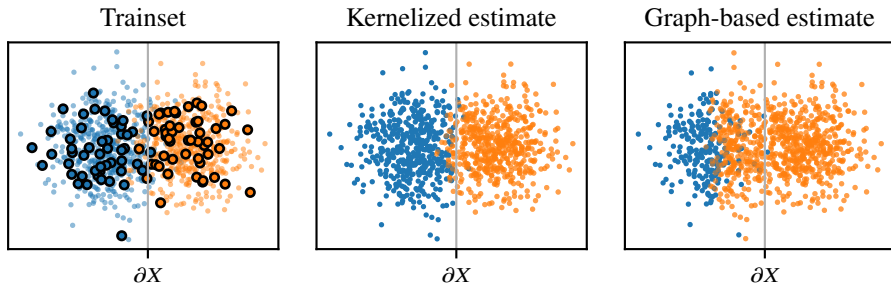

Figure 5: Setting of Figure 3 with $n = 1000$. (Left) Training set. We represent a cut of $\mathcal{X} \subset \mathbb{R}^d$ according to the two first coordinates $\{(x_1, x_2) \,|\, (x_1, x_2, \cdots, x_d) \in \mathcal{X}\}$. We have two Gaussians distribution with unit variance, one centered at $x = (0, 0, \cdots 0)$ and the other one centered at $x = (3, 0, \cdots, 0)$. One of the Gaussian distribution is associated to the blue class, the other one with the orange class. We consider $n = 1000$ unlabelled points, represented by small points, colored according to their classes, and $n_\ell = 100$ labelled points, represented in colour with black edges. (Middle) Reconstruction with our kernelized Laplacian methods. Our method uncovers correctly the structure of the problem, allowing to make a quite optimal reconstruction. The optimal decision frontier being illustrated by the grey line ∂X. (Right) Reconstruction with graph-Laplacian. The graph-Laplacian diffuses information too far away from what it should, leading to many incorrect guesses.

approximation can even be beneficial as it tends to lower the risk for overfitting as discussed by Rudi et al. [45].

B.2 Comparison with graph-based Laplacian

One the main goal of this paper is to make people drop graph-based Laplacian methods and adopt our "kernelized" technique. As such, we would like to discuss in more detail our comparison with graph-based Laplacian. In particular, we will discuss how and why we choose the hyperparameters and the setting of Figure 3.

The setting of Figure 3 is the one of Figure 5, we considered two Gaussians with unit variance and whose centers are at distance $\delta = 3$ of each other. We chose Gaussians distributions as it is a well-understood setting. We chose $\delta = 3$ so that there is an mild overlap between the two distributions. For the bandwidth parameter, we considered $\sigma_n = Cn^{-\frac{1}{d+4}} \log(n)$ as this is known to be the optimal bandwidth for graph Laplacian [24]. We chose $C = 1$ as this leads to σ_n of the order of δ. We chose $\lambda = 1$ to enforce Laplacian regularization and $\mu_n = 1/n$, as this is a classical regularization parameter in RKHS. We did not cross-validate parameters in order to be fair with graph-Laplacian that do not have as much parameters as our kernel method. We compute the error in a transductive setting, retaking the exact problem and algorithm of Zhu et al. [62]. We choose $d = 10$, as we know
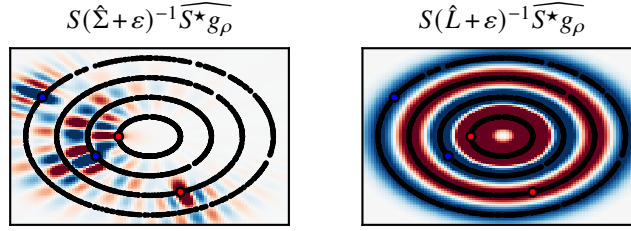
$$S(\hat{\Sigma}+\varepsilon)^{-1}\widehat{S^\star g_\rho} \qquad\qquad S(\hat{L}+\varepsilon)^{-1}\widehat{S^\star g_\rho}$$

Figure 6: Usefulness of Laplacian regularization. We illustrate the reconstruction based on our spectral filtering techniques based on the sole use of the covariance matrix Σ on the left, and on the sole use of the Laplacian matrix L on the right. We see that the covariance matrix does not capture the geometry of the problem, which contrasts with the use of Laplacian regularization.

that this is a good dimension parameter in order to illustrate the curse of dimensionality phenomenon without needing too much data.[3]

B.3 Usefulness of Laplacian regularization

It is natural to ask about the relevance of Laplacian regularization. To give convergence results, we have used Assumptions 1 and 2, which imply that g^* belongs to the RKHS $\mathcal{H}$, and we got convergence rates in $n_l^{1/2}$, which is not better than the rates we could get with pure kernel ridge regression. In particular, our algorithm can be split between an unsupervised part that learn the penalty $\left\|\mathcal{L}^{1/2}g\right\|_{L^2(\rho_\mathcal{X})}^2$ and a supervised part, that solve the problem of estimating g_λ from few labels (X_i, Y_i) given the penalty associated to $\mathcal{L}$. But the same method can be used for pure kernel ridge regression: unsupervised data could be leveraged to learn the covariance matrix Σ (6), and supervised data could be used to get $\widehat{S^\star g_\rho}$ to converge towards $S^\star g_\rho$. The same analysis would yield the same type of convergence rates. Yet the parameter σ_ℓ appearing in Theorem 1 would not be linked with the variance of $Y(I + \lambda\mathcal{L} + \lambda\mu K^{-1})^{-1}\delta_X$ but with the variance of $Y(I + \mu K^{-1})^{-1}\delta_X$. This is a key fact, the geometry of the covariance operator Σ is not supposed to be that relevant to the problem, while the one of L is. We illustrate this fact on Figure 6.

C Central Operators

The paper makes an intensive use of operators. This section aims at providing details and intuitions on those operators, in order to help the reader. In particular, we discuss on Assumptions 1 and 2 and we prove the equality in Eq. (6).

C.1 The diffusion operator

In this subsection, we discuss on the diffusion operator, and recall its basic properties.

The diffusion operator is a well-known operator in the realm of partial differential equation. Let us assume that $\rho_\mathcal{X}$ admit a smooth density $\rho_\mathcal{X}(dx) = p(x)dx$, say $p \in \mathcal{C}^2(\mathbb{R}^d)$, that cancels outside a domain $\Omega \subset \mathbb{R}^d$. Then the diffusion operator $\mathcal{L}$ can be explicitly written, for g twice differentiable, as

$$\mathcal{L}g(x) = -\Delta g(x) + \frac{1}{p(x)}\left\langle \nabla p(x), \nabla g(x) \right\rangle.$$

---

[3]Note that our consistency result Theorem 1 describes a convergence regime that applies to a vast class of problems. Such a regime usually takes place after a certain number of data (depending on the value of the constant C). Before entering this regime, describing the error of our algorithm would require more precise analysis specific to each problem instance, eventually involving tools from random matrix theory.

Table 1: Notations

Symbol	Description
$(X_i)_{i \leq n}$	n samples of input data
$(Y_i)_{i \leq n_l}$	n_l labels
ρ	Distribution of (X, Y)
g_ρ	Function to learn (1)
λ, μ	Regularization parameters
$g_\lambda, g_{\lambda,\mu}$	Biased estimates (3, 4)
$\hat{g}$	Empirical estimate (5)
$\hat{g}_p$	Empirical estimate with low-rank approximation (Algo. 1)
$\mathcal{H}$	Reproducing kernel Hilbert space
k	Reproducing kernel
S	Embedding of $\mathcal{H}$ in L^2
$S^\star$	Adjoint of S, operating from L^2 to $\mathcal{H}$
$\Sigma = S^\star S$	Covariance operator on $\mathcal{H}$
$K = SS^\star$	Equivalent of Σ on L^2
$\mathcal{L}$	Diffusion operator (*a.k.a* Laplacian)
$L = S^\star \mathcal{L} S$	Restriction of the diffusion operator to $\mathcal{H}$
g	Generic element in L^2
θ	Generic element in $\mathcal{H}$
λ_i	Generic eigen value
e_i	Generic eigen vector in L^2

This follows from the fact that for f once and g twice differentiable, using Stokes theorem,

$$
\begin{aligned}
\langle f, \mathcal{L}g \rangle_{L^2(\rho_\mathcal{X})} = \langle \nabla f, \nabla g \rangle_{L^2(\rho_\mathcal{X})} &= \langle \nabla f, p \nabla g \rangle_{L^2(\mathrm{d}x)} \\
&= \int_\mathcal{X} \mathrm{div}(f p \nabla g)\,\mathrm{d}x - \langle f, \mathrm{div}(p \nabla g) \rangle_{L^2(\mathrm{d}x)} = - \langle f, \mathrm{div}(p \nabla g) \rangle_{L^2(\mathrm{d}x)} \\
&= - \langle f, p^{-1} \mathrm{div}(p \nabla g) \rangle_{L^2(\rho_\mathcal{X})} = - \langle f, \mathrm{div}\, \nabla g + p^{-1}(\nabla p).\nabla g \rangle_{L^2(\rho_\mathcal{X})}.
\end{aligned}
$$

Note that when the distribution is uniform on Ω, the diffusion operator is exactly the opposite usual Laplacian operator Δ. As for the Laplacian case, it can be shown that under mild assumption on p, whose smoothness properties directly translates to the smoothness properties of the boundary of Ω, that the diffusion operator $\mathcal{L}$ has a compact resolvent (that is, for $\lambda \notin \mathrm{spec}(\mathcal{L})$, $(\mathcal{L} + \lambda I)^{-1}$ is compact). This is a standard result implied by a standard version of the famous Rellich-Kondrachov compactness embedding theorem: $H^2(\Omega)$ is compactly injected in $L^2(\Omega)$ whenever Ω is a bounded open with C^2-boundaries.

In such a setting, we can consider the eigenvalue decomposition of $\mathcal{L}^{-1}$, that is, $(\lambda_i, e_i) \in (\mathbb{R}_+ \times L^2)^\mathbb{N}$, with $(e_i)_{i \in \mathbb{N}}$ an orthonormal basis of L^2 and $(e_i)_{i \leq \dim \ker \mathcal{L}}$ generating the null space of $\mathcal{L}$, with the convention $\lambda_i = M$ for $i \leq \dim \ker \mathcal{L}$, with M an abstraction representing $+\infty$, and (λ_i) decreasing towards zero afterwards. This decomposition reads

$$
\mathcal{L}^{-1} = \sum_{i \in \mathbb{N}} \lambda_i e_i \otimes e_i. \tag{16}
$$

Note that the fact that all the (λ_i) are positive, is due to the fact that $\mathcal{L}^{-1}$ is the inverse of a positive self-adjoint operator. As a consequence, the diffusion operator has discrete spectrum, and can be written as

$$
\mathcal{L} = \sum_{i \in \mathbb{N}} \lambda_i^{-1} e_i \otimes e_i. \tag{17}
$$

In such a setting, the kernel-free Tikhonov regularization Eq. (3) reads

$$
g_\lambda = \sum_{i \in \mathbb{N}} \psi(\lambda_i) \left\langle g_\rho, \lambda_i^{1/2} e_i \right\rangle_{L^2} \lambda_i^{1/2} e_i, \tag{18}
$$

with $\psi : x \rightarrow (x + \lambda)^{-1}$, and the convention $M\psi(M) = 1$.

C.2 Regularity of the eigen vectors of the diffusion operator

In this subsection, we discuss on the regularity assumed in Assumption 2.

Introducing the kernel k and its associated RKHS $\mathcal{H}$ is useful when the eigen vectors of $\mathcal{L}$ can be well approximated by functions in $\mathcal{H}$. In applications, people tends to go for kernels that are translation-invariant, which implied that the RKHS $\mathcal{H}$ is made of smooth functions, could it be analytical functions (for the Gaussian kernel) or functions in H^m (for Sobolev kernels). As a consequence, we should investigate on the regularity of those eigen vectors. Indeed, if ρ derives from a Gibbs potential, that is $\rho(\mathrm{d}x) = e^{-V(x)}\,\mathrm{d}x$, the eigen vectors of $\mathcal{L}$ can be shown to inherit from the smoothness of the potential V [41]. For example if V belongs to H^m, and $H^m \subset \mathcal{H}$, we expect (e_i) to belongs to $\mathcal{H}$, thus verifying Assumption 2.

Counter-example and beyond Assumption 2.   Note that if ρ has several connected components of non-empty interiors, the null space of $\mathcal{L}$ is made of functions that are constants on each connected components of $\operatorname{supp}\rho_{\mathcal{X}}$. Those functions are not analytic. In such a setting, the Gaussian kernel is not sufficient for Assumption 2 to hold, and one should favor kernel associated with richer functional space such as the Laplace kernel or the neural tangent kernel [12]. However, as illustrated by Figure 2, e_i not belonging to $\mathcal{H}$ does not mean that e_i can not be well approximated by $\mathcal{H}$. Indeed it is well known that the approximation power of $\mathcal{H}$ for e_i can be measure in the biggest power p such that $e_i \in \operatorname{im} K^p$ [10], where $K = SS^*$. Assumption 2 corresponds to $p = 1/2$, but it should be seen as a specific instance of more generic approximation conditions.

Handling constants in RKHS.   Finally, note also that many RKHS do not contain constant functions, and therefore might not contain the constant function e_0 (although we are only looking for equality in the support of $\rho_{\mathcal{X}}$), however this specific point with e_0 can easily be circumvent either by assuming that g_ρ has zero mean, either by centering the covariance matrices Σ and $\hat{\Sigma}$ [41]. This relates with the usual technique for SVM consisting in adding a unpenalized bias [49].

C.3 Low-density separation

In this subsection, we discuss on how Assumption 1 relates to the idea of low-density separation.

Low-variation intuition.   The low-density separation supposes that the variations of g^* take place in region with low-density, so that $\left\|\mathcal{L}^{1/2}g^*\right\| / \|g^*\|$ is small. As such, using Courant-Fischer principle, Assumption 1 can be reformulated as g^* belonging to the space

$$\operatorname{Span}\{e_i\}_{i\leq r} = \underset{\substack{\mathcal{F}\subset L^2;\\ \dim\mathcal{F}=r}}{\arg\min}\,\max_{g\in\mathcal{F}}\,\frac{\left\|\mathcal{L}^{-1/2}f\right\|_{L^2}^2}{\|f\|_{L^2}^2}.$$

In other terms, Assumption 1 can be restated as g^* belonging to a finite dimensionnal space that minimizes a measure of variation given by the Dirichlet energy.

