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

Anna Scampicchio* [*]
Department of Electrical Engineering
Chalmers University of Technology, Sweden
anna.scampicchio@chalmers.se

Leonardo F. Toso*
Department of Electrical Engineering
Columbia University, USA
leonardo.toso@columbia.edu

Rahel Rickenbach
Institute for Dynamic Systems and Control
ETH Zürich, Switzerland
rrahel@ethz.ch

James Anderson
Department of Electrical Engineering
Columbia University , USA
james.anderson@columbia.edu

Melanie N. Zeilinger
Institute for Dynamic Systems and Control
ETH Zürich, Switzerland
mzeilinger@ethz.ch

ABSTRACT

A major challenge in physics-informed machine learning is to understand how the incorporation of prior domain knowledge affects learning rates when training data is not independent and identically distributed. 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 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) leveraging data-driven methods to enhance numerical solvability and accuracy of physical models given, for instance, by systems of partial differential equations; (ii) improve learning algorithm performance by including physical information in the loss function or as an 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, which play an important role especially in view of downstream tasks such as safe learning-based control (Nghiem et al., 2023; Drgona et al., 2025). However, despite the empirical success of physics-informed machine learning algorithms (Raissi et al., 2019; Rai & Sahu, 2020; Cuomo et al., 2022; Hao et al., 2023), theoretically quantifying the beneficial impact of incorporating physical information into learning algorithms is technically challenging and still an active area of research (see von Rueden et al. (2023a) and references therein).

We tackle this problem by considering a statistical learning set-up and focusing on *regularized empirical risk minimization* problems of the following form:

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

[*] Corresponding author. A. Scampicchio and L.F. Toso share first-authorship of this paper.

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* $||\hat{f} - f_\star||^2_{\mathscr{L}^2}$.Informally, our results (both in high probability and expectation) take the form given below:

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

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

The error bound established above reveals for the first time 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 with dependent data* 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). For 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 Section 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, truncated Gaussian, and 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 $\|f\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} := \langle f,\, f\rangle_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}$. Moreover, 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 Section 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. Our data consists of trajectories of length T, denoted , $\mathcal{D} \doteq \{X_t, Y_t\}_{t=0}^{T-1}$, 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, $\mathbb{E}_{W_t}\left[W_t|\mathcal{X}_{t-1}\right] = 0$ 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 $(\mathbb{R}^{d_Y}, \|\cdot\|_2)$,

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

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*:

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

Physics-informed regularization. The regularizer $\Psi(\cdot)\colon \mathscr{F} \to \mathbb{R}_{\geq 0}$ encodes available prior physical information about the "true" function $f_\star$ — in other words, $\Psi(f)$ is designed to penalize the physical inconsistency between f and 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:

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 Section 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 make the following standing assumption:

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

In case such an assumption is violated, the excess risk bound would feature an additional term quantifying the error introduced by the mis-specification of the hypothesis space $\mathscr{F}$. Such a bias error is entirely data-independent and deterministic — therefore, it is decoupled from the analysis performed in this paper. In general, tools to bound the bias error rely on *interpolation theory* (Lunardi, 2018), and some results can be found, e.g., in Cucker & Smale (2002); Cucker & Zhou (2007).

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 Section 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 Section 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.

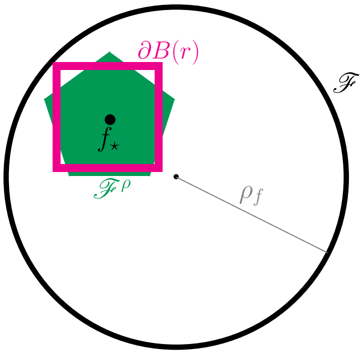

Figure 1: Visualization of the involved hypothesis spaces. Note that the set $\partial B(r) = \{f \in \mathscr{F}_\star \mid \|f\|^2_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} = r^2\}$ introduced in Section 4.1 is represented as a square to highlight the fact that the norm therein involved 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 bound 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 Section 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 $\partial B(r)$, 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, $\mathbf{M}_T[\mathscr{F}_\star^\rho]$. In particular, for every $f \in \mathscr{F}_\star^\rho$ (i.e., $f = f' - f_\star$ for some $f' \in \mathscr{F}^\rho$), one has that

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

We defer to Theorem 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 Section 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 Section F:

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

We are now in place to present the bounds for the excess risk, both in probability and in expectation.

