PHYSICS-INFORMED LEARNING UNDER MIXING: HOW PHYSICAL KNOWLEDGE SPEEDS UP LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A major challenge in physics-informed machine learning is to understand how the incorporation of prior domain knowledge affects learning rates when data are dependent. Focusing on empirical risk minimization with physics-informed regularization, we derive complexity-dependent bounds on the excess risk in probability and in expectation. We prove that, when the physical prior information is aligned, the learning rate improves from the (slow) Sobolev minimax rate to the (fast) optimal i.i.d. one without any sample-size deflation due to data dependence.

1 INTRODUCTION

Physics-informed machine learning encompasses a wide taxonomy of approaches that combine physical knowledge and learning algorithms to address two main tasks: (i) enhancing physical models (given, e.g., by systems of partial differential equations) through data-driven methods to improve their accuracy and numerical solvability; (ii) improve the learning algorithms' performance by including physical information, e.g., as additional constraint (Karniadakis et al., 2021; Meng et al., 2025). Focusing on the second class of methods, surveyed in Rai & Sahu (2020); von Rueden et al. (2023b), the resulting approaches turn out to be practically effective in terms of data efficiency, generalization capability and interpretability, especially in view of downstream tasks such as safe learning-based control (Nghiem et al., 2023; Drgona et al., 2025). However, theoretically quantifying the beneficial impact of physical information into learning algorithms is technically challenging and still an active research question (see von Rueden et al. (2023a) and references therein).

In this paper, we tackle this question by considering a statistical learning set-up and focusing on regularized empirical risk minimization problems of the following form:

$$\hat{f} = \underset{\substack{f \,\in\, \text{ball in} \\ \text{Sobolev space}}}{\arg\min} \boxed{\begin{array}{c}\text{data-fit} \\ \text{squared loss}(f)\end{array}} + \lambda_T \boxed{\begin{array}{c}\text{physics-informed} \\ \text{regularizer}(f)\end{array}}, \tag{1.1}$$

where data entering the fit term are *dependent*, derived from observations of a ground-truth nonlinear dynamical system $X_{t+1} = f_\star(X_t) + W_t$, with W_t being a sub-Gaussian noise martingale difference sequence. The regularizer in (1.1) encodes the information that the true function to be estimated, $f_\star$, approximately satisfies a known partial differential equation induced by a linear operator $\mathscr{D}$ — i.e., we have that the regularizer takes the form $\|\mathscr{D}(f)\|^2_{\mathscr{L}^2}$, and we say that *knowledge alignment* occurs if it holds that $\|\mathscr{D}(f_\star)\|^2_{\mathscr{L}^2} \simeq 0$.

The main results of this paper are *complexity-dependent* bounds — i.e., bounds that depend on $||\mathscr{D}(f_\star)||_{\mathscr{L}^2}$ (Lecué & Mendelson, 2017) — for the *excess risk* $||\hat{f} - f_\star||^2_{\mathscr{L}^2}$ in physics-informed and non-parametric learning with dependent data. Informally, our results (both in high probability and expectation) will look like this:

Theorem (Informal). For a suitable choice of the regularization parameter λ_T, for a sufficiently large number of samples T, and letting $d < 1$ be the *Sobolev minimax rate* (Ibragimov & Has'minskii, 1981; Nussbaum, 2006), it holds that

$$\text{(Excess risk)} \quad ||\hat{f} - f_\star||^2_{\mathscr{L}^2} \le C_{\mathtt{slow}} \frac{\|\mathscr{D}(f_\star)\|^{\text{some power}}_{\mathscr{L}^2}}{T^d} + C_{\mathtt{fast}} \frac{\text{noise level}}{T}.$$

Thanks to this we show that, under knowledge alignment, the regularized estimate $\hat{f}$ converges to the true, unknown function $f_\star$ at the i.i.d. rate of $\mathcal{O}(1/T)$: in other words, it behaves like classic optimal rates for i.i.d. learning *even if the data are dependent* after a suitable burn-in time.

The remainder of the paper unfolds as follows: Section 2 provides the set-up of the learning problem, introducing the *weighted, vector-valued* function spaces that will be used throughout the paper. Next, the learning problem is stated in Section 3, and in Section 4 we provide the general statement for the excess risk bounds, both in probability and in expectation. Our analysis culminates in Section 5, where we prove how knowledge alignment leads to optimal i.i.d. rates even if data are dependent. We discuss our results in juxtaposition with related works in Section 6, and present some concluding remarks in Section 7.

2 Problem set-up

This section collects preliminary concepts, defining the probability set-up of the data-generation mechanism (Section 2.1) and the involved weighted, vector-valued function spaces (Section 2.2).

2.1 Input domain and trajectory distribution

Let $\Omega \subseteq [-L, L]^{d_X} \subset \mathbb{R}^{d_X}$ be the input domain whose boundary is locally Lipschitz (Adams & Fournier, 2003, Definition 4.9). Suppose we have a horizon length T, the input trajectories denoted by $X \doteq (X_0,\, X_1,\, \cdots,\, X_{T-1})$ belong to the metric space $(\Omega^T, \{\mathcal{X}_t\}_{t=0}^{T-1}, \mathbb{P}_X)$, where $\Omega^T \doteq \bigtimes_{t=0}^{T-1}\Omega$ is the Cartesian product of the single-component input domains Ω; $\{\mathcal{X}_t\}_{t=0}^{T-1}$ is the *filtration* given by a sequence of increasing σ-algebras $\mathcal{X}_{t+1} \subset \mathcal{X}_t$ with respect to which X is *adapted* (Rogers & Williams, 2000, Chapter II.45); and $\mathbb{P}_X$ is the joint probability distribution of the input trajectory. As detailed in Appendix A.1, there exists a probability distribution associated with every component of X — we denote it by μ_t for each $t = 0, \cdots, T-1$, and we mostly work with a *known* initial distribution μ_0 for X_0 (typically, a Dirac measure centered at the observed initial state X_0). Overall, we make use of the following:

Assumption 1. Let μ_λ be the Lebesgue measure defined on $\Omega \subset \mathbb{R}^{d_X}$. For all $t = 0, \cdots, T-1$, each measure $\mu_t \colon \mathcal{X}_t \to \mathbb{R}_{\geq 0}$ is assumed to admit a density with respect to μ_λ. We denote such density by $p_t(\cdot)$, and we assume that there exist $0 < \underline{\kappa} < \overline{\kappa} < \infty$ such that, for all $t = 0, \cdots, T-1$, $\underline{\kappa} \leq p_t(\cdot) \leq \overline{\kappa}$.

Note that Assumption 1 accounts for many cases of practical relevance, such as the uniform, the truncated Gaussian and the beta distributions (Krishnamoorthy, 2016).

2.2 Spaces of functions

Space of square-integrable functions $\mathscr{L}^2$. We will focus on the Hilbert space $\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})$ of vector-valued, square-integrable functions that consist of multiple evaluations of a function $f\colon \Omega \to \mathbb{R}^{d_Y}$ along the input trajectory X. Such a space allows us to consider the trajectory X and is endowed with the inner product defined as follows: given $f, g\colon \Omega \to \mathbb{R}^{d_Y}$, we have

$$\begin{aligned}\langle f,\, g\rangle_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} &\doteq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{\mathbb{P}_X}\left[\langle f(X_t),\, g(X_t)\rangle_2\right] = \frac{1}{T}\sum_{t=0}^{T-1}\int_{\Omega^T}\langle f(X_t),\, g(X_t)\rangle_2\, d\mathbb{P}_X \\ &= \frac{1}{T}\sum_{t=0}^{T-1}\int_{\Omega}\langle f(X_t),\, g(X_t)\rangle_2\,\mu_t(dX_t), \end{aligned} \tag{2.1}$$

where $\langle\cdot,\cdot\rangle_2$ is the standard inner product defined in the Euclidean space $\mathbb{R}^{d_Y}$, and μ_t is the probability measure of the t-th component of X introduced in Section 2.1. The inner product (2.1) induces the trajectory norm $\|f\|_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}$ such that $\|f\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} = \langle f,\, f\rangle_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}$. Furthermore, it follows by construction that one can write $\|f\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{\mathbb{P}_X}[\|f(X_t)\|_2^2]$. Note in addition that, thanks to the separability of $\mathbb{R}^{d_Y}$, the vector-valued space $\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y}) = \{f\colon \Omega \to \mathbb{R}^{d_Y} \mid \|f\|_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} < \infty\}$ can be written as the direct sum $\bigoplus_{i=1}^{d_Y}\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R})$ (Conway, 2007, Chapter I.6): indeed, following (2.1), we can write

$$\|f\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} = \sum_{i=1}^{d_Y}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{\mathbb{P}_X}\left[f_i(X_t)^2\right] = \sum_{i=1}^{d_Y}\|f_i\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R})}.$$

General $\mathscr{L}^p$ spaces. In general, one can define the space $\mathscr{L}^p(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})$ for any $p \in \mathbb{Z}_{\geq 0}$ endowed with the norm $\|f\|^p_{\mathscr{L}^p(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{\mathbb{P}_X}[\|f(X_t)\|_2^p]$. Of particular interest will be the Banach space of bounded functions $\mathscr{L}^\infty(\Omega^T; \mathbb{R}^{d_Y})$ equipped with the norm $\|f\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} \doteq \sup_{x\in\Omega} \|f(x)\|_2$.