To tell the story differently, suppose that we are in a classification setting, *i.e.* $Y \in \{-1, 1\}$, and that the $\operatorname{supp}\rho_{\mathcal{X}}$ is connected. Then we know that the null space of $\mathcal{L}$ is made of constant functions. Then the first eigen vector e_2 of $\mathcal{L}$ is a function that is orthogonal to constants. Hence e_2 is a function that changes its sign and that is "balanced" in the sense that $\mathbb{E}[e_2] = 0$ – *i.e.* if $e_2(x) = \mathbb{E}_\mu[Y \,|\, X = x]$ for some measure μ, we have $\mathbb{E}_\mu[Y] = 0$, meaning that classes are "balanced". Moreover, in order to minimize $\left\|\mathcal{L}^{1/2}e_2\right\|$, the variations of e_2 should take place in low-density regions of $\mathcal{X}$.

Diffusion intuition.   Finally, as $\mathcal{L}$ is a diffusion operator, we also have an interpretation of Assumption 1 is term of diffusion. Consider (λ_i, e_i) the eigen elements of Eq. (17). The diffusion of g_ρ according the density $\rho_{\mathcal{X}}$ can be written as, for $t \in \mathbb{R}$,

$$g_t = e^{-t\mathcal{L}}g_\rho = \sum_{i\in\mathbb{N}} e^{-t\lambda_i^{-1}}\langle g_\rho, e_i\rangle e_i.$$

This diffusion will cut off the high frequencies of g_ρ that corresponds to $\langle g_\rho, e_i\rangle$ for big i, and big λ_i^{-1}. Indeed, the difference between the diffusion and the original g_ρ can be measured as

$$\|g_t - g_\rho\|_{L^2}^2 = \sum_{i\in\mathbb{N}}(e^{-t\lambda_i^{-1}} - 1)^2\langle g_\rho, e_i\rangle^2 = \sum_{i\in\mathbb{N}} t^2\lambda_i^{-2}\langle g_\rho, e_i\rangle^2 + o(t^2\lambda_i^{-2}).$$

So that assuming that g_ρ is supported on few of the first eigen vectors of $\mathcal{L}$, can be rephrased as saying that the diffusion of g_ρ does not modify it too much.

The variance σ_ℓ.   Theorem 1 shows that the need for labels depends on the variance parameter σ_ℓ^2. It is natural to wonder on how this parameters relates to the low-density hypothesis. As we discussed, this parameter is linked to the variance of $Z = Y(I + \lambda\mathcal{L})^{-1}\delta_X$. We can separate the variability of this variable due to X and the variability due to Y

$$Z = Z_X + Z_Y, \qquad \text{with} \qquad Z_X = (I + \lambda\mathcal{L})^{-1}g_\rho(X)\delta_X, \quad Z_Y = (I + \lambda\mathcal{L})^{-1}(Y - \mathbb{E}[Y \mid X])\delta_X.$$

As such we see that this variance depends on the structure of the density $\rho_\mathcal{X}$ with the variance of $(I + \lambda\mathcal{L})^{-1}\delta_X$, and the labelling noise with the variance of $(Y \mid X)$. The low-density separation does not tell us anything about the level of noise in Y or the diffusion structure linked with $\rho_\mathcal{X}$, but additional hypothesis could be made to characterize those.

C.4   Kernel operators

In this subsection, we define formally the operators S and Σ.

We now turn towards operators linked with the Hilbert space $\mathcal{H}$. Recall that for $k : \mathcal{X} \to \mathcal{X} \to \mathbb{R}$ a kernel, $\mathcal{H}$ is defined the closure of the span of the elements k_x under the scalar product $\langle k_x, k_{x'} \rangle = k(x, x')$. In particular, $\|k_x\|_\mathcal{H}^2 = k(x, x)$. $\mathcal{H}$ parameterize a vast class of function in $\mathbb{R}^\mathcal{X}$ through the mapping

$$S : \quad \begin{matrix} \mathcal{H} & \to & \mathbb{R}^\mathcal{X} \\ \theta & \to & (\langle k_x, \theta \rangle)_{x \in \mathcal{X}}. \end{matrix}$$

Under mild assupmtions, S maps $\mathcal{H}$ to a space of function belongs to L^2.

Proposition 2. *When $x \to k(x, x)$ belongs to $L^1(\rho_\mathcal{X})$, S is a continuous mapping from $\mathcal{H}$ to $L^2(\rho_\mathcal{X})$. This is particularly the case when $\rho_\mathcal{X}$ has compact support and k is continuous.*

Proof. Consider $\theta \in \mathcal{H}$, we have

$$\|S\theta\|_{L^2}^2 = \int_\mathcal{X} \langle k_x, \theta \rangle^2 \, \rho_\mathcal{X}(\mathrm{d}x) \le \int_\mathcal{X} \langle k_x, \theta \rangle_\mathcal{H}^2 \, \rho_\mathcal{X}(\mathrm{d}x) \le \int_\mathcal{X} \|k_x\|_\mathcal{H}^2 \|\theta\|_\mathcal{H}^2 \, \rho_\mathcal{X}(\mathrm{d}x)$$

$$= \|\theta\|_\mathcal{H}^2 \int_\mathcal{X} k(x, x)\rho_\mathcal{X}(\mathrm{d}x) = \|\theta\|_\mathcal{H}^2 \|x \to k(x, x)\|_{L^1}.$$

Moreover, when $\rho_\mathcal{X}$ has compact support and k is continuous, k is bounded on the support of $\rho_\mathcal{X}$ therefore $x \to k(x, x)$ belongs to L^1. $\qquad\square$

As a continuous operator from the Hilbert space $\mathcal{H}$ to the Hilbert space L^2, S is naturally associated with many linear structure. In particular its adjoint $S^\star$, but also the self-adjoint operators $K = SS^\star$ and $\Sigma = S^\star S$.

Proposition 3. *The adjoint of S is defined as*

$$S^\star : \quad \begin{matrix} L^2 & \to & \mathcal{H} \\ g & \to & \int_\mathcal{X} g(x)k_x\rho_\mathcal{X}(\mathrm{d}x) = \mathbb{E}_{X \sim \rho_\mathcal{X}}[g(X)k_X]. \end{matrix}$$

To S is associated the kernel self-adjoint operator on L^2

$$K := SS^\star : \quad \begin{matrix} L^2 & \to & L^2 \\ g & \to & (x \to \int_\mathcal{X} k(x, x')g(x')\rho_\mathcal{X}(\mathrm{d}x')), \end{matrix}$$

as well as the (not-centered) covariance on $\mathcal{H}$, $\Sigma := S^\star S = \mathbb{E}_{X \sim \rho_\mathcal{X}}[k_X \otimes k_X]$.

Proof. We shall prove the equality defining those operators. Consider $\theta \in \mathcal{H}$ and $g \in L^2$, we have

$$\langle S^\star g, \theta \rangle_\mathcal{H} = \langle g, S\theta \rangle_{L^2} = \mathbb{E}_{X \sim \rho_\mathcal{X}}[g(X) \langle k_X, \theta \rangle_\mathcal{H}] = \langle \mathbb{E}_{X \sim \rho_\mathcal{X}}[g(X)k_X], \theta \rangle_\mathcal{H}.$$

We also have, for $x \in \mathcal{X}$,

$$(SS^\star g)(x) = \langle k_x, \mathbb{E}_{X \sim \rho_\mathcal{X}}[g(X)k_X] \rangle_\mathcal{H} = \mathbb{E}_{X \sim \rho_\mathcal{X}}[g(X) \langle k_x, k_X \rangle_\mathcal{H}] = \mathbb{E}_{X \sim \rho_\mathcal{X}}[g(X)k(X, x)].$$

Finally, we have

$$S^\star S\theta = \mathbb{E}_{X \sim \rho_\mathcal{X}}[S\theta(X)k_X] = \mathbb{E}_{X \sim \rho_\mathcal{X}}[\langle \theta, k_X \rangle_\mathcal{H} k_X] = \mathbb{E}_{X \sim \rho_\mathcal{X}}[k_X \otimes k_X]\theta.$$

This provides the last of all the equalities stated above. $\qquad\square$

The functional space $\mathcal{H}$. In the main text, we have written everything in term of θ, highlighting the parametric nature of kernel methods. This made it easier to dissociate the norm on functions derived from $\mathcal{H}$ and the one derived from L^2 or H^1. In literature, people tends to keep everything in term of functions $g_\theta = S\theta$ without even mentioning the dependency in θ. Such a setting consists in considering directly the RKHS $\mathcal{H}$ whose scalar product is defined for $g, g' \in (\ker K)^\perp$ by $\langle g, g' \rangle_{\mathcal{H}} = \langle g, K^{-1}g' \rangle_{L^2}$.

C.5 Derivative operators

In this subsection, we discuss on derivative in RKHS and we define formally the operator L.

As well as evaluation maps can be represented in $\mathcal{H}$, under mild assumptions, derivative evaluation maps can be benefited of such a property. Indeed, for $g_\theta = S\theta$, $x \in \mathcal{X}$ and $u \in \mathcal{B}_{\mathcal{X}}(0,1)$ a unit vector, we have

$$\partial_u g_\theta(x) = \lim_{t \to 0} \frac{g_\theta(x+tu) - g_\theta(x)}{t} = \lim_{t \to 0} \frac{\langle \theta, k_{x+tu} \rangle_{\mathcal{H}} - \langle g, k_x \rangle_{\mathcal{H}}}{t} = \lim_{t \to 0} \left\langle \theta, \frac{k_{x+tu} - k_x}{t} \right\rangle_{\mathcal{H}}$$

As a linear combination of elements in $\mathcal{H}$, the difference quotient evaluation map $t^{-1}(k_{x+tu} - k_x)$ belongs to $\mathcal{H}$ and has a norm

$$\left\| \frac{k_{x+tu} - k_x}{t} \right\|_{\mathcal{H}}^2 = \frac{k(x+tu, x+tu) - 2k(x+tu, x) + k(x,x)}{t^2}.$$

In order for the limit when t goes to zero to belong to $\mathcal{H}$, we see the importance of k to be twice differentiable. This limit $\partial_u k_x$, whose existence is proven formally by Zhou [61], provides a derivative evaluation map in the sense that

$$\partial_u g_\theta(x) = \langle \theta, \partial_u k_x \rangle_{\mathcal{H}}.$$

From this equality, we derive that $\partial_{1i} k(x, x') = \langle k_{x'}, \partial_i k_x \rangle$, and recursively that $\langle \partial_i k_x, \partial_j k_{x'} \rangle = \partial_{1i}\partial_{2j}k(x, x')$.

Similarly to the operator S, we can introduce the operators Z_i for $i \in [1, d]$, defined as

$$Z_i : \begin{array}{ccc} \mathcal{H} & \to & \mathbb{R}^{\mathcal{X}} \\ \theta & \to & (\langle \partial_i k_x, \theta \rangle_{\mathcal{H}})_{i \leq d} \end{array}.$$

Once again, under mild assumptions, $\operatorname{im} Z_i$ inherit from an Hilbertian structure.

Proposition 4. *When $x \to \partial_{1i}\partial_{2i}k(x, x)$ belongs to $L^1(\rho_{\mathcal{X}})$, Z_i is a continuous mapping from $\mathcal{H}$ to $L^2(\rho_{\mathcal{X}})$. This is particularly the case when $\rho_{\mathcal{X}}$ has compact support and k is twice differentiable with continuous derivatives.*

Proof. Consider $\theta \in \mathcal{H}$, similarly to before, we have

$$\|Z\theta\|_{L^2}^2 = \int_{\mathcal{X}} \langle \partial_i k_x, \theta \rangle^2 \rho_{\mathcal{X}}(\mathrm{d}x) \leq \|\theta\|_{\mathcal{H}}^2 \int_{\mathcal{X}} \|\partial_i k_x\|_{\mathcal{H}}^2 \rho_{\mathcal{X}}(\mathrm{d}x) = \|\theta\|_{\mathcal{H}}^2 \|x \to \partial_{1,i}\partial_{2,i}k(x,x)\|_{L^1}.$$

Moreover, when $\rho_{\mathcal{X}}$ has compact support and $\partial_{1,i}\partial_{2,i}k$ is continuous, $\partial_{1,i}\partial_{2,i}k$ is bounded on the support of $\rho_{\mathcal{X}}$ therefore $x \to \partial_{1,i}\partial_{2,i}k$ belongs to L^1. $\qquad\square$

Among the linear operator that can be build from Z_i, in the theoretical part of this paper, we are mainly interested in $Z_i^\star Z_i$. In the empirical part however, we might be interested in $Z_i Z_j^\star$ as well as $Z_i S^\star$ as it might appear in Algorithm 1 (where Z_n has to be understood as the empirical version of $Z = [Z_1; \cdots; Z_d]$).

Proposition 5. *The energy Dirichlet on $\mathcal{H}$ can be represented through the operator*

$$S^\star \mathcal{L} S = \sum_{i=1}^d Z_i^\star Z_i = \sum_{i=1}^d \mathbb{E}_{X \sim \rho_{\mathcal{X}}}[\partial_i k_X \otimes \partial_i k_X].$$

Proof. Let $\theta \in \mathcal{H}$ and $g_\theta = S\theta$, we have

$$\langle g_\theta, \mathcal{L}g_\theta \rangle_{L^2} = \langle \theta, S^\star \mathcal{L}S\theta \rangle_{\mathcal{H}} = \mathbb{E}_{X \sim \rho_{\mathcal{X}}} \left[\|\nabla g_\theta(X)\|^2 \right] = \sum_{i=1}^d \mathbb{E}_{X \sim \rho_{\mathcal{X}}} \left[(\partial_i g_\theta(X))^2 \right]$$

$$= \sum_{i=1}^d \mathbb{E}_{X \sim \rho_{\mathcal{X}}} \left[\langle \partial_i k_X, \theta \rangle_{\mathcal{H}}^2 \right] = \sum_{i=1}^d \|Z_i \theta\|_{L^2}^2 = \left\langle \theta, \sum_{i=1}^d Z_i^\star Z_i \theta \right\rangle_{\mathcal{H}}$$

$$= \sum_{i=1}^d \mathbb{E}_{X \sim \rho_{\mathcal{X}}} [\langle \theta, (\partial_i k_X \otimes \partial_i k_X)\theta \rangle_{\mathcal{H}}] = \left\langle \theta, \sum_{i=1}^d \mathbb{E}_{X \sim \rho_{\mathcal{X}}} \left[\partial_i k_X \otimes \partial_i k_X \right] \theta \right\rangle_{\mathcal{H}}.$$

Since the three operators are self-adjoint and they all represent the same quadratic form, they are equals. $\qquad\square$

C.6 Relation between Σ and L

In this subsection, we discuss on the relation between Σ and L and show that we can expect the existence of $a \in (1 - 2/d, 1]$ and $c > 0$ such that $L \preceq c\Sigma^a$.

