Fast Learning Rates for Physics-Informed Kernel Methods

Luc Brogat-Motte
SAIL Unit, IIT, Genoa, Italy
luc.brogatmotte@iit.it

Joachim Bona-Pellissier
MaLGa Center, DIBRIS, UniGE
joachim.bona@edu.unige.it

Giacomo Meanti
MaLGa Center, DIBRIS, UniGE
giacomo.meanti@edu.unige.it

Lorenzo Rosasco
SAIL Unit, IIT, Genoa, Italy
MaLGa Center, DIBRIS, UniGE
lorenzo.rosasco@unige.it

Abstract

In physics-informed machine learning, a target function u^* is learned from noisy value observations $y_i = u^*(x_i) + \varepsilon_i$, together with differential information, given either by noisy observations $d_j = (Du^*)(z_j) + \xi_j$ or by a known physical constraint $Du^* = v$. We consider the setting where D is a linear differential operator and analyze a physics-informed kernel estimator $\hat{u}$ combining n value observations and m differential observations. In this context, we ask how much can differential information improve predictions, and how does this improvement depend quantitatively on n, m, and D. We prove finite-sample bounds, supported by numerical simulations, revealing a two-regime structure for the prediction error. When m is limited, the rate depends jointly on n and m; when m exceeds a problem-dependent threshold, the rate saturates and matches the oracle rate obtained when the perfect constraint $D\hat{u} = Du^*$ is imposed. Examples are discussed for Sobolev spaces which are reproducing kernel Hilbert spaces and include partial Laplacian constraints on the torus and gradient observations on bounded domains. These examples illustrate the range of possible learning rate improvements – from the standard nonparametric $n^{-1/4}$ to the parametric rate $n^{-1/2}$. Finally, we derive physically consistent rates in a stronger norm that jointly controls the errors in $\hat{u}$ and $D\hat{u}$.

1 Introduction

Supervised learning algorithms seek to estimate an unknown target function from noisy observations of its values. In many applications arising in physics, engineering, and scientific computing, however, one has access not only to value observations, but also to physical information involving derivatives or other differential quantities associated with the target function [1]. This additional information is often specified through functional constraints, such as gradients, Partial Differential Equation (PDE) residuals, or conservation laws [2–4]. Concrete examples include learning interatomic potentials, where energies and forces of atomic systems provide coupled value and gradient information through $F = -\nabla E$ [5, 6]; 3D reconstruction problems, where surface normals give differential information about the underlying shape [7, 8]; or computational cardiology, where physical models involve differential constraints from fluid dynamics [9, 10].

In this paper we investigate physics-informed machine learning from the point of view of statistical learning theory. While there exists a comprehensive theory describing the statistical performance of learning from value observations [11, 12], much less is known about how differential information affects learning rates and sample complexity. We are particularly interested in quantitative results

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

of how differential information improves prediction: what is the improvement it can provide? how many differential observations are needed for this gain to appear?

We study this question for target functions in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, which makes precise analyses feasible. We assume that there exists a target function $u^* \in \mathcal{H}$ and that, in addition to noisy value observations

$$y_i = u^*(x_i) + \varepsilon_i, \qquad i = 1, \dots, n,$$

also noisy differential quantities are available

$$d_j = (Du^*)(z_j) + \xi_j, \qquad j = 1, \dots, m,$$

where D is a linear differential operator.

Our contributions are as follows:

- We derive finite-sample bounds that quantitatively describe how the prediction error depends on sample sizes n, m, differential operator D, and regularization parameters λ, γ.
- We introduce a novel capacity assumption for physics-informed learning. This assumption decomposes the learning problem into a component that can be learned from differential data and a residual component not seen by D. It then quantifies *how fast* the first component can be learned from physics data and *how much* complexity remains in the residual one. We illustrate both quantities with two interpretable Sobolev RKHS examples, showing how they depend on the choice of D.
- Under this assumption, we obtain learning rates by deriving the best choice of λ and γ. The rates exhibit two regimes: a differential-limited regime, where the error decreases jointly with n and m, and a saturation regime, where the rate matches the one obtained when the physical constraint is known exactly. In the Sobolev space examples, the saturated prediction rates range from $n^{-1/4}$ to the parametric rate $n^{-1/2}$. Simulations further illustrate the predicted improvements and two-regime behavior.
- Finally, we show that the same estimator is physically consistent beyond value prediction. We prove learning rates in a stronger norm that jointly controls the errors in $\hat{u}$ and $D\hat{u}$, ruling out accurate value prediction without learning the differential structure.

1.1 Related work

Settings. Learning from both function values and derivatives (or, more generally, differential operators) is classically known as *Hermite-Birkhoff* interpolation, whose study in a statistical setting dates back at least to the seminal work of Kimeldorf and Wahba [13]. This setting has recently resurfaced in machine learning under the names Sobolev training [14] and physics-informed learning [15]. Incorporating differential information, whether in discrete or continuous form, can serve various purposes. Differential operators can be used to regularize the learning procedure: penalties of the form $\int (Du(x))^2 dx$ have been well studied for spline smoothing [16], inverse problems [17, 18], and machine learning [19]. Differential operators can also carry information about the data distribution. This is the case in semi-supervised learning, where the operator D is learned from unlabeled data [20–24]. Finally, differential constraints can come from the physical knowledge of the object of study itself, as is often the case in scientific applications [1]. Typically taking the form of a PDE, such constraints provide complementary data, either on the same domain as the value measurements [25] or on a different domain. A classical example of the latter is the one where Du^* is known inside a domain Ω while u^* is known only on the boundary $\partial\Omega$, which is well studied in the PDE literature [26].

Methods. Deep learning-based algorithms to tackle problems with differential constraints include the Deep Ritz Method [27], neural operators [28, 29], Universal Differential Equations [4]. Closest to our approach are Physics-Informed Neural Networks (PINNs) [2], which minimize a loss function similar to ours. Kernel methods have also been considered to tackle the problems described above, from splines [13] to RBF collocation [30–33], semi-supervised learning [22] and, more recently, in scientific machine learning applications [34–38]. Our estimator belongs to this kernel-based family: it combines the usual kernel ridge loss on values with an empirical loss on differential observations $Du(z_j)$. An orthogonal class of methods that is commonly used to estimate PDE solutions are finite elements/volumes [39, 25].

Theory. Many convergence results exist for learning PDE solutions with kernels [40, 33, 41, 42] or neural networks [43–46]. In the settings considered, data observations are only available on the boundary of the input domain (and not inside the domain). Such settings differ from ours and are closer to classical numerical methods for PDEs such as finite differences or finite element methods. Instead, we consider settings in which the value and differential observations lie in the same domain, and discuss the existing results for such settings, which are commonly of two types. A first type is concerned with physical consistency: showing that a physics-informed estimator $\hat{u}$ converges to the target u^* in a physically consistent, stronger than L^2 sense [47–49]. In Section 3.4, we provide a result of this nature with learning rates in a norm which controls both $\hat{u}$ and $D\hat{u}$. In particular, Doumèche et al. [47] shows, for a broad class of differential operators that PINNs are physically consistent. For kernel estimators, Shi et al. [48] and ul Abdeen et al. [49] show that learning with gradients allows to obtain convergence rates in the H^1 norm. A second type of result focuses on how the physics-informed penalty benefits the L^2 convergence rates, which we show in Section 3.3. Previously, Fisher et al. [50] studied the impact of gradient information on L^2 performance using a random feature model, and highlighted the existence of a regime where incorporating gradient information may hurt the prediction accuracy. Arnone et al. [51] investigated a regression estimator regularized with a 2-dimensional elliptic PDE, when u^* belongs to the Sobolev space H^2. Their estimator converges at least at the standard rate for H^2 functions ($n^{-2/3}$), and can reach the faster $n^{-4/5}$ rate when the PDE constraint is satisfied by the target. Doumèche et al. [37] reformulated empirical risk minimization with homogeneous linear differential constraints as kernel regression in a physics-informed RKHS. Their analysis yields theoretical evidence that physical constraints can improve convergence rates, characterized through the reduced effective dimension of the new kernel. Such improvements are illustrated for the setting of $D = \partial_x$ with homogeneous constraints ($Du^* = 0$). Their estimator is based on the full physical constraint, which corresponds to the case where the number m of physical measurements tends to infinity in the present paper. As a consequence, their estimator is not computable in closed form and must be approximated. This is why subsequent works [52, 53] focus on approximations and fast implementations based on Fourier decompositions. Overall, these theoretical results either study an ideal estimator for which continuous constraints are enforced [51, 37], or tie the number of value and differential data points ($m = n$) [48–50]. In contrast, the present work provides results which hold for arbitrary finite values of m and n, explicitly dealing with approximately enforced constraints. Our main result is to characterize the entire finite-m transition: how quickly differential observations improve prediction before saturation, how much improvement is attainable, how many observations are required to approach the oracle regime, and how the operator separately determines the visible and invisible complexities. To our knowledge, this joint dependence on arbitrary finite n and m has not previously been characterized.

Finally, we must mention works which characterized the statistical performance of kernel regression [11, 54, 55], which we extend to handle physical constraints. They introduce the key assumptions (source and capacity conditions) under which rates can be derived and the notion of *effective dimension* which controls the difficulty of the learning problem.

2 Physics-informed kernel regression

2.1 Problem setup

We now give a formal description of the learning problem. Let $\mathcal{X}$ be a measurable space and let $u^* : \mathcal{X} \to \mathbb{R}$ denote the target function. We will consider linear differential operators D of order $s \geq 1$, which take the form

$$D = \sum_{|\alpha| \leq s} c_\alpha \partial^\alpha,$$

with coefficients $c_\alpha \in \mathbb{R}$. We will assume (in (A2)) that Du^* is well defined. Let ρ and ρ_D be distributions over $\mathcal{X}$. We consider **value** and **differential** observations

$$\begin{aligned} y_i &= u^*(x_i) + \varepsilon_i, & x_i &\overset{i.i.d.}{\sim} \rho, & i &= 1, \dots, n, \\ d_j &= (Du^*)(z_j) + \xi_j, & z_j &\overset{i.i.d.}{\sim} \rho_D, & j &= 1, \dots, m, \end{aligned}$$

where $(\varepsilon_i)_i$ and $(\xi_j)_j$ are real-valued noise variables. In general, ρ and ρ_D can be different, reflecting the fact that value and differential information may be available on different regions of the domain.

2.2 Physics informed kernel estimator

Let $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ be an RKHS with associated kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Given parameters $\lambda > 0$ and $\gamma > 0$, the regularized empirical risk is defined, for $u \in \mathcal{H}$, as

$$\widehat{R}(u) = \frac{1}{n}\sum_{i=1}^{n}\big(u(x_i) - y_i\big)^2 + \gamma\frac{1}{m}\sum_{j=1}^{m}\big(Du(z_j) - d_j\big)^2 + \lambda\|u\|_{\mathcal{H}}^2, \tag{1}$$

and the Physics Informed Kernel MethodS (PIKS, [38]) estimator is the result of an optimization procedure in $\mathcal{H}$

$$\hat{u} := \operatorname*{argmin}_{u\in\mathcal{H}} \widehat{R}(u).$$

A difference from the estimator defined in [38] is the introduction of the scaling parameter γ. The parameter λ controls the regularity of the estimator, while γ controls the relative weight assigned to derivative observations. Setting $\gamma = 0$ recovers standard kernel ridge regression with value observations.

Since $\lambda > 0$, $\widehat{R}$ is a strictly convex quadratic functional on $\mathcal{H}$, hence the minimizer $\hat{u}$ exists and is unique. Moreover, the estimator admits a finite-dimensional representer form. In particular,

$$\hat{u}(\cdot) = \sum_{i=1}^{n}\alpha_i k(x_i,\cdot) + \sum_{j=1}^{m}\beta_j D_1 k(z_j,\cdot),$$

where the coefficients $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^m$ are the solutions to a linear system involving the value, mixed, and differential Gram matrices. The explicit system is given in Appendix A.1.

3 Main results

In this section, we present learning rates for PIKS and quantify the impact of derivative information. After listing the assumptions in Section 3.1, we state a general finite sample result in Section 3.2, from which, in Section 3.3, we derive learning rates, emphasizing the existence of two regimes. In Section 3.4 we study the rates in a stronger norm involving both u and Du.

3.1 Assumptions

Assumptions (A1) to (A5) are standard and ensure that derivative evaluations are well defined in the RKHS and that both value and derivative observations are generated by the same underlying target function. The main novel assumption is (A6), which quantifies how much of the function space is visible through derivative information.

Assumption (A1) (Linear differential operator)**.** *We consider linear differential operators D of the form $D = \sum_{|\alpha|\le s} c_\alpha \partial^\alpha$ for some integer $s \ge 1$ and coefficients $c_\alpha \in \mathbb{R}$.*

Typical examples include directional derivatives, Laplacians, and more general linear partial differential operators. The order s determines the degree of smoothness required in the kernel such that derivative evaluations are well defined.

Assumption (A2) (Kernel regularity)**.** *We assume that the kernel k is of class $C^{s,s}$ on $\mathcal{X}\times\mathcal{X}$, meaning that all mixed derivatives $\partial_x^\alpha\partial_{x'}^\beta k(x,x')$ with $|\alpha|,|\beta| \le s$ exist and are continuous.*

Under assumptions (A1) and (A2), the differential operator D can be applied to the kernel and gives rise to well-defined derivative representers in the RKHS. More precisely, let $\phi(x) := k(x,\cdot)$ be the RKHS feature map. Since derivative evaluations are continuous linear functionals on $\mathcal{H}$ [56, Corollary 4.36], and defining $D\phi(x) := Dk(x,\cdot)$, where D acts on the first argument of k, one has $D\phi(x) \in \mathcal{H}$ and the reproducing property holds on both the kernel and its derivatives:

$$u(x) = \langle u, \phi(x)\rangle_{\mathcal{H}}, \qquad Du(z) = \langle u, D\phi(z)\rangle_{\mathcal{H}}.$$

Assumption (A3) (Boundedness of features and outputs)**.** *There exist constants $\kappa_V, \kappa_D, L_V, L_D > 0$ such that, almost surely,*

$$\|\phi(x)\|_{\mathcal{H}} \le \kappa_V, \qquad \|D\phi(z)\|_{\mathcal{H}} \le \kappa_D, \qquad |y| \le L_V, \qquad |d| \le L_D.$$

Such boundedness assumptions are standard in statistical analyses of kernel methods and are satisfied under assumptions (A1) and (A2) when $\mathcal{X}$ is compact.

Assumption (A4) (Attainability)**.** *We assume that the regression function $u^*(x) := \mathbb{E}[y \mid x]$ belongs to $\mathcal{H}$. Equivalently, there exists $w^* \in \mathcal{H}$ such that*

$$u^*(x) = \langle w^*, \phi(x) \rangle_{\mathcal{H}}.$$

This is a standard assumption for well-specified problems: it states that the target function belongs to the RKHS induced by the kernel, so it can be represented by the model class under consideration.

Assumption (A5) (Consistency of derivative observations)**.** *We assume that the derivative observations are unbiased measurements of the derivative of the regression function, in the sense that*

$$\mathbb{E}[d \mid z] = Du^*(z).$$

This assumption states that, when Du^* is known, it can be interpreted as having a physical constraint which is well-specified. It is also referred to as using a *perfect* measure ρ_D [48].

To quantify both the effective dimension of the hypothesis space and its alignment with value and derivative observations, we introduce the following covariance operators

$$\Sigma := \mathbb{E}[\phi(x) \otimes \phi(x)], \qquad \Sigma_D := \mathbb{E}[D\phi(z) \otimes D\phi(z)],$$

where, for $a, b \in \mathcal{H}$, the rank-one operator $a \otimes b : \mathcal{H} \to \mathcal{H}$ is defined as $(a \otimes b)w := \langle w, b \rangle_{\mathcal{H}}\, a$. For self-adjoint operators A, B, we write $A \preceq B$ when $B - A$ is positive semi-definite.

Assumption (A6) (Value-derivative capacity decomposition)**.** *Assume that there exist positive semi-definite operators $\Sigma_1, \Sigma_2 \succeq 0$ such that*

$$\Sigma = \Sigma_1 + \Sigma_2,$$

and that there exist constants $c_1, c_2 > 0$ and $\alpha, r \in [0, 1]$ such that, for all $\lambda \in (0, 1]$,

$$\mathrm{Tr}\big(\Sigma_1(\Sigma_1 + \lambda I)^{-1}\big) \;\le\; c_1\, \lambda^{-\alpha}, \qquad \mathrm{Tr}\big((\Sigma_D + \lambda I)^{-1}\Sigma_2\big) \;\le\; c_2\, \lambda^{-r}. \tag{2}$$

This assumption decomposes the geometry of function values into directions that are *invisible*, or only weakly visible, to the operator D (captured by Σ_1) and directions that are *detectable* through differential observations (captured by Σ_2). The exponent α governs the effective dimension of the invisible component, while r measures how well the detectable component is separated from the null-space of D. In the limiting case $\Sigma_1 = \Sigma$ and $\Sigma_2 = 0$, the assumption reduces to the usual effective-dimension condition of standard kernel regression. Together, these conditions formalize how differential information can reduce the statistical complexity of learning function values via capacity reduction, depending on interactions between kernel, differential operator, and data distribution.