Sobolev space $\mathscr{H}^s$. Another fundamental function space derived from $\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})$ is the multi-output, weighted *Sobolev space* of order $s \in \mathbb{Z}_{\geq 0}$, which is defined as follows:

$$\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y}) \doteq \left\{ f \in \mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y}) \mid \|f\|_{\mathscr{H}^s(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} < \infty \right\},$$

where the norm is induced by the inner product

$$\langle f,\, g\rangle_{\mathscr{H}^s(\Omega,\mathbb{P}_X;\mathbb{R}^{d_Y})} \doteq \sum_{|\alpha|\leq s} \langle D^\alpha f,\, D^\alpha g\rangle_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})},$$

with $D^\alpha f$ being the differential given by the multi-index $\alpha \doteq (\alpha_1, \cdots, \alpha_{d_X})$ of non-negative integers with order $|\alpha| \doteq \sum_{i=1}^{d_X} \alpha_i$, i.e., $D^\alpha \doteq \partial^{|\alpha|} f / \partial x_1^{\alpha_1} \cdots \partial x_{d_X}^{\alpha_{d_X}}$. Regarding the order of the Sobolev spaces we will consider, we will rely on the following:

Assumption 2. The order s of $\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})$ is a non-negative integer that satisfies $s \geq 2d_X$.

Finally, note that also the space $\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})$ admits the representation as the direct sum $\bigoplus_{i=1}^{d_Y} \mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R})$ thanks to the separability of $\mathbb{R}^{d_Y}$. This allows us to extend key results of scalar Sobolev spaces to our vector-valued ones, as detailed in Appendix B. In particular, we show that the Sobolev Imbedding Theorem (Adams & Fournier, 2003, Theorem 4.12) holds in our set-up, which will provide the necessary structure for the hypothesis space involved in the learning problem.

3 Problem Statement

Measurement model. Assume to collect T data points, $\mathcal{D} \doteq \{X_t, Y_t\}_{t=0}^{T-1}$, generated according to the measurement model

$$Y_t \doteq X_{t+1} = f_\star(X_t) + W_t, \tag{3.1}$$

where the noise sequence satisfies the following:

Assumption 3. The additive noise $\{W_t\}_{t\in\mathbb{Z}_{\geq 0}}$ is a martingale difference sequence with respect to the filtration $\{\mathcal{X}_t\}_{t\in\mathbb{Z}_{\geq 0}}$: thus, $\mathbb{E}_{W_t}[W_t|\mathcal{X}_{t-1}] = 0$ for all $t = 0, \cdots, T-1$. Moreover, each W_t is also assumed to be σ_W^2-conditionally sub-Gaussian given $\mathcal{X}_{t-1}$: i.e., it holds that, for every $\xi \in \mathbb{R}$ and every u in the unit sphere in $(\mathbb{R}^{d_Y}, \|\cdot\|_2)$,

$$\mathbb{E}\left[\exp\left\{\xi \langle W_t,\, u\rangle_2\right\} \mid \mathcal{X}_{t-1}\right] \leq \exp\left\{\frac{\xi^2\sigma_W^2}{2}\right\}. \tag{3.2}$$

The learning problem. In general, the learning problem can be stated as that of minimizing the *excess risk* $\|\hat{f} - f_\star\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}$, searching for the estimate $\hat{f}$ within a chosen *hypothesis space* $\mathscr{F}$ (which we specify later). However, since the underlying probability measures are unknown, the amount of data in $\mathcal{D}$ is finite and the hypothesis space $\mathscr{F}$ might be large, the estimate $\hat{f}$ is typically computed through *(regularized) empirical risk minimization*:

$$\hat{f} \doteq \operatorname*{arg\,min}_{f\in\mathscr{F}} \frac{1}{T}\sum_{t=0}^{T-1} \|Y_t - f(X_t)\|_2^2 + \lambda_T \Psi(f). \tag{3.3}$$

Focus on the physics-informed regularizer. In the set-up of our interest, the regularizer $\Psi(\cdot)\colon \mathscr{F} \to \mathbb{R}_{\geq 0}$ encodes available prior physical information on the "true" function $f_\star$ — in other words, $\Psi(f)$ penalizes the physical inconsistency of f with respect to the prior on $f_\star$. Such physical information is conveyed by the fact that $f_\star$ is assumed to approximately satisfy a known partial differential equation given by the linear operator $\mathscr{D}\colon \mathscr{H}^s(\Omega, \mu_\lambda; \mathbb{R}^{d_Y}) \to \mathscr{L}^2(\Omega, \mu_\lambda; \mathbb{R}^{d_Y})$. Such an operator is defined component-wise as

$$[\mathscr{D}(f)]_i \doteq \sum_{|\alpha|\leq s} p_{i,\alpha} D^\alpha f_i \text{ for all } i = 1, \cdots, d_Y, \tag{3.4}$$

where each $p_{i,\alpha} \colon \Omega \to \mathbb{R}$ is a bounded function — therefore, if we denote by p the collection of all $p_{i,\alpha}$, then we have that $\|p\|_\infty$ is finite. To describe the regularity of the differential operator in (3.4), we make the following:

Assumption 4. The differential operator $\mathscr{D}(f)$ is *elliptic* — that is, for all $i = 1, \cdots, d_Y$ and any $\xi \in \mathbb{R}^{d_X} \setminus \{0\}$, it holds that $\sum_{|\alpha|=s} p_{i,\alpha} \xi_1^{\alpha_1} \cdots \xi_{d_X}^{\alpha_{d_X}} \neq 0$.

Elliptic partial differential equations abound in practical applications, as they can be seen as generalizations of the Laplace and Poisson operators (Evans, 2010, Chapter 6). The differential operator $\mathscr{D}$ enters the definition of the regularizer in (3.3), where we have

$$\Psi(f) \doteq \|\mathscr{D}(f)\|^2_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})}, \tag{3.5}$$

which is a 2-proper regularizer (Lecué & Mendelson, 2017, Assumption 1.1) — see Appendix E for the definition and further insights.

Hypothesis space. Let us now focus on the hypothesis space $\mathscr{F}$. We consider it as the ball of radius ρ_f in the Sobolev space, i.e.,

$$\mathscr{F} \doteq \left\{ f \in \mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y}) \mid \|f\|_{\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} \leq \rho_f \right\}. \tag{3.6}$$

Alternatively, as pointed out in (Cucker & Zhou, 2007, Theorem 8.21)), one could write the cost in (3.3) as $\frac{1}{T} \sum_{t=0}^{T-1} (Y_t - f(X_t))^2 + \tilde{\lambda}_T \|f\|^2_{\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} + \lambda_T \Psi(f)$, and the minimization would be performed for $f \in \mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})$, thanks to the equivalence yielding $\rho_f = \rho_f(\tilde{\lambda}_T)$. In this paper, we will rely on the following:

Assumption 5. The hypothesis space $\mathscr{F}$ contains the unknown function to be estimated, $f_\star$.