Informal capacity considerations. We want to compare Σ and L, as $L \preceq c\Sigma^a$ with the biggest a possible. This depend on how fast the eigen values are decreasing, which is linked to the entropy numbers of those two compact operators. Those entropy numbers are linked with the capacity of the functional spaces $\left\{ g \in L^2 \,\middle|\, \left\| K^{-1/2}g \right\|_{L^2} < \infty \right\}$ and $\left\{ g \in L^2 \,\middle|\, \left\| K^{-1/2}\mathcal{L}^{-1/2}g \right\|_{L^2} < \infty \right\}$. The first space is the reproducing kernel Hilbert space linked with k, the second space is, roughly speaking, a space of function whose integral belongs to the first space. As such, if the first space is $\mathcal{H}^m$, the second is $\mathcal{H}^{m-1}$, and we can consider $a = (m-1)/m$. Because we are considering kernel, we have $m > d/2$ (this to make sure that the evaluation functionals $L_X : f \to f(x)$ are continuous), so that $a > 1 - 2/d$. Without trying to make those "algebraic" considerations more formal, we will give an example on the torus.

Translation-invariant kernel and Fourier transform. Consider $L^2([0,1]^d, \mathrm{d}x)$ the space of periodic functions in dimension d, square integrable against the Lebesgue measure on $[0,1]^d$. For simplicity, we will suppose that $\rho_{\mathcal{X}}$ is the Lebesgue measure on $[0,1]^d$. Consider a translation invariant kernel

$$k(x,y) = q(x-y) \qquad \text{for } q : \mathbb{R}^d \to \mathbb{R} \text{ that is one periodic.}$$

In this setting, the operator K, operating on L^2, is the convolution by q, that is

$$K : \begin{array}{ccc} L^2 & \to & L^2 \\ g & \to & q * g \end{array}, \qquad \text{hence} \qquad \widehat{Kg} = \hat{q}\hat{g}.$$

Where we have used the fact that convolutions can be represented by a product in Fourier. Note that, from Böchner theorem, we know that k being positive definite implies that the Fourier transform of q exists and is not negative. Let us define the Fourier coefficient and the inverse Fourier transform as

$$\forall \omega \in \mathbb{Z}^d, \quad \hat{g}(\omega) = \int_{[0,1]^d} g(x)e^{-2i\pi\omega^\top x} \, \mathrm{d}x, \quad \text{and} \quad \forall x \in [0,1]^d, \quad g(x) = \sum_{\omega \in \mathbb{Z}^d} e^{2i\pi\omega^\top x} \hat{g}(\omega).$$

K being a convolution operator, it is diagonalizable with eigen elements $(\hat{q}(\omega), x \to e^{2i\pi\omega^\top x})_{\omega \in \mathbb{Z}^d}$. From this, we can explicit many of our abstract operators. First of all, using Perceval's theorem,

$$\|g\|_{\mathcal{H}}^2 = \left\langle g, K^{-1}g \right\rangle_{L^2} = \sum_{\omega \in \mathbb{Z}^d} \frac{|\hat{g}(\omega)|^2}{\hat{q}(\omega)}.$$

Hence we can parametrize $\mathcal{H}$ with $(\theta_\omega)_{\omega \in \mathbb{Z}^d} \in \mathbb{C}^{\mathbb{Z}^d}$ and the ℓ^2-metric, where $\theta_\omega = \hat{g}(\omega)/\sqrt{\hat{q}(\omega)}$ and

$$(S\theta)(x) = g_\theta(x) = \sum_{\omega \in \mathbb{Z}^d} e^{2i\pi\omega^\top x} \sqrt{\hat{q}(\omega)} \theta_\omega.$$

Note that this is not the usual parameterization of $\mathcal{H}$ by elements $\theta \in \mathcal{H}$ as $(\mathbb{C}^{\mathbb{Z}^d}, \ell^2)$ is not a space of functions. However, such a parametrization of $\mathcal{H}$ does not change any of the precedent algebraic considerations on the operators S, Σ, K, and L.

Diffusion operator and Fourier transform.   As well as convolution operators are well represented in Fourier, derivation operators are. Indeed, when g is regular, we have

$$\left\| \mathcal{L}^{1/2} g \right\|_{L^2}^2 = \|\nabla g\|_{L^2}^2 = \sum_{j=1}^d \|\partial_j g\|_{L^2}^2 = \sum_{j=1}^d \sum_{\omega \in \mathbb{Z}^d} \omega_j^2 |\hat{g}(\omega)|^2 .$$

As a consequence, using the expression of $S\theta$, we have

$$\Sigma \theta = \sum_{\omega \in \mathbb{Z}^d} \hat{q}(\omega) \theta_\omega, \qquad \text{while} \qquad L\theta = \sum_{\omega \in \mathbb{Z}^d} \|\omega\|_2^2 \hat{q}(\omega) \theta_\omega, \quad \text{where} \quad \|\omega\|_2^2 = \sum_{j=1}^d \omega_i^2.$$

With this parameterization, the eigen elements of Σ are $(\hat{q}(\omega), \delta_\omega)_{\omega \in \mathbb{Z}^d}$ while the one of L are $(\|\omega\|_2^2 \hat{q}(\omega), \delta_\omega)_{\omega \in \mathbb{Z}^d}$.

Eigen value decay comparison.   Hence, having $L \preceq c\Sigma^a$ is equivalent to having $\|\omega\|_2^2 \hat{q}(\omega) \leq c\hat{q}(\omega)^a$. Now suppose that the decay of $\hat{q}$ is governed by

$$c_1(1 + \sigma^{-1} \|\omega\|_2^2)^{-m} \leq \hat{q}(\omega) \leq c_2(1 + \sigma^{-1} \|\omega\|_2^2)^{-m},$$

for two constants $c_1, c_2 > 0$. In particular, this is verified for Matérn kernels, corresponding to the fractional Sobolev space H^m, and for the Laplace kernel with $m = (d+1)/2$, which reads $k(x, y) = \exp(-\sigma^{-1} \|x - y\|)$. The Gaussian kernel could be seen as $m = +\infty$ as it has exponential decay. With such a decay we have, assuming without restrictions that we are in one dimension

$$\omega^2 \hat{q}(\omega) \leq c_2 \omega^2 (1 + \sigma^{-1} \omega^2)^{-m} \leq c_2 \sigma (1 + \sigma^{-1} \omega^2)^{-(m-1)} \leq c_1^{\frac{m}{m-1}} c_2 \sigma \hat{q}(\omega)^{\frac{m-1}{m}}.$$

In other terms, we can consider $c = c_1^{\frac{m}{m-1}} c_2 \sigma$ and $a = (m-1)/m$. Assuming that q is square-integrable, so is $\hat{q}$, which implies that $2m > d$. As a consequence, we do have $a > 1 - 2/d$. Note that this reasoning could be extended to the case where $\rho_{\mathcal{X}}$ has a density against the Lebesgue measure, that is bounded above and below away from zero.

D   Spectral decomposition

In this section, we recall facts on spectral regularization, before proving Proposition 1 and extending it to the case $\mu = 0$.

D.1   Generalized singular value with matrices

Generalized singular value decomposition.   Let $A \in \mathbb{R}^{m1 \times n}$ and $B \in \mathbb{R}^{m_2 \times n}$ be two matrices. There exists $U \in \mathbb{R}^{m_1 \times m_1}$, $V \in \mathbb{R}^{m_2 \times m_2}$ two orthogonal matrices, $c \in \mathbb{R}^{m_1 \times r}$ and $s \in \mathbb{R}^{m_2 \times r}$ be two 1-diagonal matrices such that $c^\top c + s^\top s = I_r$, and $H \in \mathbb{R}^{n \times r}$ a non-singular matrix such that

$$A = UcH^{-1}, \qquad B = VsH^{-1}.$$

To be more precise c is such that only entries $c_{ii} = \cos(\theta_i)$ for $i < \min(r, m_2)$ are non-zeros and s such that only entries $s_{m_1-i,r-i} = \sin(\theta_{r-i})$ for $i < \min(r, m_1)$ are non-zeros, with $\theta_i \in [-\pi/2, \pi/2]$ an angle. Here, c stands for cosine, s for sinus and r for rank.

Link with generalized eigenvalue problem.   As well as the singular value of A is linked with the eigen value of $A^\top A$, the generalized singular value decomposition of $[A; B]$ is linked with the generalized eigenvalue problem linked with $(A^\top A, B^\top B)$. Indeed, we have

$$A^\top A = H^{-\top} c^\top c H^{-1}, \qquad B^\top B = H^{-\top} s^\top s H^{-1}.$$

In particular, with (e_i) the canonical basis of $\mathbb{R}^r$, and h_i the i-th column of H, we get

$$H^\top A^\top A h_i = \cos(\theta_i)^2 e_i = \tan(\theta_i)^{-2} \sin(\theta_i)^2 e_i = \tan(\theta_i)^{-2} H^\top B^\top B h_i.$$

From which we deduce that, since $\operatorname{im} A \cup \operatorname{im} B \subset \operatorname{im} H^\top$,

$$A^\top A h_i = \tan(\theta_i)^{-2} B^\top B h_i, \qquad h_j^\top B^\top B h_i = \sin(\theta_i)^2 \mathbf{1}_{i=j}.$$

So if we denote by $f_i = |\sin(\theta_i)|^{-1} h_i$ and $\lambda_i = \tan(\theta_i)^{-2}$, assuming $\lambda_i \neq 0$ for all $i \leq r$ (which corresponds to $\ker B \subset \ker A$), $(\lambda_i)_{i \leq r}, (f_i)_{i \leq r}$ provide the generalized eigenvalue decomposition of $(A^\top A, B^\top B)$ in the sense that

$$A^\top A f_i = \lambda_i B^\top B f_i, \qquad f_j^\top B^\top B f_i = \mathbf{1}_{i=j}, \qquad f_j^\top A^\top A f_i = \lambda_i \mathbf{1}_{i=j}.$$

D.2 Tikhonov regularization

Define the Tikhonov regularization

$$x_\lambda = \arg\min_{x \in \mathbb{R}^n} \|Ax - b\|^2 + \lambda \|Bx\|^2 .$$

When this problem is well-defined, the solution is defined as

$$x_\lambda = (A^\top A + \lambda B^\top B)^\dagger A^\top b.$$

With the generalized singular value decomposition of A and B, we have

$$A^\top A + \lambda B^\top B = H^{-\top} \gamma_\lambda H^{-1}, \quad \text{with} \quad \gamma_\lambda = c^\top c + \lambda s^\top s.$$

Using the fact that $A^\top b = H^{-\top} c^\top U^\top b$, we get

$$x_\lambda = H \gamma_\lambda^{-1} c^\top U^\top b = \left(\sum_{i=1}^r \frac{\cos(\theta_i)}{\cos(\theta_i)^2 + \lambda \sin(\theta_i)^2} h_i \otimes u_i \right) b.$$

Now, we would like to replace c_{ii}, s_{ii}, h_i and u_i with quantities that depend on λ_i, f_i and A. To do so recall that $AH = Uc$, therefore $\cos(\theta_i)u_i = Ah_i$, and recall that $h_i = \sin(\theta_i)f_i$ and $\lambda_i = \cos(\theta_i)^2 / \sin(\theta_i)^2$. Inputting those equality in the last expression of x_λ we get

$$x_\lambda = \left(\sum_{i=1}^r \frac{\sin(\theta_i)^2}{\cos(\theta_i)^2 + \lambda \sin(\theta_i)^2} f_i \otimes Af_i \right) b = \left(\sum_{i=1}^r \frac{1}{\lambda_i + \lambda} f_i \otimes Af_i \right) b.$$

Finally,

$$b_\lambda = Ax_\lambda = \sum_{i=1}^r \psi(\lambda_i) \langle Af_i, b \rangle Af_i, \qquad \text{where} \qquad \psi(x) = \frac{1}{x + \lambda}.$$

D.3 Extension to operators

To end the proof of Proposition 1, we should prove that we can apply the generalized eigenvalue decomposition to operators. We will only prove that it is possible for $(\Sigma, L + \mu)$ based on simple considerations.

Proposition 6. *When k is continuous and* $\operatorname{supp} \rho_\mathcal{X}$ *is bounded, Σ is a compact operator.*

Proof. We have $\Sigma = \mathbb{E}[k_X \otimes k_X]$ and $\|k_x\| = k(x, x)$. Since k is continuous and $\operatorname{supp} \rho_\mathcal{X}$ is compact, for $x \in \operatorname{supp} \rho_\mathcal{X}$, $k(x, x)$ is bounded. Hence Σ is a nuclear operator, hence trace class and compact. □

Proposition 7. *When k is twice differentiable with continuous derivative, and* $\operatorname{supp} \rho_\mathcal{X}$ *is compact, L is a compact operator. As a consequence, L has a compact spectrum, and has a pseudo inverse that we will denote, with a slight abuse of notation, by L^{-1}.*

Proof. The proof is similar to the one showing that Σ is compact, based on the fact that $L = \sum_{i=1}^d \mathbb{E}[\partial_i k_X \otimes \partial_i k_X]$, and $\|\partial_i k_X\|^2 = \partial_{1i}\partial_{2i}k(x, x)$. □

Proposition 8. *When Σ is compact, for all $\mu > 0$, $(L + \mu)^{-1/2}\Sigma(L + \mu)^{-1/2}$ is a compact operator.*

Proof. The proof is straightforward

$$\operatorname{Tr}((L+\mu)^{-1/2}\Sigma(L+\mu)^{-1/2}) = \operatorname{Tr}(\Sigma(L+\mu)^{-1}) \le \left\|(L+\mu)^{-1}\right\|_{\text{op}} \operatorname{Tr}(\Sigma) \le \mu^{-1} \operatorname{Tr}(\Sigma) < +\infty.$$

Therefore the operator is trace class, hence compact. □

Proposition 9. *For any $\mu > 0$, the generalized eigen value decomposition of $(\Sigma, L + \mu)$ as defined in Proposition 1 exists.*

Proof. Using the spectral theorem, since $(L + \mu)^{-1/2}\Sigma(L + \mu)^{-1/2}$ is positive self-adjoint compact operator, there exists (ξ_i) a basis of $\mathcal{H}$ and $(\lambda_i) \in \mathbb{R}_+$ a decreasing sequence (note that $\ker(L + mu) = \ker \Sigma = \{0\}$), such that

$$(L + \mu)^{-1/2}\Sigma(L + \mu)^{-1/2} = \sum_{i \in \mathbb{N}} \lambda_i \xi_i \otimes \xi_i.$$

Taking $\theta_i = (L + \mu)^{-1/2}\xi_i$, we get $\Sigma\theta_i = \lambda_i L\theta_i$. Because (ξ_i) generates $\mathcal{H}$, and $(L + \mu)^{-1/2}$ is bijective (since L is compact, $(L + \mu)^{-1}$ is coercive), $((L + \mu)^{-1/2}\xi_i)$ generates $\mathcal{H}$. $\square$

Proposition 1 follows from prior discussion on Tikhonov regularization extended to infinite summations.