Remark 1 (Determining the capacity exponents)**.** *The exponents α and r in Assumption (A6) depend on the kernel, the differential operator, and the sampling distributions. One approach is to derive the effective-dimension bounds from spectral properties of the covariance operators and inequalities controlling the L^2 norm of u by that of Du, up to the nullspace of D. Section 4 illustrates this approach using Fourier analysis for the partial Laplacian and the Poincaré inequality for gradients. Alternatively, the exponents could be estimated empirically through the scaling of $\mathrm{Tr}(\widehat{\Sigma}(\widehat{\Sigma}+\gamma\widehat{\Sigma}_D+\lambda I)^{-1})$ in λ and γ, whose population counterpart is bounded under Assumption (A6) by $\lambda^{-\alpha} + \gamma^{r-1}\lambda^{-r}$. Establishing statistical guarantees for such estimates remains open. In practice, PIKS does not require these exponents as inputs, and λ and γ can be selected by cross-validation.*

3.2 Finite-sample bounds

Theorem 3.1 (High-probability finite-sample bounds for PIKS)**.** *Under assumptions (A1) to (A5), let $\lambda > 0$, $\gamma > 0$ and let $\delta \in (0, 1/2)$. Define*

$$\Sigma_\gamma := \Sigma + \gamma\Sigma_D, \qquad \mathcal{B}(\lambda, \gamma) := \big\|(\Sigma_\gamma + \lambda I)^{-1/2} w^*\big\|_{\mathcal{H}},$$

and the (value and derivative) capacity terms

$$\mathcal{N}_V(\lambda, \gamma) := \big\|(\Sigma_\gamma + \lambda I)^{-1/2}\Sigma^{1/2}\big\|_{\mathrm{HS}}, \qquad \mathcal{N}_D(\lambda, \gamma) := \big\|(\Sigma_\gamma + \lambda I)^{-1/2}\Sigma_D^{1/2}\big\|_{\mathrm{HS}}.$$

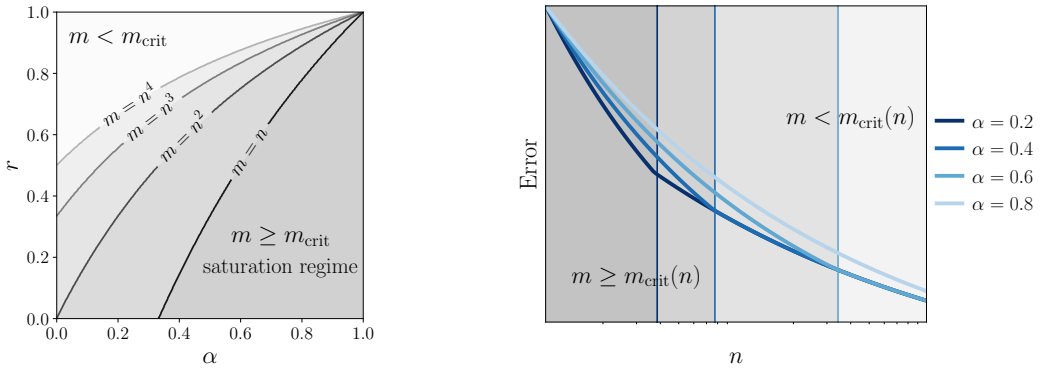

Figure 1: (*left*) Rate-regime changes in the (α, r) plane. On the lower-right of each contour line the problem parameters are such that m is larger than the critical threshold. (*right*) Log-linear plot of the expected error as a function of n and α for fixed m, r. For smaller α a larger m/n ratio is required to be in the fast saturated regime.

If

$$\frac{36}{n}\log\frac{2n}{\delta} \;\leq\; \lambda \;\leq\; \|\Sigma\|_\infty \quad \textit{and} \quad \frac{36}{m}\log\frac{2m}{\delta} \;\leq\; \lambda\gamma^{-1} \;\leq\; \|\Sigma_D\|_\infty, \tag{3}$$

then there exists a constant $c > 0$, depending only on $\kappa_V, \kappa_D, L_V, L_D$ and $\|u^*\|_{\mathcal{H}}$, but not on $\lambda, \gamma, n, m, \delta$, such that, with probability at least $1 - 2\delta$,

$$\mathbb{E}\left[(\hat{u}(x) - u^*(x))^2\right]^{\frac{1}{2}} \leq c\log\frac{4}{\delta}\left[\frac{1}{\sqrt{\lambda}}\left(\frac{1}{n} + \frac{\gamma}{m}\right) + \frac{\mathcal{N}_V(\lambda,\gamma)}{\sqrt{n}} + \gamma\frac{\mathcal{N}_D(\lambda,\gamma)}{\sqrt{m}} + \lambda\mathcal{B}(\lambda,\gamma)\right]. \tag{4}$$

Sketch of the proof. The proof stems from decomposing the error into separate approximation and estimation terms, and controlling the stochastic parts through concentration inequalities for covariance operators in Hilbert spaces. The main novelty compared to the classical kernel ridge regression analysis lies in handling of the derivative observations. We introduce the covariance operator $\Sigma_\gamma = \Sigma + \gamma\Sigma_D$, and decompose all deviation terms into value and derivative contributions while carefully tracking their dependence on γ, n, and m. Once these additional decompositions are established, the remainder of the argument closely parallels the standard KRR proof [11]. The full proof is given in Theorem A.9. □

The bound in Theorem 3.1 separates the contributions of value observations, derivative observations, and regularization bias. The quantities $\mathcal{N}_V(\lambda,\gamma)$ and $\mathcal{N}_D(\lambda,\gamma)$ play the role of effective dimensions associated with the value and derivative components of the problem. They control the corresponding variance contributions, while the bias term $\lambda\mathcal{B}(\lambda,\gamma)$ captures the approximation error induced by regularization. The contributions of $\mathcal{N}_V(\lambda,\gamma)$ and $\mathcal{N}_D(\lambda,\gamma)$ interact through the shared parameters λ and γ. Increasing γ places more weight on the derivative constraints, which reduces the effective complexity of the value component by shrinking the set of admissible functions in directions that are observable through D, but also amplifies the variance contribution coming from noisy derivative observations. Setting $\gamma = 0$ recovers the standard KRR bound [11]. Setting $m = \infty$ instead recovers the rates of Doumèche et al. [37].

3.3 Learning rate acceleration from differential information

We now instantiate the finite-sample bound under assumption (A6) and optimize over the regularization parameters λ and γ to precisely characterize the learning rates of the PIKS estimator.

Corollary 3.1 (Learning rates for PIKS)**.** *Under assumptions (A1) to (A6), there exist choices of regularization parameters $\lambda = \lambda(n,m)$ and $\gamma = \gamma(n,m)$, given explicitly in the proof and which depend polynomially on n, m, $\log n$, and $\log m$, such that, with high probability,*

$$\mathbb{E}\left[(\hat{u}(x) - u^*(x))^2\right]^{1/2} \;\lesssim\; \begin{cases} n^{-\frac{1}{4}}m^{-\frac{1-r}{8}}, & m \lesssim m_{\text{crit}}, \\ n^{-\frac{1}{2(1+\alpha)}}, & m \gtrsim m_{\text{crit}}, \end{cases}$$

where $m_{\mathrm{crit}} \asymp n^{\frac{2(1-\alpha)}{(1+\alpha)(1-r)}}$.

The proof is given in Corollary A.1. Here and throughout, $a(n,m) \lesssim b(n,m)$ means that $a(n,m) \leq c\, b(n,m)$ for some $c > 0$ independent of n and m, up to logarithmic factors in n, m, and δ^{-1} (the high-probability parameter).

The error bound reveals two regimes depending on the amount of differential samples m, relative to the number of function-value samples n as depicted also in Fig. 1:

- *Differential-limited regime ($m \lesssim m_{\mathrm{crit}}$).* With low m, the error decreases as both n and m increase. In particular as the number of differential observations m increases, the estimation error along directions captured by Σ_2 decreases. The rate of decrease depends on how well the Σ_2 directions are aligned with the differential covariance Σ_D. The exponent r quantifies this effect, with smaller values of r leading to faster decay of the error with respect to m.
- *Saturation threshold ($m \asymp m_{\mathrm{crit}}$).* Increasing m, a threshold is reached where differential and value contributions to the error are of the same order:
$$n^{-\frac{1}{4}} m^{-\frac{1-r}{8}} \asymp n^{-\frac{1}{2(1+\alpha)}},$$
which yields $m_{\mathrm{crit}} \asymp n^{\frac{2(1-\alpha)}{(1+\alpha)(1-r)}}$. The dependence of m_{crit} on α and r follows from this relation: smaller values of r lead to a faster decay of the differential term in m, reducing the required m, while smaller values of α lead to a faster decay of the value term in n, so that a larger m is required to balance the two contributions.
- *Saturation regime ($m \gtrsim m_{\mathrm{crit}}$).* Once m exceeds the threshold, the rate saturates: increasing the number of differential observations only reduces lower-order terms, but it no longer improves the leading error term. At this point, the error on the differential-accessible component is below the error driven by Σ_1, which captures directions that are either in the null-space of D, and therefore invisible to differential observations, or only weakly visible through them. These directions can be estimated only with value samples. In this regime, the leading rate matches the physical oracle rate obtained when the differential information is known exactly (see Appendix A.8).

For fixed n, these regimes provide guidance for differential-data acquisition. The exponent α determines the attainable gain, while r controls how quickly this gain is approached as m increases. The threshold m_{crit} provides an order-of-magnitude target for the differential-data budget. Beyond this threshold, additional observations increase computational cost without improving the leading learning rate.

Remark 2 (Value-derivative source decomposition)**.** *Our rates assume $u^* \in \mathcal{H}$ and focus on the capacity reduction induced by differential information. A natural extension would be to introduce a source-side counterpart of Assumption (A6), characterizing the alignment of the target with directions that are visible and invisible to differential observations. This could provide finer bounds on $\lambda\mathcal{B}(\lambda,\gamma)$ as a function of both λ and γ, leading to learning rates that reflect both the structure of the target and the capacity reduction. We leave this extension to future work.*

3.4 Learning rates in physically consistent norms

In many physics-informed settings, respecting the differential structure is as important as minimizing prediction error. For instance, in learning interatomic potentials, predicting differential quantities such as atomic forces is essential for molecular dynamics simulations [57]. However, improved learning rates for function values do not by themselves ensure that $\hat{u}$ captures this structure. Indeed, convergence in value alone does not control derivatives: on any bounded domain, $u_\varepsilon(x) = \varepsilon \sin(\varepsilon^{-2}x)$ satisfies $\|u_\varepsilon\|_{L^2} \to 0$ as $\varepsilon \to 0$, whereas for $D = \partial_x$, $Du_\varepsilon(x) = \varepsilon^{-1}\cos(\varepsilon^{-2}x)$ and $\|Du_\varepsilon\|_{L^2} \to \infty$. We therefore study convergence in a stronger L^2-type norm that jointly controls function-value and differential errors.

Corollary 3.2 (Learning rates in the physically consistent norm)**.** *Under assumptions (A1) to (A6) there exist choices of regularization parameters $\lambda = \lambda(n,m)$ and $\gamma = \gamma(n,m)$, given explicitly in the proof with $\gamma \geq 1$ such that, with high probability,*

$$\mathbb{E}\left[(\hat{u}(x) - u^*(x))^2 + (D\hat{u}(z) - Du^*(z))^2\right]^{1/2} \lesssim \begin{cases} n^{-\frac{1}{4}} m^{-\frac{1-r}{8}} + m^{-1/4}, & m \lesssim m_{\mathrm{crit}} \\ n^{-\frac{1}{2(1+\alpha)}} + m^{-1/4}, & m \gtrsim m_{\mathrm{crit}}, \end{cases}$$

where $m_{\text{crit}} \asymp n^{\frac{2(1-\alpha)}{(1+\alpha)(1-r)}}$.

The proof is given in Corollary A.2. The stronger norm bounds rule out the possibility that the PIKS estimator has a good prediction accuracy on u^* while failing to accurately learn Du^*. This can be interpreted as a form of physical consistency: the learned function respects both the values and the differential structure of the target. Compared with the value-only rates in Corollary 3.1, this stronger guarantee incurs an additional term $m^{-1/4}$, which depends only on the number of derivative observations and therefore cannot be reduced by increasing n alone. This is a reasonable price to pay in settings where differential predictions are themselves of interest or are used in downstream tasks.

4 Examples

Here we illustrate the improved rates of Corollary 3.1 in concrete learning settings. We focus in Section 4.1 on Sobolev spaces on the torus where Laplacian information can provide large benefits, and in Section 4.2 on bounded domains, which are particularly relevant for their connection to PDEs.

4.1 Example 1: Sobolev spaces on the torus

We define the periodic Matérn kernel on $\mathbb{T}^d = [0,1]^d$ with Fourier expansion

$$k(x - x') = \sum_{k \in \mathbb{Z}^d} \mu_k \, e^{2\pi i k \cdot (x - x')}, \qquad \mu_k \asymp (1 + |k|^2)^{-s},$$

for some $s > \frac{d}{2} + 2$, where $\asymp$ denotes the equality up to positive multiplicative constants. The associated RKHS $\mathcal{H}$ consists of functions $f(x) = \sum_{k \in \mathbb{Z}^d} c_k \, e^{2\pi i k \cdot x}$ with norm

$$\|f\|_{\mathcal{H}}^2 = \sum_{k \in \mathbb{Z}^d} \frac{|c_k|^2}{\mu_k} \asymp \sum_{k \in \mathbb{Z}^d} (1 + |k|^2)^s |c_k|^2,$$

where $c_k \in \mathbb{C}$ are the Fourier coefficients. $\mathcal{H}$ is norm-equivalent to the Sobolev space $H^s(\mathbb{T}^d)$.

Proposition 4.1 (Capacity decomposition for Sobolev spaces)**.** *Let $\mathcal{X} = \mathbb{T}^d$, let k be a periodic Matérn kernel of smoothness $s > \frac{d}{2} + 2$, and let $D = \sum_{i \in S} \partial_{z_i}^2$ be a partial Laplacian with $S \subseteq \{1, \dots, d\}$. Assume that value and differential samples are drawn uniformly on $\mathbb{T}^d$. Then the value covariance Σ admits a decomposition $\Sigma = \Sigma_1 + \Sigma_2$ satisfying assumption (A6), with exponents*

$$\alpha = \frac{d - |S|}{2s}, \qquad r = \frac{d}{2s}.$$

Proposition 4.1 provides a functional-analytic interpretation of assumption (A6). Combining it with Corollary 3.1 gives explicit learning rates for this Sobolev example. When $m = 0$, we recover standard KRR for estimating a function in $H^s(\mathbb{T}^d)$ from value observations, with rate $n^{-s/(2s+d)}$. As m increases, differential observations progressively constrain the directions outside the nullspace of the partial Laplacian D. Once m exceeds the saturation threshold, the learning rate is governed by the remaining component, which lies in the nullspace of D and depends only on $d_0 = d - |S|$ variables. In this regime, we obtain

$$\mathbb{E}\left[(\hat{u}(x) - u^*(x))^2\right]^{1/2} \lesssim n^{-\frac{s}{2s+d-|S|}}.$$

The saturated rate thus matches the minimax rate for Sobolev regression on $\mathbb{T}^{d_0}$, improving over standard KRR through the reduction from d to d_0 variables. In particular, the full Laplacian ($d_0 = 0$) yields the parametric rate $n^{-1/2}$. This makes explicit how differential information reduces the dimensionality of the problem.

4.2 Example 2: Gradients on bounded domain

Consider now the case when D is the gradient operator ∇, and the RKHS $\mathcal{H}$ is a Sobolev space of smoothness $s > \frac{d}{2} + 1$ on domain $\mathcal{X} \subset \mathbb{R}^d$. We consider the reproducing kernel k associated with $\mathcal{H}$ to be a Matérn kernel of smoothness s [58].

Note that, while in Sections 2 and 3 the analysis is limited to the case of scalar-valued D, it still holds for vector-valued operators. Denoting by D_i each scalar component of D, we decompose the differential-data covariance as $\Sigma_D := \sum_{\ell=1}^{k} \Sigma_{D_\ell}$, where $\Sigma_{D_\ell} := \mathbb{E}[D_\ell\phi(z) \otimes D_\ell\phi(z)]$. Provided this new Σ_D satisfies assumption (A6) (which is unchanged), Corollary 3.1 still holds. We give more details on the extension to the vector-valued case in Appendix A.7.

When $D = \nabla$, we have $\Sigma_D = \mathbb{E}\left[\sum_{i=1}^{d} \partial_i\phi(z) \otimes \partial_i\phi(z)\right]$. The effect of learning with differential constraints is shown in the following proposition.

Proposition 4.2. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a bounded, connected and Lipschitz domain. Fix $s > \frac{d}{2} + 1$. Let $\mathcal{H} = H^s(\Omega)$ and let $D = \nabla$ be the gradient operator. Then the value covariance operator Σ admits a decomposition $\Sigma = \Sigma_1 + \Sigma_2$ satisfying assumption (A6), with exponents*

$$\alpha = 0, \qquad r = \frac{d}{2(s-1)}.$$

The *saturated* rate from Corollary 3.1 is obtained as long as $m \geq m_{crit} = n^{\frac{2(s-1)}{(s-1)-d/2}} \geq n^2$. Then,

$$\mathbb{E}\left[(\hat{u}(x) - u^*(x))^2\right]^{1/2} \lesssim n^{-\frac{1}{2}},$$

which proves that the gradient information allows to obtain the parametric rate.

Remark 3 (Determining the capacity exponents)**.** *The examples above illustrate how the exponents α and r in Assumption (A6) can be derived from properties of the kernel, differential operator, and sampling distributions. For more general operators, analogous bounds could be obtained from coercivity or regularity estimates controlling u through Du, up to the nullspace of D. Such estimates may depend on the domain and boundary conditions. Alternatively, the exponents could be investigated empirically through the scaling of $\mathrm{Tr}(\widehat{\Sigma}(\widehat{\Sigma} + \gamma\widehat{\Sigma}_D + \lambda I)^{-1})$ with λ and γ. Establishing guarantees for such estimates remains an open question. In practice, PIKS does not require these exponents as inputs, and λ and γ can be selected by cross-validation.*

5 Numerical experiments

We show how the different rate-regimes of Corollary 3.1 look like in practice through simulations on two learning problems. We will show i) the *error saturation effect* when m increases and ii) the effect of decreasing the capacity of Σ_1 (decreasing α) via the partial Laplacian. More details on the implementation are available in Appendix C.