The case in which such an assumption is violated is dealt with in the literature on *approximation theory* – see, e.g., Cucker & Smale (2002); Cucker & Zhou (2007); however, these discussions are beyond the scope of this paper.

Additionally, we will also consider the *effective hypothesis space* induced by the regularizer, namely

$$\mathscr{F}^\rho = \left\{ f \in \mathscr{F} \,\middle|\, \Psi(f - f_\star) \leq \rho \right\}. \tag{3.7}$$

For a visualization of these hypothesis spaces, please refer to Figure 1. Finally, we will sometimes simplify notation by considering the shifted hypothesis space $\mathcal{H}_\star \doteq \mathcal{H} - f_\star = \{f - f_\star \mid f \in \mathcal{H}\}$, with $\mathcal{H}$ being for instance $\mathscr{F}$ or $\mathscr{F}^\rho$.

Modelling sample dependence in trajectories. Finally, we assume regularity in the trajectory X given by the following one-sided exponential inequality (Samson, 2000):

Assumption 6. The trajectory X governed by the law $\mathbb{P}_X$ in the hypothesis class $\mathscr{F}$ is S-persistent for some $S \in [1, \infty)$. Specifically, for every $\xi \geq 0$ and every $f \in \mathscr{F}$, we have that

$$\mathbb{E}\left[\exp\left(-\xi \sum_{t=0}^{T-1} \|f(X_t)\|_2^2\right)\right] \leq \exp\left(-\xi \sum_{t=0}^{T-1} \mathbb{E}\left[\|f(X_t)\|_2^2\right] + \frac{\xi^2 S}{2} \sum_{t=0}^{T-1} \mathbb{E}\left[\|f(X_t)\|_2^4\right]\right).$$

Typically, S is expressed in terms of the *dependence matrix* of X (see Appendix A.2 for its definition), and such a parameter attains higher values the more dependent X_t is on its past. In general, S might depend on T; however, in this paper we will focus on the case in which S is a constant: as pointed out in (Samson, 2000, Section 2), this is a rather weak condition satisfied by a large class of Markov chains and of ϕ-mixing processes — see Appendix A.2 for more details.

Contribution. Our results demonstrate that the physics-informed regularization in the empirical risk minimization problem (3.3) can speed-up the learning even in presence of dependent data. In particular, we derive complexity-dependent bounds for the excess risk $||\hat{f} - f_\star||^2_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})}$, both in probability and in expectation, for learning under mixing, and prove that the rate of the excess risk matches the one from i.i.d learning in presence of knowledge alignment. Therefore, our results theoretically quantify the beneficial impact of physical knowledge in learning algorithms, even in the challenging scenario of learning with dependent data.

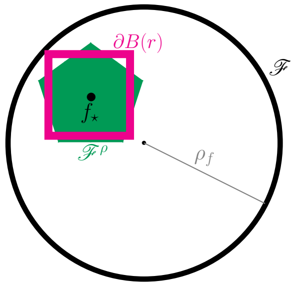

Figure 1: Visualization of the involved hypothesis spaces. Note that the set $\partial B(r) = \{f \in \mathscr{F}_\star \mid \|f\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} = r^2\}$ introduced in Section 4.1 is represented as a square to highlight the fact that the norm therein involved is different to the one defining $\mathscr{F}$ (3.6). Similarly, we represented $\mathscr{F}^\rho$ (3.7) as a convex set that is not necessarily a ball in the Sobolev norm.

4 Error bounds

We now present the bounds for the excess risk, both in probability and in expectation. We start in Section 4.1 by conveying the underlying ideas that lead to those results, and then provide the result in probability (Section 4.2) and in expectation (Section 4.3). These results will be further analyzed in Section 5 to obtain our main claims on the convergence rate of learning with physics-informed regularization. Before proceeding, we emphasize that the excess risk is a random quantity depending on the distribution of the input sequence X and of the noises $\{W_t\}_{t=0}^{T-1}$: therefore, often we will simply write $\mathbb{P}$ and $\mathbb{E}$ instead of $\mathbb{P}_{\mathbb{P}_X,W}$ and $\mathbb{E}_{\mathbb{P}_X,W}$ to streamline notation.

4.1 The idea

The main idea consists of identifying an event according to which, with high probability and for some parameter θ,

$$\|f - f_\star\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} \leq \frac{\theta}{T}\sum_{t=0}^{T-1} \|f(X_t) - f_\star(X_t)\|_2^2 . \tag{4.1}$$

This kind of one-sided concentration inequality was studied for the i.i.d. setting in Mendelson (2014), to which we defer for a thorough discussion. The proof that (4.1) holds with high probability in the i.i.d. case is given in Mendelson (2014) thanks to the *small-ball condition*, which is a rather weak assumption from a statistical point of view: see the discussion after Assumption 1.2 in Lecué & Mendelson (2017), together with its interpretation in terms of identifiability. In our data-dependent setting, the small-ball condition will be imposed by (C,α)-hypercontractivity with $\alpha = 2$ (see Appendix D.2), and we show that it holds in the set $\partial B(r) \doteq \{f \in \mathscr{F} \mid \|f - f_\star\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} = r^2\}$ for any fixed $r > 0$. Therefore, the probability level of the event in (4.1) will be controlled by the radius r. We present a visualization of $B(r)$, together with all the hypothesis spaces, in Figure 1.

Crucially, inequality (4.1) allows us to shift the analysis of the excess risk to that of its empirical version. The next step consists then in upper-bounding the latter (i.e., the right-hand side in (4.1)) by the *martingale offset complexity* of the effective hypothesis space $\mathbf{M}_T[\mathscr{F}^\rho_\star]$. In particular, for every $f \in \mathscr{F}^\rho_\star$ (i.e., $f = f' - f_\star$ for some $f' \in \mathscr{F}^\rho$), one has that

$$\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2 \leq \sup_{f\in\mathscr{F}^\rho_\star}\frac{1}{T}\sum_{t=0}^{T-1} 4\langle W_t,\, f(X_t)\rangle_2 - \|f(X_t)\|_2^2 \doteq \mathbf{M}_T[\mathscr{F}^\rho_\star] . \tag{4.2}$$

We defer to Lemma G.1 for a derivation of such an inequality. Along the lines of Liang et al. (2015), we would like to stress that the term $\|f(X_t)\|_2^2$ in the right-hand side introduces a self-normalizing effect that compensates the fluctuations of the term $\langle W_t,\, f(X_t)\rangle_2$. This fact is key in making the martingale offset complexity *not depend on mixing*, as discussed in Section 5. One can provide bounds in probability and in expectation for the martingale offset complexity (see Appendix G), and these will play a key role in the excess risk bounds that we present in the remainder of the section and further discuss in Section 5.

Before presenting the aforementioned bounds, let us formally introduce the *lower isometry event*, which is the complement of (4.1), whose probability we bound in Appendix F:

$$\mathcal{A}_r \doteq \sup_{f\in\mathscr{F}^\rho_\star\setminus B(r)}\left\{\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2 - \frac{1}{\theta}\|f\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} \leq 0\right\}.$$

4.2 Result in probability

Theorem 4.1. *Let Assumptions 1 to 3, 5 and 6 hold. Consider a parameter $\theta > 8$, and let $\hat{f}$ be the solution of the estimation problem* (3.3) *with $\lambda_T > 0$, and let the radius ρ defining the effective hypothesis class $\mathscr{F}^\rho$ be such that $\rho \geq 10\Psi(f_\star)$. Then, on the event*

$$\mathcal{A}_r^\complement \cap \left\{ \lambda_T \geq \frac{40}{3\rho} \mathbf{M}_T \left[\mathscr{F}^\rho\right] \right\}$$

we have that

$$\left\| \hat{f} - f_\star \right\|^2_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} \leq \theta \mathbf{M}_T \left[\mathscr{F}^\rho\right] + 2\lambda_T \Psi(f_\star) + r^2. \tag{4.3}$$

Proof. (Sketch). The proof follows Lecué & Mendelson (2017); Ziemann & Tu (2022) and it consists in characterizing the scenarios that lead to the event $\mathcal{A}_r^\complement$, showing that the case for which $\hat{f} \in \mathscr{F} \setminus \mathscr{F}^\rho$ cannot occur for λ_T sufficiently large. The detailed proof is given in Appendix H.1. □