D.4 The case $\mu = 0$

When $\mu = 0$, Eq. (7) should be seen as the rewritting of Eq. (18) based on the RKHS $\mathcal{G} = \mathrm{im}\, S$. This can only be done when the eigen vectors of $\mathcal{L}$ appearing in Eq. (17) belongs to $\mathcal{G} = \mathrm{im}\, S = \mathrm{im}\, K^{1/2}$, which is exactly what Assumption 2 provides. In such a setting, we can find $(\theta_i) \in \mathcal{H}^{\mathbb{N}}$ to write $\lambda_i^{1/2}e_i = S\theta_i$ as soon as $\lambda_i \neq 0$ (write $Me_i = S\theta_i$ for M an abstraction representing $+\infty$ when $\lambda_i = 0$, handling the potential fact that $\ker B \not\subset \ker A$), we get $\theta_i S^*S\theta_j = \lambda_i \mathbf{1}_{i=j}$, and $L\theta_i = \lambda_i^{-1}\Sigma\theta_i$, and we can extend Proposition 1 to the case $\mu = 0$, with

$$g_\lambda = \sum_{i \in \mathbb{N}} \psi(\lambda_i)\langle S^\star g_\rho, \theta_i\rangle S\theta_i, \tag{19}$$

where we handle the null space of $\mathcal{L}$ with the equality $M\psi(M) = 1$, verified by M our abstraction representing $+\infty$, so that $\psi(M)\langle S^\star g_\rho, \theta_i\rangle S\theta_i = \langle g_\rho, e_i\rangle e_i$ as soon as $\lambda_i = 0$.

Beyond Assumption 2. Assumption 2 could be made generic by considering the biggest $(p_i) \in \mathbb{R}_+^{\mathbb{N}}$ such that $K^{-p_i}e_i$ belongs to L^2, and rewriting Eq. (19) under the form $g_\lambda = \sum_{i \in \mathbb{N}} \psi(\lambda_i)\langle (S_0 K^{p_i})^\star g_\rho, \theta_i\rangle S_0 K^{p_i}\theta_i$, with $\theta_i = \lambda_i^{-1/2}S_0^{-1}K^{-p_i}\theta_i$ and $S_0 = K^{-1/2}S$ the isomorphism between $\mathcal{H}$ and L^2 (assuming that S is dense in L^2). Such an assumption would describe all situations from no assumption ($p_i = 0$ for all i), Assumption 2 ($p_i = 1/2$ for all i) to even more optimistic assumptions ($p_i \geq 1$ for all i).

E Consistency analysis

This section is devoted to the proof of Theorem 1. The proof is based on Eqs. (7) and (19), and splits the error of $\|g_\rho - \hat{g}_p\|_{L^2}^2$ into several components linked with how well we approximate $S^\star g_\rho$, and how well we approximate the eigenvalue decomposition (λ_i, θ_i) of (Σ, L).

E.1 Sketch and understanding of the proof

In this subsection, we explain how work the proofs for consistency theorems such as Theorem 1.

Let us define the mapping $G : \mathcal{H} \times \mathcal{C} \to L^2$ with $\mathcal{C}$ the set of pairs of self-adjoint operators on $\mathcal{H}$ that admit a generalized eigen value decomposition, as

$$G(\theta, (A, B)) = \sum_{i \in \mathbb{N}} \psi(\lambda_i)\langle \theta, \theta_i\rangle S\theta_i \qquad \text{with} \qquad (\lambda_i, \theta_i) \in \mathrm{GEVD}(A, B). \tag{20}$$

$G(\theta, (A, B)) \in L^2$ corresponds to writing $\theta \in \mathcal{H}$ in the basis associated with the generalized eigen value decomposition (GEVD) of (A, B).

Proposition 10. *Under Assumptions 1 and 2, and with ψ defined in Theorem 1*

$$g_\lambda = G(S^\star g_\rho, (\Sigma, L)), \qquad and \qquad \hat{g}_p = G(S^{\hat{\star}}g_\rho, (P\Sigma P, P\hat{L}P + \mu P)),$$

with P the projection matrix from $\mathcal{H}$ to $\mathrm{Span}\,\{k_{X_i}\}_{i \leq p}$.

Proof. This is a direct application of Assumptions 1, 2, Eq. (19) and Algorithm 1. □

The main point of the proof is to relate g_ρ to $\hat{g}_p$. To do so, we will use several functions in L^2 generated by G. We detail our steps in Table 2. Table 2 gives a first answer to the two questions asked in the opening of Section 4. The number of unlabelled data controls the convergence of the operators $(P\hat{\Sigma}P, P\hat{L}P + \mu P)$ towards $(\Sigma, L + \mu)$. The number of labelled data controls the convergence of the vector $\widehat{S^\star g_\rho}$ towards $S^\star g_\rho$. Priors on the structure of the problem, such as source and approximation conditions, control the convergence of the bias estimate $g_{\lambda,\mu}$ towards g_ρ. Furthermore, a more precise study reveals that the concentration of operators are related to efficient dimension [10] and are accelerated by capacity assumptions on the functional space whose norm is $\|g\| = \left\|(\mathcal{L} + K^{-1})^{1/2}g\right\|_{L^2}$, and that the concentration of the vector $\widehat{S^\star g_\rho}$ is accelerated by assumptions on moments of the variable $Y(I + \lambda\mathcal{L})^{-1}\delta_X$ (inheriting randomness from $(X, Y) \sim \rho$).

Table 2: Steps in the consistency analysis

Estimate	Vector	Property of convergence	Basis
$\hat{g}_p$	$\widehat{S^\star g_\rho}$		$(P\hat{\Sigma}P, P\hat{L}P + \mu P)$
		Low-rank approximation [45]	$\downarrow$
$\hat{g}$	$\widehat{S^\star g_\rho}$		$(\hat{\Sigma}, \hat{L} + \mu)$
		Concentration for self-adjoint operators [36]	$\downarrow$
g_{n_ℓ}	$\widehat{S^\star g_\rho}$		$(\Sigma, L + \mu)$
		Concentration for vector in Hilbert space [58]	$\downarrow$
$g_{\lambda,\mu}$	$S^\star g_\rho$		$(\Sigma, L + \mu)$
		Bias controlled with source condition [10, 30]	$\downarrow$
g_λ	$S^\star g_\rho$		(Σ, L)
$\downarrow$		Bias controlled with source condition [10, 30]	
g_ρ			

Control of biases. We begin our study in a downward fashion regarding Table 2. Indeed, for Tikhonov regularization Eq. (3), we can show that for $q \in [0, 1]$,

$$\|g_\lambda - g_\rho\|_{L^2} \leq \lambda^q \|\mathcal{L}^q g_\rho\|_{L^2}.$$

Meaning that if we have the source condition $g_\rho \in \operatorname{im} \mathcal{L}^q$ (which is a condition on how fast $(\langle g_\rho, e_i\rangle)_{i\in\mathbb{N}}$ decreases compared to $(\lambda_i)_{i\in\mathbb{N}}$ for (λ_i, e_i) the eigen value decomposition of $\mathcal{L}^{-1}$), the rates of convergence of this term when n goes to infinity is controlled by the regularization scheme λ_n^q.

In a similar fashion to the kernel-free bias above, for $(q_i) \in (0, 1)^{\mathbb{N}}$, we can have

$$\|g_{\lambda,\mu} - g_\lambda\|_{L^2}^2 \leq 2 \sum_{i\in\mathbb{N}} \lambda^{2q_i} \mu^{2q_i} \left(\frac{\lambda_i}{\lambda + \lambda_i}\right)^2 |\langle e_i, g_\rho\rangle|^2 \left\|K^{-q_i}e_i\right\|_{L^2}^2.$$

This shows explicitly the usefulness of controlling at the same time how g_ρ is supported on the eigen spaces of $\mathcal{L}$ and how the eigen vectors are well approximated by the RKHS $\mathcal{H}$, which can be read in the value of (q_i) such that all $e_i \in \operatorname{im} K^{q_i}$.

Vector concentration. Let us now switch to concentration of $\widehat{S^\star g_\rho} = n_\ell^{-1} \sum_{i=1}^{n_\ell} Y_i k_{X_i}$ towards $S^\star g_\rho = \mathbb{E}_{(X,Y)\sim\rho}[Y k_X]$, it will allow to control $\|g_{n_\ell} - g_{\lambda,\mu}\|_{L^2}^2$ with the notations appearing in Table 2. Note how this convergence should be measured in term of the reconstruction error

$$\left\| \sum_{i\in\mathbb{N}} \psi(\lambda_{i,\mu}) \left\langle S^\star g_\rho - \widehat{S^\star g_\rho}, \theta_{i,\mu} \right\rangle S\theta_{i,\mu} \right\|_{L^2}.$$

This error might behave in a must better fashion than the L^2 error between $SS^\star g_\rho$ and $S\widehat{S^\star g_\rho}$. In particular, on Figure 1, we can consider $\psi(\lambda_{i,\mu}) = 0$ for $i > 4$, and we might have $\langle Y k_X, \theta_{i,\mu} \rangle = Y\mathbf{1}_{x\in C_i}$, for $i \leq 4$ and $(X,Y) \in \mathcal{X} \times \mathcal{Y}$, where C_i is the i-th innermost circle. In this setting, when all four $(Y \mid X \in C_i)$ are deterministic, we only need one labelled point per circle to clear the reconstruction error. Based on concentration results in Hilbert space, when $|Y|$ is bounded by a constant $c_{\mathcal{Y}}$, and $x \to k(x,x)$ by a constant κ^2, we have, with $\mathcal{D}_{n_\ell} \sim \rho^{\otimes n_\ell}$ the dataset generating the labelled data

$$\mathbb{E}_{\mathcal{D}_{n_\ell}} \left[\|g_{n_\ell} - g_{\lambda,\mu}\|_{L^2}^2 \right] \leq 2\sigma_\ell^2 (\mu\lambda n_\ell)^{-1} + \frac{4}{9} c_{\mathcal{Y}}^2 \kappa^2 (\mu\lambda n_\ell^2)^{-1}.$$

where $\sigma_\ell^2 \leq c_{\mathcal{Y}}^2 \operatorname{Tr}(\Sigma)$ is a variance parameter to relate to the variance of $Y(I + \lambda\mathcal{L})^{-1}\delta_X$ (where the randomness is inherited from $(X,Y) \sim \rho$). The fact that the need for labelled data depends on the variance of (X,Y) after being diffused through $\mathcal{L}$ is coherent with the results obtained by Lelarge and Miolane [29] in the specific case of a mixture of two Gaussians.

Basis concentration. We are left with the comparison of g_{n_ℓ}, which is the filtering of $\widehat{S^\star g_\rho}$ with the operators $(\Sigma, L + \mu)$, and $\hat{g}_p$, which is the filtering of the same vector with the operators $(P\hat{\Sigma}P, P\hat{L}P + \mu P)$. As the number of samples grows towards infinity, we know that $(P\hat{\Sigma}P, P\hat{L}P + \mu P)$ will converge in operator norm towards $(\Sigma, L + \mu)$. Yet, how to leverage this property to quantify the convergence of $\hat{g}_p$ towards g_{n_ℓ}? Let us write $(\lambda_i, \theta_i) = \operatorname{GEVD}(\Sigma, L + \mu)$, and $(\lambda_i', \theta_i') = \operatorname{GEVD}(P\hat{\Sigma}P, P\hat{L}P + \mu P)$, we have

$$\|\hat{g}_p - g_{n_\ell}\|_{L^2} = \left\| \sum_{i\in\mathbb{N}} \psi(\lambda_i) \left\langle \theta_i, \hat{\theta}_\rho \right\rangle S\theta_i - \psi(\lambda_i') \left\langle \theta_i', \hat{\theta}_\rho \right\rangle S\theta_i' \right\| \qquad \text{with} \qquad \hat{\theta}_\rho = \widehat{S^\star g_\rho}.$$

The generic study of this quantity requires to control eigenspaces one by one. Note that we expect convergence of eigenspaces to depend on gaps between eigenvalues. However, when considering Tikhonov regularization, this quantity can be written under a simpler form. In particular, the concentration of operators is controlled, up to few leftovers, through the quantity $\left\| (\Sigma + \lambda L + \mu\lambda)^{-1/2}((\Sigma - \hat{\Sigma}) + \lambda(L - \hat{L}))(\Sigma + \lambda L + \mu\lambda)^{-1/2} \right\|_{\operatorname{op}}$ where $\|\cdot\|_{\operatorname{op}}$ designs the operator norm. In this setting, the low-rank approximation is controlled through $\left\| (\Sigma + \lambda L)^{1/2}(I - P) \right\|_{\operatorname{op}}$, and when $L \preceq c\Sigma^\alpha$, this term can be controlled by $\left\| \Sigma^{1/2}(I - P) \right\|_{\operatorname{op}} + \lambda^{1/2} \left\| \Sigma^{1/2}(I - P) \right\|_{\operatorname{op}}^\alpha$ which can be controlled based on the work of Rudi et al. [45].

E.2 Risk decomposition

In this subsection, we decompose the risk appearing in Theorem 1.

E.2.1 Control of biases

We begin by splitting the error $\|g_\rho - \hat{g}_p\|_{L^2}$ between a bias term due to the regularization parameters and a variance term due to the data. With the notation of Table 2,

$$\|g_\rho - \hat{g}_p\|_{L^2} \leq \|g_\rho - g_\lambda\|_{L^2} + \|g_\lambda - g_{\lambda,\mu}\|_{L^2} + \|g_{\lambda,\mu} - \hat{g}_p\|_{L^2}. \tag{21}$$

We will control the first two terms here, and the last term in the following subsections.