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*

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

we have that

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

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 Section H.1. □

4.3 Result in expectation

Theorem 4.2. *Let Assumptions 1 to 3, 5 and 6 hold. Having the set $\partial B(r)$ that is $(C(r), 2)$-hypercontractive, let $\mathscr{F}_r$ be a $r/\sqrt{\theta}$-cover in the infinity norm of $\partial B(r)$, and let $\mathcal{N}_\infty\left(\partial B(r), \frac{r}{\sqrt{\theta}}\right)$ be the associated covering number. 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 $f \in \mathscr{F}$, we have*

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

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 Section 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 Section 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 $||\hat{f} - f_\star||^2_{\mathscr{L}^2((\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y}))}$, whose detailed proofs are deferred to Section I. Building upon Theorems 4.1 and 4.2, we also obtain sufficient conditions on the parameter λ_T and on the minimal sample size T for the bounds to hold. 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:

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

where $C_{\mathtt{slow}}$ is a constant that depends on $s, d_X, d_Y, \sigma_W^2, \sqrt{\log(1/\delta)}$, and $C_{\mathtt{fast}}$ 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 Section I.1, where the values of all of the involved constants are 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:

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

where $C_{\texttt{slow}}$ and $C_{\texttt{fast}}$ 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 Section 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. Statistical learning theory (Vapnik, 1998; Cucker & Smale, 2002) offers powerful tools to analyze the theoretical performance of nonparametric learning algorithms. Within such a framework, two main streams to derive learning rates have been developed, 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; Cuomo et al., 2022; von Rueden et al., 2023b; Hao et al., 2023). 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 Section J, the excess risk for $\hat{f}'$ behaves, both in probability and in expectation, in the following way (informally):

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

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 in practice 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 (Section B).

Numerical illustration. 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 Section K, we consider the dynamics of a unicycle robot described by the differential equations

$$\begin{cases} \dot{x}_1(t) = \nu(t) \cos \vartheta(t), \\ \dot{x}_2(t) = \nu(t) \sin \vartheta(t), \\ \dot{\vartheta}(t) = \omega(t), \end{cases}$$

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.

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 $\mathcal{O}(T^{-0.646})$. In contrast, incorporating physics-informed regularization yields a markedly faster decay, with an empirical slope of approximately $\mathcal{O}(T^{-0.993})$, 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 the theory in Section 5 – especially the result in expectation presented in Theorem 5.2.

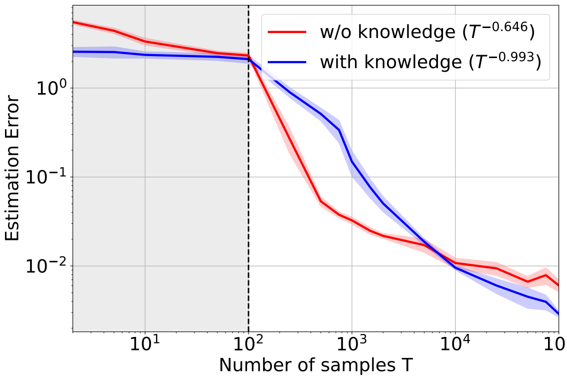

Figure 2: Log-log plot of the empirical excess risk (estimation error) with respect to the number of samples T for the unicycle dynamics. Each curve is obtained by averaging over 20 independent random realizations of the training data, with solid lines indicating the mean estimation error and shaded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

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 analysis is set in the general framework of statistical learning, and applies to a wide range of approaches such as kernel-based methods and (deep) neural networks with physics-informed loss functions. 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.

Acknowledgments

Anna Scampicchio acknowledges that this work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Leonardo F. Toso thanks Ingvar Ziemann for an instructive discussion during the early stage of this work. Leonardo F. Toso is funded by the Center for AI and Responsible Financial Innovation (CAIRFI) Fellowship and by the Columbia Presidential Fellowship. James Anderson is partially funded by NSF grants ECCS 2144634 and 2231350.

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.

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, 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.

Andrea Caponnetto and Ernesto 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.

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.

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):88, July 2022. ISSN 1573-7691. doi: 10.1007/s10915-022-01939-z. URL https://doi.org/10.1007/s10915-022-01939-z.

Victor De La Peña and Evarist Giné. *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.