4.3 Result in expectation

Theorem 4.2. *Let Assumptions 1 to 3, 5 and 6 hold. Consider a parameter $\theta > 8$, a radius $r > 0$, and let $\mathscr{F}_r$ be a $r/\sqrt{\theta}$-cover in the infinity norm of $\partial B(r)$ that is $(C(r), 2)$-hypercontractive. Consider the regularized empirical risk minimization problem in* (3.3) *with regularization parameter satisfying $\lambda_T \geq \frac{40}{3\rho} \mathbb{E}_W \left[\mathbf{M}_T \left[\mathscr{F}^\rho\right]\right]$, where $\rho \geq 10\Psi(f_\star)$. Then, letting B be the positive constant such that $\|f\|_{\mathscr{L}^\infty(\Omega^T; \mathbb{R}^{d_Y})} \leq B$ for all $f \in \mathscr{F}$, the estimate $\hat{f}$ satisfies*

$$\mathbb{E}\left[\left\| \hat{f} - f_\star \right\|^2_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} \right] \leq 4B^2 \mathcal{N}_\infty \left(\partial B(r), \frac{r}{\sqrt{\theta}} \right) \exp\left\{ -\frac{8T}{\theta^2 C_r S} \right\} + \theta \mathbb{E}\left[\mathbf{M}_T \left[\mathscr{F}^\rho\right]\right] + \lambda_T \Psi(f_\star) + r^2.$$

Proof. (Sketch). The idea consists in decomposing the expected value according to the lower-isometry event $\mathcal{A}_r$ and its complement: informally, we would write $\mathbb{E}\left[\text{excess risk}\right] = \mathbb{E}\left[\text{excess risk} \cap \mathcal{A}_r\right] + \mathbb{E}\left[\text{excess risk} \cap \mathcal{A}_r^\complement\right]$. The first term would then be bounded thanks to S-persistence, $(C, 2)$-hypercontractivity and B-boundedness, which allow us to quantify the probability of the lower-isometry event $\mathcal{A}_r$ (see Appendix F). The bound for the second term is derived along the lines of the proof of Theorem 4.1. The full details are presented in Appendix H.2. □

Overall, our analysis deploys the concepts of S-persistence and (C, α)-hypercontractivity to adapt the small-ball argument of Mendelson (2014) to the data-dependent case. Thanks to this construction, we can identify the lower-isometry event, which enables the derivation of our bounds depending on the martingale offset complexity, the ground-truth regularizer $\Psi(f_\star)$ and the critical radius r. In the next section, we will characterize the behavior of these terms to obtain the desired convergence rates for physics-informed learning.

5 Convergence rates

We finally provide our main results in terms of convergence rates for the excess risk, whose detailed proofs are deferred to Appendix I. Throughout this section, we will denote by $d = {}^{2s}/_{2s+d_X}$ the Sobolev minimax rate, and $d' = {}^{2d_X}/_{2s+d_X}$.

5.1 Bound in probability

Theorem 5.1. *Let Assumptions 1 to 6 hold, and let $\hat{f}$ be the solution of* (3.3)*. Fix a probability of failure $\delta \in (0, 1)$, and assume the regularization parameter λ_T satisfies*

$$\lambda_T \geq \frac{4}{3T^d} \left[\frac{C_I \sigma_W^{1+d}}{\Psi(f_\star)^{1-\frac{d'}{4}}} + \frac{(C_{II} + C_{IV})\sigma_W^{2d}}{\Psi(f_\star)^{1-\frac{d'}{2}}} + \frac{C_{III} \sigma_W^2 \log(1/\delta)}{\Psi(f_\star)} \right],$$

where C_I, C_{II}, C_{III} and C_{IV} are constants depending only on s, d_X, d_Y and $\sqrt{\log(1/\delta)}$. If the number of samples T satisfies

$$T \geq \frac{\theta^2 C_h S}{8}\left[C_M\left(\frac{1}{r}\right)^{\frac{6d_X}{2s-d_X}}\log\left(1+C_L\left(\frac{1}{r}\right)^{\frac{4s-d_X}{2s-d_X}}\right)+\left(\frac{1}{r}\right)^{\frac{4d_X}{2s-d_X}}\log(1/\delta)\right]$$

for $r^2 = \lambda_T\Psi(f_\star) + \sigma_W^2/T$ and C_h, C_M, C_L being uniform constants depending on $\rho_f, \overline{\kappa}, \theta, s, d_X$ and Ω, then, with probability at least $1-6\delta$, the excess risk enjoys the following convergence rate:

$$\left\|\hat{f} - f_\star\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} \leq C_{slow}\frac{\max\left\{\Psi(f_\star)^{d'/4}, \Psi(f_\star)^{d'/2}\right\}}{T^d} + C_{fast}\frac{\sigma_W^2\log(1/\delta)}{T},$$

where C_{slow} is a constant that depends on $s, d_X, d_Y, \sigma_W^2, \sqrt{\log(1/\delta)}$, and C_{fast} is a constant that depends on s, d_X, d_Y.

Proof. (Sketch). The result builds upon the bound in probability on the excess risk of Theorem 4.1, and its crux consists in conveniently setting the values for the critical radius r, the radius ρ of the effective hypothesis class $\mathscr{F}^\rho$, and the regularization parameter λ_T. This allows us to rewrite the excess risk bound (4.3) in terms of the martingale offset complexity, which can in turn be bounded according to (Ziemann, 2022, Theorem 4.2.2) (see Theorem G.2 for its detailed proof). Finally, the characterization of the burn-in follows from the probability of the lower-isometry event. The full proof is reported in Appendix I.1, where the value of all of the involved constants is given. □

5.2 Bound in expectation

Theorem 5.2. *Let Assumptions 1 to 6 hold, and let $\hat{f}$ be the solution of* (3.3) *with regularization parameter λ_T satisfying*

$$\lambda_T \geq \frac{4(C_I + C_{II})(\sigma_W^2)^d}{3T\Psi(f_\star)^{1-\frac{d'}{2}}},$$

where C_I and C_{II} are constants depending only on s, d_X and d_Y. If T satisfies

$$T \geq \frac{\theta^2 C_h S}{8}\left(\frac{1}{r}\right)^{\frac{4d_X}{2s-d_X}}\left[C_M\left(\frac{1}{r}\right)^{\frac{2d_X}{2s-d_X}}\log\left(4B^2\left(1+C_L\left(\frac{1}{r}\right)^{\frac{4s-d_X}{2s-d_X}}\right)\right)+\log\left(\frac{\sigma_W^2}{T}\right)\right],$$

where B is such that $\|f\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} \leq B$ for all $f \in \mathscr{F}$ and C_M, C_h, C_L are constants depending on $\rho_f, \overline{\kappa}, \theta, s, d_X$ and Ω, then the excess risk enjoys the following convergence rate:

$$\mathbb{E}\left[\left\|\hat{f} - f_\star\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}\right] \leq C_{slow}\frac{\Psi(f_\star)^{d'/2}}{T^d} + C_{fast}\frac{\sigma_W^2\log(1/\delta)}{T},$$

where C_{slow} and C_{fast} are constants that depend on s, d_X, d_Y and σ_W^2.

Proof. (Sketch). Similarly to Theorem 5.1, one starts from Theorem 4.2 to set the values for ρ and λ_T, and then deploys the bound on the expected martingale offset complexity of (Ziemann, 2022, Theorem 3.2.1) (see Theorem G.3 for its detailed proof). Ultimately, the claim is obtained by suitably choosing the critical radius r and accordingly characterizing the lower-isometry event probability, leading to the expression for the burn-in. The detailed proof can be found in Appendix I.2. □

Notably, our analysis allows us to transfer the contribution of data dependence from the excess risk bound to the burn-in time condition. Moreover, our bounds feature a fast, i.i.d.-like term ($\mathcal{O}(T^{-1})$) and a slower Sobolev rate term ($\mathcal{O}(T^{-d})$) that becomes annihilated when $\Psi(f_\star) \simeq 0$: this proves that, under knowledge alignment, the learning rate speeds up to $\mathcal{O}(T^{-1})$ even if data are dependent.