Proposition 11 (Bias in λ). *Under Assumption 1*

$$\|g_\lambda - g_\rho\|_{L^2} \leq \lambda \|\mathcal{L}g_\rho\|_{L^2}. \tag{22}$$

Proof. Based on the definition of $g_\lambda = (I + \lambda\mathcal{L})^{-1}g_\rho$, we have

$$g_\lambda - g_\rho = ((I + \lambda\mathcal{L})^{-1} - I)g_\rho = -\lambda\mathcal{L}g_\rho.$$

Because g_ρ is supported on the first eigen vectors of the Laplacian (Assumption 1), we have $g_\rho = \sum_{i=1}^r \langle g_\rho, e_i \rangle e_i$, with e_i the eigen vector of $\mathcal{L}$ appearing in Eq. (17), and

$$\|\mathcal{L}g_\rho\|_{L^2}^2 = \left\| \sum_{i=1}^r \lambda_i^{-1} \langle g_\rho, e_i \rangle e_i \right\|_{L^2}^2 = \sum_{i=1}^r \lambda_i^{-2} \langle g_\rho, e_i \rangle^2 \leq \lambda_r^{-2} \|g_\rho\|_{L^2}^2 < +\infty.$$

This ends the proof of this proposition. $\qquad\square$

Proposition 12 (Bias in μ). *Under Assumptions 1 and 2, we have*

$$\|g_{\lambda,\mu} - g_\lambda\|_{L^2}^2 \leq \lambda\mu c_a^2 \|g_\rho\|_{L^2}^2, \qquad with \qquad c_a^2 = \sum_{i=1}^r \left\| K^{-1/2}e_i \right\|_{L^2}^2 = \sum_{i=1}^r \|e_i\|_{\mathcal{H}}^2. \qquad (23)$$

Proof. Before diving into the proof, recall that the RKHS norm penalization can be written as $\|g\|_{\mathcal{H}} = \|K^{-1/2}g\|_{L^2}$. Using the fact that $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, we have

$$\begin{aligned}
g_{\lambda,\mu} - g_\lambda &= ((I + \lambda\mathcal{L} + \mu\lambda K^{-1})^{-1} - (I + \lambda\mathcal{L})^{-1})g_\rho \\
&= -(I + \lambda\mathcal{L} + \mu\lambda K^{-1})^{-1}\lambda\mu K^{-1}(I + \lambda\mathcal{L})^{-1}g_\rho \\
&= -(\lambda\mu)^{1/2}(I + \lambda\mathcal{L} + \mu\lambda K^{-1})^{-1/2}(I + \lambda\mathcal{L} + \mu\lambda K^{-1})^{-1/2} \\
&\qquad \cdots \times (\lambda\mu K^{-1})^{1/2}K^{-1/2}(I + \lambda\mathcal{L})^{-1}g_\rho.
\end{aligned}$$

As a consequence,

$$\|g_{\lambda,\mu} - g_\lambda\|_{L^2}^2 \leq \lambda\mu \left\| K^{-1/2}(I + \lambda\mathcal{L})^{-1}g_\rho \right\|_{L^2}^2,$$

where we used the fact that $I + \lambda\mathcal{L} + \mu\lambda K^{-1} \succeq I$, so that $\left\|(I + \lambda\mathcal{L} + \mu\lambda K^{-1})^{-1/2}\right\|_{\mathrm{op}} \leq 1$ (with $\|\cdot\|_{\mathrm{op}}$ the operator norm), and that

$$\begin{aligned}
\left\|(I + \lambda\mathcal{L} + \mu\lambda K^{-1})^{-1/2}(\lambda\mu K^{-1})^{1/2}\right\|_{\mathrm{op}}^2 &= \lambda\mu \left\| K^{-1/2}(I + \lambda\mathcal{L} + \mu\lambda K^{-1})^{-1}K^{-1/2} \right\|_{\mathrm{op}} \\
&= \lambda\mu \left\| (K + \lambda K^{1/2}\mathcal{L}K^{1/2} + \mu\lambda)^{-1} \right\|_{\mathrm{op}} \leq 1.
\end{aligned}$$

We continue the proof with

$$\begin{aligned}
\left\| K^{-1/2}(I + \lambda\mathcal{L})^{-1}g_\rho \right\| &= \left\| \sum_{i=1}^r \frac{\lambda_i}{\lambda + \lambda_i} \langle g_\rho, e_i \rangle K^{-1/2}e_i \right\| \leq \sum_{i=1}^r \frac{\lambda_i}{\lambda + \lambda_i} |\langle g_\rho, e_i \rangle| \left\| K^{-1/2}e_i \right\| \\
&\leq \sum_{i=1}^r |\langle g_\rho, e_i \rangle| \left\| K^{-1/2}e_i \right\|_{L^2} \leq \|g_\rho\|_{L^2} \left(\sum_{i \leq r} \left\| K^{-1/2}e_i \right\|_{L^2}^2 \right)^{1/2}.
\end{aligned}$$

Putting all the pieces together ends the proof. $\qquad\square$

E.2.2 Vector concentration

We are left with the study of the variance $\|\hat{g}_p - g_{\lambda,\mu}\|$. To ease derivations, we denote $C = \Sigma + \lambda L$, $\hat{C} = \hat{\Sigma} + \lambda\hat{L}$, $\theta_\rho = S^\star g_\rho$, $\hat{\theta}_\rho = \widehat{S^\star g_\rho}$ and P the projection from $\mathcal{H}$ to $\mathrm{Span}\{k_{X_i}\}_{i \leq p}$. We have, for Tikhonov regularization

$$\begin{aligned}
\|\hat{g}_p - g_{\lambda,\mu}\|_{L^2} &= \left\| S\left(P(P\hat{C}P + \lambda\mu)^{-1}P\hat{\theta}_\rho - (C + \lambda\mu)^{-1}\theta_\rho \right) \right\|_{L^2} \\
&= \left\| \Sigma^{1/2}\left(P(P\hat{C}P + \lambda\mu)^{-1}P\hat{\theta}_\rho - (C + \lambda\mu)^{-1}\theta_\rho \right) \right\|_{\mathcal{H}}.
\end{aligned}$$

We begin by isolating the dependency to labelled data

$$\|\hat{g}_p - g_{\lambda,\mu}\|_{L^2} \leq \left\|\Sigma^{1/2}P(P\hat{C}P + \lambda\mu)^{-1}(\hat{\theta}_\rho - \theta_\rho)\right\|_{\mathcal{H}}$$
$$\cdots + \left\|\Sigma^{1/2}\left(P(P\hat{C}P + \lambda\mu)^{-1}\theta_\rho - (C + \lambda\mu)^{-1}\theta_\rho\right)\right\|_{\mathcal{H}}. \tag{24}$$

We will control the first term here, and the second term in the following subsection.

Lemma 13 (Vector term). *When* $\left\|(C + \lambda\mu)^{-1/2}(\hat{C} - C)(C + \lambda\mu)^{-1/2}\right\|_{\mathrm{op}} \leq 1/2$, *we have*

$$\left\|\Sigma^{1/2}P(P\hat{C}P + \lambda\mu)^{-1}P(\hat{\theta}_\rho - \theta_\rho)\right\|_{\mathcal{H}} \leq 2\left\|(C + \lambda\mu)^{-1/2}(\hat{\theta}_\rho - \theta_\rho)\right\|_{\mathcal{H}}. \tag{25}$$

Proof. We begin with the splitting

$$\left\|\Sigma^{1/2}P(P\hat{C}P + \lambda\mu)^{-1}P(\hat{\theta}_\rho - \theta_\rho)\right\|_{\mathcal{H}} \leq \left\|\Sigma^{1/2}P(P\hat{C}P + \lambda\mu)^{-1}P(C + \lambda\mu)^{1/2}\right\|_{\mathrm{op}}$$
$$\cdots \times \left\|(C + \lambda\mu)^{-1/2}(\hat{\theta}_\rho - \theta_\rho)\right\|_{\mathcal{H}}.$$

The first term will concentrate towards a matrix smaller than identity, while the second term concentrates towards zero. We can make those considerations more formal. Following basic properties with the Löwner order on operators, we have

$$(C + \lambda\mu)^{-1/2}(C - \hat{C})(C + \lambda\mu)^{-1/2} \preceq t$$
$$\Rightarrow \quad \hat{C} \succeq (1 - t)C - t\lambda\mu$$
$$\Rightarrow \quad P\hat{C}P \succeq (1 - t)PCP - t\lambda\mu P \succeq (1 - t)PCP - t\lambda\mu$$
$$\Rightarrow \quad P\hat{C}P + \lambda\mu \succeq (1 - t)(PCP + \lambda\mu)$$
$$\Rightarrow \quad (P\hat{C}P + \lambda\mu)^{-1} \preceq (1 - t)^{-1}(PCP + \lambda\mu)^{-1}$$
$$\Rightarrow \quad (C + \lambda\mu)^{1/2}P(P\hat{C}P + \lambda\mu)^{-1}P(C + \lambda\mu)^{1/2}$$
$$\preceq (1 - t)^{-1}(C + \lambda\mu)^{1/2}P(PCP + \lambda\mu)^{-1}P(C + \lambda\mu)^{1/2} \preceq (1 - t)^{-1},$$

where we have used the fact that the last operator is a projection. As a consequence, for any $t \in (0, 1)$, we have

$$\left\|(C + \lambda\mu)^{-1/2}(\hat{C} - C)(C + \lambda\mu)^{-1/2}\right\|_{\mathrm{op}} \leq t$$
$$\Rightarrow \quad \left\|(C + \lambda\mu)^{1/2}P(P\hat{C}P + \lambda\mu)^{-1}(C + \lambda\mu)^{1/2}\right\|_{\mathrm{op}} \leq (1 - t)^{-1}.$$
$$\Rightarrow \quad \left\|\Sigma^{1/2}P(P\hat{C}P + \lambda\mu)^{-1}P(C + \lambda\mu)^{1/2}\right\|_{\mathrm{op}} \leq (1 - t)^{-1}.$$

Where the last implication follows from the fact that $C + \lambda\mu = \Sigma + \lambda L + \lambda\mu \succeq \Sigma$. $\square$

E.2.3 Basis concentration

We are left with the study of the basis concentration with the number of unlabelled data.

Lemma 14 (Basis term). *When* $\left\|(C + \lambda\mu)^{-1/2}(\hat{C} - C)(C + \lambda\mu)^{-1/2}\right\|_{\mathrm{op}} \leq 1/2$, *we have*

$$\left\|\Sigma^{1/2}(P(P\hat{C}P + \lambda\mu)^{-1} - (C + \lambda\mu)^{-1})\theta_\rho\right\|_{\mathcal{H}}$$
$$\leq 3\left\|C^{1/2}(I - P)\right\|_{\mathrm{op}}\|g_\lambda\|_{\mathcal{H}} + 2\left\|(C + \lambda\mu)^{-1/2}(\hat{C} - C)(C + \lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}. \tag{26}$$

Notice that Assumptions 1 and 2 imply $\|g_\lambda\|_{\mathcal{H}} \leq c_a\|g_\rho\|_{L^2} < +\infty$.

Proof. First of all, using that $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, notice that

$$\left\|\Sigma^{1/2}(P(P\hat{C}P + \lambda\mu)^{-1} - (C + \lambda\mu)^{-1})\theta_\rho\right\|_{\mathcal{H}}$$

$$= \left\|\Sigma^{1/2}P(P\hat{C}P + \lambda\mu)^{-1}P(C - \hat{C}P)(C + \lambda\mu)^{-1}\theta_\rho - \Sigma^{1/2}(I - P)(C + \lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}$$

$$\leq \left\|\Sigma^{1/2}P(P\hat{C}P + \lambda\mu)^{-1}P(\hat{C}P - C)(C + \lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}} + \left\|\Sigma^{1/2}(I - P)(C + \lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}$$

$$\leq \left\|\Sigma^{1/2}P(P\hat{C}P + \lambda\mu)^{-1}P(C + \lambda\mu)^{1/2}\right\|_{\mathrm{op}}\left\|(C + \lambda\mu)^{-1/2}P(\hat{C}P - C)(C + \lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}$$

$$\cdots + \left\|\Sigma^{1/2}(I - P)\right\|_{\mathrm{op}}\left\|(C + \lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}.$$

Because $\Sigma \preceq \Sigma + \lambda L = C$, we have $\left\|\Sigma^{1/2}(I - P)\right\|_{\mathrm{op}} \leq \left\|C^{1/2}(I - P)\right\|_{\mathrm{op}}$, and we also have

$$\left\|(C + \lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}} \leq \left\|C^{-1}\theta_\rho\right\|_{\mathcal{H}} = \left\|K^{-1/2}SC^{-1}\theta_\rho\right\|_{\mathcal{H}} = \left\|K^{-1/2}g_\lambda\right\|_{L^2} = \left\|g_\lambda\right\|_{\mathcal{H}}.$$

Recall, that, for any $t \in (0, 1)$, we have already shown that

$$\left\|(C + \lambda\mu)^{-1/2}(\hat{C} - C)(C + \lambda\mu)^{-1/2}\right\|_{\mathrm{op}} \leq t$$

$$\Rightarrow \quad \left\|\Sigma^{1/2}P(P\hat{C}P + \lambda\mu)^{-1}P(C + \lambda\mu)^{1/2}\right\|_{\mathrm{op}} \leq (1 - t)^{-1}.$$

We are left with one last term to work on

$$\left\|(C + \lambda\mu)^{-1/2}P(\hat{C}P - C)(C + \lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}} \leq \left\|(C + \lambda\mu)^{-1/2}P(\hat{C} - C)P(C + \lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}$$

$$\cdots + \left\|(C + \lambda\mu)^{-1/2}C(I - P)(C + \lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}.$$

We control the first term with the fact for A, B, C three self-adjoint operators and x a vector we have

$$\|APBPCx\| = \|APBPCx \otimes xCPBPA\|_{\mathrm{op}}^{1/2},$$

and that

$$PCx \otimes xCP \preceq Cx \otimes xC$$

$$\Rightarrow \quad PBPCx \otimes xCPBP \preceq BPCx \otimes xCPB \preceq BCx \otimes xCB$$

$$\Rightarrow \quad APBPCx \otimes xCPBPA \preceq ABCx \otimes xCBA,$$

so that

$$\left\|(C + \lambda\mu)^{-1/2}P(\hat{C} - C)P(C + \lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}} \leq \left\|(C + \lambda\mu)^{-1/2}(\hat{C} - C)(C + \lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}.$$

We control the second term with

$$\left\|(C + \lambda\mu)^{-1/2}C(I - P)(C + \lambda\mu)^{-1}\theta_\rho\right\|$$

$$\leq \left\|(C + \lambda\mu)^{-1/2}C^{1/2}\right\|\left\|C^{1/2}(I - P)\right\|\left\|(C + \lambda\mu)^{-1}\theta_\rho\right\|.$$

Using that $(C + \lambda\mu)^{-1/2}C^{1/2} \preceq I$, we can add up everything to get the lemma.