Palahenedi H. Diananda and Maurice 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.

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].

David E. Edmunds and Hans 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.

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

Vladimir Gol'dshtein and Alexander Ukhlov. Weighted Sobolev spaces and embedding theorems. *Transactions of the American Mathematical Society*, 361, 2009. URL http://arxiv.org/abs/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.

Zhongkai Hao, Songming Liu, Yichi Zhang, Chengyang Ying, Yao Feng, Hang Su, and Jun Zhu. Physics-Informed Machine Learning: A Survey on Problems, Methods and Applications, March 2023. URL http://arxiv.org/abs/2211.08064. arXiv:2211.08064 [cs].

Theodore 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.

Ildar 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.

Ildar A. Ibragimov and Rafail 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.

Rafael José Jr Iorio 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.

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

Kalimuthu 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.

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.

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. In *JMLR: Workshop and Conference Proceedings*, volume 1, June 2015. doi: 10.48550/arXiv.1502.06134. URL http://arxiv.org/abs/1502.06134. arXiv:1502.06134 [stat].

Alessandra Lunardi. *Interpolation Theory*. Scuola Normale Superiore, Pisa, 2018. ISBN 978-88-7642-639-1 978-88-7642-638-4. doi: 10.1007/978-88-7642-638-4. URL http://link.springer.com/10.1007/978-88-7642-638-4.

Katalin 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. URL https://onlinelibrary.wiley.com/doi/abs/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/.

M. Raissi, P. Perdikaris, and G. 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, February 2019. ISSN 0021-9991. doi: 10.1016/j.jcp.2018.10.045. URL https://www.sciencedirect.com/science/article/pii/S0021999118307125.

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.

Chris 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.

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.

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, 2005. ISBN 978-3-540-24518-6. 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.

Vladimir N. Vapnik. *Statistical Learning Theory*. Adaptive and Cognitive Dynamic Systems: Signal Processing, Learning, Communications and Control. Wiley edition, 1998. URL https://www.wiley.com/en-us/Statistical+Learning+Theory-p-9780471030034.

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.

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/ann als-of-statistics/volume-27/issue-5/Information-theoretic-determination-of-minimax-rates-of-convergence/10.1214/aos/1017939142.full.

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.

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.

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. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, February 2022. doi: 10.48550/arXiv.2202.08311. 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.

• Technical appendix •

In the following sections we provide the derivations of all of the results stated in the paper. This technical appendix is structured as follows:

Section A provides the necessary results for the probability set-up of the estimation problem. Specifically, we construct the marginal probability measures stated in Assumption 1 (Section A.1); we discuss the meaning of the S-persistence given in Assumption 6 and its relation to data dependence (Section A.2); and derive some useful properties of sub-Gaussian random vectors that will be useful in the martingale offset complexity bounds (Section A.3).

Section B constructs the auxiliary set-up of weighted and vector-valued Sobolev spaces needed to define the hypothesis spaces for the empirical risk minimization problem (3.3). First, we extend the Sobolev imbedding theorem to the weighted, vector-valued case (Section B.1). Next, we present definitions and key results of periodic Sobolev spaces (Section B.2), and show that the Sobolev space of interest, $\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})$ is imbedded in one of them. Such a construction will be leveraged in the proof of $(C(r), 2)$-hyper-contractivity of $\partial B(r)$ in Section D.2.

Section C focuses on bounds for covering numbers of convex sets of vector-valued Sobolev spaces. By deploying the direct-sum structure of the Sobolev space $\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})$ elucidated in Section 2.2, we first show how the covering of the multi-dimensional set can be obtained from the covering of the one-dimensional counterpart (Section C.1). We then use such a result to extend classic results on covering numbers, culminating in the bound for the covering number of the effective hypothesis space $\mathscr{F}^\rho$ (Section C.2).

Section D shows key properties of the hypothesis spaces $\mathscr{F}$ and $\mathscr{F}^\rho$, such as convexity and B-boundedness (Section D.1), and $(C(r), 2)$-hypercontractivity of $\partial B(r)$ (Section D.2). We also show how the trajectory hypercontractivity condition enforces the small-ball property.

Section E focuses on the physics-informed regularizer and its properties. In particular, we show that the physics-informed regularizer is 2-proper (Section E.1), and prove an inequality on the difference $\Psi(\hat{f}) - \Psi(f_\star)$ that will be useful in the proofs of Theorems 4.1 and 4.2 (Section E.2).