6 Related work and discussion

General statistical learning framework. The general theory of statistical learning rates has developed along two main streams, as identified by Fischer & Steinwart (2020). The first relies on the spectral analysis of integral operators in reproducing kernel Hilbert spaces (Smale & Zhou, 2007;

Caponnetto & De Vito, 2007; Steinwart et al., 2009), while the second builds on empirical process techniques and the small-ball method (Mendelson, 2014; 2018; Lecué & Mendelson, 2017). Our work belongs to the latter stream, adapting the small-ball method to the *dependent-data* case along the lines of the localization analysis of Ziemann & Tu (2022).

Learning rates for dependent data. A common approach to handle dependence is through *blocking* techniques (Yu, 1994; Sancetta, 2021), where the trajectory is divided into blocks of length k so that consecutive blocks can be treated as independent. However, this deflates the effective sample size, leading to suboptimal rates. Similar rates appear also in Steinwart & Christmann (2009); Zou et al. (2009); Agarwal & Duchi (2012); Kuznetsov & Mohri (2017), and Nagaraj et al. (2020) shows that such a deflation in a worst-case agnostic model set-up is unavoidable. To contrast this phenomenon, a significant line of work has studied learning under dependent data *without regularization*. In the linear setting, Simchowitz et al. (2018) and Nagaraj et al. (2020) established sample complexity bounds for system identification and stochastic gradient descent. Moreover, Roy et al. (2021) extended the small-ball method to dependent processes, but without using one-sided concentration, leading to slower rates. Similar slower-rate phenomena also appear in Ziemann et al. (2022). More recently, Ziemann & Tu (2022) proposed an adaptation of the small-ball method and offset complexity technique of Liang et al. (2015) to obtain optimal rates for nonlinear settings. Our work builds upon this line of thought, extending the analysis to *physics-informed regularization*. However, the results in this paper are not a mere adaptation: the physics-informed regularizer introduces additional challenges, such as characterizing the entropy numbers of the effective hypothesis class (e.g., under ellipticity, non-trivial nullspaces of the operator, and boundary conditions), determining trajectory hypercontractivity and working with weighted, vector-valued Sobolev spaces.

Theoretical analysis of physics-informed machine learning. Our work belongs to the branch of physics-informed machine learning that aims at enhancing learning algorithms with available physical knowledge — a class of models also known as *hybrid modeling* (Rai & Sahu, 2020; von Rueden et al., 2023b). To the best of the authors' knowledge, results aimed at quantifying the beneficial impact of physical priors in learning algorithms are von Rueden et al. (2023a) and Doumèche et al. (2024). The present paper is very similar in spirit to the latter work in the way complexity-dependent rates are derived, but crucially deals with non-i.i.d. data and presents bounds for the excess risk not only just in expectation, but also in probability. We further summarize related work in Table 1.

Table 1: Comparison of convergence rates for non-parametric regression with and without regularization. The rate from Ziemann & Tu (2022) follows from its Corollary 4.1 with $q = d_X/s$ under the metric entropy bound $\log \mathcal{N}_\infty(\mathscr{F}, \varepsilon) \sim (1/\varepsilon)^q$. The rate from Lecué & Mendelson (2017) follows from its Lemma 2.1 assuming $r^2(\rho) \sim \sigma_W^2 T^{-1}$, with $\lambda_T \sim T^{-d}$.

Work	Hypothesis class	Data	Regularization	Assumption	Rate
Nussbaum (2006)	$\mathscr{L}^2$ Sobolev space	i.i.d.	✗	σ_W^2-Gaussian, $d_X = 1$	$\sigma_W^2 T^{-2s/(2s+1)}$
Farahmand & Szepesvári (2012)	General Sobolev space	non-i.i.d.	✗	Exponential mixing, $d_Y = 1$	$T^{-d}\log(T)$
Lecué & Mendelson (2017)	General	i.i.d.	Proper regularizer	σ_W^2-sub-Gaussian, $d_Y = 1$	$\Psi(f_\star)T^{-d} + \sigma_W^2 T^{-1}$
Ziemann & Tu (2022)	General (not too large)	non.i.i.d.	✗	σ_W^2-sub-Gaussian	$\sigma_W^2 T^{-d}$
Doumèche et al. (2024)	Periodic Sobolev space	i.i.d.	Physics-informed	σ_W^2-sub-Gamma, $d_Y = 1$	$\Psi(f_\star)T^{-d} + \sigma_W^2 T^{-1}$
Our work	$\mathscr{L}^2$ Sobolev space	non-i.i.d.	Physics-informed	σ_W^2-sub-Gaussian, $s \geq 2d_X$	$\Psi(f_\star)^{d'/2}T^{-d} + \sigma_W^2 T^{-1}$

Quantifying the impact of knowledge alignment. We now showcase the impact of knowledge alignment $\Psi(f_\star) \simeq 0$ in contrast with the rates of empirical risk minimization *without regularization* — i.e., considering $\hat{f}'$ as the solution of (3.3) when $\lambda_T = 0$. As shown in detail in Appendix J, the excess risk for $\hat{f}'$ behaves, both in probability and in expectation, in the following way (informally):

$$\text{(Excess risk)} \quad ||\hat{f}' - f_\star||^2_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} \leq \frac{C'_{\texttt{slow}}}{T^d} + C'_{\texttt{fast}} \frac{\sigma_W^2}{T}.$$

We can notice how, for the result without regularization, the term decaying according to the Sobolev rate is not modulated by any design parameter (as happened with $\Psi(f_\star)$ in Theorems 5.1 and 5.2), and is thus the dominant term dictating the slow Sobolev convergence rate of the excess risk.

On the behavior of λ_T. It is worth emphasizing that, in both the expectation and probability analyses, the condition on the regularization parameter depends on $1/\Psi(f_\star)^\beta$ for some $\beta > 1$.

This condition reflects the well-known regularization-complexity trade-off: as the hypothesis class is restricted (i.e., as ρ becomes small), one must increase λ_T to compensate for the reduced richness of the class and the potentially higher sensitivity to noise or variance, as discussed in (Lecué & Mendelson, 2017, Section 2) and also displayed in (Doumèche et al., 2024, Theorem 5.3). Even if such a phenomenon prevents us from considering the case $\Psi(f_\star) = 0$, our bounds still capture the (practical) annihilation of the Sobolev rate term in presence of knowledge alignment. Finally, as pointed out in Doumèche et al. (2024), even if λ_T depends on the unknown $\Psi(f_\star)$, it can still be estimated via, e.g., cross-validation (Wahba, 1990).

On the burn-in condition and the Sobolev order s. In Theorems 5.1 and 5.2, the burn-in time scales as $(1/r)^{6d_X/2s-d_X}$, and r in turn scales as $T^{-1/2}$. Therefore, to ensure well-posedness of the burn-in time condition, we have to impose that $3d_X/2s-d_X \leq 1$, which yields Assumption 2. Thus, our results come at the price of a stronger requirement on s with respect to the standard $s \geq d_X/2$ needed, e.g., for the Sobolev imbedding theorem (Appendix B).

Numerical experiment. We complement our theoretical analysis with an example showcasing the benefit of prior domain knowledge in learning a nonlinear dynamical system. In this experiment, whose full details can be found in Appendix K, we consider the dynamics of a unicycle robot described by the differential equations $\dot{x}_1(t) = \nu(t)\cos\vartheta(t), \quad \dot{x}_2(t) = \nu(t)\sin\vartheta(t), \quad \dot{\vartheta}(t) = \omega(t)$, where $(x_1, x_2) \in \mathbb{R}^2$ is the position of the robot on the plane, $\vartheta \in [0, \pi/2]$ is the orientation angle, and (ν, ω) are the translational and angular velocities, respectively. The physical information we want to incorporate is that the velocity has no lateral component, enforcing the non-slip behavior of the unicycle kinematics. Such a constraint is embedded in the learning problem (3.3) as a (discretized) $\mathscr{L}^2$-regularization term, and we perform estimation by deploying a multilayer perceptron with two hidden layers featuring 64 nodes and ReLU activation functions.