Saturation effect on Matérn data. We sample synthetic data on the 2d disk according to the following function, letting x_{supp} be a support point in the domain and $p := \|x - x_{\mathrm{supp}}\|$

$$u^*(x) = \left(1 - \|x\|^2\right)\left[\left(1 + \sqrt{5}p + (5/3)p^2\right)\exp\left(-\sqrt{5}p\right)\right].$$

u^* is in H^s with $s = 3$, and we will use the corresponding Matérn ($\nu = 5/2$) kernel. In Fig. 2 we show that, for fixed n and increasing m, the error rates switch between the two regimes of Corollary 3.1: an initial phase in which $m < m_{\mathrm{crit}}$ and the error decreases as m grows, followed by a phase in which m has grown beyond the threshold and the error has reached a *saturation* with respect to m in which additional data does not improve accuracy. Here we used $D = \nabla$ and plotted empirical rates using PIKS next to the theoretical results from Section 4.2. Despite some differences – notably the experimental curves saturate at similar values of m – due to the unknown problem-dependent constants affecting the rates, the saturation effect is clear as well as the dependence of the initial rate on both n and m.

Partial Laplacian. We now consider Sobolev spaces on the torus with partial Laplacian information. Unlike the example in Section 4.1, we approximate the infinite-dimensional Fourier construction by truncating to a finite number of frequencies. With $\mathcal{X} = \mathbb{T}^4$, we randomly sample frequency vectors $k_\ell \in \mathbb{Z}^4$ such that a single dimension is active (non-zero) at a time. Writing $\phi_\ell(x) = \sqrt{2}\cos(2\pi\langle k_\ell, x\rangle)$ for the feature-map, target function and kernel are defined as

$$u^*(x) = u_0 + \sum_{\ell=1}^{F} c_\ell\phi_\ell(x), \qquad k(x, x') = 1 + \sum_{\ell=1}^{F} \mu_\ell\phi_\ell(x)\phi_\ell(x'),$$

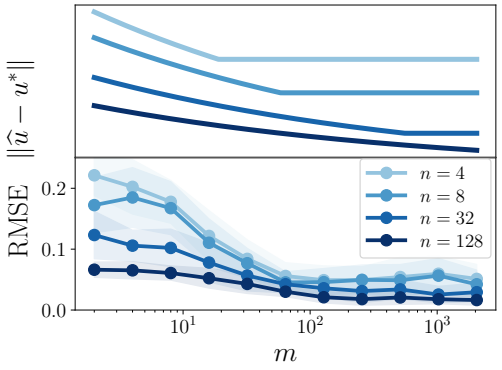

Figure 2: Theoretical and experimental rates for fixed n, increasing m showing the saturation effect with gradient information on Matérn data.

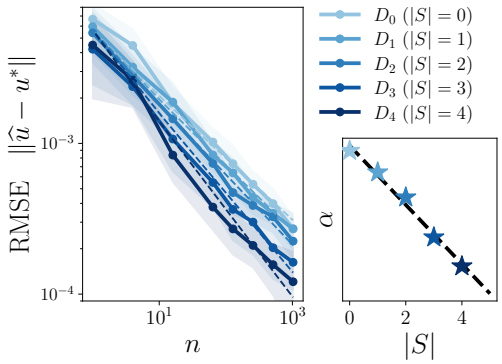

Figure 3: Test error rates increasing n and partial Laplacian dimensions $|S|$. Dashed lines show a best fit for the rates of Proposition 4.1. On the right are the inferred α exponents.

with F frequency coefficients c_ℓ and μ_ℓ which depend on a smoothness parameter of the problem. k_ℓ are sampled such that u^* decomposes into frequency blocks $u^*(x) = u_0 + u_1(x) + u_2(x) + u_3(x) + u_4(x)$, each of which only depends on a single dimension of x and cannot be observed unless the partial Laplacian $D_s = \sum_{i=1}^{s} \partial^2_{z_i}$ includes that specific dimension. This creates a simple setting in which increasing the number of dimensions included in the partial Laplacian order s, increases the amount of information which can be learned with the differential data. Moreover the Sobolev capacity decomposition from Proposition 4.1 remains valid after truncation, with constants independent of F.

In Fig. 3 we observe the learning rate in n, as the type of differential data changes: from having access to function data only (D_0) to having access to the full Laplacian (D_4). The number of differential points m is set as a constant factor of n. Best fits are obtained for the theoretical rates of Theorem 3.1 to obtain the exponent term of the error rate in n. Despite the high variance of the experiments, due to the noise ε, ξ and the random sampling of points x_i, z_j, it is evident that with each increase in the amount of differential information $|S|$ the learning rate increases. From the fitted curves we can infer the values of α which – as predicted by Proposition 4.1 – have a linear relationship with the number of partial Laplacian dimensions (denoted by $|S|$), despite the mismatch caused by the finite-dimensional setting.

6 Conclusion and research directions

In this paper, we provide a precise characterization of how kernel regression benefits from learning with both function-value and differential observations. Our analysis quantifies how the prediction error depends jointly on the number of value samples, the number of differential samples, and the structure of the underlying differential operator. The resulting rates reveal two regimes: a differential-limited regime, in which the error decreases with both types of samples, and a saturation regime, in which the leading rate matches the physical oracle rate attainable with exact differential information. We also establish convergence in a stronger norm that controls both function-value and differential errors. Several questions remain open: are these rates minimax optimal? Can similar gains be obtained for nonlinear differential operators or misspecified physical constraints? Is it possible to preserve the statistical gains with approximate algorithms which are more efficient?

Acknowledgements

This material is based upon work supported by the European Commission (Horizon Europe grant ELIAS 101120237), and the Ministry of Education, University and Research (FARE grant ML4IP R205T7J2KP). GM is partly supported by ERC starting project MALIN (Horizon Europe grant 101117133). This work represents only the view of the authors. The European Commission and the other organisations are not responsible for any use that may be made of the information it contains. The authors declare no competing interests.

References

[1] Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3), 2022.

[2] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[3] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6), 2021.

[4] Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal differential equations for scientific machine learning. *arXiv preprint arXiv:2001.04385*, 2020.

[5] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14):146401, 2007.

[6] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.*, 104, 2010. doi: 10.1103/PhysRevLett.104.136403.

[7] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3569–3579. 2020.

[8] Aarya Patel, Hamid Laga, and Ojaswa Sharma. Normal-guided detail-preserving neural implicit function for high-fidelity 3D surface reconstruction. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 8(1), 2025. doi: https://doi.org/10.1145/3728293.

[9] Alfio Quarteroni, Paola Gervasio, and Francesco Regazzoni. Combining physics-based and data-driven models: advancing the frontiers of research with scientific machine learning. *Mathematical Models and Methods in Applied Sciences*, 35(4):905–1071, 2025. doi: 10.1142/S0218202525500125.

[10] Tommaso Botarelli, Marco Fanfani, Paolo Nesi, and Lorenzo Pinelli. Using physics-informed neural networks for solving navier-stokes equations in fluid dynamic complex scenarios. *Engineering Applications of Artificial Intelligence*, 148, 2025. doi: https://doi.org/10.1016/j.engappai.2025.110347.

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

[12] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.

[13] George Kimeldorf and Grace Wahba. Some results on Tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

[14] Wojciech M Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. *Advances in neural information processing systems*, 30, 2017.

[15] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations, 2017.

[16] Grace Wahba. *Spline models for observational data*. SIAM, 1990.

[17] Martin Hanke. Regularization with differential operators: an iterative approach. *Numerical functional analysis and optimization*, 13(5-6):523–540, 1992.

[18] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[19] T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247(4945), 1990. doi: 10.1126/science.247.4945.978.

[20] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.

[21] Dengyong Zhou and Bernhard Schölkopf. Regularization on discrete spaces. In *Joint Pattern Recognition Symposium*, pages 361–368. Springer, 2005.

[22] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(11), 2006.

[23] Dejan Slepcev and Matthew Thorpe. Analysis of p-Laplacian regularization in semisupervised learning. *SIAM Journal on Mathematical Analysis*, 51(3):2085–2120, 2019.

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

[25] Laura M Sangalli. Spatial regression with partial differential equation regularisation. *International Statistical Review*, 89(3):505–531, 2021.

[26] Lawrence C Evans. *Partial differential equations*, volume 19. American mathematical society, 2010.

[27] Bing Yu et al. The deep Ritz method: a deep learning-based numerical algorithm for solving variational problems. *Communications in Mathematics and Statistics*, 6(1), 2018.

[28] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature machine intelligence*, 3(3), 2021.

[29] Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=c8P9NQVtmnO`.

[30] Edward J Kansa. Multiquadrics—a scattered data approximation scheme with applications to computational fluid-dynamics—ii solutions to parabolic, hyperbolic and elliptic partial differential equations. *Computers & mathematics with applications*, 19(8-9):147–161, 1990.

[31] Wu Zongmin. Hermite-Birkhoff interpolation of scattered data by radial basis functions. *Approximation Theory and its Applications*, 8(2):1–10, 1992.

[32] Gregory E Fasshauer. Solving partial differential equations by collocation with radial basis functions. In *Proceedings of Chamonix*, volume 1997, pages 1–8, 1996.

[33] Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

[34] Houman Owhadi. Bayesian numerical homogenization. *Multiscale Modeling & Simulation*, 13(3):812–828, 2015.

[35] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Machine learning of linear differential equations using Gaussian processes. *Journal of Computational Physics*, 348:683–693, 2017.

[36] Yifan Chen, Bamdad Hosseini, Houman Owhadi, and Andrew M Stuart. Solving and learning nonlinear PDEs with Gaussian processes. *Journal of Computational Physics*, 447:110668, 2021.

[37] Nathan Doumèche, Francis Bach, Gérard Biau, and Claire Boyer. Physics-informed machine learning as a kernel method. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1399–1450. PMLR, 2024.

[38] Joachim Bona-Pellissier, Giacomo Meanti, Matteo Santacesaria, and Lorenzo Rosasco. Piks: Universal physics-informed kernel methods. *arXiv preprint arXiv:2607.27062*, 2026.

[39] Laura Azzimonti, Laura M Sangalli, Piercesare Secchi, Maurizio Domanin, and Fabio Nobile. Blood flow velocity field estimation via spatial regression with PDE penalization. *Journal of the American Statistical Association*, 110(511):1057–1071, 2015.

[40] Carsten Franke and Robert Schaback. Solving partial differential equations by collocation using radial basis functions. *Applied Mathematics and computation*, 93(1):73–82, 1998.

[41] Robert Schaback and Holger Wendland. Kernel techniques: from machine learning to meshless methods. *Acta numerica*, 15, 2006.

[42] Pau Batlle, Yifan Chen, Bamdad Hosseini, Houman Owhadi, and Andrew M Stuart. Error analysis of kernel/GP methods for nonlinear and parametric PDEs. *Journal of Computational Physics*, 520, 2025.

[43] Yeonjong Shin, Jerome Darbon, and George Em Karniadakis. On the convergence of physics informed neural networks for linear second-order elliptic and parabolic type PDEs. *Communications in Computational Physics*, 28(5), 2020.

[44] Siddhartha Mishra and Roberto Molinaro. Estimates on the generalization error of physics-informed neural networks for approximating PDEs. *IMA Journal of Numerical Analysis*, 43(1):1–43, 2023.

[45] Yeonjong Shin, Zhongqiang Zhang, and George Em Karniadakis. Error estimates of residual minimization using neural networks for linear PDEs. *Journal of Machine Learning for Modeling and Computing*, 4(4), 2023.

[46] Tim De Ryck, Ameya D Jagtap, and Siddhartha Mishra. Error estimates for physics-informed neural networks approximating the Navier–Stokes equations. *IMA Journal of Numerical Analysis*, 44(1):83–119, 2024.

[47] Nathan Doumèche, Gérard Biau, and Claire Boyer. On the convergence of PINNs. *Bernoulli*, 31(3):2127 – 2151, 2025. doi: 10.3150/24-BEJ1799.

[48] Lei Shi, Xin Guo, and Ding-Xuan Zhou. Hermite learning with gradient data. *Journal of computational and applied mathematics*, 233(11):3046–3059, 2010.

[49] Zain ul Abdeen, Ruoxi Jia, Vassilis Kekatos, and Ming Jin. A theoretical analysis of using gradient data for Sobolev training in RKHS. *IFAC-PapersOnLine*, 56(2), 2023. doi: https://doi.org/10.1016/j.ifacol.2023.10.1491. 22nd IFAC World Congress.

[50] Katharine E Fisher, Matthew TC Li, Youssef Marzouk, and Timo Schorlepp. Precise asymptotic analysis of Sobolev training for random feature models. *arXiv preprint arXiv:2511.03050*, 2025.

[51] Eleonora Arnone, Alois Kneip, Fabio Nobile, and Laura M Sangalli. Some first results on the consistency of spatial regression with partial differential equation regularization. *Statistica Sinica*, 32(1):209–238, 2022.

[52] Nathan Doumèche, Francis Bach, Gérard Biau, and Claire Boyer. Physics-informed kernel learning. *Journal of Machine Learning Research*, 26(124):1–39, 2025.

[53] Nathan Doumèche, Francis Bach, Gérard Biau, and Claire Boyer. Fast kernel methods: Sobolev, physics-informed, and additive models. *arXiv preprint arXiv:2509.02649*, 2025.

[54] Ingo Steinwart, Don R Hush, Clint Scovel, et al. Optimal rates for regularized least squares regression. In *COLT*, pages 79–93, 2009.

[55] Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(205):1–38, 2020.

[56] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[57] Frank Noé, Alexandre Tkatchenko, Klaus-Robert Müller, and Cecilia Clementi. Machine learning for molecular simulation. *Annual review of physical chemistry*, 71(1), 2020.

[58] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[59] Alessandro Rudi, Guillermo D Canas, and Lorenzo Rosasco. On the sample complexity of subspace learning. *Advances in Neural Information Processing Systems*, 26, 2013.

[60] Luc Brogat-Motte, Riccardo Bonalli, and Alessandro Rudi. Learning controlled stochastic differential equations. *arXiv preprint arXiv:2411.01982*, 2024.

A Proofs

This section provides the proofs of the main results. We first derive the closed-form expressions of the empirical and population minimizers in Section A.1, then establish an error decomposition in Section A.2, derive high-probability bounds for its constituent terms in Section A.3, and finally combine these results to obtain finite-sample guarantees and optimized learning rates in Sections A.4 and A.5.

A.1 Closed forms

We derive here the closed-form expressions of the empirical minimizer of the regularized empirical risk $\widehat{R}$ introduced in Section 2,

$$\widehat{R}(u) = \frac{1}{n}\sum_{i=1}^{n}\big(u(x_i) - y_i\big)^2 + \gamma\,\frac{1}{m}\sum_{j=1}^{m}\big(Du(z_j) - d_j\big)^2 + \lambda\|u\|_{\mathcal{H}}^2,$$

and of the minimizer of its population counterpart

$$R_{\lambda,\gamma}(u) = \mathbb{E}\big[(u(x) - y)^2\big] + \gamma\,\mathbb{E}\big[(Du(z) - d)^2\big] + \lambda\|u\|_{\mathcal{H}}^2,$$

where the expectations are taken over $x \sim \rho$, $z \sim \rho_D$, and the corresponding observation noises. The empirical minimizer is given in operator form in Lemma A.1 and in finite-dimensional form in Lemma A.2. The population minimizer is given in Lemma A.3.