Section F provides the bound for the probability of the lower-isometry event $\mathcal{A}_r$ introduced in Section 4.1. We first show an ancillary inequality linking hypercontractivity and S-persistence (Section F.1) and then derive the main lower-isometry bound result (Section F.2), also presenting its corollary that will be useful in characterizing the burn-ins in Section 5.

Section G provides the full derivation of the bounds for the martingale offset complexity, which play a key role in the results of Sections 4 and 5. We first show the inequality that underpins the definition of martingale offset complexity (Section G.1), and then prove its bounds, both in probability (Section G.2) and in expectation (Section G.3).

Section H contains the proofs of the excess bound rates, namely of Theorem 4.1 (Section H.1) and of Theorem 4.2 (Section H.2).

Section I collects the proof of the convergence rates results of Section 5, specifically of Theorem 5.1 (Section I.1) and of Theorem 5.2 (Section I.2).

Section J provides the corollaries of the results given in Section 5 dealing with empirical risk minimization without regularization, which we use in the comparison performed in Section 6. Specifically, we derive the result in probability (Section J.1) and in expectation (Section J.2).

Section K presents the full details of the numerical experiment set-up outlined in the discussion reported in Section 6 (Section K.1), together with an additional experiment involving the Poisson equation (Section K.2).

A Probability measure and stochastics set-up

This section collects all the ancillary results concerning the probability space $(\Omega^T, \{\mathcal{X}_t\}_{t=0}^{T-1}, \mathbb{P}_X)$, the inter-sample dependence in trajectories X belonging to such a space, and the noise sequence we are considering. In particular, in Section A.1 we specify the marginal distributions presented in Assumption 1; then, in Section A.2 we discuss the S-persistence condition in Assumption 6, showing how S quantifies the degree of dependence between data samples separated in time; finally, in Section A.3 we present some useful ancillary results on the second statistical moment of the sub-Gaussian random vectors given in Assumption 3.

A.1 On the construction of probability measures

We now characterize the probability measures μ_t, defined for each $t = 0, \dots, T-1$, associated with each term of the input trajectory X.

The classic set-up involves *independent* samples. In this situation, the σ-algebra on Ω^T is given by the tensor product of the single σ-algebras $\mathcal{X}_t$. Moreover, by construction each component X_t of the trajectory X has a distribution μ_t, and the resulting probability space is $(\Omega^T, \otimes_{t=0}^{T-1}\mathcal{X}_t, \prod_{t=0}^{T-1}\mu_t)$ — see, e.g., (Halmos, 1950, Chapter VII) and (Billingsley, 2012, Section 18).

We now detail the case with *dependent data* building upon the results in (Çinlar, 2011, Chapter I.6). We are in the situation in which the transition between X_{t-1} and X_t for all $t = 1, \dots, T$ is described by a map from $(\Omega, \mathcal{X}_{t-1})$ to $(\Omega, \mathcal{X}_t)$. Such a map is called *transition kernel* $\mathcal{K}_t(\cdot,\cdot) : \Omega \times \mathcal{X}_t \to \mathbb{R}_{\geq 0}$ and is such that $x_{t-1} \mapsto \mathcal{K}_t(x_t, A)$ is $\mathcal{X}_{t-1}$-measurable for every set $A \in \mathcal{X}_t$, and $A \mapsto \mathcal{K}_t(x_{t-1}, A)$ is a measure on $(\Omega, \mathcal{X}_t)$ for every $x_{t-1} \in \Omega$. Before proving the main result in Theorem A.3, we recall two key results:

Lemma A.1. *Let $(E, \mathcal{E})$ be a measurable space, and let L be a functional mapping the space of non-negative measurable functions defined on $\mathcal{E}$ to $\mathbb{R}_{\geq 0}$. Then there exists a unique measure ν on $(E, \mathcal{E})$ such that $L(g) = \nu g$ for any function g in the domain of L if and only if*

1. *$g = 0$ implies $L(g) = 0$;*
2. *for any $\mathfrak{a}, \mathfrak{b} \in \mathbb{R}_{\geq 0}$ and any g, g' in the domain of L, $L(\mathfrak{a}g + \mathfrak{b}g') = \mathfrak{a}L(g) + \mathfrak{b}L(g')$;*
3. *for any increasing sequence $\{g_n\}_n \nearrow g$ we have that $L(g_n) \nearrow L(g)$.*