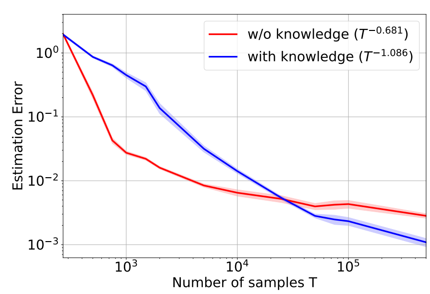

Figure 2: Log-log plot of the empirical excess risk (estimation error) with respect to the number of samples T for the unicycle dynamics after the burn-in period. Each curve is obtained by averaging over 20 independent random realizations of the training data, with solid lines indicating the mean estimation error and shaded regions denoting 95% confidence intervals.

The experiment, whose results are displayed in Figure 2, compares the empirical rates obtained with and without physics-informed regularization. We can notice that both estimators eventually return an accurate model for the ground-truth dynamics. However, without physics knowledge the rate of decay of the estimation error is relatively slow, with an empirical slope of approximately $\mathcal{O}(T^{-0.681})$. In contrast, incorporating physics-informed regularization yields a markedly faster decay, with an empirical slope of approximately $\mathcal{O}(T^{-1.086})$, as the model is explicitly constrained by the domain knowledge that unicycle dynamics do not admit lateral velocity. This experiment demonstrates how embedding physics-based operators into the training objective leads to provable improvements in sample efficiency, consistent with our theoretical trends predicted in Section 5 – especially the result in expectation presented in Theorem 5.2.

7 Conclusions

This work focused on vector-valued function estimation from dependent data, and studied the excess risk of the estimate $\hat{f}$ obtained through regularized empirical risk minimization, where regularization is induced by physical knowledge (namely, that the unknown function approximately satisfies a partial differential equation). The main message of this work is that knowledge alignment (i.e., the regularizer is approximately zero when evaluated at the ground-truth function $f_\star$) allows to speed up the learning rate from the slow, Sobolev rate $\mathcal{O}(T^{-d})$, with $d = 2s/2s+d_X < 1$, to the fast, optimal i.i.d. one $\mathcal{O}(T^{-1})$. Taken together, our results provide the first convergence rates for physics-informed learning under dependent data that avoid the sample-size deflation inherent to blocking techniques, and reveal a transition from Sobolev minimax rates to fast i.i.d.-optimal rates through knowledge alignment. This bridges classical statistical learning theory, physics-informed regularization, and learning with dependent data.

REFERENCES

Robert Adams and John Fournier. *Sobolev Spaces*. Academic Press, 2003.

Alekh Agarwal and John C. Duchi. The Generalization Ability of Online Algorithms for Dependent Data, June 2012. URL http://arxiv.org/abs/1110.2529. arXiv:1110.2529 [stat].

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US, Boston, MA, 2004. ISBN 978-1-4613-4792-7 978-1-4419-9096-9. doi: 10.1007/978-1-4419-9096-9. URL http://link.springer.com/10.1007/978-1-4419-9096-9.

Patrick Billingsley. *Probability and Measure*. John Wiley and Sons, anniversary edition, 2012.

Vladimir I. Bogachev and Oleg G. Smolyanov. The Fourier Transform and Sobolev Spaces. In Vladimir I. Bogachev and Oleg G. Smolyanov (eds.), *Real and Functional Analysis*, pp. 397–432. Springer International Publishing, Cham, 2020. ISBN 978-3-030-38219-3. doi: 10.1007/978-3-030-38219-3_9. URL https://doi.org/10.1007/978-3-030-38219-3_9.

Adam Bowers and Nigel J. Kalton. *An Introductory Course in Functional Analysis*. Universitext. Springer, New York, NY, 2014. ISBN 978-1-4939-1944-4 978-1-4939-1945-1. doi: 10.1007/978-1-4939-1945-1. URL https://link.springer.com/10.1007/978-1-4939-1945-1. ISSN: 0172-5939, 2191-6675.

Richard C. Bradley. Basic Properties of Strong Mixing Conditions. In Ernst Eberlein and Murad S. Taqqu (eds.), *Dependence in Probability and Statistics: A Survey of Recent Results*, pp. 165–192. Birkhäuser, Boston, MA, 1986. ISBN 978-1-4615-8162-8. doi: 10.1007/978-1-4615-8162-8_8. URL https://doi.org/10.1007/978-1-4615-8162-8_8.

Richard C. Bradley. Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. *Probability Surveys*, 2(none), January 2005. ISSN 1549-5787. doi: 10.1214/154957805100000104. URL http://arxiv.org/abs/math/0511078. arXiv:math/0511078.

Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, New York, NY, 2011. ISBN 978-0-387-70913-0 978-0-387-70914-7. doi: 10.1007/978-0-387-70914-7. URL https://link.springer.com/10.1007/978-0-387-70914-7.

Robert Bush and Frederick Mosteller. A Stochastic Model with Applications to Learning. *The Annals of Mathematical Statistics*, 1953. URL https://www.jstor.org/stable/2236781?seq=1.

A. Caponnetto and E. De Vito. Optimal Rates for the Regularized Least-Squares Algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, July 2007. ISSN 1615-3383. doi: 10.1007/s10208-006-0196-8. URL https://doi.org/10.1007/s10208-006-0196-8.

Seng-Kee Chua. On Weighted Sobolev Spaces. *Canadian Journal of Mathematics*, 48(3):527–541, June 1996. ISSN 0008-414X, 1496-4279. doi: 10.4153/CJM-1996-027-5. URL https://www.cambridge.org/core/journals/canadian-journal-of-mathematics/article/on-weighted-sobolev-spaces/4EB5795BBCA448EBC767B7E05BF6D187.

John B. Conway. *A Course in Functional Analysis*. Graduate Texts in Mathematics. Springer, New York, NY, 2007. ISBN 978-1-4419-3092-7 978-1-4757-4383-8. doi: 10.1007/978-1-4757-4383-8. URL http://link.springer.com/10.1007/978-1-4757-4383-8. ISSN: 0072-5285.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2002.

Felipe Cucker and Ding Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2007. doi: 10.1017/CBO9780511618796.

Victor De La Pena and Evarist Gine. *Decoupling: From Dependence to Independence*. Probability and its Applications. Springer, New York, NY, 1999. ISBN 978-1-4612-6808-6 978-1-4612-0537-1. doi: 10.1007/978-1-4612-0537-1. URL http://link.springer.com/10.1007/978-1-4612-0537-1.

P. H. Diananda and M. S. Bartlett. Some probability limit theorems with statistical applications. *Mathematical Proceedings of the Cambridge Philosophical Society*, 49(2):239–246, April 1953. ISSN 1469-8064, 0305-0041. doi: 10.1017/S0305004100028334. URL https://www.cambridge.org/core/journals/mathematical-proceedings-of-the-cambridge-philosophical-society/article/some-probability-limit-theorems-with-statistical-applications/3FD6E7D20E03C8FD10B877CD9ADB3B1F.

Paul Doukhan. *Mixing*, volume 85 of *Lecture Notes in Statistics*. Springer, New York, NY, 1994. ISBN 978-0-387-94214-8 978-1-4612-2642-0. doi: 10.1007/978-1-4612-2642-0. URL http://link.springer.com/10.1007/978-1-4612-2642-0.

Nathan Doumèche, Francis Bach, Gérard Biau, and Claire Boyer. Physics-informed machine learning as a kernel method. In *Proceedings of Thirty Seventh Conference on Learning Theory*, pp. 1399–1450. PMLR, June 2024. URL https://proceedings.mlr.press/v247/doumeche24a.html. ISSN: 2640-3498.

Jan Drgona, Truong X. Nghiem, Thomas Beckers, Mahyar Fazlyab, Enrique Mallada, Colin Jones, Draguna Vrabie, Steven L. Brunton, and Rolf Findeisen. Safe Physics-Informed Machine Learning for Dynamics and Control, June 2025. URL http://arxiv.org/abs/2504.12952. arXiv:2504.12952 [eess].

D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 1996. ISBN 978-0-521-56036-8. doi: 10.1017/CBO9780511662201. URL https://www.cambridge.org/core/books/function-spaces-entropy-numbers-differential-operators/386A287CACFD61C15A8C1021A5A9E6CD.

Lawrence C. Evans. *Partial Differential Equations*. American Mathematical Soc., 2010. ISBN 978-0-8218-4974-3. Google-Books-ID: Xnu0o_EJrCQC.