For the part concerning $\|g_\lambda\|_{\mathcal{H}}$, notice that

$$\|g_\lambda\|_{\mathcal{H}} = \left\|K^{-1/2}g_\lambda\right\|_{L^2} = \left\|\sum_{i=1}^{r}\frac{\lambda_i}{\lambda_i + \lambda}\langle g_\rho, e_i\rangle K^{-1/2}e_i\right\|_{L^2} \leq \sum_{i=1}^{r}|g_\rho|\,e_i\left\|K^{-1/2}e_i\right\|$$

$$\leq \|g_\rho\|_{L^2}\left(\sum_{i=1}^{d}\left\|K^{-1/2}e_i\right\|_{L^2}^2\right)^{1/2} = c_a\|g_\rho\|_{L^2},$$

with c_a defined as before. $\qquad\square$

E.2.4 Conclusion

Based on the last subsections, we have proved the following proposition.

Proposition 15 (Risk decomposition). *Under the Assumptions 1 and 2, when* $\left\|(C + \lambda\mu)^{-1/2}(\hat{C} - C)(C + \lambda\mu)^{-1/2}\right\| \leq 1/2$,

$$
\|\hat{g}_p - g_\rho\|_{L^2}^2 \leq 4\lambda^2 \|\mathcal{L}g_\rho\|_{L^2}^2 + 4\lambda\mu c_a^2 \|g_\rho\|_{L^2}^2 + 8 \left\|(C + \lambda\mu)^{-1/2}(\hat{\theta}_\rho - \theta_\rho)\right\|_{\mathcal{H}}^2
$$
$$
\cdots + 12 c_a^2 \left\|C^{1/2}(I - P)\right\|_{\text{op}}^2 \|g_\rho\|_{L^2}^2 + 8 \left\|(C + \lambda\mu)^{-1/2}(\hat{C} - C)(C + \lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}^2. \tag{27}
$$

We are left with the quantification of the different convergences when the number of labelled and unlabelled data grows towards infinity. We will quantify those convergences based on concentration inequalities.

E.3 Probabilistic inequalities

In this subsection, we bound each term appearing in Eq. (27) based on concentration inequalities.

E.3.1 Vector concentration

The concentration of $\hat{\theta}_\rho = \widehat{S^\star g_\rho}$ towards $\theta_\rho = S^\star g_\rho$ is controlled through Bernstein inequality.

Theorem 2 (Concentration in Hilbert space [58]). *Let denote by $\mathcal{A}$ a Hilbert space and by (ξ_i) a sequence of independent random vectors in $\mathcal{A}$ such that $\mathbb{E}[\xi_i] = 0$, that are bounded by a constant M, with finite variance $\sigma^2 = \mathbb{E}[\sum_{i=1}^{n} \|\xi_i\|^2]$. For any $t > 0$,*

$$
\mathbb{P}(\left\|\sum_{i=1}^{n} \xi_i\right\| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2 + 2tM/3}\right).
$$

Proposition 16 (Vector concentration). *When $|Y|$ is bounded by a constant $c_{\mathcal{Y}}$, and $x \to k(x,x)$ by a constant κ^2, we have, with $\mathcal{D}_{n_\ell} \sim \rho^{\otimes n_\ell}$ the dataset generating the labelled data*

$$
\mathbb{P}_{\mathcal{D}_{n_\ell}} \left(\left\|(C + \lambda\mu)^{-1/2}(\hat{\theta}_\rho - \theta_\rho)\right\|_{\mathcal{H}} \geq t\right) \leq 2\exp\left(-\frac{n_\ell t^2}{2\sigma_\ell^2(\mu\lambda)^{-1} + 2tc_{\mathcal{Y}}(\mu\lambda)^{-1/2}\kappa/3}\right), \tag{28}
$$

where $\sigma_\ell^2 \leq c_{\mathcal{Y}}^2 \text{Tr}(\Sigma)$ is a variance parameter to relate with the variance of $Y(I + \lambda\mathcal{L})^{-1}\delta_X$ (where the randomness is inherited from $(X, Y) \sim \rho$).

Proof. Recall that

$$
(C + \lambda\mu)^{-1/2}(\hat{\theta}_\rho - \theta_\rho) = (\Sigma + \lambda L + \lambda\mu)^{-1/2}(n_\ell^{-1}\sum_{i=1}^{n_\ell} Y_i k_{X_i} - \mathbb{E}_\rho[Y k_X])
$$

We want to apply Bernstein inequality to the vector $\xi_i = (\Sigma + \lambda L + \mu\lambda)^{-1/2} Y_i k_{X_i}$, after centering it. Let us denote by $c_{\mathcal{Y}}$ a bound on $|Y|$, $c_{\mathcal{Y}} \in \mathbb{R}$ exists since we have supposed ρ of compact support. We have

$$
\sigma^2 = \mathbb{E}[\sum_{i=1}^{n_\ell} \|\xi_i - \mathbb{E}[\xi_i]\|^2] = n_\ell \mathbb{E}[\|\xi - \mathbb{E}[\xi_i]\|^2] \leq n_\ell \mathbb{E}[\|\xi\|^2]
$$
$$
= n_\ell \mathbb{E}_{(X,Y)\sim\rho} \left[Y^2 \left\langle k_X, (\Sigma + \lambda L + \mu\lambda)^{-1} k_X\right\rangle\right]
$$
$$
\leq n_\ell c_{\mathcal{Y}}^2 \mathbb{E}_{X\sim\rho_{\mathcal{X}}} \left[\left\langle k_X, (\Sigma + \lambda L + \mu\lambda)^{-1} k_X\right\rangle\right]
$$
$$
= n_\ell c_{\mathcal{Y}}^2 \text{Tr}\left((\Sigma + \lambda L + \mu\lambda)^{-1}\Sigma\right)
$$
$$
\leq n_\ell c_{\mathcal{Y}}^2 \text{Tr}(\Sigma) \left\|(\Sigma + \lambda L + \mu\lambda)^{-1}\right\|_{\text{op}} \leq n_\ell c_{\mathcal{Y}}^2 \text{Tr}(\Sigma)(\mu\lambda)^{-1}.
$$

Note that we have proceed with a generic upper bound, but we expect this variance, which is related to the variance of $Y(I + \lambda\mathcal{L} + \lambda\mu K^{-1})^{-1}\delta_X$ to be potentially much smaller – if we remove the term in P the vector concentration is the concentration of the vector $S(S^*S + \lambda S^\star \mathcal{L}S + \lambda\mu)^{-1}Y k_X \simeq$

$K^{1/2}(K + \lambda K^{1/2} \mathcal{L} K^{1/2} + \lambda\mu)^{-1} K^{1/2} S^{-\star} Y k_X = (I + \lambda L + \lambda\mu K^{-1})^{-1} Y S^{-\star} k_X \simeq (I + \lambda L + \lambda\mu K^{-1})^{-1} Y \delta_X$. To capture this fact, we will write $\sigma^2 \leq n_\ell \sigma_\ell^2 (\mu\lambda)^{-1}$, with $\sigma_\ell = c_{\mathcal{Y}} \operatorname{Tr}(\Sigma)^{1/2}$ in our analysis, but potentially much smaller under refined hypothesis and in practice. Similarly to the bound on the variance, we have

$$\|\xi - \mathbb{E}[\xi]\| \leq \|\xi\| = \left\| (\Sigma + \lambda L + \mu\lambda)^{-1/2} Y_i k_{X_i} \right\| \leq (\mu\lambda)^{-1/2} c_{\mathcal{Y}} \kappa,$$

with κ an upper bound on $k(x,x)^{1/2}$ for $x \in \operatorname{supp} \rho_{\mathcal{X}}$. As a consequence, applying Bernstein concentration inequality, we get, for any $t > 0$,

$$\mathbb{P}_{\mathcal{D}_{n_\ell}} \left(\left\| n_\ell^{-1} \sum_{i=1}^{n_\ell} \xi_i - E[\xi_i] \right\| \geq t \right) \leq 2 \exp \left(-\frac{n_\ell t^2}{2\sigma_\ell^2 (\mu\lambda)^{-1} + 2t c_{\mathcal{Y}} (\mu\lambda)^{-1/2} \kappa/3} \right).$$

This ends the proof. $\qquad\square$

E.3.2 Operator concentration

The convergence of $\hat{C}$ towards C is controlled with Bernstein inequality for self-adjoint operators.

Theorem 3 (Bernstein inequality for self-adjoint [36])**.** *Let $\mathcal{A}$ be a separable Hilbert space, and (ξ_i) a sequence of independent random self-adjoint operators operators on $\mathcal{A}$ Assume that (ξ_i) are bounded by $M \in \mathbb{R}$, in the sense that, almost everywhere, $\|\xi\|_{\mathrm{op}} < M$, and have a finite variance $\sigma^2 = \left\| \sum_{i=1}^n \mathbb{E}[\xi_i^2] \right\|_{\mathrm{op}}$. For any $t > 0$,*

$$\mathbb{P} \left(\left\| \sum_{i=1}^n (\xi_i - \mathbb{E}[\xi_i]) \right\|_{\mathrm{op}} > t \right)$$
$$\leq 2 \left(1 + 6 \frac{\sigma^2 + Mt/3}{t^2} \right) \frac{\operatorname{Tr}\left(\sum_{i=1}^n \mathbb{E}[\xi_i^2] \right)}{\left\| \sum_{i=1}^n \mathbb{E}[\xi_i^2] \right\|_{\mathrm{op}}} \exp \left(-\frac{t^2}{2\sigma^2 + 2tM/3} \right).$$

Proposition 17 (Operator concentration)**.** *When $x \to k(x,x)$ is bounded by κ^2 a nd $x \to \partial_{1,j}\partial_{2,j}k(x,x)$ is bounded by κ_j^2, we have*

$$\mathbb{P}_{\mathcal{D}_n} \left(\left\| (C + \mu\lambda)^{-\frac{1}{2}} (C - \hat{C})(C + \mu\lambda)^{-\frac{1}{2}} \right\|_{\mathrm{op}} > 1/2 \right) \leq \left(2 + 56 \frac{\kappa^2 + \lambda \sum_{i=1}^d \kappa_j^2}{\lambda\mu n} \right)$$
$$\cdots \times (1 + \lambda\mu \|C\|_{\mathrm{op}}^{-1}) \frac{\kappa^2 + \lambda \sum_{j=1}^d \kappa_j^2}{\lambda\mu} \exp \left(-\frac{\lambda\mu n}{10 \left(\kappa^2 + \lambda \sum_{j=1}^d \kappa_j^2 \right)} \right). \tag{29}$$

Proof. We want to apply the precedent concentration inequality to

$$\xi_i = (\Sigma + \lambda L + \lambda\mu)^{-1/2} \left(k_{X_i} \otimes k_{X_i} + \lambda \sum_{j=1}^d \partial_j k_{X_i} \otimes \partial_j k_{X_i} \right)(\Sigma + \lambda L + \lambda\mu)^{-1/2},$$

since we have, based on the fact that $C = \Sigma + \lambda L$ and that $\Sigma = \mathbb{E}[k_X \otimes k_X]$ and $L = \mathbb{E}[\sum_{j=1}^n \partial_j k_X \otimes \partial_j k_X]$,

$$\left\| (C + \lambda\mu)^{-1/2} (\hat{C} - C)(C + \lambda\mu)^{-1/2} \right\|_{\mathrm{op}} = n^{-1} \left\| \sum_{i=1}^n \xi_i - \mathbb{E}[\xi_i] \right\|_{\mathrm{op}}.$$

32

We bound ξ with

$$\|\xi\|_{\mathrm{op}} = \left\| (C + \mu\lambda)^{-\frac{1}{2}} \left(k_X \otimes k_X + \lambda \sum_{j=1}^{d} \partial_j k_X \otimes \partial_j k_X \right) (C + \mu\lambda)^{-\frac{1}{2}} \right\|_{\mathrm{op}}$$

$$\leq \mathrm{Tr} \left((C + \mu\lambda)^{-\frac{1}{2}} \left(k_X \otimes k_X + \lambda \sum_{j=1}^{d} \partial_j k_X \otimes \partial_j k_X \right) (C + \mu\lambda)^{-\frac{1}{2}} \right)$$

$$= \mathrm{Tr} \left((C + \mu\lambda)^{-\frac{1}{2}} k_X \otimes k_X (C + \mu\lambda)^{-\frac{1}{2}} \right)$$

$$\cdots + \lambda \sum_{j=1}^{d} \mathrm{Tr} \left((C + \mu\lambda)^{-\frac{1}{2}} \partial_j k_X \otimes \partial_j k_X (C + \mu\lambda)^{-\frac{1}{2}} \right)$$

$$= \left\| (C + \mu\lambda)^{-\frac{1}{2}} k_X \right\|_{\mathcal{H}}^2 + \lambda \sum_{j=1}^{d} \left\| (C + \mu\lambda)^{-\frac{1}{2}} \partial_j k_X \right\|_{\mathcal{H}}^2 \leq (\lambda\mu)^{-1} \left(\kappa^2 + \lambda \sum_{j=1}^{d} \kappa_j^2 \right).$$

With κ^2 an upper bound on the kernel k and κ_j^2 an upper bound on $\partial_{1,j}\partial_{2,j}k$. For the variance we have, using Löwner order,

$$\mathbb{E}[\xi^2] \preceq \sup_{X \in \mathcal{X}} \|\xi(X)\|_{\mathrm{op}} \mathbb{E}[\xi] \preceq (\lambda\mu)^{-1} \left(\kappa^2 + \lambda \sum_{j=1}^{d} \kappa_j^2 \right) \mathbb{E}[\xi]$$

$$= (\lambda\mu)^{-1} \left(\kappa^2 + \lambda \sum_{j=1}^{d} \kappa_j^2 \right) (C + \lambda)^{-1} C \preceq (\lambda\mu)^{-1} \left(\kappa^2 + \lambda \sum_{j=1}^{d} \kappa_j^2 \right).$$

Therefore, we get for any $t > 0$,

$$\mathbb{P}_{\mathcal{D}_n} \left(\left\| (C + \mu\lambda)^{-\frac{1}{2}} (C - \hat{C}) (C + \mu\lambda)^{-\frac{1}{2}} \right\|_{\mathrm{op}} > t \right)$$

$$\leq 2 \left(1 + 6 \frac{(\kappa^2 + \lambda \sum_{i=1}^{d} \kappa_j^2)(1 + t/3)}{\lambda\mu n t^2} \right) \frac{\|C\|_{\mathrm{op}} + \lambda\mu}{\|C\|_{\mathrm{op}}} \mathrm{Tr} \left((C + \lambda)^{-1} C \right)$$

$$\cdots \times \exp \left(-\frac{n t^2}{2(\lambda\mu)^{-1} \left(\kappa^2 + \lambda \sum_{j=1}^{d} \kappa_j^2 \right)(1 + t/3)} \right).$$

Remark that

$$\mathrm{Tr} \left((C + \mu\lambda)^{-1} C \right) \leq \left\| (C + \mu\lambda)^{-1} \right\|_{\mathrm{op}} \mathrm{Tr}(C) \leq (\lambda\mu)^{-1} (\kappa^2 + \lambda \sum_{j=1}^{d} \kappa_j^2).$$

Taking $t = 1/2$ ends the lemma. $\qquad \square$

E.3.3 Basis concentration

Similarly we could control $\left\| (C + \lambda\mu)^{-1/2} (\hat{C} - C)(C + \lambda\mu)^{-1} \theta_\rho \right\|_{\mathcal{H}}$ by using concentration of self-adjoint, yet this will lead to laxer bounds, than using concentration on vectors.