Proof. We defer the interested reader to (Çinlar, 2011, Theorem 4.21). □

Lemma A.2. *Let $\mathcal{K}_\tau(\cdot,\cdot)$ be a transition kernel from $(\Omega, \mathcal{X}_{\tau-1})$ to $(\Omega, \mathcal{X}_\tau)$, and $\mathcal{K}_{\tau+1}(\cdot,\cdot)$ be a transition kernel from $(\Omega, \mathcal{X}_\tau)$ to $(\Omega, \mathcal{X}_{\tau+1})$. Then, their product is the transition kernel $\mathcal{K}_\tau\mathcal{K}_{\tau+1}$ from $(\Omega, \mathcal{X}_{\tau-1})$ to $(\Omega, \mathcal{X}_{\tau+1})$ such that*

$$\mathcal{K}_\tau\mathcal{K}_{\tau+1}(x_{\tau-1}, A) = \int_\Omega \mathcal{K}_\tau(x_{\tau-1}, dx_\tau)\,\mathcal{K}_{\tau+1}(x_\tau, A) \quad \textit{for } x_{\tau-1} \in \Omega, A \in \mathcal{X}_{\tau+1}.$$

Proof. It follows directly from Theorem A.1. □

We are now ready to state the existence and uniqueness of the probability measures associated with each component X_t of the input trajectory X.

Theorem A.3. *Let $g \colon \Omega \to \mathbb{R}_{\geq 0}$, and assume that there exists a probability measure μ_0 associated with the first component of the input trajectory X. Then, for each $t = 1, \dots, T-1$ there exists a unique probability measure such that*

$$\int_{\Omega^T} g(X_t) d\mathbb{P}_X = \int_\Omega g(X_t) d\mu_t.$$

Proof. We are considering

$$\int_{\Omega^T} g(X_t) d\mathbb{P}_X = \int_\Omega \mu_0(dX_0) \cdots \int_\Omega \mathcal{K}_t(X_{t-1}, dX_t)\, g(X_t) \cdots \int_\Omega \mathcal{K}_T(X_{T-1}, dX_T). \quad \text{(A.1)}$$

By an iterative application of Theorem A.2 to Equation (A.1), the contribution of the transition kernels $\mathcal{K}_{t+1}(\cdot,\cdot), \cdots \mathcal{K}_T(\cdot,\cdot)$ integrates to 1. Furthermore, we can apply again Theorem A.2 to the kernels $\mathcal{K}_1(\cdot,\cdot), \cdots, \mathcal{K}_t(\cdot,\cdot)$ and obtain the composed kernel $\bar{\mathcal{K}}(\cdot,\cdot)$ such that (A.1) can be re-written as

$$\int_{\Omega^T} g(X_t) d\mathbb{P}_X = \int_\Omega \mu_0(dX_0) \int_X \bar{\mathcal{K}}(X_0, dX_t)\, g(X_t).$$

It can be shown, along the lines of (Çinlar, 2011, Theorem 6.3), that the right-hand side of the equation above satisfies (A.1), thus proving the claim. □

Note that this theorem holds also for the independent-measures case, where each μ_t is the t-th marginal of $\mathbb{P}_X$ and can be computed relying on Fubini's Theorem.

A.2 On S-persistence and data dependence

We now focus on the concept of S-persistence (see Assumption 6) and on how it relates to the dependence of the samples in the trajectory X.
We first start by recalling the definition of S-persistence. The tuple $(\mathscr{F}, \mathbb{P}_X)$ is S-persistent if, for every $\xi \geq 0$ and $f \in \mathscr{F}$, we have that

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

The parameter S is related to the "degree of dependence" of the samples within the trajectory X: this is typically quantified in terms of the *dependence matrix* (also defined *mixing matrix* in Paulin (2018)), which we now define.