Amir-massoud Farahmand and Csaba Szepesvári. Regularized least-squares regression: Learning from a β-mixing sequence. *Journal of Statistical Planning and Inference*, 142(2):493–505, February 2012. ISSN 0378-3758. doi: 10.1016/j.jspi.2011.08.007. URL https://www.sciencedirect.com/science/article/pii/S0378375811003181.

Douglas Farenick. *Fundamentals of Functional Analysis*. Universitext. Springer International Publishing, Cham, 2016. ISBN 978-3-319-45631-7 978-3-319-45633-1. doi: 10.1007/978-3-319-45633-1. URL http://link.springer.com/10.1007/978-3-319-45633-1. ISSN: 0172-5939, 2191-6675.

Simon Fischer and Ingo Steinwart. Sobolev Norm Learning Rates for Regularized Least-Squares Algorithms. *Journal of Machine Learning Research*, 2020.

V. Gol'dshtein and A. Ukhlov. Weighted Sobolev spaces and embedding theorems, September 2007. URL http://arxiv.org/abs/math/0703725. arXiv:math/0703725.

Pierre Grisvard. *Elliptic Problems in Nonsmooth Domains*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, January 2011. ISBN 978-1-61197-202-3. doi: 10.1137/1.9781611972030. URL https://epubs.siam.org/doi/book/10.1137/1.9781611972030.

Ying Guo, P.L. Bartlett, J. Shawe-Taylor, and R.C. Williamson. Covering numbers for support vector machines. *IEEE Transactions on Information Theory*, 48(1):239–250, January 2002. ISSN 1557-9654. doi: 10.1109/18.971752. URL https://ieeexplore.ieee.org/document/971752.

Joachim Gwinner and Ernst Peter Stephan. A Fourier Series Approach. In Joachim Gwinner and Ernst Peter Stephan (eds.), *Advanced Boundary Element Methods: Treatment of Boundary Value, Transmission and Contact Problems*, pp. 43–62. Springer International Publishing, Cham, 2018. ISBN 978-3-319-92001-6. doi: 10.1007/978-3-319-92001-6_3. URL https://doi.org/10.1007/978-3-319-92001-6_3.

Paul R. Halmos. *Measure Theory*. Graduate Texts in Mathematics. Springer, New York, NY, 1950. ISBN 978-1-4684-9442-6 978-1-4684-9440-2. doi: 10.1007/978-1-4684-9440-2. URL http://link.springer.com/10.1007/978-1-4684-9440-2. ISSN: 0072-5285, 2197-5612.

T. E. Harris. On chains of infinite order. *Pacific Journal of Mathematics*, 5(S1):707–724, January 1955. ISSN 0030-8730. URL https://projecteuclid.org/journals/pacific-journal-of-mathematics/volume-5/issue-S1/On-chains-of-infinite-order/pjm/1171984831.full. Publisher: Pacific Journal of Mathematics, A Non-profit Corporation.

I. A. Ibragimov. Some Limit Theorems for Stationary Processes. *Theory of Probability & Its Applications*, 7(4):349–382, January 1962. ISSN 0040-585X. doi: 10.1137/1107036. URL https://epubs.siam.org/doi/abs/10.1137/1107036. Publisher: Society for Industrial and Applied Mathematics.

I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation*. Springer, New York, NY, 1981. ISBN 978-1-4899-0029-6 978-1-4899-0027-2. doi: 10.1007/978-1-4899-0027-2. URL http://link.springer.com/10.1007/978-1-4899-0027-2.

Jr Iorio, Rafael José and Valéria de Magalhães Iorio. *Fourier Analysis and Partial Differential Equations*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 2001. ISBN 978-0-521-62116-8. doi: 10.1017/CBO9780511623745. URL https://www.cambridge.org/core/books/fourier-analysis-and-partial-differential-equations/39312A08B4D4F25F65F39581D229285B.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, June 2021. ISSN 2522-5820. doi: 10.1038/s42254-021-00314-5. URL https://www.nature.com/articles/s42254-021-00314-5. Publisher: Nature Publishing Group.

Tero Kilpelainen. Weighted Sobolev spaces and capacity. *Annales Academiæ Scientiarum Fennicæ*, 19:95–113, 1994.

K. Krishnamoorthy. *Handbook of Statistical Distributions with Applications*. Chapman and Hall/CRC, New York, 2 edition, January 2016. ISBN 978-0-429-15581-9. doi: 10.1201/b19191.

Alois Kufner. *Weighted Sobolev Spaces*. Wiley, July 1985. ISBN 978-0-471-90367-3. Google-Books-ID: V1mqAAAAIAAJ.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, January 2017. ISSN 1573-0565. doi: 10.1007/s10994-016-5588-2. URL https://doi.org/10.1007/s10994-016-5588-2.

John Lamperti and Patrick Suppes. Chains of infinite order and their application to learning theory. *Pacific Journal of Mathematics*, 9(3):739–754, January 1959. ISSN 0030-8730. URL https://projecteuclid.org/journals/pacific-journal-of-mathematics/volume-9/issue-3/Chains-of-infinite-order-and-their-application-to-learning-theory/pjm/1103039115.full. Publisher: Pacific Journal of Mathematics, A Non-profit Corporation.

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method II: complexity dependent error rates. *Journal of Machine Learning Research*, 18(146):1–48, 2017. ISSN 1533-7928. URL http://jmlr.org/papers/v18/16-422.html.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with Square Loss: Localization through Offset Rademacher Complexity, June 2015. URL http://arxiv.org/abs/1502.06134. arXiv:1502.06134 [stat].

K. Marton. A measure concentration inequality for contracting markov chains. *Geometric & Functional Analysis GAFA*, 6(3):556–571, May 1996. ISSN 1420-8970. doi: 10.1007/BF02249263. URL https://doi.org/10.1007/BF02249263.

Shahar Mendelson. Learning without concentration. In *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pp. 25–39, Barcelona, Spain, 2014. PMLR.

Shahar Mendelson. Learning without concentration for general loss functions. *Probability Theory and Related Fields*, 171(1):459–502, June 2018. ISSN 1432-2064. doi: 10.1007/s00440-017-0784-y. URL https://doi.org/10.1007/s00440-017-0784-y.

Chuizheng Meng, Sam Griesemer, Defu Cao, Sungyong Seo, and Yan Liu. When physics meets machine learning: a survey of physics-informed machine learning. *Machine Learning for Computational Science and Engineering*, 1(1):20, May 2025. ISSN 3005-1436. doi: 10.1007/s44379-025-00016-0. URL https://doi.org/10.1007/s44379-025-00016-0.

Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2 edition, 2009. ISBN 978-0-521-73182-9. doi: 10.1017/CBO9780511626630. URL https://www.cambridge.org/core/books/markov-chains-and-stochastic-stability/E2B82BFB409CD2F7D67AFC5390C565EC.

Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least Squares Regression with Markovian Data: Fundamental Limits and Algorithms. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16666–16676. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c22abfa379f38b5b0411bc11fa9bf92f-Abstract.html.

Truong X. Nghiem, Ján Drgoňa, Colin Jones, Zoltan Nagy, Roland Schwan, Biswadip Dey, Ankush Chakrabarty, Stefano Di Cairano, Joel A. Paulson, and Andrea Carron. Physics-informed machine learning for modeling and control of dynamical systems. In *2023 American Control Conference (ACC)*, pp. 3735–3750. IEEE, 2023. URL https://ieeexplore.ieee.org/abstract/document/10155901/.

Michael Nussbaum. Minimax Risk, Pinsker Bound for. In *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Ltd, 2006. ISBN 978-0-471-66719-3. doi: 10.1002/0471667196.ess1098.pub2. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/0471667196.ess1098.pub2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471667196.ess1098.pub2.

Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods, November 2018. URL http://arxiv.org/abs/1212.2015. arXiv:1212.2015 [math].

Johanna Penteker. Sobolev Spaces. Lecture Notes, Institute of Analysis, Johannes Kepler University Linz, 2015.