Proposition 18 (Basis concentration). *When* $x \to k(x,x)$ *is bounded by* κ^2, $x \to \partial_{1,j}\partial_{2,j}k(x,x)$ *is bounded by* κ_j^2, *with Assumptions 1 and 2, we have*

$$\mathbb{P}_{\mathcal{D}_n} \left(\left\| (C + \mu\lambda)^{-\frac{1}{2}} (C - \hat{C})(C + \mu\lambda)^{-1} \theta_\rho \right\|_{\mathcal{H}} > t \right) \leq 2 \exp \left(-\frac{\mu\lambda n t^2}{2 c_1 (c_1 + \lambda^{1/2}\mu^{1/2} t/3)} \right).$$

(30)

with $c_1 = (\kappa^2 + \lambda \sum_{i=1}^{d} \kappa_j^2) c_a \|g_\rho\|_{L^2}$.

Proof. We want to apply Bernstein concentration inequality to the vectors

$$\xi_i = (C + \mu\lambda)^{-1/2} \left(k_{X_i} \otimes k_{X_i} + \lambda \sum_{j=1}^{d} \partial_j k_{X_i} \otimes \partial_j k_{X_i} \right) (C + \lambda\mu)^{-1} \theta_\rho,$$

since

$$\left\| (C + \mu\lambda)^{-\frac{1}{2}} (C - \hat{C})(C + \mu\lambda)^{-1} \theta_\rho \right\|_{\mathcal{H}} = n^{-1} \left\| \sum_{i=1}^{n} \xi_i - \mathbb{E}[\xi_i] \right\|_{\mathcal{H}}.$$

We bound ξ, reusing prior derivations, with

$$\|\xi_i\|_{\mathcal{H}} = \left\| (C + \mu\lambda)^{-1/2} \left(k_{X_i} \otimes k_{X_i} + \lambda \sum_{j=1}^{d} \partial_j k_{X_i} \otimes \partial_j k_{X_i} \right) (C + \lambda\mu)^{-1} \theta_\rho \right\|_{\mathcal{H}}$$

$$\leq \left\| (C + \mu\lambda)^{-1/2} \right\|_{\mathrm{op}} \left\| \left(k_{X_i} \otimes k_{X_i} + \lambda \sum_{j=1}^{d} \partial_j k_{X_i} \otimes \partial_j k_{X_i} \right) \right\|_{\mathrm{op}} \left\| (C + \mu\lambda)^{-1} \theta_\rho \right\|_{\mathcal{H}}.$$

$$\leq (\mu\lambda)^{-1/2} (\kappa^2 + \lambda \sum_{i=1}^{d} \kappa_j^2) c_a \|g_\rho\|_{L^2}.$$

For the variance, we have, similarly to prior derivations,

$$\mathbb{E}[\|\xi\|^2] \leq \sup_{X \in \mathcal{X}} \left\| k_X \otimes k_X + \lambda \sum_{j=1}^{d} \partial_j k_X \otimes \partial_j k_X \right\|_{\mathrm{op}} \left\| (C + \lambda\mu)^{-1} \theta_\rho \right\|^2$$

$$\cdots \times \mathbb{E}\left[\left\| (C + \mu\lambda)^{-1} \left(k_X \otimes k_X + \lambda \sum_{j=1}^{d} \partial_j k_{X_i} \otimes \partial_j k_X \right) \right\|_{\mathrm{op}} \right]$$

$$\leq \left(\kappa^2 + \lambda \sum_{i=1}^{d} \kappa_i^2 \right) c_a^2 \|g_\rho\|_{L^2}^2$$

$$\cdots \mathbb{E}\left[\left\| (C + \mu\lambda)^{-1} k_X \otimes k_X \right\|_{\mathrm{op}} + \lambda \sum_{j=1}^{d} \left\| (C + \mu\lambda)^{-1} \partial_j k_{X_i} \otimes \partial_j k_X \right\|_{\mathrm{op}} \right]$$

$$= \left(\kappa^2 + \lambda \sum_{i=1}^{d} \kappa_i^2 \right) c_a^2 \|g_\rho\|_{L^2}^2 \operatorname{Tr}\left((C + \mu\lambda)^{-1} C \right)$$

$$\leq (\lambda\mu)^{-1} \left(\kappa^2 + \lambda \sum_{i=1}^{d} \kappa_i^2 \right)^2 c_a^2 \|g_\rho\|_{L^2}^2.$$

As a consequence, using Bernstein inequality,

$$\mathbb{P}\left(n^{-1} \left\| \sum_{i=1}^{n} \xi_i - \mathbb{E}[\xi_i] \right\| > t \right) \leq 2 \exp\left(-\frac{\mu\lambda n t^2}{2c_1(c_1 + \lambda^{1/2}\mu^{1/2}t/3)}, \right)$$

with $c_1 = (\kappa^2 + \lambda \sum_{i=1}^{d} \kappa_j^2) c_a \|g_\rho\|_{L^2}$. Note that we have bound naively the variable ξ and its variance, but we have shown how appears $\sup_{X \in \mathcal{X}} \left\| (C + \lambda\mu)^{-1} k_X \right\| + \lambda \sum_{i=1}^{d} \left\| (C + \lambda\mu)^{-1} \partial_i k_X \right\|$ and $\operatorname{Tr}((C + \lambda\mu)^{-1} C)$, which under interpolation and capacity assumptions but be controlled in a better fashion. $\square$

E.3.4 Low-rank approximation

We now switch to Nyström approximation.

Proposition 19 (Low-rank approximation). *When $x \to k(x, x)$ is bounded by κ^2, for any $p \in \mathbb{N}$ and $t > 0$, we have*

$$\mathbb{P}_{\mathcal{D}_p}\left(\left\|(I - P)\Sigma^{1/2}\right\|^2 > t\right) \leq \left(2 + \frac{116\kappa^2}{tp}\right)(2 + t\|\Sigma\|_{\mathrm{op}}^{-1})\frac{\kappa^2}{t}\exp\left(-\frac{pt}{10\kappa^2}\right),$$

Proof. Reusing Proposition 3 of Rudi et al. [45], for any $\gamma > 0$, we have, with P the projection on $\mathrm{Span}\{k_{X_i}\}_{i \leq p}$ and $\hat{\Sigma} = p^{-1}\sum_{i=1}^{p} k_{X_i} \otimes k_{X_i}$,

$$\left\|(I - P)\Sigma^{1/2}\right\|^2 \leq \gamma\left\|(\hat{\Sigma} + \gamma)^{-1/2}\Sigma^{1/2}\right\|_{\mathrm{op}}^2 \leq \gamma\left\|\Sigma^{1/2}(\hat{\Sigma} + \gamma)^{-1}\Sigma^{1/2}\right\|_{\mathrm{op}}.$$

As a consequence, skipping derivations that can be retaken from our precedent proofs,

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}_p}\left(\left\|(I - P)\Sigma^{1/2}\right\|^2 > t\right) &\leq \inf_{\gamma > 0} \mathbb{P}_{\mathcal{D}_p}\left(\gamma\left\|\Sigma^{1/2}(\hat{\Sigma} + \gamma)^{-1}\Sigma^{1/2}\right\|_{\mathrm{op}} > t\right) \\
&\leq \inf_{\gamma > 0} \mathbb{P}_{\mathcal{D}_p}\left(\left\|(\Sigma + \gamma)^{-1/2}(\hat{\Sigma} - \Sigma)(\Sigma + \gamma)^{-1/2}\right\|_{\mathrm{op}} > (1 - \gamma t^{-1})\right) \\
&\leq \inf_{\gamma > 0}\left(2 + 56\frac{\kappa^2}{\gamma p}\right)(1 + \gamma\|\Sigma\|_{\mathrm{op}}^{-1})\frac{\kappa^2}{\gamma}\exp\left(-\frac{p\gamma u^2}{2\kappa^2(1 + u/3)}\right),
\end{aligned}$$

with $u = (1 - \gamma t^{-1})$. Taking $\gamma = t/2$, this term is simplified as

$$\mathbb{P}_{\mathcal{D}_p}\left(\left\|(I - P)\Sigma^{1/2}\right\|^2 > t\right) \leq \left(2 + 116\frac{\kappa^2}{tp}\right)(2 + t\|\Sigma\|_{\mathrm{op}}^{-1})\frac{\kappa^2}{t}\exp\left(-\frac{pt}{10\kappa^2}\right),$$

which is the object of this proposition. $\qquad\square$

Lemma 20. *When $L \leq c_d\Sigma^a$, we have*

$$\left\|(I - P)C^{1/2}\right\|_{\mathrm{op}}^2 \leq \left\|(I - P)\Sigma^{1/2}\right\|_{\mathrm{op}}^2 + c_d\lambda\left\|(I - P)\Sigma^{1/2}\right\|_{\mathrm{op}}^{2a}. \tag{31}$$

Proof. This follows from the fact that

$$\begin{aligned}
\left\|C^{1/2}(I - P)\right\|_{\mathrm{op}}^2 &= \|(I - P)C(I - P)\|_{\mathrm{op}} = \|(I - P)(\Sigma + \lambda L)(I - P)\|_{\mathrm{op}} \\
&\leq \|(I - P)\Sigma(I - P)\|_{\mathrm{op}} + \lambda\|(I - P)L(I - P)\|_{\mathrm{op}} \\
&\leq \|(I - P)\Sigma(I - P)\|_{\mathrm{op}} + \lambda c_d\|(I - P)\Sigma^a(I - P)\|_{\mathrm{op}} \\
&= \left\|(I - P)\Sigma^{1/2}\right\|_{\mathrm{op}}^2 + \lambda c_d\left\|(I - P)\Sigma^{a/2}\right\|_{\mathrm{op}}^2 \\
&= \left\|(I - P)\Sigma^{1/2}\right\|_{\mathrm{op}}^2 + \lambda c_d\left\|(I - P)^a\Sigma^{a/2}\right\|_{\mathrm{op}}^2 \\
&\leq \left\|(I - P)\Sigma^{1/2}\right\|_{\mathrm{op}}^2 + \lambda c_d\left\|(I - P)\Sigma^{1/2}\right\|_{\mathrm{op}}^{2a},
\end{aligned}$$

where we used the fact that $(I - P)^a = (I - P)$ and that $\|A^s B^s\| \leq \|AB\|^s$ for $s \in [0, 1]$ and A, B positive self-adjoint. $\qquad\square$

E.4 Averaged excess of risk - Ending the proof

Based on the precedent excess of risk decomposition, and precedent concentration inequalities, we have all the elements to derive convergence rates of our algorithm. We will enunciate this convergence in term of the averaged excess of risk of $\mathbb{E}_{\mathcal{D}_n}\left[\|\hat{g}_p - g_\rho\|_{L^2}^2\right]$.

Lemma 21. *Under Assumptions 1 and 2,*

$$\mathbb{E}_{\mathcal{D}_n}\left[\|\hat{g}_p - g_\rho\|_{L^2}^2\right] \leq 4c_{\mathcal{Y}}^2\, \mathbb{P}\left(\left\|(C + \lambda\mu)^{-1/2}(\hat{C} - C)(C + \lambda\mu)^{-1/2}\right\| \leq 1/2\right)$$

$$\cdots + 4\lambda^2\, \|\mathcal{L}g_\rho\|_{L^2}^2 + 4\lambda\mu c_a^2\, \|g_\rho\|_{L^2}^2$$

$$\cdots + 8\,\mathbb{E}_{\mathcal{D}_n}\left[\left\|(C + \lambda\mu)^{-1/2}(\hat{\theta}_\rho - \theta_\rho)\right\|_{\mathcal{H}}^2\right] + 12c_a^2\, \|g_\rho\|_{L^2}^2\, \mathbb{E}_{\mathcal{D}_n}\left[\left\|C^{1/2}(I - P)\right\|_{\text{op}}^2\right] \quad (32)$$

$$\cdots + 8\,\mathbb{E}_{\mathcal{D}_n}\left[\left\|(C + \lambda\mu)^{-1/2}(\hat{C} - C)(C + \lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}^2\right].$$

Proof. We proceed using the fact that $\mathbb{E}[X] = \mathbb{E}[X \mid {}^c\!A]\,\mathbb{P}({}^c\!A) + \mathbb{E}[X \mid A]\,\mathbb{P}(A) \leq \sup X\,\mathbb{P}({}^c\!A) + \mathbb{E}[X \mid A]\,\mathbb{P}(A)$, with $A = \left\{\mathcal{D}_n \,\middle|\, \left\|(C + \lambda\mu)^{-1/2}(\hat{C} - C)(C + \lambda\mu)^{-1/2}\right\| \leq 1/2\right\}$,

$$\mathbb{E}_{\mathcal{D}_n}\left[\|\hat{g}_p - g_\rho\|_{L^2}^2\right] \leq \sup_{\mathcal{D}_n} \|\hat{g}_p - g_\rho\|^2\, \mathbb{P}\left({}^c\!A\right) + \mathbb{E}_{\mathcal{D}_n}\left[\|\hat{g}_p - g_\rho\|^2 \,\middle|\, A\right]\mathbb{P}(A).$$