Lemma A.1 (Empirical solution)**.** *Let $\lambda > 0$ and $\gamma > 0$. Define*

$$\widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\otimes\phi(x_i), \qquad \widehat{\Sigma}_D := \frac{1}{m}\sum_{j=1}^{m}D\phi(z_j)\otimes D\phi(z_j),$$

$$\widehat{V} := \frac{1}{n}\sum_{i=1}^{n}y_i\,\phi(x_i), \qquad \widehat{V}_D := \frac{1}{m}\sum_{j=1}^{m}d_j\,D\phi(z_j).$$

Then $\widehat{R}$ has a unique minimizer $\hat{u} \in \mathcal{H}$ of the form

$$\hat{u}(x) = \langle \widehat{w}, \phi(x)\rangle_{\mathcal{H}}, \qquad \widehat{w} = \big(\widehat{\Sigma} + \gamma\widehat{\Sigma}_D + \lambda I\big)^{-1}\big(\widehat{V} + \gamma\widehat{V}_D\big).$$

Proof. Using the representation $u(x) = \langle w, \phi(x)\rangle_{\mathcal{H}}$, $Du(z) = \langle w, D\phi(z)\rangle_{\mathcal{H}}$, we write $\widehat{R}(u) = \widehat{R}(w)$ as

$$\widehat{R}(w) = \frac{1}{n}\sum_{i=1}^{n}\big(\langle w, \phi(x_i)\rangle_{\mathcal{H}} - y_i\big)^2 + \gamma\,\frac{1}{m}\sum_{j=1}^{m}\big(\langle w, D\phi(z_j)\rangle_{\mathcal{H}} - d_j\big)^2 + \lambda\langle w, w\rangle_{\mathcal{H}}.$$

Expanding the least-squares terms in $\mathcal{H}$, we obtain for the value part

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\big(\langle w, \phi(x_i)\rangle - y_i\big)^2 &= \frac{1}{n}\sum_{i=1}^{n}\Big(\langle w, \phi(x_i)\rangle^2 - 2y_i\langle w, \phi(x_i)\rangle + y_i^2\Big)\\ &= \left\langle w, \Big(\tfrac{1}{n}\sum_{i=1}^{n}\phi(x_i)\otimes\phi(x_i)\Big)w\right\rangle_{\mathcal{H}} - 2\left\langle \tfrac{1}{n}\sum_{i=1}^{n}y_i\phi(x_i),\, w\right\rangle_{\mathcal{H}} + \text{const}\\ &= \langle w, \widehat{\Sigma}w\rangle_{\mathcal{H}} - 2\langle \widehat{V}, w\rangle_{\mathcal{H}} + \text{const},\end{aligned}$$

and for the derivative part

$$\gamma\,\frac{1}{m}\sum_{j=1}^{m}\big(\langle w, D\phi(z_j)\rangle - d_j\big)^2 = \gamma\Big(\langle w, \widehat{\Sigma}_D w\rangle_{\mathcal{H}} - 2\langle \widehat{V}_D, w\rangle_{\mathcal{H}} + \text{const}\Big).$$

Adding the regularization term $\lambda\langle w, w\rangle_{\mathcal{H}}$, we get

$$\widehat{R}(w) = \big\langle w, (\widehat{\Sigma} + \gamma\widehat{\Sigma}_D + \lambda I)w\big\rangle_{\mathcal{H}} - 2\big\langle \widehat{V} + \gamma\widehat{V}_D, w\big\rangle_{\mathcal{H}} + \text{const}.$$

Since $\widehat{\Sigma}$ and $\widehat{\Sigma}_D$ are positive self-adjoint and $\lambda > 0$, the operator $\widehat{\Sigma} + \gamma\widehat{\Sigma}_D + \lambda I$ is strictly positive. Thus $\widehat{R}(w)$ is a strictly convex quadratic functional of w, whose unique minimizer $\widehat{w}$ is characterized by the normal equation

$$(\widehat{\Sigma} + \gamma\widehat{\Sigma}_D + \lambda I)\,\widehat{w} = \widehat{V} + \gamma\widehat{V}_D,$$

which yields

$$\widehat{w} = \big(\widehat{\Sigma} + \gamma\widehat{\Sigma}_D + \lambda I\big)^{-1}\big(\widehat{V} + \gamma\widehat{V}_D\big).$$

The corresponding minimizer in function space is $\hat{u}(x) = \langle \widehat{w}, \phi(x)\rangle_{\mathcal{H}}$. □

Lemma A.2 (Finite-dimensional form of the PIKS estimator)**.** *Let $\lambda > 0$ and $\gamma > 0$. Define the Gram matrices*

$$(K_{XX})_{ii'} = k(x_i, x_{i'}), \qquad (K_{XZ})_{ij} = D_1 k(z_j, x_i),$$
$$(K_{ZX})_{ji} = D_2 k(x_i, z_j), \qquad (K^D_{ZZ})_{jj'} = D_1 D_2 k(z_j, z_{j'}).$$

Then the PIKS estimator can be computed as

$$\hat{u}(\cdot) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\alpha_i k(x_i,\cdot) + \sqrt{\frac{\gamma}{m}}\sum_{j=1}^{m}\beta_j D_1 k(z_j,\cdot), \qquad \theta = \begin{bmatrix}\alpha\\ \beta\end{bmatrix},$$

where $\theta \in \mathbb{R}^{n+m}$ solves

$$\left(\begin{bmatrix}\frac{1}{n}K_{XX} & \sqrt{\frac{\gamma}{nm}}\,K_{XZ}\\ \sqrt{\frac{\gamma}{nm}}\,K_{ZX} & \frac{\gamma}{m}K^D_{ZZ}\end{bmatrix} + \lambda I\right)\theta = \begin{bmatrix} y/\sqrt{n}\\ \sqrt{\gamma}\,d/\sqrt{m}\end{bmatrix}.$$

Proof. Let $(e_i)_{i=1}^n$ and $(f_j)_{j=1}^m$ denote the canonical bases of $\mathbb{R}^n$ and $\mathbb{R}^m$, and define

$$\Phi e_i = k(x_i,\cdot), \qquad D\Phi f_j = D_1 k(z_j,\cdot).$$

Set

$$\Gamma_\gamma = \begin{bmatrix}\frac{1}{\sqrt{n}}\Phi & \sqrt{\frac{\gamma}{m}}D\Phi\end{bmatrix}, \qquad b_\gamma = \begin{bmatrix} y/\sqrt{n}\\ \sqrt{\gamma}\,d/\sqrt{m}\end{bmatrix}.$$

Then

$$\widehat{R}(u) = \|\Gamma_\gamma^* u - b_\gamma\|_2^2 + \lambda\|u\|_{\mathcal{H}}^2.$$

The optimality condition gives

$$(\Gamma_\gamma\Gamma_\gamma^* + \lambda I)\hat{u} = \Gamma_\gamma b_\gamma.$$

Using

$$(\Gamma_\gamma\Gamma_\gamma^* + \lambda I)^{-1}\Gamma_\gamma = \Gamma_\gamma(\Gamma_\gamma^*\Gamma_\gamma + \lambda I)^{-1},$$

we get

$$\hat{u} = \Gamma_\gamma\theta, \qquad \theta = (\Gamma_\gamma^*\Gamma_\gamma + \lambda I)^{-1}b_\gamma.$$

Finally,

$$\Gamma_\gamma^*\Gamma_\gamma = \begin{bmatrix}\frac{1}{n}K_{XX} & \sqrt{\frac{\gamma}{nm}}K_{XZ}\\ \sqrt{\frac{\gamma}{nm}}K_{ZX} & \frac{\gamma}{m}K^D_{ZZ}\end{bmatrix},$$

which gives the claimed system. Moreover, since

$$\hat{u} = \Gamma_\gamma\theta = \frac{1}{\sqrt{n}}\Phi\alpha + \sqrt{\frac{\gamma}{m}}D\Phi\beta,$$

we obtain

$$\hat{u}(\cdot) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\alpha_i k(x_i,\cdot) + \sqrt{\frac{\gamma}{m}}\sum_{j=1}^{m}\beta_j D_1 k(z_j,\cdot).$$

□

Lemma A.3 (Population solution)**.** *Let $x \sim \rho$ and $z \sim \rho_D$, and let the expectations involving y and d be taken with respect to the joint laws induced by the observation models. Define*

$$\Sigma := \mathbb{E}_{x\sim\rho}\big[\phi(x) \otimes \phi(x)\big], \qquad \Sigma_D := \mathbb{E}_{z\sim\rho_D}\big[D\phi(z) \otimes D\phi(z)\big],$$
$$V := \mathbb{E}\big[y\,\phi(x)\big], \qquad V_D := \mathbb{E}\big[d\,D\phi(z)\big].$$

Then $R_{\lambda,\gamma}$ has a unique minimizer $u_{\lambda,\gamma} \in \mathcal{H}$ of the form

$$u_{\lambda,\gamma}(x) = \langle w_{\lambda,\gamma}, \phi(x)\rangle_{\mathcal{H}}, \qquad w_{\lambda,\gamma} = \big(\Sigma + \gamma\Sigma_D + \lambda I\big)^{-1}(V + \gamma V_D).$$

Proof. The derivations are the same as in the empirical case, (Lemma A.1) with empirical averages replaced by expectations and $\widehat{\Sigma}, \widehat{\Sigma}_D, \widehat{V}, \widehat{V}_D$ replaced by Σ, Σ_D, V, V_D, and with the same quadratic expansion of $R_{\lambda,\gamma}(w)$.

□

A.2 Error decomposition

We next establish a standard error decomposition adapted to our setting. The idea is to separate the error into a bias term and an estimation term, controlled by the deviations of the empirical covariance operators from their population counterparts.

Lemma A.4 (Error decomposition)**.** *Let*

$$\Sigma_\gamma := \Sigma + \gamma\Sigma_D, \qquad \widehat{\Sigma}_\gamma := \widehat{\Sigma} + \gamma\widehat{\Sigma}_D, \qquad V_\gamma := V + \gamma V_D, \qquad \widehat{V}_\gamma := \widehat{V} + \gamma\widehat{V}_D,$$

and define

$$w_{\lambda,\gamma} := (\Sigma_\gamma + \lambda I)^{-1}V_\gamma, \qquad \widehat{w} := (\widehat{\Sigma}_\gamma + \lambda I)^{-1}\widehat{V}_\gamma, \qquad u^*(x) = \langle w^*, \phi(x)\rangle_{\mathcal{H}}.$$

Assume Assumption (A4) holds. Define the quantities

$$\Delta_{\Sigma,1} := \|(\Sigma_\gamma + \lambda I)^{1/2}(\widehat{\Sigma}_\gamma + \lambda I)^{-1/2}\|$$
$$\Delta_{\Sigma,2} := \big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_\gamma - \widehat{\Sigma}_\gamma)\big\|$$
$$\Delta_V := \big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\widehat{V}_\gamma - V_\gamma)\big\|$$
$$\mathcal{B}(\lambda,\gamma) := \big\|(\Sigma_\gamma + \lambda I)^{-1/2}w^*\big\|.$$

Then

$$\|\Sigma^{1/2}(\widehat{w} - w^*)\| \le \Delta_{\Sigma,1}^2(\Delta_V + \big\|w^*\big\|\Delta_{\Sigma,2}) + \lambda\mathcal{B}(\lambda,\gamma).$$

Proof. We first split the error into an estimation and a bias term:

$$\|\Sigma^{1/2}(\widehat{w} - w^*)\| \le \|\Sigma^{1/2}(\widehat{w} - w_{\lambda,\gamma})\| + \|\Sigma^{1/2}(w_{\lambda,\gamma} - w^*)\|.$$

We now bound each term.

Bias term. Under the Assumption (A4) we have

$$V = \Sigma w^*, \qquad V_D = \Sigma_D w^*,$$

hence

$$V_\gamma = V + \gamma V_D = (\Sigma + \gamma\Sigma_D)w^* = \Sigma_\gamma w^*.$$

Therefore

$$\begin{aligned} w_{\lambda,\gamma} - w^* &= (\Sigma_\gamma + \lambda I)^{-1}V_\gamma - w^* \\ &= (\Sigma_\gamma + \lambda I)^{-1}\Sigma_\gamma w^* - w^* \\ &= \big[(\Sigma_\gamma + \lambda I)^{-1}\Sigma_\gamma - I\big]w^* \\ &= -\lambda(\Sigma_\gamma + \lambda I)^{-1}w^*. \end{aligned}$$

As a consequence,

$$\|\Sigma^{1/2}(w_{\lambda,\gamma} - w^*)\| = \lambda\big\|\Sigma^{1/2}(\Sigma_\gamma + \lambda I)^{-1}w^*\big\| \le \lambda\big\|(\Sigma_\gamma + \lambda I)^{-1/2}w^*\big\|,$$

where we used $\|\Sigma^{1/2}(\Sigma_\gamma + \lambda I)^{-1/2}\| \leq 1$ since $\Sigma \preccurlyeq \Sigma_\gamma + \lambda I$.

Estimation term. We have

$$\widehat{w} - w_{\lambda,\gamma} = (\widehat{\Sigma}_\gamma + \lambda I)^{-1}(\widehat{V}_\gamma - V_\gamma) + \Big[(\widehat{\Sigma}_\gamma + \lambda I)^{-1} - (\Sigma_\gamma + \lambda I)^{-1}\Big] V_\gamma,$$

and, using $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ with $A = \widehat{\Sigma}_\gamma + \lambda I$, $B = \Sigma_\gamma + \lambda I$,

$$(\widehat{\Sigma}_\gamma + \lambda I)^{-1} - (\Sigma_\gamma + \lambda I)^{-1} = (\widehat{\Sigma}_\gamma + \lambda I)^{-1}(\Sigma_\gamma - \widehat{\Sigma}_\gamma)(\Sigma_\gamma + \lambda I)^{-1}.$$

Multiplying by $\Sigma^{1/2}$ and inserting $(\Sigma_\gamma + \lambda I)^{\pm 1/2}$, we obtain

$$\begin{aligned}\|\Sigma^{1/2}(\widehat{w} - w_{\lambda,\gamma})\| \leq \big\|\Sigma^{1/2}(\widehat{\Sigma}_\gamma + \lambda I)^{-1/2}\big\| \, \big\|(\widehat{\Sigma}_\gamma + \lambda I)^{-1/2}(\Sigma_\gamma + \lambda I)^{1/2}\big\| \\ \times \Big[\big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\widehat{V}_\gamma - V_\gamma)\big\| \\ + \big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_\gamma - \widehat{\Sigma}_\gamma)\big\| \, \big\|(\Sigma_\gamma + \lambda I)^{-1} V_\gamma\big\|\Big].\end{aligned}$$

Then,

$$\begin{aligned}\big\|\Sigma^{1/2}(\widehat{\Sigma}_\gamma + \lambda I)^{-1/2}\big\| &= \big\|\Sigma^{1/2}(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_\gamma + \lambda I)^{1/2}(\widehat{\Sigma}_\gamma + \lambda I)^{-1/2}\big\| \\ &\leq \big\|\Sigma^{1/2}(\Sigma_\gamma + \lambda I)^{-1/2}\big\| \, \big\|(\Sigma_\gamma + \lambda I)^{1/2}(\widehat{\Sigma}_\gamma + \lambda I)^{-1/2}\big\| \\ &\leq \big\|(\Sigma_\gamma + \lambda I)^{1/2}(\widehat{\Sigma}_\gamma + \lambda I)^{-1/2}\big\|,\end{aligned}$$

where we used again that $\|\Sigma^{1/2}(\Sigma_\gamma + \lambda I)^{-1/2}\| \leq 1$.

Moreover, using $V_\gamma = \Sigma_\gamma w^*$, we have

$$\begin{aligned}\big\|(\Sigma_\gamma + \lambda I)^{-1} V_\gamma\big\| &= \big\|(\Sigma_\gamma + \lambda I)^{-1}\Sigma_\gamma w^*\big\| \\ &\leq \big\|w^*\big\|.\end{aligned}$$

Conclusion. We define

$$\begin{aligned}\Delta_{\Sigma,1} &:= \big\|(\Sigma_\gamma + \lambda I)^{1/2}(\widehat{\Sigma}_\gamma + \lambda I)^{-1/2}\big\| \\ \Delta_{\Sigma,2} &:= \big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_\gamma - \widehat{\Sigma}_\gamma)\big\| \\ \Delta_V &:= \big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\widehat{V}_\gamma - V_\gamma)\big\| \\ \mathcal{B}(\lambda,\gamma) &:= \big\|(\Sigma_\gamma + \lambda I)^{-1/2} w^*\big\|.\end{aligned}$$

Collecting the terms, we obtain

$$\|\Sigma^{1/2}(\widehat{w} - w_{\lambda,\gamma})\| \;\leq\; \Delta_{\Sigma,1}^2(\Delta_V + \big\|w^*\big\|\Delta_{\Sigma,2}),$$

and

$$\|\Sigma^{1/2}(w_{\lambda,\gamma} - w^*)\| \;\leq\; \lambda\mathcal{B}(\lambda,\gamma),$$

which together yields

$$\|\Sigma^{1/2}(\widehat{w} - w^*)\| \leq \Delta_{\Sigma,1}^2(\Delta_V + \big\|w^*\big\|\Delta_{\Sigma,2}) + \lambda\mathcal{B}(\lambda,\gamma).$$

□

A.3 High-probability bounds for the error decomposition terms $\Delta_{\Sigma,1}$, $\Delta_{\Sigma,2}$, and Δ_V

We next control the random quantities appearing in the error decomposition of Lemma A.4. Compared with the standard proof for kernel ridge regression, the only additional difficulty is the presence of the derivative term and the need to keep track of its dependence on γ. This leads us to decompose both covariance and moment deviations into value and derivative contributions. The remainder of the argument then follows the standard kernel ridge regression proof strategy: we derive high-probability bounds for Δ_V and $\Delta_{\Sigma,2}$, and control the multiplicative factor $\Delta_{\Sigma,1}$ through the auxiliary normalized covariance deviation

$$\Delta_{\Sigma,3} := \big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_\gamma - \widehat{\Sigma}_\gamma)(\Sigma_\gamma + \lambda I)^{-1/2}\big\|.$$

The key point is that when $\Delta_{\Sigma,3} < 1$, the empirical covariance operator remains well conditioned relative to $\Sigma_\gamma + \lambda I$, which yields a bound on $\Delta_{\Sigma,1}$.