Rahul Rai and Chandan K. Sahu. Driven by Data or Derived Through Physics? A Review of Hybrid Physics Guided Machine Learning Techniques With Cyber-Physical System (CPS) Focus. *IEEE Access*, 8:71050–71073, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.2987324. URL https://ieeexplore.ieee.org/document/9064519/.

Michael Renardy and Robert Rogers. *An Introduction to Partial Differential Equations*, volume 13 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 2004. ISBN 978-0-387-00444-0. doi: 10.1007/b97427. URL http://link.springer.com/10.1007/b97427.

L. C. G. Rogers and David Williams. *Diffusions, Markov Processes, and Martingales: Volume 1: Foundations*, volume 1 of *Cambridge Mathematical Library*. Cambridge University Press, Cambridge, 2 edition, 2000. ISBN 978-0-521-77594-6. doi: 10.1017/CBO9781107590120. URL https://www.cambridge.org/core/books/diffusions-markov-processes-and-martingales/188B6A2BAABAF735E61796C3CD18114B.

Abhishek Roy, Krishnakumar Balasubramanian, and Murat A. Erdogdu. On Empirical Risk Minimization with Dependent and Heavy-Tailed Data, September 2021. URL http://arxiv.org/abs/2109.02224. arXiv:2109.02224 [math].

Julien Royer. A brief introduction to Sobolev spaces and applications, 2020. URL https://www.math.univ-toulouse.fr/~jroyer/TD/2020-21-M1/M1-Ch5.pdf.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 1976. ISBN 978-0-07-085613-4.

Paul-Marie Samson. Concentration of measure inequalities for Markov chains and Φ-mixing processes. *The Annals of Probability*, 28(1):416–461, January 2000. ISSN 0091-1798, 2168-894X. doi: 10.1214/aop/1019160125. URL https://projecteuclid.org/journals/annals-of-probability/volume-28/issue-1/Concentration-of-measure-inequalities-for-Markov-chains-and-Phi-mixing/10.1214/aop/1019160125.full. Publisher: Institute of Mathematical Statistics.

Alessio Sancetta. Estimation in Reproducing Kernel Hilbert Spaces With Dependent Data. *IEEE Transactions on Information Theory*, 67(3):1782–1795, March 2021. ISSN 1557-9654. doi: 10.1109/TIT.2020.3045290. URL https://ieeexplore.ieee.org/document/9296271/?arnumber=9296271. Conference Name: IEEE Transactions on Information Theory.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I. Jordan, and Benjamin Recht. Learning Without Mixing: Towards A Sharp Analysis of Linear System Identification. In *Proceedings of the 31st Conference On Learning Theory*, pp. 439–473. PMLR, July 2018. URL https://proceedings.mlr.press/v75/simchowitz18a.html. ISSN: 2640-3498.

Steve Smale and Ding-Xuan Zhou. Learning Theory Estimates via Integral Operators and Their Approximations. *Constructive Approximation*, 26(2):153–172, August 2007. ISSN 1432-0940. doi: 10.1007/s00365-006-0659-y. URL https://doi.org/10.1007/s00365-006-0659-y.

Elias M. Stein. *Singular Integrals and Differentiability Properties of Functions*. Princeton University Press, 1970. ISBN 978-1-4008-8388-2. doi: 10.1515/9781400883882. URL https://www.degruyterbrill.com/document/doi/10.1515/9781400883882/html.

Ingo Steinwart and Andreas Christmann. Fast Learning from Non-i.i.d. Observations. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://papers.nips.cc/paper/2009/hash/a89cf525e1d9f04d16ce31165e139a4b-Abstract.html.

Ingo Steinwart, D. Hush, and C. Scovel. Optimal Rates for Regularized Least Squares Regression. 2009. URL https://www.semanticscholar.org/paper/Optimal-Rates-for-Regularized-Least-Squares-Steinwart-Hush/1dc0f2c3068eb4b56a7208b0cd3e42f8b79e5660.

Michel Talagrand. *The Generic Chaining*. Springer Monographs in Mathematics. Springer-Verlag, Berlin/Heidelberg, 2005. ISBN 978-3-540-24518-6. doi: 10.1007/3-540-27499-5. URL http://link.springer.com/10.1007/3-540-27499-5.

Michael E. Taylor. *Partial Differential Equations I: Basic Theory*, volume 115 of *Applied Mathematical Sciences*. Springer International Publishing, Cham, 2023. ISBN 978-3-031-33858-8 978-3-031-33859-5. doi: 10.1007/978-3-031-33859-5. URL https://link.springer.com/10.1007/978-3-031-33859-5.

Roger Temam. *Navier-Stokes Equations and Nonlinear Functional Analysis*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, January 1995. ISBN 978-0-89871-340-4. doi: 10.1137/1.9781611970050. URL https://epubs.siam.org/doi/book/10.1137/1.9781611970050.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2024.

Laura von Rueden, Jochen Garcke, and Christian Bauckhage. How Does Knowledge Injection Help in Informed Machine Learning? In *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, June 2023a. doi: 10.1109/IJCNN54540.2023.10191994. URL https://ieeexplore.ieee.org/document/10191994/?arnumber=10191994. ISSN: 2161-4407.

Laura von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, Michal Walczak, Jochen Garcke, Christian Bauckhage, and Jannis Schuecker. Informed Machine Learning – A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):614–633, January 2023b. ISSN 1558-2191. doi: 10.1109/TKDE.2021.3079836. URL https://ieeexplore.ieee.org/stampPDF/getPDF.jsp?arnumber=9429985. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Grace Wahba. *Spline Models for Observational Data*. SIAM, September 1990. ISBN 978-0-89871-244-5.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2019. ISBN 978-1-108-49802-9. doi: 10.1017/9781108627771. URL https://www.cambridge.org/core/books/highdimensional-statistics/8A91ECEEC38F46DAB53E9FF8757C7A4E.

Jianjun Wang, Hua Huang, Zhangtao Luo, and Baili Chen. Estimation of Covering Number in Learning Theory. In *2009 Fifth International Conference on Semantics, Knowledge and Grid*, pp. 388–391, October 2009. doi: 10.1109/SKG.2009.27. URL https://ieeexplore.ieee.org/document/5370097/.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, October 1999. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1017939142. URL https://projecteuclid.org/journals/annals-of-statistics/volume-27/issue-5/Information-theoretic-determination-of-minimax-rates-of-convergence/10.1214/aos/1017939142.full. Publisher: Institute of Mathematical Statistics.

Bin Yu. Rates of Convergence for Empirical Processes of Stationary Mixing Sequences. *The Annals of Probability*, 22(1):94–116, 1994. ISSN 0091-1798. URL https://www.jstor.org/stable/2244496. Publisher: Institute of Mathematical Statistics.

Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, September 2002. ISSN 0885-064X. doi: 10.1006/jcom.2002.0635. URL https://www.sciencedirect.com/science/article/pii/S0885064X02906357.

Ding-Xuan Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49(7):1743–1752, July 2003. ISSN 1557-9654. doi: 10.1109/TIT.2003.813564. URL https://ieeexplore.ieee.org/document/1207372. Conference Name: IEEE Transactions on Information Theory.

Ingvar Ziemann. *Statistical Learning, Dynamics and Control : Fast Rates and Fundamental Limits for Square Loss*. PhD thesis, KTH Royal Institute of Technology, 2022. URL https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-320345. Publisher: KTH Royal Institute of Technology.

Ingvar Ziemann and Stephen Tu. Learning with little mixing. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*. Curran Associates, Inc., 2022. doi: 10.48550/arXiv.2206.08269. URL http://arxiv.org/abs/2206.08269. arXiv:2206.08269 [cs].

Ingvar Ziemann, Henrik Sandberg, and Nikolai Matni. Single Trajectory Nonparametric Learning of Nonlinear Dynamics, February 2022. URL http://arxiv.org/abs/2202.08311. arXiv:2202.08311 [cs].

Bin Zou, Luoqing Li, and Zongben Xu. The generalization performance of ERM algorithm with strongly mixing observations. *Machine Learning*, 75(3):275–295, June 2009. ISSN 1573-0565. doi: 10.1007/s10994-009-5104-z. URL https://doi.org/10.1007/s10994-009-5104-z.

Erhan Çinlar. *Probability and Stochastics*. Springer International Publishing, New York, 2011.