When Y is bounded by $c_{\mathcal{Y}}$, because g_ρ is a convex combination of Y, we know that $\|g_\rho\|_{L^2} \leq c_{\mathcal{Y}}$, as a consequence, we can clip $\hat{g}_p$ to $[-c_{\mathcal{Y}}, c_{\mathcal{Y}}]$, which will only improve the estimation of g_ρ, as a consequence, we can consider the clipping estimate for which we have $\sup_{\mathcal{D}_n} \|\hat{g}_p - g_\rho\|_{L^2}^2 \leq 4c_{\mathcal{Y}}^2$. Regarding the second part, we have already decomposed the risk under the event $A = \left\{\mathcal{D}_n \,\middle|\, \left\|(C + \lambda\mu)^{-1/2}(\hat{C} - C)(C + \lambda\mu)^{-1/2}\right\| \leq 1/2\right\}$. As a consequence, we have

$$\mathbb{E}_{\mathcal{D}_n}\left[\|\hat{g}_p - g_\rho\|_{L^2}^2\right] \leq 4c_{\mathcal{Y}}^2\, \mathbb{P}({}^c\!A) + 4\lambda^2\, \|\mathcal{L}g_\rho\|_{L^2}^2\, \mathbb{P}(A) + 4\lambda\mu c_a^2\, \|g_\rho\|_{L^2}^2\, \mathbb{P}(A)$$

$$\cdots + 8\,\mathbb{E}_{\mathcal{D}_n}\left[\left\|(C + \lambda\mu)^{-1/2}(\hat{\theta}_\rho - \theta_\rho)\right\|_{\mathcal{H}}^2 \,\middle|\, A\right]\mathbb{P}(A)$$

$$\cdots + 12c_a^2\, \|g_\rho\|_{L^2}^2\, \mathbb{E}_{\mathcal{D}_n}\left[\left\|C^{1/2}(I - P)\right\|_{\text{op}}^2 \,\middle|\, A\right]\mathbb{P}(A)$$

$$\cdots + 8\,\mathbb{E}_{\mathcal{D}_n}\left[\left\|(C + \lambda\mu)^{-1/2}(\hat{C} - C)(C + \lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}^2 \,\middle|\, A\right]\mathbb{P}(A).$$

To control the conditional expectation, we use that, when X is positive

$$\mathbb{E}\left[X \mid A\right] P(A) = \mathbb{E}[X] - \mathbb{E}\left[X \mid {}^c\!A\right]\mathbb{P}({}^c\!A) \leq \mathbb{E}[X].$$

This ends the proof. $\qquad\square$

Based on deviation inequalities, we can control expectations based on the equality, for X positive, $\mathbb{E}[X] = \int_0^{+\infty} \mathbb{P}(X > t)\, \mathrm{d}t$.

Lemma 22. *In the setting of the paper,*

$$\mathbb{E}_{\mathcal{D}_n}\left[\left\|(C + \lambda\mu)^{-1/2}(\hat{\theta}_\rho - \theta_\rho)\right\|_{\mathcal{H}}^2\right] \leq 8\sigma_\ell^2(n_\ell\mu\lambda)^{-1} + 8c_{\mathcal{Y}}^2\kappa^2(n_\ell^2\mu\lambda)^{-1}. \quad (33)$$

Proof. First, recall that

$$\mathbb{P}\left(\left\|(C + \lambda\mu)^{-1/2}(\hat{\theta}_\rho - \theta_\rho)\right\|_{\mathcal{H}} > t\right) \leq 2\exp\left(-\frac{n_\ell t^2}{2\sigma_\ell^2(\mu\lambda)^{-1} + 2tc_{\mathcal{Y}}(\mu\lambda)^{-1/2}\kappa/3}\right)$$

$$\leq 2\exp\left(-\frac{n_\ell t^2}{2\max\left(2\sigma_\ell^2(\mu\lambda)^{-1}, 2tc_{\mathcal{Y}}(\mu\lambda)^{-1/2}\kappa/3\right)}\right)$$

$$\leq 2\exp\left(-\frac{n_\ell\mu\lambda t^2}{4\sigma_\ell^2}\right) + 2\exp\left(-\frac{3n_\ell\mu^{1/2}\lambda^{1/2}t}{4c_{\mathcal{Y}}\kappa}\right).$$

As a consequence

$$\mathbb{E}\left[\left\|(C+\lambda\mu)^{-1/2}(\hat{\theta}_\rho-\theta_\rho)\right\|_{\mathcal{H}}^2\right] = \int_0^{+\infty}\mathbb{P}\left(\left\|(C+\lambda\mu)^{-1/2}(\hat{\theta}_\rho-\theta_\rho)\right\|_{\mathcal{H}}^2 > t\right)\mathrm{d}t$$

$$\le 2\int\exp\left(-\frac{n_\ell\mu\lambda t}{4\sigma_\ell^2}\right)\mathrm{d}t + 2\int\exp\left(-\frac{3n_\ell\mu^{1/2}\lambda^{1/2}t^{1/2}}{4c_\mathcal{Y}\kappa}\right)\mathrm{d}t.$$

$$= 8\sigma_\ell^2(n_\ell\mu\lambda)^{-1} + \frac{64c_\mathcal{Y}^2\kappa^2}{9}(n_\ell^2\mu\lambda)^{-1}.$$

This is the result stated in the lemma. $\qquad\square$

Lemma 23. *In the setting of the paper,*

$$\mathbb{E}_{\mathcal{D}_n}\left[\left\|(C+\lambda\mu)^{-1/2}(\hat{C}-C)(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}^2\right] \le 8(\kappa^2+\lambda\partial\kappa^2)^2 c_a^2\left\|g_\rho\right\|_{L^2}^2$$
$$\cdots\times\left((\mu\lambda n)^{-1}+(\mu\lambda n^2)^{-1}\right),\tag{34}$$

with $\partial\kappa^2 = \sum_{i=1}^d\kappa_i^2$.

Proof. Let us denote by A the quantity $\left\|(C+\lambda\mu)^{-1/2}(\hat{C}-C)(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}$, and $\partial\kappa^2 = \sum_{i=1}^d\kappa_j^2$. Recall that

$$\mathbb{P}\left(A>t\right) \le 2\exp\left(-\frac{\mu\lambda n t^2}{2c_1(c_1+\lambda^{1/2}\mu^{1/2}t/3)}\right)$$

$$\le 2\exp\left(-\frac{\mu\lambda n t^2}{4c_1^2}\right) + 2\exp\left(-\frac{3(\mu\lambda)^{1/2}nt}{4c_1}\right).$$

We conclude the proof similarly to the precedent lemma. $\qquad\square$

Lemma 24. *Under Assumption 3,*

$$\mathbb{E}_{\mathcal{D}_n}\left[\left\|C^{1/2}(I-P)\right\|_{\mathrm{op}}^2\right] \le \left(\frac{10\kappa^2\log(p)}{p} + \frac{10^a\kappa^{2a}c_d\lambda\log(p)^a}{p^a}\right)$$
$$\cdots\times\left(1+\frac{2\kappa^2}{\|\Sigma\|_{\mathrm{op}}\log(p)}\left(1+\frac{6}{\log(p)}\right)\left(\frac{1}{p}+\frac{1}{5\log(p)}\right)\right).\tag{35}$$

Proof. Once again, this result comes from integration of the tail bound obtained on $\left\|C^{1/2}(I-P)\right\|_{\mathrm{op}}^2$ through the one we have on $\left\|\Sigma^{1/2}(I-P)\right\|_{\mathrm{op}}^2$ and the fact that $\left\|C^{1/2}(I-P)\right\|_{\mathrm{op}}^2 \le \left\|\Sigma^{1/2}(I-P)\right\|_{\mathrm{op}}^2 + c_d\lambda\left\|\Sigma^{1/2}(I-P)\right\|_{\mathrm{op}}^{2a}$. For any $a,b>0$, we have

$$\mathbb{E}_{\mathcal{D}_n}\left[\left\|\Sigma^{1/2}(I-P)\right\|_{\mathrm{op}}^2\right] = \int_0^\infty\mathbb{P}_{\mathcal{D}_n}\left(\left\|\Sigma^{1/2}(I-P)\right\|_{\mathrm{op}}^2 > t\right)\mathrm{d}t$$

$$\le \int_0^\infty\min\left\{1, 2\kappa^2\|\Sigma\|_{\mathrm{op}}^{-1}\left(1+\frac{58\kappa^2}{tp}\right)\left(1+\frac{2\kappa^2}{t}\right)\exp\left(-\frac{pt}{10\kappa^2}\right)\right\}\mathrm{d}t$$

$$= \frac{10\kappa^2 a}{p}\int_0^\infty\min\left\{1, 2\kappa^2\|\Sigma\|_{\mathrm{op}}^{-1}\left(1+\frac{58}{10au}\right)\left(1+\frac{p}{5au}\right)\exp\left(-au\right)\right\}\mathrm{d}u$$

$$\le \frac{10\kappa^2 a}{p}\left(b+\int_b^\infty 2\kappa^2\|\Sigma\|_{\mathrm{op}}^{-1}\left(1+\frac{6}{au}\right)\left(1+\frac{p}{5au}\right)\exp\left(-au\right)\mathrm{d}u\right)$$

$$\le \frac{10\kappa^2}{p}\left(ab+2\kappa^2\|\Sigma\|_{\mathrm{op}}^{-1}\left(1+\frac{6}{ab}\right)\left(1+\frac{p}{5ab}\right)\exp\left(-ab\right)\right).$$

This last quantity is optimized for $ab = \log(p)$, which leads to the first part of lemma. Similarly

$$\mathbb{E}_{\mathcal{D}_n}\left[\left\|\Sigma^{1/2}(I-P)\right\|_{\text{op}}^{2a}\right] = \int_0^\infty \mathbb{P}_{\mathcal{D}_n}\left(\left\|\Sigma^{1/2}(I-P)\right\|_{\text{op}}^{2a} > t\right)\mathrm{d}t$$

$$= \int_0^\infty \mathbb{P}_{\mathcal{D}_n}\left(\left\|\Sigma^{1/2}(I-P)\right\|_{\text{op}}^2 > t^{1/a}\right)\mathrm{d}t$$

$$\leq \int_0^\infty \min\left\{1, 2\kappa^2 \|\Sigma\|_{\text{op}}^{-1}\left(1 + \frac{58\kappa^2}{t^{1/a}p}\right)\left(1 + \frac{2\kappa^2}{t^{1/a}}\right)\exp\left(-\frac{pt^{1/a}}{10\kappa^2}\right)\right\}\mathrm{d}t$$

$$= \frac{10^a\kappa^{2a}ac^a}{p^a}\int_0^\infty \min\left\{u^{a-1}, 2\kappa^2 \|\Sigma\|_{\text{op}}^{-1}\left(1 + \frac{58}{10cu}\right)\left(1 + \frac{p}{5cu}\right)\frac{1}{u^{1-a}}\exp\left(-cu\right)\right\}\mathrm{d}u$$

$$\leq \frac{10^a\kappa^{2a}ac^a}{p^a}\left(\frac{b^a}{a} + \int_b^\infty 2\kappa^2 \|\Sigma\|_{\text{op}}^{-1}\left(1 + \frac{6}{cu}\right)\left(1 + \frac{p}{5cu}\right)\frac{1}{u^{1-a}}\exp\left(-cu\right)\mathrm{d}u\right)$$

$$\leq \frac{10^a\kappa^{2a}}{p^a}\left((cb)^a + 2\kappa^2 \|\Sigma\|_{\text{op}}^{-1}\left(1 + \frac{6}{cb}\right)\left(1 + \frac{p}{5cb}\right)\frac{1}{(cb)^{1-a}}\exp\left(-cb\right)\right).$$

Once again this is optimized for $cb = \log(p)$. $\qquad\square$

Remark 25 (Leverage scores). *Out of simplicity, we only present low rank approximation with random subsampling. Yet, we can improve the result by considering subsampling based on leverage scores. If we consider the Gaussian kernel, $Sk_x \in L^2$ can be thought as a function that is a little bump around $x \in \mathcal{X}$. In essence, subsampling based on leverage scores, consists in representing the solution on a subsampled sequence $(k_{X_i})_{i\in I}$ where the X_i are far from one another so that the bump functions (Sk_{X_i}) can approximate a maximum of functions. [45] shows that with leverage scores, we can take $p = (\mu\lambda)^\gamma \log(n)$, with γ linked with the capacity of the RKHS linked with the kernel k.*

If we add all derivations, we have derived the following theorem.

Theorem 4. *Under Assumptions 1, 2 and 3,*

$$\mathbb{E}_{\mathcal{D}_n}\left[\|\hat{g} - g_\rho\|_{L^2}^2\right]$$

$$\leq 8c_{\mathcal{Y}}^2\left(1 + 28\frac{\kappa^2 + \lambda\partial\kappa^2}{\lambda\mu n}\right)(1 + \lambda\mu \|C\|_{\text{op}}^{-1})\frac{\kappa^2 + \lambda\partial\kappa^2}{\lambda\mu}\exp\left(-\frac{\lambda\mu n}{10(\kappa^2 + \lambda\partial\kappa^2)}\right)$$

$$\cdots + 4\lambda^2 \|\mathcal{L}g_\rho\|_{L^2}^2 + 4\lambda\mu c_a^2 \|g_\rho\|_{L^2}^2 + 64\sigma_\ell^2(n_\ell\mu\lambda)^{-1} + 57c_{\mathcal{Y}}^2\kappa^2(n_\ell^2\mu\lambda)^{-1}$$

$$\cdots + 64(\kappa^2 + \lambda\partial\kappa^2)^2 c_a^2 \|g_\rho\|_{L^2}^2 (\mu\lambda n)^{-1} + 57(\kappa^2 + \lambda\partial\kappa^2)^2 c_a^2 \|g_\rho\|_{L^2}^2 (\mu\lambda n^2)^{-1} \qquad (36)$$

$$\cdots + 12c_a^2 \|g_\rho\|_{L^2}^2\left(\frac{10\kappa^2 \log(p)}{p} + \frac{10^a\kappa^{2a}c_d\lambda \log(p)^a}{p^a}\right)$$

$$\cdots \times \left(1 + \frac{2\kappa^2}{\|\Sigma\|_{\text{op}} \log(p)}\left(1 + \frac{6}{\log(p)}\right)\left(\frac{1}{p} + \frac{1}{5\log(p)}\right)\right).$$

where $c_{\mathcal{Y}}$ is an upper bound on Y, κ^2 is an upper bound on $x \to k(x,x)$, $\partial\kappa^2 = \sum_{i=1}^d \kappa_i^2$ with κ_i^2 a bound on $x \to \partial_{1i}\partial_{2i}k_{x_i}$, c_d and a the constants appearing in Assumption 3, c_a a constant such that $\|g\|_{\mathcal{H}} \leq c_a \|g\|_{L^2}$ and $\sigma_\ell^2 \leq c_{\mathcal{Y}}^2\kappa^2$ a variance parameter linked with the variance of $Y(I+\lambda\mathcal{L})^{-1}\delta_X$.

Theorem 1 is a corollary of this theorem.