Lemma A.5 (High-probability bound on $\Delta_{\Sigma,3}$)**.** *We define*

$$\Delta_{\Sigma,3} := \big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_\gamma - \widehat{\Sigma}_\gamma)(\Sigma_\gamma + \lambda I)^{-1/2}\big\|.$$

Let $\delta \in (0,1)$ *and* $\lambda > 0$ *satisfy*

$$\frac{36}{n}\log\frac{n}{\delta} \;\le\; \lambda \;\le\; \|\Sigma\|_\infty, \qquad \frac{36}{m}\log\frac{m}{\delta} \;\le\; \lambda\gamma^{-1} \;\le\; \|\Sigma_D\|_\infty.$$

Then, with probability at least $1 - 2\delta$*,*

$$\Delta_{\Sigma,3} \;\le\; \tfrac{1}{2}.$$

Proof. We first relate $\Delta_{\Sigma,3}$ to the value and derivative parts separately. Using the decomposition

$$\Sigma_\gamma - \widehat{\Sigma}_\gamma = (\Sigma - \widehat{\Sigma}) + \gamma(\Sigma_D - \widehat{\Sigma}_D),$$

and the triangle inequality, we obtain

$$\begin{aligned}\Delta_{\Sigma,3} &= \Big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_\gamma - \widehat{\Sigma}_\gamma)(\Sigma_\gamma + \lambda I)^{-1/2}\Big\| \\ &\le \Big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma - \widehat{\Sigma})(\Sigma_\gamma + \lambda I)^{-1/2}\Big\| + \gamma\Big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_D - \widehat{\Sigma}_D)(\Sigma_\gamma + \lambda I)^{-1/2}\Big\|.\end{aligned}$$

For the first term, write

$$(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma - \widehat{\Sigma})(\Sigma_\gamma + \lambda I)^{-1/2} = A\,(\Sigma + \lambda I)^{-1/2}(\Sigma - \widehat{\Sigma})(\Sigma + \lambda I)^{-1/2}\,A^*,$$

where

$$A := (\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma + \lambda I)^{1/2}.$$

From $\Sigma \preccurlyeq \Sigma_\gamma$ we have $(\Sigma_\gamma + \lambda I)^{-1} \preccurlyeq (\Sigma + \lambda I)^{-1}$, hence $\|A\| \le 1$. Therefore

$$\Big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma - \widehat{\Sigma})(\Sigma_\gamma + \lambda I)^{-1/2}\Big\| \le \big\|(\Sigma + \lambda I)^{-1/2}(\Sigma - \widehat{\Sigma})(\Sigma + \lambda I)^{-1/2}\big\|.$$

For the second term we argue analogously, but comparing with $\gamma\Sigma_D$. Using $\gamma\Sigma_D \preccurlyeq \Sigma_\gamma$ we obtain

$$\Big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_D - \widehat{\Sigma}_D)(\Sigma_\gamma + \lambda I)^{-1/2}\Big\| \le \big\|(\gamma\Sigma_D + \lambda I)^{-1/2}(\Sigma_D - \widehat{\Sigma}_D)(\gamma\Sigma_D + \lambda I)^{-1/2}\big\|,$$

and hence

$$\Delta_{\Sigma,3} \;\le\; \big\|(\Sigma+\lambda I)^{-1/2}(\Sigma-\widehat{\Sigma})(\Sigma+\lambda I)^{-1/2}\big\| + \gamma\big\|(\gamma\Sigma_D+\lambda I)^{-1/2}(\Sigma_D-\widehat{\Sigma}_D)(\gamma\Sigma_D+\lambda I)^{-1/2}\big\|.$$

We now control each term probabilistically using Lemma 3.6 in [59]. Although the lemma yields a bound of $1/2$, this constant can be reduced by rescaling the regularization parameter and applying the same proof. In particular, applying the proof of Lemma 3.6 with target bound $1/4$, and using the admissibility conditions

$$\frac{36}{n}\log\frac{n}{\delta} \;\le\; \lambda \;\le\; \|\Sigma\|_\infty, \qquad \gamma\frac{36}{m}\log\frac{m}{\delta} \;\le\; \lambda \;\le\; \gamma\|\Sigma_D\|_\infty,$$

we obtain that, with probability at least $1 - \delta$,

$$\big\|(\Sigma + \lambda I)^{-1/2}(\Sigma - \widehat{\Sigma})(\Sigma + \lambda I)^{-1/2}\big\| \;<\; \tfrac{1}{4},$$

and, again with probability at least $1 - \delta$,

$$\gamma\big\|(\gamma\Sigma_D + \lambda I)^{-1/2}(\Sigma_D - \widehat{\Sigma}_D)(\gamma\Sigma_D + \lambda I)^{-1/2}\big\| \;<\; \tfrac{1}{4}.$$

Here the improvement from $\frac{1}{2}$ to $\frac{1}{4}$ comes from running the same concentration proof with the smaller target threshold $\frac{1}{4}$, at the cost of replacing the constant 9 in the lower admissibility condition by 36.

By a union bound, both events hold simultaneously with probability at least $1 - 2\delta$, and on this event we have

$$\Delta_{\Sigma,3} \;\le\; \tfrac{1}{4} + \tfrac{1}{4} \;=\; \tfrac{1}{2}.$$

This concludes the proof. □

Lemma A.6 (Bound on $\Delta_{\Sigma,1}$)**.** *Recall*

$$\Delta_{\Sigma,1} := \big\|(\Sigma_\gamma + \lambda I)^{1/2}(\widehat{\Sigma}_\gamma + \lambda I)^{-1/2}\big\|.$$

We have

$$\Delta_{\Sigma,1} \le (1 - \Delta_{\Sigma,3})^{-1/2}.$$

Proof. We have

$$\Delta_{\Sigma,1}^2 = \big\|(\Sigma_\gamma + \lambda I)^{1/2}(\widehat{\Sigma}_\gamma + \lambda I)^{-1}(\Sigma_\gamma + \lambda I)^{1/2}\big\|.$$

Define

$$\widehat{B} := (\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_\gamma - \widehat{\Sigma}_\gamma)(\Sigma_\gamma + \lambda I)^{-1/2}.$$

By definition of $\Delta_{\Sigma,3}$,

$$\|\widehat{B}\| = \Delta_{\Sigma,3}.$$

We can rewrite

$$\widehat{\Sigma}_\gamma + \lambda I = \Sigma_\gamma + \lambda I - (\Sigma_\gamma - \widehat{\Sigma}_\gamma) = (\Sigma_\gamma + \lambda I)^{1/2}(I - \widehat{B})(\Sigma_\gamma + \lambda I)^{1/2},$$

so

$$(\widehat{\Sigma}_\gamma + \lambda I)^{-1} = (\Sigma_\gamma + \lambda I)^{-1/2}(I - \widehat{B})^{-1}(\Sigma_\gamma + \lambda I)^{-1/2}.$$

Plugging this into the expression for $\Delta_{\Sigma,1}^2$ gives

$$\Delta_{\Sigma,1}^2 = \|(I - \widehat{B})^{-1}\|.$$

If $\|\widehat{B}\| = \Delta_{\Sigma,3} < 1$, then the spectrum of $\widehat{B}$ lies in $[-\Delta_{\Sigma,3}, \Delta_{\Sigma,3}]$, hence the spectrum of $(I - \widehat{B})^{-1}$ is contained in $\{(1-\mu)^{-1} : |\mu| \le \Delta_{\Sigma,3}\}$, and therefore

$$\Delta_{\Sigma,1}^2 = \|(I - \widehat{B})^{-1}\| \le (1 - \|\widehat{B}\|)^{-1} = (1 - \Delta_{\Sigma,3})^{-1}.$$

□

Lemma A.7 (High-probability bound on Δ_V)**.** *Recall*

$$\Delta_V := \big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\widehat{V}_\gamma - V_\gamma)\big\|.$$

Then

$$\Delta_V \;\le\; \big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\widehat{V} - V)\big\| \;+\; \gamma\,\big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\widehat{V}_D - V_D)\big\|.$$

For any $\delta \in (0,1)$, with probability at least $1 - 2\delta$,

$$\begin{aligned}\Delta_V \;&\le\; 2\,\beta_V\,\frac{\log(2/\delta)}{n} \;+\; \sqrt{2}\,\sigma_V\,\sqrt{\frac{\log(2/\delta)}{n}} \\ &\quad + \gamma\left(2\,\beta_D\,\frac{\log(2/\delta)}{m} + \sqrt{2}\,\sigma_D\,\sqrt{\frac{\log(2/\delta)}{m}}\right),\end{aligned}$$

where the explicit constants are

$$\begin{aligned}\beta_V &:= 2\,\lambda^{-1/2} L_V \kappa_V, \qquad & \sigma_V^2 &:= L_V^2 \kappa_V^2 \big\|(\Sigma_\gamma + \lambda I)^{-1/2}\Sigma^{1/2}\big\|_{\mathrm{HS}}^2, \\ \beta_D &:= 2\,\lambda^{-1/2} L_D \kappa_D, \qquad & \sigma_D^2 &:= L_D^2 \kappa_D^2 \big\|(\Sigma_\gamma + \lambda I)^{-1/2}\Sigma_D^{1/2}\big\|_{\mathrm{HS}}^2.\end{aligned}$$

Proof. The decomposition

$$\widehat{V}_\gamma - V_\gamma = (\widehat{V} - V) + \gamma(\widehat{V}_D - V_D)$$

immediately gives

$$\Delta_V \le \big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\widehat{V} - V)\big\| + \gamma\big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\widehat{V}_D - V_D)\big\|.$$

Value part. Write

$$(\Sigma_\gamma + \lambda I)^{-1/2}(\widehat{V} - V) = \frac{1}{n}\sum_{i=1}^{n} M(x_i, y_i), \qquad M(x,y) := (\Sigma_\gamma + \lambda I)^{-1/2}(y\phi(x) - V).$$

Since $|y| \le L_V$ and $\|\phi(x)\| \le \kappa_V$,

$$\|y\phi(x) - V\| \le L_V\kappa_V + \|V\| \le 2L_V\kappa_V,$$

hence

$$\|M(x,y)\| \le \|(\Sigma_\gamma + \lambda I)^{-1/2}\|\, 2L_V\kappa_V \le 2\lambda^{-1/2}L_V\kappa_V =: \beta_V.$$

The second moment satisfies

$$\mathbb{E}\|M(x,y)\|^2 \le L_V^2\kappa_V^2\|(\Sigma_\gamma + \lambda I)^{-1/2}\Sigma^{1/2}\|_{\mathrm{HS}}^2 =: \sigma_V^2.$$

Applying Hilbert-space Bernstein (Prop. 7.15 of [60]) gives, with probability at least $1-\delta$,

$$\big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\widehat{V} - V)\big\| \le 2\beta_V\,\frac{\log(2/\delta)}{n} + \sqrt{2}\,\sigma_V\sqrt{\frac{\log(2/\delta)}{n}}.$$

Derivative part. Exactly the same argument with

$$M_D(z,d) := (\Sigma_\gamma + \lambda I)^{-1/2}(d\,D\phi(z) - V_D)$$

yields

$$\|M_D(z,d)\| \le 2\lambda^{-1/2}L_D\kappa_D =: \beta_D,$$

and

$$\sigma_D^2 = L_D^2\kappa_D^2\|(\Sigma_\gamma + \lambda I)^{-1/2}\Sigma_D^{1/2}\|_{\mathrm{HS}}^2.$$

Bernstein again gives, with probability at least $1-\delta$,

$$\big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\widehat{V}_D - V_D)\big\| \le 2\,\beta_D\,\frac{\log(2/\delta)}{m} + \sqrt{2}\,\sigma_D\sqrt{\frac{\log(2/\delta)}{m}}.$$

Conclusion. A union bound yields the stated bound with probability at least $1-2\delta$. □

Lemma A.8 (High-probability bound on $\Delta_{\Sigma,2}$)**.** *Recall*

$$\Delta_{\Sigma,2} := \big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_\gamma - \widehat{\Sigma}_\gamma)\big\|, \qquad \Sigma_\gamma = \Sigma + \gamma\Sigma_D, \quad \widehat{\Sigma}_\gamma = \widehat{\Sigma} + \gamma\widehat{\Sigma}_D.$$

Then

$$\Delta_{\Sigma,2} \;\le\; \big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma - \widehat{\Sigma})\big\| \;+\; \gamma\,\big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_D - \widehat{\Sigma}_D)\big\|.$$

Moreover, for any $\delta \in (0,1)$, with probability at least $1-2\delta$,

$$\Delta_{\Sigma,2} \;\le\; 2\,\beta_\Sigma\,\frac{\log(2/\delta)}{n} \;+\; \sqrt{2}\,\sigma_\Sigma\sqrt{\frac{\log(2/\delta)}{n}} \;+\; \gamma\left(2\,\beta_{\Sigma_D}\,\frac{\log(2/\delta)}{m} \;+\; \sqrt{2}\,\sigma_{\Sigma_D}\sqrt{\frac{\log(2/\delta)}{m}}\right),$$

where

$$\beta_\Sigma := \lambda^{-1/2}\,2\kappa_V^2, \qquad \sigma_\Sigma^2 := 4\kappa_V^2\,\big\|(\Sigma_\gamma + \lambda I)^{-1/2}\Sigma^{1/2}\big\|_{\mathrm{HS}}^2,$$
$$\beta_{\Sigma_D} := \lambda^{-1/2}\,2\kappa_D^2, \qquad \sigma_{\Sigma_D}^2 := 4\kappa_D^2\,\big\|(\Sigma_\gamma + \lambda I)^{-1/2}\Sigma_D^{1/2}\big\|_{\mathrm{HS}}^2.$$

Proof. The decomposition

$$\Sigma_\gamma - \widehat{\Sigma}_\gamma = (\Sigma - \widehat{\Sigma}) + \gamma(\Sigma_D - \widehat{\Sigma}_D)$$

and the triangle inequality give

$$\begin{aligned}\Delta_{\Sigma,2} &= \big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_\gamma - \widehat{\Sigma}_\gamma)\big\| \\ &\le \big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma - \widehat{\Sigma})\big\| + \gamma\,\big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_D - \widehat{\Sigma}_D)\big\|.\end{aligned}$$

We now bound these two terms separately, with the same Hilbert-space Bernstein argument as in Lemma A.7, but now in the Hilbert space of Hilbert-Schmidt operators and with $y\phi(x)$ replaced by $\phi(x) \otimes \phi(x)$.

Value part. Recall

$$\Sigma = \mathbb{E}[\phi(x) \otimes \phi(x)], \qquad \widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) \otimes \phi(x_i).$$

Define

$$U(x) := \Sigma - \phi(x) \otimes \phi(x), \qquad M_\Sigma(x) := (\Sigma_\gamma + \lambda I)^{-1/2} U(x).$$

Then $\mathbb{E}[U(x)] = 0$, and

$$(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma - \widehat{\Sigma}) = \frac{1}{n}\sum_{i=1}^{n} M_\Sigma(x_i).$$

We view $M_\Sigma(x)$ as a random element of the Hilbert space of Hilbert-Schmidt operators. We first bound its norm. Using the boundedness $\|\phi(x)\| \le \kappa_V$ we have

$$\|\phi(x) \otimes \phi(x)\|_{\mathrm{HS}} = \|\phi(x)\|^2 \le \kappa_V^2, \qquad \|\Sigma\|_{\mathrm{HS}} \le \mathbb{E}\|\phi(x) \otimes \phi(x)\|_{\mathrm{HS}} \le \kappa_V^2.$$

Thus

$$\|U(x)\|_{\mathrm{HS}} \le \|\phi(x) \otimes \phi(x)\|_{\mathrm{HS}} + \|\Sigma\|_{\mathrm{HS}} \le 2\kappa_V^2,$$

and since $\|(\Sigma_\gamma + \lambda I)^{-1/2}\| \le \lambda^{-1/2}$,

$$\|M_\Sigma(x)\|_{\mathrm{HS}} \le \lambda^{-1/2}\, 2\kappa_V^2 = \beta_\Sigma.$$

Next, we bound the second moment. Using that $\Sigma_\gamma + \lambda I$ is self-adjoint and positive, we have

$$\|M_\Sigma(x)\|_{\mathrm{HS}}^2 = \big\|(\Sigma_\gamma + \lambda I)^{-1/2} U(x)\big\|_{\mathrm{HS}}^2 = \mathrm{Tr}\Big(U(x)(\Sigma_\gamma + \lambda I)^{-1} U(x)\Big).$$

Take expectation and use linearity of the trace:

$$\mathbb{E}\|M_\Sigma(x)\|_{\mathrm{HS}}^2 = \mathrm{Tr}\Big((\Sigma_\gamma + \lambda I)^{-1}\, \mathbb{E}\big[U(x)^2\big]\Big).$$

We now bound $\mathbb{E}[U(x)^2]$ in terms of Σ. First note that

$$U(x)^2 = \big(\Sigma - \phi(x) \otimes \phi(x)\big)^2 \preccurlyeq 2\Sigma^2 + 2\big(\phi(x) \otimes \phi(x)\big)^2,$$

by the elementary operator inequality $(A - B)^2 = A^2 + B^2 - AB - BA \preccurlyeq 2A^2 + 2B^2$ for self-adjoint A, B.

Moreover,

$$\big(\phi(x) \otimes \phi(x)\big)^2 = \|\phi(x)\|^2\, \phi(x) \otimes \phi(x) \preccurlyeq \kappa_V^2\, \phi(x) \otimes \phi(x),$$

and $\Sigma^2 \preccurlyeq \|\Sigma\|\, \Sigma \preccurlyeq \kappa_V^2\, \Sigma$, since $\|\Sigma\| \le \kappa_V^2$ for a bounded feature map. Combining these,

$$U(x)^2 \preccurlyeq 2\kappa_V^2\, \phi(x) \otimes \phi(x) + 2\kappa_V^2\, \Sigma.$$

Taking expectations yields

$$\mathbb{E}[U(x)^2] \preccurlyeq 2\kappa_V^2\, \mathbb{E}[\phi(x) \otimes \phi(x)] + 2\kappa_V^2\, \Sigma = 4\kappa_V^2\, \Sigma.$$

Plugging this into the expression for the second moment, we get

$$\begin{aligned}
\mathbb{E}\|M_\Sigma(x)\|_{\mathrm{HS}}^2 &\le \mathrm{Tr}\Big((\Sigma_\gamma + \lambda I)^{-1}\, 4\kappa_V^2 \Sigma\Big) \\
&= 4\kappa_V^2\, \mathrm{Tr}\Big((\Sigma_\gamma + \lambda I)^{-1/2} \Sigma (\Sigma_\gamma + \lambda I)^{-1/2}\Big) \\
&= 4\kappa_V^2\, \big\|(\Sigma_\gamma + \lambda I)^{-1/2} \Sigma^{1/2}\big\|_{\mathrm{HS}}^2.
\end{aligned}$$

Thus we set

$$\sigma_\Sigma^2 := 4\kappa_V^2\, \big\|(\Sigma_\gamma + \lambda I)^{-1/2} \Sigma^{1/2}\big\|_{\mathrm{HS}}^2$$

and obtain

$$\mathbb{E}\|M_\Sigma(x)\|_{\mathrm{HS}}^2 \le \sigma_\Sigma^2.$$

We can now apply the Hilbert-space Bernstein inequality (Proposition 7.15 in [60]) in the Hilbert space of Hilbert-Schmidt operators to the empirical average $\frac{1}{n}\sum_{i=1}^{n} M_\Sigma(x_i)$. This gives: for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma - \widehat{\Sigma})\big\|_{\mathrm{HS}} \;\le\; 2\,\beta_\Sigma\, \frac{\log(2/\delta)}{n} + \sqrt{2}\,\sigma_\Sigma \sqrt{\frac{\log(2/\delta)}{n}}.$$

Since $\|T\| \le \|T\|_{\mathrm{HS}}$ for every Hilbert-Schmidt operator T, the same bound holds for

$$\big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma - \widehat{\Sigma})\big\|.$$

Derivative part. The argument is identical with

$$\Sigma_D = \mathbb{E}[D\phi(z) \otimes D\phi(z)], \qquad \widehat{\Sigma}_D = \frac{1}{m}\sum_{j=1}^{m} D\phi(z_j) \otimes D\phi(z_j).$$

Define

$$U_D(z) := \Sigma_D - D\phi(z) \otimes D\phi(z), \qquad M_{\Sigma_D}(z) := (\Sigma_\gamma + \lambda I)^{-1/2} U_D(z).$$

Then

$$(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_D - \widehat{\Sigma}_D) = \frac{1}{m}\sum_{j=1}^{m} M_{\Sigma_D}(z_j),$$

with $\mathbb{E}[M_{\Sigma_D}(z)] = 0$.

Using $\|D\phi(z)\| \le \kappa_D$ we obtain

$$\|U_D(z)\|_{\mathrm{HS}} \le 2\kappa_D^2, \qquad \|M_{\Sigma_D}(z)\|_{\mathrm{HS}} \le \lambda^{-1/2} 2\kappa_D^2 = \beta_{\Sigma_D},$$

and, by the same operator inequalities as above,

$$\mathbb{E}\|M_{\Sigma_D}(z)\|_{\mathrm{HS}}^2 \le 4\kappa_D^2 \left\|(\Sigma_\gamma + \lambda I)^{-1/2}\Sigma_D^{1/2}\right\|_{\mathrm{HS}}^2 =: \sigma_{\Sigma_D}^2.$$

Another application of the Hilbert-space Bernstein inequality in the Hilbert space of Hilbert-Schmidt operators yields, with probability at least $1 - \delta$,

$$\left\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_D - \widehat{\Sigma}_D)\right\|_{\mathrm{HS}} \;\le\; 2\,\beta_{\Sigma_D}\,\frac{\log(2/\delta)}{m} + \sqrt{2}\,\sigma_{\Sigma_D}\sqrt{\frac{\log(2/\delta)}{m}}.$$

Again the same bound holds for

$$\left\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_D - \widehat{\Sigma}_D)\right\|.$$

Conclusion. We have shown that, with probability at least $1 - \delta$,

$$\left\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma - \widehat{\Sigma})\right\| \le 2\,\beta_\Sigma\,\frac{\log(2/\delta)}{n} + \sqrt{2}\,\sigma_\Sigma\sqrt{\frac{\log(2/\delta)}{n}},$$

and, with probability at least $1 - \delta$,

$$\left\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_D - \widehat{\Sigma}_D)\right\| \le 2\,\beta_{\Sigma_D}\,\frac{\log(2/\delta)}{m} + \sqrt{2}\,\sigma_{\Sigma_D}\sqrt{\frac{\log(2/\delta)}{m}}.$$

By a union bound, both events hold simultaneously with probability at least $1 - 2\delta$, and combining these with the initial decomposition of $\Delta_{\Sigma,2}$ yields the claimed high-probability bound. □

A.4 Finite-sample bounds

Combining Lemma A.4 with the concentration results of the previous subsection yields the following finite-sample bound restated from Theorem 3.1.

Theorem A.9 (High-probability finite-sample bounds for PIKS)**.** *Under assumptions (A1) to (A5), let $\lambda > 0$, $\gamma > 0$ and let $\delta \in (0, 1/2)$. Define*

$$\Sigma_\gamma := \Sigma + \gamma\Sigma_D, \qquad \mathcal{B}(\lambda, \gamma) := \left\|(\Sigma_\gamma + \lambda I)^{-1/2} w^*\right\|_{\mathcal{H}},$$

and the (value and derivative) capacity terms

$$\mathcal{N}_V(\lambda, \gamma) := \left\|(\Sigma_\gamma + \lambda I)^{-1/2}\Sigma^{1/2}\right\|_{\mathrm{HS}}, \qquad \mathcal{N}_D(\lambda, \gamma) := \left\|(\Sigma_\gamma + \lambda I)^{-1/2}\Sigma_D^{1/2}\right\|_{\mathrm{HS}}.$$

If

$$\frac{36}{n}\log\frac{2n}{\delta} \;\le\; \lambda \;\le\; \|\Sigma\|_\infty \quad \textit{and} \quad \frac{36}{m}\log\frac{2m}{\delta} \;\le\; \lambda\gamma^{-1} \;\le\; \|\Sigma_D\|_\infty, \tag{5}$$

then there exists a constant $c > 0$, depending only on $\kappa_V, \kappa_D, L_V, L_D$ and $\|w^*\|$, but not on $\lambda, \gamma, n, m, \delta$, such that, with probability at least $1 - 2\delta$,

$$\mathbb{E}\left[\left(\hat{u}(x) - u^*(x)\right)^2\right]^{\frac{1}{2}} \le c\log\frac{4}{\delta}\left[\frac{1}{\sqrt{\lambda}}\left(\frac{1}{n} + \frac{\gamma}{m}\right) + \frac{\mathcal{N}_V(\lambda,\gamma)}{\sqrt{n}} + \gamma\frac{\mathcal{N}_D(\lambda,\gamma)}{\sqrt{m}} + \lambda\mathcal{B}(\lambda,\gamma)\right]. \tag{6}$$

Proof. From the error decomposition Lemma A.4, we have

$$\|\Sigma^{1/2}(\widehat{w} - w^*)\| \;\le\; \Delta_{\Sigma,1}^2\big(\Delta_V + \|w^*\|\Delta_{\Sigma,2}\big) \;+\; \lambda\,\mathcal{B}(\lambda,\gamma),$$

where

$$\Delta_{\Sigma,1} := \big\|(\Sigma_\gamma + \lambda I)^{1/2}(\widehat{\Sigma}_\gamma + \lambda I)^{-1/2}\big\|, \quad \Delta_{\Sigma,2} := \big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_\gamma - \widehat{\Sigma}_\gamma)\big\|,$$
$$\Delta_V := \big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\widehat{V}_\gamma - V_\gamma)\big\|.$$

Step 1: bound on $\Delta_{\Sigma,1}$. Recall

$$\Delta_{\Sigma,3} := \big\|(\Sigma_\gamma + \lambda I)^{-1/2}(\Sigma_\gamma - \widehat{\Sigma}_\gamma)(\Sigma_\gamma + \lambda I)^{-1/2}\big\|.$$

Applying Lemma A.5 with confidence parameter η, and the condition (3), we obtain an event of probability at least $1 - 2\eta$ on which $\Delta_{\Sigma,3} \le 1/2$. On this event, Lemma A.6 yields

$$\Delta_{\Sigma,1}^2 \le (1 - \Delta_{\Sigma,3})^{-1} \le 2.$$

Step 2: bounds on Δ_V *and* $\Delta_{\Sigma,2}$. We use Lemmas A.7 and A.8.

From Lemma A.7, for any $\eta \in (0,1)$, with probability at least $1 - 2\eta$,

$$\Delta_V \;\le\; \underbrace{\Big(2\beta_V\,\frac{\log(2/\eta)}{n} + \sqrt{2}\sigma_V\,\sqrt{\frac{\log(2/\eta)}{n}}\Big)}_{\text{value part}} + \gamma\,\underbrace{\Big(2\beta_D\,\frac{\log(2/\eta)}{m} + \sqrt{2}\sigma_D\,\sqrt{\frac{\log(2/\eta)}{m}}\Big)}_{\text{derivative part}},$$

with

$$\beta_V = \lambda^{-1/2} 2 L_V \kappa_V, \qquad \sigma_V^2 = L_V^2\,\kappa_V^2\big\|(\Sigma_\gamma + \lambda I)^{-1/2}\Sigma^{1/2}\big\|_{\text{HS}}^2,$$
$$\beta_D = \lambda^{-1/2} 2 L_D \kappa_D, \qquad \sigma_D^2 = L_D^2\,\kappa_D^2\big\|(\Sigma_\gamma + \lambda I)^{-1/2}\Sigma_D^{1/2}\big\|_{\text{HS}}^2.$$

Hence there exist constants $a_V, b_V, a_D, b_D > 0$ depending only on (κ_V, L_V) and (κ_D, L_D) such that

$$\Delta_V \;\le\; \lambda^{-1/2}\Big(a_V\,\frac{\log(2/\eta)}{n} + \gamma\,a_D\,\frac{\log(2/\eta)}{m}\Big) + \Big(b_V\,\mathcal{N}_V(\lambda,\gamma)\,\sqrt{\frac{\log(2/\eta)}{n}} + \gamma\,b_D\,\mathcal{N}_D(\lambda,\gamma)\,\sqrt{\frac{\log(2/\eta)}{m}}\Big).$$

Similarly, from Lemma A.8, for the same confidence parameter, with probability at least $1 - 2\eta$,

$$\Delta_{\Sigma,2} \;\le\; \lambda^{-1/2}\Big(\tilde{a}_V\,\frac{\log(2/\eta)}{n} + \gamma\,\tilde{a}_D\,\frac{\log(2/\eta)}{m}\Big) + \Big(\tilde{b}_V\,\mathcal{N}_V(\lambda,\gamma)\,\sqrt{\frac{\log(2/\eta)}{n}} + \gamma\,\tilde{b}_D\,\mathcal{N}_D(\lambda,\gamma)\,\sqrt{\frac{\log(2/\eta)}{m}}\Big),$$

for some constants $\tilde{a}_V, \tilde{a}_D, \tilde{b}_V, \tilde{b}_D > 0$ depending only on κ_V and κ_D.

Combining these two bounds, we obtain constants $A_1, A_2, B_1, B_2 > 0$, depending only on $(\kappa_V, \kappa_D, L_V, L_D)$, such that

$$\Delta_V \;\le\; \lambda^{-1/2}\Big(A_1\,\frac{\log(2/\eta)}{n} + A_2\,\gamma\,\frac{\log(2/\eta)}{m}\Big) + \Big(B_1\,\mathcal{N}_V(\lambda,\gamma)\,\sqrt{\frac{\log(2/\eta)}{n}} + B_2\,\gamma\,\mathcal{N}_D(\lambda,\gamma)\,\sqrt{\frac{\log(2/\eta)}{m}}\Big), \tag{7}$$

and

$$\Delta_{\Sigma,2} \;\le\; \lambda^{-1/2}\Big(A_1\,\frac{\log(2/\eta)}{n} + A_2\,\gamma\,\frac{\log(2/\eta)}{m}\Big) + \Big(B_1\,\mathcal{N}_V(\lambda,\gamma)\,\sqrt{\frac{\log(2/\eta)}{n}} + B_2\,\gamma\,\mathcal{N}_D(\lambda,\gamma)\,\sqrt{\frac{\log(2/\eta)}{m}}\Big), \tag{8}$$

where we have simply taken maxima of the corresponding constants from Δ_V and $\Delta_{\Sigma,2}$.

Step 3: combine all bounds. Set $\eta := \delta/3$ and apply Lemmas A.5, A.7, and A.8 with confidence parameter η. Each holds with probability at least $1 - 2\eta$, so by a union bound, with probability at least

$$1 - 3 \cdot 2\eta \;=\; 1 - 2\delta,$$

we simultaneously have

$$\Delta_{\Sigma,1}^2 \le 2, \quad \text{and the bounds (7)-(8) for } \Delta_V, \Delta_{\Sigma,2}.$$

On this event,

$$\begin{aligned}\|\Sigma^{1/2}(\widehat{w}-w^*)\| &\le \Delta_{\Sigma,1}^2\big(\Delta_V + \|w^*\|\Delta_{\Sigma,2}\big) + \lambda\,\mathcal{B}(\lambda,\gamma)\\ &\le 2\big(\Delta_V + \|w^*\|\Delta_{\Sigma,2}\big) + \lambda\,\mathcal{B}(\lambda,\gamma).\end{aligned}$$

Substituting the bounds (7)-(8), we obtain

$$\begin{aligned}\|\Sigma^{1/2}(\widehat{w}-w^*)\| \le 2(1+\|w^*\|)\Bigg[&\lambda^{-1/2}\Big(A_1\,\frac{\log(2/\eta)}{n} + A_2\,\gamma\,\frac{\log(2/\eta)}{m}\Big)\\ &+\Big(B_1\,\mathcal{N}_V(\lambda,\gamma)\sqrt{\frac{\log(2/\eta)}{n}} + B_2\,\gamma\,\mathcal{N}_D(\lambda,\gamma)\sqrt{\frac{\log(2/\eta)}{m}}\Big)\Bigg] + \lambda\,\mathcal{B}(\lambda,\gamma).\end{aligned}$$

Since $\eta=\delta/3$, we have $\log(2/\eta)=\log(6/\delta)\le c_1\log(4/\delta)$ for some constant $c_1>0$. Absorbing the constants $2(1+\|w^*\|)A_1, 2(1+\|w^*\|)A_2, 2(1+\|w^*\|)B_1, 2(1+\|w^*\|)B_2$ and $\sqrt{c_1}$ into a single constant $c>0$ depending only on $\kappa_V,\kappa_D,L_V,L_D$ and $\|w^*\|$, we rewrite the bound as

$$\|\Sigma^{1/2}(\widehat{w}-w^*)\| \;\le\; c\Bigg[\lambda^{-1/2}\Big(\frac{\log(4/\delta)}{n}+\gamma\,\frac{\log(4/\delta)}{m}\Big)+\Big(\frac{\mathcal{N}_V(\lambda,\gamma)}{\sqrt{n}}+\gamma\,\frac{\mathcal{N}_D(\lambda,\gamma)}{\sqrt{m}}\Big)\sqrt{\log\frac{4}{\delta}}+\lambda\,\mathcal{B}(\lambda,\gamma)\Bigg].$$

Now recall the identity $\|\Sigma^{1/2}(\widehat{w}-w^*)\| = \mathbb{E}\left[(\widehat{u}(x)-u^*(x))^2\right]^{\frac{1}{2}}$, and since $\log(4/\delta)\ge 1$ we can factor it out to obtain the slightly simplified bound of Equation (4)

$$\mathbb{E}\left[(\hat{u}(x)-u^*(x))^2\right]^{\frac{1}{2}} \le c\log\frac{4}{\delta}\left[\frac{1}{\sqrt{\lambda}}\left(\frac{1}{n}+\frac{\gamma}{m}\right)+\frac{\mathcal{N}_V(\lambda,\gamma)}{\sqrt{n}}+\gamma\frac{\mathcal{N}_D(\lambda,\gamma)}{\sqrt{m}}+\lambda\mathcal{B}(\lambda,\gamma)\right].$$

□

A.5 Learning rate acceleration from differential information

We now instantiate the finite-sample bound under the Assumption (A6) and optimize over λ,γ to obtain Corollary 3.1.

Corollary A.1 (Learning rates for PIKS)**.** *Under assumptions (A1) to (A6), there exist choices of regularization parameters $\lambda=\lambda(n,m)$ and $\gamma=\gamma(n,m)$ given explicitly in the proof which depend polynomially on n, m, $\log n$, and $\log m$, such that, with high probability,*

$$\mathbb{E}\left[(\widehat{u}(x)-u^*(x))^2\right]^{\frac{1}{2}} \;\lesssim\; \begin{cases} n^{-\frac{1}{4}}m^{-\frac{1-r}{8}}, & m\lesssim m_{\mathrm{crit}},\\ n^{-\frac{1}{2(1+\alpha)}}, & m\gtrsim m_{\mathrm{crit}},\end{cases}$$

where $m_{\mathrm{crit}} \asymp n^{\frac{2(1-\alpha)}{(1+\alpha)(1-r)}}$.

Proof. The proof proceeds in four steps.

Step 1: Capacity bounds. As a consequence of Assumption (A6), using $\Sigma_\gamma+\lambda I\succeq\Sigma_1+\lambda I$ and $\Sigma_\gamma+\lambda I\succeq\gamma\Sigma_D+\lambda I$, we obtain

$$\mathcal{N}_V(\lambda,\gamma)^2 = \mathrm{Tr}\big(\Sigma_1(\Sigma_\gamma+\lambda I)^{-1}\big)+\mathrm{Tr}\big(\Sigma_2(\Sigma_\gamma+\lambda I)^{-1}\big) \;\lesssim\; \lambda^{-\alpha}+\gamma^{r-1}\lambda^{-r}. \tag{9}$$

Moreover, since $\Sigma_\gamma+\lambda I\succeq\lambda I$,

$$\mathcal{N}_D(\lambda,\gamma)^2 \;\le\; \lambda^{-1}\,\mathrm{Tr}(\Sigma_D) \;\lesssim\; \lambda^{-1}.$$

Step 2: Error bound. From Assumption (A4), the model is well specified, i.e. $\|w^*\|_{\mathcal{H}}<+\infty$. Then

$$\mathcal{B}(\lambda,\gamma)^2 = \langle w^*,(\Sigma_\gamma+\lambda I)^{-1}w^*\rangle \;\le\; \lambda^{-1}\|w^*\|_{\mathcal{H}}^2, \qquad \lambda\,\mathcal{B}(\lambda,\gamma)\;\lesssim\;\sqrt{\lambda}.$$

Throughout the remainder of the proof, we restrict to regularization parameters satisfying

$$\max\left\{\frac{\log n}{n},\,\gamma\,\frac{\log m}{m}\right\} \;\lesssim\; \lambda \;\lesssim\; \gamma.$$

This ensures the admissibility condition of Theorem 3.1. Under this regime, using $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$, the finite-sample bound simplifies to

$$\|\Sigma^{1/2}(\widehat{w} - w^*)\| \;\lesssim\; \frac{\lambda^{-\alpha/2}}{\sqrt{n}} \;+\; \frac{\gamma^{\frac{r-1}{2}}\lambda^{-r/2}}{\sqrt{n}} \;+\; \frac{\gamma\,\lambda^{-1/2}}{\sqrt{m}} \;+\; \sqrt{\lambda}. \tag{10}$$

This bound makes explicit how derivative information can reduce the effective value complexity through the aligned component Σ_2, while simultaneously introducing a variance cost that grows with γ through the derivative observations.

Step 3: Optimization over γ. For fixed λ, only the middle two terms in (10) depend on γ. Increasing γ reduces the variance associated with estimating the function values in directions constrained by the derivative operator (by shrinking the contribution of the corresponding eigenspaces), but simultaneously amplifies the variance coming from the noisy estimation of derivative information. The optimal choice of γ is obtained by balancing these two effects by solving

$$\frac{\gamma^{\frac{r-1}{2}}\lambda^{-r/2}}{\sqrt{n}} \;=\; \frac{\gamma\,\lambda^{-1/2}}{\sqrt{m}}.$$

This is equivalent to

$$\gamma^{\frac{r-3}{2}} \;=\; \Big(\frac{n}{m}\Big)^{1/2}\lambda^{\frac{r-1}{2}},$$

which yields

$$\gamma^\star(\lambda) \;\asymp\; \Big(\frac{m}{n}\Big)^{\frac{1}{3-r}}\lambda^{\frac{1-r}{3-r}}. \tag{11}$$

Substituting (11) into (10), the two γ–dependent terms become equal and we obtain

$$\|\Sigma^{1/2}(\widehat{w} - w^*)\| \;\lesssim\; A(\lambda) \;+\; B(\lambda) \;+\; \sqrt{\lambda}, \tag{12}$$

where

$$A(\lambda) := \frac{\lambda^{-\alpha/2}}{\sqrt{n}}, \qquad B(\lambda) := m^{-\frac{1-r}{2(3-r)}}\,n^{-\frac{1}{3-r}}\,\lambda^{-\frac{1+r}{2(3-r)}}.$$

Substituting $\gamma^\star(\lambda) \asymp (m/n)^{\frac{1}{3-r}}\lambda^{\frac{1-r}{3-r}}$ into the working regime $\max\{\log n/n, \gamma\log m/m\} \lesssim \lambda \lesssim \gamma$ yields the equivalent λ-only conditions

$$\max\!\left\{\frac{\log n}{n},\; \Big(\frac{m}{n}\Big)^{\frac{1}{2}}\Big(\frac{\log m}{m}\Big)^{\frac{3-r}{2}}\right\} \;\lesssim\; \lambda \;\lesssim\; \sqrt{\frac{m}{n}}.$$

Step 4: Optimization over λ and regime analysis. We choose λ by balancing the bias $\sqrt{\lambda}$ with one of the variance terms in (12).

(i) Balance with $A(\lambda)$. We solve

$$\sqrt{\lambda} \;=\; \frac{\lambda^{-\alpha/2}}{\sqrt{n}},$$

which gives

$$\lambda_A^\star \;\asymp\; n^{-\frac{1}{1+\alpha}}, \qquad \sqrt{\lambda_A^\star} \;\asymp\; A(\lambda_A^\star) \;\asymp\; n^{-\frac{1}{2(1+\alpha)}}. \tag{13}$$

(ii) Balance with $B(\lambda)$. We solve

$$\sqrt{\lambda} \;=\; m^{-\frac{1-r}{2(3-r)}}\,n^{-\frac{1}{3-r}}\,\lambda^{-\frac{1+r}{2(3-r)}},$$

which gives

$$\lambda_B^\star \;\asymp\; m^{-\frac{1-r}{4}}\,n^{-\frac{1}{2}}, \qquad \sqrt{\lambda_B^\star} \;\asymp\; B(\lambda_B^\star) \;\asymp\; n^{-\frac{1}{4}}m^{-\frac{1-r}{8}}. \tag{14}$$

Validity of the regimes. The bound (12) is controlled by the dominant variance term at the chosen λ.

Saturation regime. We choose $\lambda = \lambda_A^\star$ when

$$B(\lambda_A^\star) \lesssim A(\lambda_A^\star).$$

Substituting $\lambda_A^\star \asymp n^{-1/(1+\alpha)}$ into $B(\lambda)$ gives

$$B(\lambda_A^\star) \asymp m^{-\frac{1-r}{2(3-r)}} n^{-\frac{1+2\alpha-r}{2(1+\alpha)(3-r)}}.$$

Hence

$$B(\lambda_A^\star) \lesssim A(\lambda_A^\star)$$

is equivalent to

$$m \gtrsim n^{\frac{2(1-\alpha)}{(1+\alpha)(1-r)}}.$$

Derivative-limited regime. We choose $\lambda = \lambda_B^\star$ when

$$A(\lambda_B^\star) \lesssim B(\lambda_B^\star).$$

which yields the same transition value for m.

Moreover, to ensure our working regime condition, we take the regularization parameter to be clipped as

$$\lambda := \max\left\{\lambda^\star,\ \frac{\log n}{n},\ \gamma^\star(\lambda^\star)\frac{\log m}{m}\right\},$$

where $\lambda^\star \in \{\lambda_A^\star, \lambda_B^\star\}$ is the optimizer in the corresponding regime. One can check that this choice of λ satisfies the admissibility conditions, including the upper condition $\lambda \lesssim \gamma$, in both regimes. Since the variance terms are nonincreasing in λ, clipping only increases the bias term, yielding an additional lower-order contribution of order $\sqrt{\log n/n} + \sqrt{\gamma \log m/m}$.

Resulting rate. Let

$$m_{\mathrm{crit}} \asymp n^{\frac{2(1-\alpha)}{(1+\alpha)(1-r)}}.$$

Recalling the identity $\|\Sigma^{1/2}(\widehat{w} - w^*)\| = \mathbb{E}\left[(\widehat{u}(x) - u^*(x))^2\right]^{\frac{1}{2}}$, the optimized learning rate is

$$\mathbb{E}\left[(\widehat{u}(x) - u^*(x))^2\right]^{\frac{1}{2}} \lesssim \begin{cases} n^{-\frac{1}{4}} m^{-\frac{1-r}{8}}, & m \lesssim m_{\mathrm{crit}}, \\ n^{-\frac{1}{2(1+\alpha)}}, & m \gtrsim m_{\mathrm{crit}}. \end{cases}$$

□

A.6 Learning rates in physically consistent norms

We finally show that the estimator also converges in the stronger norm induced by $\Sigma + \Sigma_D$, which jointly controls the prediction error and the error on the differential quantities (see Corollary 3.2 in the main text).

Corollary A.2 (Learning rates in the physically consistent norm)**.** *Under Assumptions (A1) to (A6), there exist choices of regularization parameters $\lambda = \lambda(n, m)$ and $\gamma = \gamma(n, m)$ given explicitly in the proof with $\gamma \geq 1$ such that, with high probability,*

$$\mathbb{E}\left[(\widehat{u}(x) - u^*(x))^2 + (D\widehat{u}(z) - Du^*(z))^2\right]^{1/2} \lesssim \begin{cases} n^{-\frac{1}{4}} m^{-\frac{1-r}{8}} + m^{-1/4}, & m \lesssim m_{\mathrm{crit}}, \\ n^{-\frac{1}{2(1+\alpha)}} + m^{-1/4}, & m \gtrsim m_{\mathrm{crit}}, \end{cases}$$

where

$$m_{\mathrm{crit}} \asymp n^{\frac{2(1-\alpha)}{(1+\alpha)(1-r)}}.$$

Proof. Since $\gamma \geq 1$, we have

$$\Sigma_\gamma = \Sigma + \gamma\Sigma_D \succeq \Sigma + \Sigma_D,$$

and therefore

$$\|(\Sigma + \Sigma_D)^{1/2}(\widehat{w} - w^*)\| \leq \|\Sigma_\gamma^{1/2}(\widehat{w} - w^*)\|.$$

Finite-sample bound. Repeating the proof of Theorem 3.1 with $\Sigma^{1/2}$ replaced by $\Sigma_\gamma^{1/2}$ yields the same finite-sample bound, since both the estimation and bias terms are handled exactly as before, but using $\|\Sigma_\gamma^{1/2}(\Sigma_\gamma + \lambda I)^{-1/2}\| \le 1$ instead of $\|\Sigma^{1/2}(\Sigma_\gamma + \lambda I)^{-1/2}\| \le 1$. Then, the same argument as in the proof of Corollary 3.1 gives

$$\|\Sigma_\gamma^{1/2}(\widehat{w} - w^*)\| \;\lesssim\; \frac{\lambda^{-\alpha/2}}{\sqrt{n}} + \frac{\gamma^{\frac{r-1}{2}}\lambda^{-r/2}}{\sqrt{n}} + \frac{\gamma\,\lambda^{-1/2}}{\sqrt{m}} + \sqrt{\lambda},$$

up to logarithmic factors and lower-order terms.

The optimization over γ is identical to that of Corollary 3.1, except that we now impose the constraint $\gamma \ge 1$. Let

$$\gamma^\star(\lambda) \;\asymp\; \Big(\frac{m}{n}\Big)^{\frac{1}{3-r}} \lambda^{\frac{1-r}{3-r}}$$

denote the unconstrained optimizer from Corollary 3.1. Under the constraint $\gamma \ge 1$, we therefore choose

$$\gamma(\lambda) \asymp \max\{1, \gamma^\star(\lambda)\}.$$

Optimizing γ and λ. When $\gamma^\star(\lambda) \ge 1$, we recover exactly the same bound as in Corollary 3.1. When $\gamma^\star(\lambda) < 1$, the constraint is active and we set $\gamma = 1$, which yields the additional term

$$\frac{\lambda^{-1/2}}{\sqrt{m}}.$$

Balancing this term with the bias term $\sqrt{\lambda}$ gives

$$\lambda \asymp m^{-1/2},$$

and therefore an additional contribution of order $m^{-1/4}$.

Therefore, the optimal regularization parameter is

$$\lambda \;\asymp\; \max\{\lambda_A^\star, \lambda_B^\star, m^{-1/2}\},$$

where $\lambda_A^\star$ and $\lambda_B^\star$ are the optimizers obtained in the proof of Corollary 3.1. Using the identity

$$\|(\Sigma + \Sigma_D)^{1/2}(\widehat{w} - w^*)\| = \mathbb{E}\left[(\widehat{u}(x) - u^*(x))^2 + (D\widehat{u}(z) - Du^*(z))^2\right]^{1/2},$$

this yields

$$\mathbb{E}\left[(\widehat{u}(x) - u^*(x))^2 + (D\widehat{u}(z) - Du^*(z))^2\right]^{1/2} \;\lesssim\; \begin{cases} n^{-\frac{1}{4}} m^{-\frac{1-r}{8}} + m^{-1/4}, & m \lesssim m_{\mathrm{crit}}, \\ n^{-\frac{1}{2(1+\alpha)}} + m^{-1/4}, & m \gtrsim m_{\mathrm{crit}}, \end{cases}$$

as claimed. □

A.7 Adaptation to the vector-valued case

The setting of this paper considers scalar-valued differential operators, for simplicity of the argument. The adaptation to vector-valued operator is straightforward as we detail in this section.

Consider a vector-valued differential operator $D = (D_1, \dots, D_k)$, where each D_i is of the form given in Assumption (A1). Then, for any $u \in \mathcal{H}$ and any $x \in \mathcal{X}$, $Du(x)$ is a vector in $\mathbb{R}^k$. The differential data points d_j are also in $\mathbb{R}^k$ and we replace the risk (1) by the following one:

$$\widehat{R}(u) = \frac{1}{n}\sum_{i=1}^{n} \big(u(x_i) - y_i\big)^2 + \gamma\,\frac{1}{m}\sum_{j=1}^{m} \|Du(z_j) - d_j\|_2^2 + \lambda\|u\|_{\mathcal{H}}^2. \tag{15}$$

We can define component-wise feature maps $D_\ell\phi(z)$, and define the differential covariance operator as

$$\Sigma_D := \mathbb{E}_{z\sim\rho_D}\left[\sum_{\ell=1}^{k} D_\ell\phi(z) \otimes D_\ell\phi(z)\right]. \tag{16}$$

The key point is that the analysis is carried out using covariance operators Σ and Σ_D, and the key assumption, Assumption (A6), is an abstract condition on Σ and Σ_D, regardless of their expressions. As such, even if going from scalar-valued to vector-valued changes the expression of the covariance operator Σ_D as we just saw, as long as one guarantees that Assumption (A6) is satisfied, the rest of the analysis holds. The other assumptions must be adapted in a straightforward way: the smoothness s required in Assumption (A2) is the maximum of the orders of each D_ℓ, and for Assumption (A3) we need the boundedness of each feature $\|D_\ell \phi(z)\|_{\mathcal{H}}$.

For the proofs, we also adapt $V_D, \widehat{\Sigma}_D, \widehat{V}_D$ by summing over ℓ:

$$V_D := \mathbb{E}\left[\sum_{\ell=1}^{k} d_\ell \, D_\ell \phi(z)\right]$$

$$\widehat{\Sigma}_D := \frac{1}{m}\sum_{j=1}^{m}\sum_{\ell=1}^{k} D_\ell \phi(z_j) \otimes D_\ell \phi(z_j)$$

$$\widehat{V}_D := \frac{1}{m}\sum_{j=1}^{m}\sum_{\ell=1}^{k} d_{j,\ell} \, D_\ell \phi(z_j)$$

Then, the concentration inequalities are still valid and the proofs are the same.

A.8 Learning rates with exact physical constraint

We derive the PIKS rate in the idealized setting where the differential information is known exactly. In this physical oracle setting, the value observations remain empirical, while the differential part of the risk is replaced by its population counterpart:

$$\widehat{R}_{\mathrm{or}}(u) = \frac{1}{n}\sum_{i=1}^{n}\big(u(x_i) - y_i\big)^2 + \gamma \, \mathbb{E}_{z\sim\rho_D}\big[(Du(z) - Du^*(z))^2\big] + \lambda\|u\|_{\mathcal{H}}^2.$$

Equivalently, in operator form, this replaces $\widehat{\Sigma}_D$ and $\widehat{V}_D$ by their population counterparts Σ_D and V_D.

Corollary A.3 (Oracle learning rate)**.** *Assume Assumptions (A1)-(A6) hold, with $r < 1$. Then there exist choices of $\lambda = \lambda(n)$ and $\gamma = \gamma(n)$ such that, with high probability,*

$$\|\Sigma^{1/2}(\widehat{w}_{\mathrm{or}} - w^*)\| \;\lesssim\; n^{-\frac{1}{2(1+\alpha)}}.$$

Proof. The oracle estimator has the closed form

$$\widehat{w}_{\mathrm{or}} = (\widehat{\Sigma} + \gamma\Sigma_D + \lambda I)^{-1}(\widehat{V} + \gamma V_D).$$

Let

$$\widehat{\Sigma}_\gamma^{\mathrm{or}} := \widehat{\Sigma} + \gamma\Sigma_D, \qquad \widehat{V}_\gamma^{\mathrm{or}} := \widehat{V} + \gamma V_D, \qquad \Sigma_\gamma := \Sigma + \gamma\Sigma_D.$$

Under Assumptions (A4) and (A5),

$$V = \Sigma w^*, \qquad V_D = \Sigma_D w^*, \qquad V + \gamma V_D = \Sigma_\gamma w^*.$$

The proof of Theorem 3.1 applies verbatim, except that the differential part is no longer empirical. Hence the only stochastic deviations are

$$\widehat{\Sigma} - \Sigma, \qquad \widehat{V} - V,$$

and all terms involving $\widehat{\Sigma}_D - \Sigma_D$ and $\widehat{V}_D - V_D$ disappear. Up to logarithmic factors,

$$\|\Sigma^{1/2}(\widehat{w}_{\mathrm{or}} - w^*)\| \;\lesssim\; \frac{1}{n\sqrt{\lambda}} + \frac{\mathcal{N}_V(\lambda,\gamma)}{\sqrt{n}} + \lambda\mathcal{B}(\lambda,\gamma),$$

where

$$\mathcal{N}_V(\lambda,\gamma) = \big\|(\Sigma_\gamma + \lambda I)^{-1/2}\Sigma^{1/2}\big\|_{\mathrm{HS}}, \qquad \mathcal{B}(\lambda,\gamma) = \big\|(\Sigma_\gamma + \lambda I)^{-1/2} w^*\big\|.$$

By Assumption (A6), as in the proof of Corollary 3.1,

$$\mathcal{N}_V(\lambda,\gamma)^2 \;\lesssim\; \lambda^{-\alpha} + \gamma^{r-1}\lambda^{-r}.$$

Moreover, by attainability,

$$\lambda\mathcal{B}(\lambda,\gamma) \lesssim \sqrt{\lambda}.$$

Therefore,

$$\|\Sigma^{1/2}(\widehat{w}_{\mathrm{or}} - w^*)\| \;\lesssim\; \frac{1}{n\sqrt{\lambda}} + \frac{\lambda^{-\alpha/2}}{\sqrt{n}} + \frac{\gamma^{\frac{r-1}{2}}\lambda^{-r/2}}{\sqrt{n}} + \sqrt{\lambda}.$$

Since the differential information is known at the population level, there is no derivative-sample variance term increasing with γ. We choose γ large enough so that

$$\gamma^{r-1}\lambda^{-r} \;\lesssim\; \lambda^{-\alpha}.$$

Equivalently, when $r > \alpha$, it suffices to take

$$\gamma \;\gtrsim\; \lambda^{-\frac{r-\alpha}{1-r}},$$

while when $r \le \alpha$, any $\gamma \gtrsim 1$ is enough. With this choice,

$$\|\Sigma^{1/2}(\widehat{w}_{\mathrm{or}} - w^*)\| \;\lesssim\; \frac{1}{n\sqrt{\lambda}} + \frac{\lambda^{-\alpha/2}}{\sqrt{n}} + \sqrt{\lambda}.$$

Balancing the dominant variance term with the bias,

$$\frac{\lambda^{-\alpha/2}}{\sqrt{n}} = \sqrt{\lambda},$$

gives

$$\lambda \asymp n^{-\frac{1}{1+\alpha}}.$$

For this choice,

$$\frac{\lambda^{-\alpha/2}}{\sqrt{n}} \asymp \sqrt{\lambda} \asymp n^{-\frac{1}{2(1+\alpha)}},$$

while

$$\frac{1}{n\sqrt{\lambda}} = n^{-1+\frac{1}{2(1+\alpha)}}$$

is lower order since $\alpha > 0$. Hence

$$\|\Sigma^{1/2}(\widehat{w}_{\mathrm{or}} - w^*)\| \;\lesssim\; n^{-\frac{1}{2(1+\alpha)}}.$$

□

B Examples

We now illustrate Assumption (A6) on concrete examples where the decomposition can be computed explicitly.

B.1 Partial Laplacian and periodic Sobolev RKHS

Lemma B.1 (Capacity decomposition for Sobolev spaces)**.** *Let $\mathcal{X} = \mathbb{T}^d$, let k be a Matérn kernel of smoothness $s > d/2$, and let $D = \sum_{i\in S}\partial_{x_i}^2$ be a partial Laplacian with $S \subseteq \{1,\dots,d\}$. Assume that value and differential samples are drawn uniformly on $\mathbb{T}^d$.*

Then the value covariance operator Σ admits a decomposition $\Sigma = \Sigma_1 + \Sigma_2$ satisfying Assumption (A6), with exponents

$$\alpha = \frac{d - |S|}{2s}, \qquad r = \frac{d}{2s}.$$

Proof. **Step 1: diagonalization in the Fourier basis.** Let D be a constant-coefficient differential operator of order q, of the form

$$D = \sum_{|\alpha| \le q} c_\alpha \, \partial^\alpha, \qquad \partial^\alpha = \partial_{x_1}^{\alpha_1} \cdots \partial_{x_d}^{\alpha_d},$$

with real coefficients c_α. Its Fourier symbol is given by

$$P(k) = \sum_{|\alpha| \le q} c_\alpha \, (2\pi i k)^\alpha, \qquad k \in \mathbb{Z}^d,$$

so that $De^{2\pi i k \cdot x} = P(k)\, e^{2\pi i k \cdot x}$.

Since the kernel k is translation invariant and both x and z are sampled uniformly on $\mathbb{T}^d$, the covariance operators

$$\Sigma = \mathbb{E}[\phi(x) \otimes \phi(x)], \qquad \Sigma_D = \mathbb{E}[D\phi(z) \otimes D\phi(z)]$$

are convolution operators and are therefore diagonal in the Fourier basis $\{e_k(x) := e^{2\pi i k \cdot x}\}_{k \in \mathbb{Z}^d}$. The corresponding eigenvalues are

$$\sigma_k = \mu_k, \qquad \tau_k = |P(k)|^2 \, \mu_k.$$

Step 2: partial Laplacian and visible/invisible frequencies. For $D = \sum_{i \in S} \partial_{x_i}^2$, the symbol is

$$P(k) = -(2\pi)^2 \|k_S\|^2, \qquad \text{so} \qquad \tau_k = (2\pi)^4 \|k_S\|^4 \, \mu_k,$$

where k_S denotes the restriction of k to coordinates in S. Define the index sets

$$\mathcal{Z} := \{k \in \mathbb{Z}^d : \; k_S = 0\}, \qquad \mathcal{Z}^c := \mathbb{Z}^d \setminus \mathcal{Z}.$$

Thus $\mathcal{Z}$ consists of frequencies *invisible* to D (since $P(k) = 0$), whereas $\mathcal{Z}^c$ corresponds to frequencies *detectable* by D.

Step 3: the decomposition $\Sigma = \Sigma_1 + \Sigma_2$. Define

$$\Sigma_1 := \sum_{k \in \mathcal{Z}} \mu_k \, e_k \otimes e_k, \qquad \Sigma_2 := \sum_{k \in \mathcal{Z}^c} \mu_k \, e_k \otimes e_k.$$

Then $\Sigma = \Sigma_1 + \Sigma_2$ and $\Sigma_1, \Sigma_2 \succeq 0$.

Step 4: capacity of the invisible component Σ_1. Let $d_0 = d - |S|$ and write $k_{-S} \in \mathbb{Z}^{d-S}$ for the subvector of coordinates outside S. The map $k_{-S} \mapsto (k_{-S}, 0_S)$ is a bijection $\mathbb{Z}^{d-S} \to \mathcal{Z}$ and $|(k_{-S}, 0_S)| = |k_{-S}|$, hence

$$\sum_{k \in \mathcal{Z}} \frac{\mu_k}{\mu_k + \lambda} = \sum_{k_{-S} \in \mathbb{Z}^{d-S}} \frac{\mu_{(k_{-S}, 0_S)}}{\mu_{(k_{-S}, 0_S)} + \lambda}, \qquad \mu_{(k_{-S}, 0_S)} \asymp (1 + |k_{-S}|^2)^{-s}.$$

Therefore, by a standard comparison between lattice sums and integrals, for all $\lambda \in (0, 1]$,

$$\operatorname{Tr}\big(\Sigma_1 (\Sigma_1 + \lambda I)^{-1}\big) = \sum_{k \in \mathcal{Z}} \frac{\mu_k}{\mu_k + \lambda} \; \lesssim \; \lambda^{-\frac{d_0}{2s}}.$$

This verifies the first trace condition with exponent $\alpha = d_0/(2s)$.

Step 5: alignment of Σ_2 with Σ_D. On $\mathcal{Z}^c$ we have $\|k_S\| \ge 1$ and hence $\tau_k = (2\pi)^4 \|k_S\|^4 \mu_k \ge c\, \mu_k$ for some $c > 0$. Therefore, for all $t \in (0, 1]$,

$$\operatorname{Tr}\big((\Sigma_D + tI)^{-1} \Sigma_2\big) = \sum_{k \in \mathcal{Z}^c} \frac{\mu_k}{\tau_k + t} \; \le \; \sum_{k \in \mathcal{Z}^c} \frac{\mu_k}{c\mu_k + t}.$$

Using $\mu_k \asymp (1 + |k|^2)^{-s}$ and standard comparison between lattice sums and integrals gives

$$\sum_{k \in \mathbb{Z}^d} \frac{\mu_k}{\mu_k + t} \; \lesssim \; t^{-\frac{d}{2s}}, \qquad t \in (0, 1],$$

and hence

$$\operatorname{Tr}\big((\Sigma_D + tI)^{-1} \Sigma_2\big) \; \lesssim \; t^{-\frac{d}{2s}}.$$

This verifies the second trace condition with exponent $r = d/(2s)$.

Combining Steps 3-5 yields Assumption (A6) with the claimed exponents. □

B.2 Gradients

Let $\Omega \subset \mathbb{R}^d$ be a bounded, connected, Lipschitz domain. Fix $s > \frac{d}{2} + 1$. Let $\mathcal{H} = H^s(\Omega)$, with norm $\|\cdot\|_{\mathcal{H}}$ equivalent to the standard $H^s(\Omega)$ norm. We denote by K the reproducing kernel associated to $\mathcal{H}$, and by $\phi : \Omega \to \mathcal{H}$ the feature map defined, for all $x \in \Omega$, by $\phi(x) = K(x, \cdot) \in \mathcal{H}$. Choose $\rho = \rho_D$ the uniform distribution on Ω.

B.2.1 Covariance operator decomposition

Definition of Σ Define the covariance operator

$$\Sigma := \mathbb{E}_{x\sim\rho}[\phi(x) \otimes \phi(x)].$$

We have, for all $u \in \mathcal{H}$,

$$\langle \Sigma u, u\rangle_{\mathcal{H}} = \|u\|^2_{L^2(\rho)}. \tag{17}$$

Definition of Σ_D Define the gradients covariance operator

$$\Sigma_D = \mathbb{E}_{z\sim\rho_D}\left[\sum_{i=1}^{d} \partial_i\phi(z) \otimes \partial_i\phi(z)\right].$$

We have, for all $u \in \mathcal{H}$,

$$\langle \Sigma_D u, u\rangle_{\mathcal{H}} = \|\nabla u\|^2_{L^2(\rho)}. \tag{18}$$

Definition of Σ_1 Denote by $\phi_\Omega = \mathbb{E}_{x\sim\rho}[\phi(x)] \in \mathcal{H}$ the representer of the averaging functional, so that

$$\langle u, \phi_\Omega\rangle_{\mathcal{H}} = \mathbb{E}_{x\sim\rho}[u(x)] =: u_\Omega.$$

Denote

$$\Sigma_1 = \phi_\Omega \otimes \phi_\Omega.$$

Then for all $u \in \mathcal{H}$,

$$\langle \Sigma_1 u, u\rangle_{\mathcal{H}} = u^2_\Omega. \tag{19}$$

Definition of Σ_2 Next, define the bounded operator

$$A : \mathcal{H} \to L^2(\rho), \qquad Au := u - u_\Omega,$$

Set

$$\Sigma_2 := A^*A,$$

so that for all $u \in \mathcal{H}$,

$$\langle \Sigma_2 u, u\rangle_{\mathcal{H}} = \|u - u_\Omega\|^2_{L^2(\rho)} = \mathrm{Var}_{x\sim\rho}(u(x)). \tag{20}$$

Proposition B.2 (Decomposition of Σ)**.** *One has*

$$\Sigma = \Sigma_1 + \Sigma_2.$$

Proof. Let $u \in \mathcal{H}$. By the variance decomposition formula,

$$\mathbb{E}_{x\sim\rho}\big[u(x)^2\big] = \Big(\mathbb{E}_{x\sim\rho}[u(x)]\Big)^2 + \mathrm{Var}_{x\sim\rho}(u(x)).$$

Equivalently,

$$\|u\|^2_{L^2(\rho)} = u^2_\Omega + \|u - u_\Omega\|^2_{L^2(\rho)} = u^2_\Omega + \mathrm{Var}_{x\sim\rho}(u(x)).$$

Using (17), (19), and (20), we obtain

$$\langle \Sigma u, u\rangle_{\mathcal{H}} = \langle \Sigma_1 u, u\rangle_{\mathcal{H}} + \langle \Sigma_2 u, u\rangle_{\mathcal{H}}$$

for every $u \in \mathcal{H}$.

Since Σ, Σ_1, and Σ_2 are bounded self-adjoint operators on $\mathcal{H}$, equality of their quadratic forms implies, by polarization, that

$$\Sigma = \Sigma_1 + \Sigma_2.$$

□

B.2.2 Computation of the coefficients

Let us first bound the effective dimension

$$\mathcal{N}_1(\lambda) := \operatorname{Tr}\left(\Sigma_1(\Sigma_1 + \lambda I)^{-1}\right).$$

Since Σ_1 has rank one, the effective dimension $\mathcal{N}_1(\lambda)$ is uniformly bounded. In particular, it satisfies

$$\mathcal{N}_1(\lambda) \le C_\alpha \lambda^{-\alpha}, \qquad \lambda \in (0, 1],$$

with

$$\boxed{\alpha = 0}.$$

Let us now bound

$$\operatorname{Tr}\big((\Sigma_D + \lambda I)^{-1}\Sigma_2\big).$$

Since Ω is bounded, connected, and Lipschitz, the Poincaré–Wirtinger inequality [26] holds: there exists $C > 0$ such that, for every $u \in H^1(\Omega)$,

$$\|u - u_\Omega\|_{L^2(\rho)} \le C\|\nabla u\|_{L^2(\rho)}.$$

Therefore, for every $u \in \mathcal{H}$,

$$\begin{aligned}\langle \Sigma_2 u, u\rangle_{\mathcal{H}} &= \|u - u_\Omega\|^2_{L^2(\rho)} \\ &\le C^2\|\nabla u\|^2_{L^2(\rho)} \\ &= C^2\langle \Sigma_D u, u\rangle_{\mathcal{H}}.\end{aligned}$$

Equivalently, in the Loewner order,

$$0 \preceq \Sigma_2 \preceq C^2\Sigma_D.$$

It follows that

$$\operatorname{Tr}\big((\Sigma_D + \lambda I)^{-1}\Sigma_2\big) \le C^2 \operatorname{Tr}\big(\Sigma_D(\Sigma_D + \lambda I)^{-1}\big).$$

We now bound the last trace. Let $(\eta_j)_{j\ge1}$ be the nonzero eigenvalues of Σ_D, arranged in nonincreasing order. Since Σ_D is the covariance operator associated with the map

$$u \longmapsto \nabla u, \qquad H^s(\Omega) \to L^2(\rho; \mathbb{R}^d),$$

the standard eigenvalue estimate for Sobolev embeddings on bounded Lipschitz domains gives

$$\eta_j \le C_D j^{-2(s-1)/d}.$$

Set

$$\beta := \frac{2(s-1)}{d}.$$

Because $s > \frac{d}{2} + 1$, we have $\beta > 1$. Hence, for $0 < \lambda \le 1$,

$$\begin{aligned}\operatorname{Tr}\big(\Sigma_D(\Sigma_D + \lambda I)^{-1}\big) &= \sum_{j\ge1} \frac{\eta_j}{\eta_j + \lambda} \\ &\le \sum_{j\ge1} \frac{C_D j^{-\beta}}{C_D j^{-\beta} + \lambda} \\ &\lesssim \lambda^{-1/\beta}.\end{aligned}$$

Since $1/\beta = d/(2(s-1))$, we obtain

$$\operatorname{Tr}\big((\Sigma_D + \lambda I)^{-1}\Sigma_2\big) \lesssim \lambda^{-d/(2(s-1))}.$$

Thus one may take

$$\boxed{r = \frac{d}{2(s-1)}}.$$

In particular, since $s > \frac{d}{2} + 1$, one has $r < 1$.

C Detailed experimental setup

In this section we provide more details about the setup used for the experiments of Section 5.

Implementation To maximize the flexibility of our system, we implemented the PIKS estimator using the Jax framework, which allows to efficiently compute arbitrary derivatives. While very flexible, any automatic differentiation framework does introduce some computational overhead and requires particular care when implementing kernel functions. For example, Matérn kernels require differentiating through a square-root which is numerically unstable at 0. We introduce a small additive nugget term to ensure stability everywhere. The kernel solver is a straightforward implementation of the equations in Theorem A.2 which yield a space complexity of $O((n+m)^2)$ and time complexity of $O((n+m)^3)$.

Bounded domain gradient example We repeated the experiment 10 times to obtain standard deviations reported in Fig. 2. We set the variance on the differential data to be low in order to avoid having to scale m too much before seeing a saturation effect. All data points were uniformly sampled on a unit disk with 1000 samples used for validation (to select hyperparameters) and 10000 for computing the test error.

Partial Laplacian example We give a few more details on the kernel and target functions for the example on Sobolev spaces on the torus with the partial Laplacian operator. Here we used $\mathcal{X} = \mathbb{T}^4$. We sampled uniformly at random $F = 64$ frequency vectors $k_\ell \in \mathbb{Z}^4$ such that for $\ell \in [0, 15]$, $k_\ell \sim [\mathcal{U}([-8, 8]\backslash\{0\}), 0, 0, 0]$; for $\ell \in [16, 31]$, $k_\ell \sim [0, \mathcal{U}([-8, 8]\backslash\{0\})]$ and so on (only one dimension active for each vector). For feature map $\phi_\ell(x) = \sqrt{2}\cos(2\pi\langle k_\ell, x\rangle)$, target function and kernel are defined as

$$u^*(x) = u_0 + \sum_{\ell=1}^{F} c_\ell \phi_\ell(x), \qquad k(x, x') = 1 + \sum_{\ell=1}^{F} \mu_\ell \phi_\ell(x)\phi_\ell(x'),$$

with the smoothness coefficients

$$\mu_\ell \asymp (1 + \|k_\ell\|^2)^{-\beta}, \quad c_\ell \asymp (1 + \|k_\ell\|^2)^{-\delta}, \quad \beta > 1, \quad \delta > \frac{1}{2}.$$

In particular, we set $\beta = \delta = 4.1$ to have a well-specified problem in a smooth enough space. We use additive Gaussian noise, with standard-deviation equal to 10% to the range of the data. This was done because the different data components have vastly different numerical ranges, and setting a fixed variance would have introduced artifacts in the results.

The best-fit lines of Fig. 3 are obtained by least-squares regression of the experimental data against the function an^b to find the exponential coefficient b. Then, using Corollary 3.1, we get $\alpha = -\frac{1}{2b} - 1$.

Hyperparameters The two hyperparameters of PIKS (λ and γ), as well as the hyperparameters of the kernel (notably the length-scale of the Matérn kernel) were determined by a coarse grid-search using small, noisy validation sets to estimate their performance at generalization time.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, we precisely state the claims at the beginning of the paper and structure the remaining sections accordingly.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The assumptions are stated and clearly commented with regards to the setting in which they apply. For example, a strong assumption we make is that we consider linear differential operators D (which we make explicit from the abstract). This can be understood as a clear limitation of our work. In the conclusions we reiterate the limitations of our work as avenues for potential extensions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We dedicate Section 3.1 to discuss our assumptions. The proofs are completely contained in the appendix and are correct to the best of our knowledge.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our experiments use synthetically generated data. We provide all details about how to compute our kernel-based models, as well as the data generating procedures in the main text and in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code is not publicly released. However, the paper provides the mathematical formulation of the estimator and details of the synthetic data generation and experimental procedures in Sections 2 and 5 and Appendix C.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Details in appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Error bars present in plots.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendix C describes the implementation and computational complexity of the estimator, but detailed hardware specifications, memory usage, and execution times are not reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We read the code of ethics and comply with it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: We believe the paper to have limited negative societal impact: it refers to theoretical understanding of a problem, more than its practical applications.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No risks posed.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: No existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human subjects were involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: N/A

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: No LLMs involved.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.