Let $\mathcal{X}_{i:j}$ be the σ-field generated by the truncated input sequence $\{X_t\}_{t=i}^j$, which we represent as $X_{i:j}$. Additionally, denote with $\|\cdot\|_{TV}$ the total variation norm between two probability measures on $\mathcal{X}_{i:j}$: specifically, such a metric is defined as $\|\nu_1 - \nu_2\|_{TV} \doteq \sup_{A \in \mathcal{X}_{i:j}} |\nu_1(A) - \nu_2(A)|$ for any couple of measures ν_1 and ν_2 defined on the σ-algebra $\mathcal{X}_{i:j}$. The dependence matrix $\Gamma(\mathbb{P}_X)$ is a lower-triangular matrix whose (i,j) element, for $i, j = 1, \cdots, T$, is given by

$$[\Gamma(\mathbb{P}_X)]_{i,j} \doteq \begin{cases} \sqrt{2 \sup_{A \in \mathcal{X}_{0:T-i}} \left\|\mathbb{P}_{X_{j:T-1}}(\cdot|A) - \mathbb{P}_{X_{j:T-1}}\right\|_{TV}} & (i < j) \\ 1 & (i = j) \\ 0 & (i > j) \end{cases} \tag{A.2}$$

Such a matrix provides a measure of dependence through its induced 2-norm, which is $\|\Gamma(\mathbb{P}_X)\|_2 = \arg\inf_{a>0}\{\|\Gamma(\mathbb{P}_X)v\|_2 \leq a\|v\|_2,\, v \in \mathbb{R}^T\}$: thus, we have that S-persistence is ruled by $S = \|\Gamma(\mathbb{P}_X)v\|_2$. If the stochastic process has independent samples, then it holds that $\|\Gamma(\mathbb{P}_X)\|_2 = 1$; on the other hand, if the process is fully dependent, then we have that $\|\Gamma(\mathbb{P}_X)\|_2$ grows linearly in T. We will focus on scenarios in which $\|\Gamma(\mathbb{P}_X)\|_2$ is a constant: as elucidated in (Samson, 2000, Section II), these are the following.

(a) *uniformly ergodic Markov chains*. In these Markov chains, the transition kernels $\mathcal{K}_t(\cdot,\cdot) = \mathcal{K}(\cdot,\cdot)$ are time-homogeneous, and there exists an invariant distribution $\widetilde{\pi}$ such that, for every initial condition x, there exists a rate $\mathfrak{r} < 1$ and a constant $\mathfrak{A} > 0$ such that $\left\|\mathcal{K}(x,\cdot)^t - \widetilde{\pi}(\cdot)\right\|_{TV} \leq \mathfrak{A}\mathfrak{r}^t$ (Meyn & Tweedie, 2009, Section 16.2.1). Another characterization of uniformly ergodic Markov chains is given by the Doeblin condition, and in (Doukhan, 1994, Section 2.4, Theorem 1) uniform ergodicity is proven also for non-homogeneous kernels. In general, Markov chains satisfying uniform ergodicity are given, for instance, by linear and stable auto-regressive models $X_{t+1} = FX_t + W_t$: indeed, these are T-chains (Meyn & Tweedie, 2009, Proposition 6.3.5), and the latter that are uniformly ergodic (Meyn & Tweedie, 2009, Theorem 16.2.5). Another notable example is given by nonlinear state-space models $X_{t+1} = F_t(X_t, U_0, \cdots, U_t, W_k)$ of the form presented in (Meyn & Tweedie, 2009, Section 2.2.2) with control model $F_t(U_0, \cdots, U_t)$ that is stable in the sense of Lagrange (Meyn & Tweedie, 2009, Section 16.2.3).

(b) *contracting Markov chains*. A weaker condition imposed on the behavior of the transition kernels is given by Marton (1996), where a weaker form of Doeblin's condition states that $\sup_{x_1,x_2\in\Omega} \|\mathcal{K}_t(x_1,) - \mathcal{K}_t(x_2,)\|_{TV} < 1$ for all t. Markov chains satisfying such a condition do not have to be time-homogeneous and are called *contracting* in Marton (1996). As argued in (Samson, 2000, Equation (2.8)), this kind of Markov chain also leads to a $\|\Gamma(\mathbb{P}_X)\|_2$ that does not depend on T.

(c) *ϕ-mixing processes.* A more general way of characterizing ergodicity without assuming the Markov property of the process is given by means of *mixing processes* reviewed, e.g., in Bradley (1986; 2005) and (Doukhan, 1994, Chapter 1). Here the focus is on ϕ-mixing processes, thoroughly studied in Ibragimov (1962), which can lead to $\|\Gamma(\mathbb{P}_X)\|_2 = \mathcal{O}(1)$. These kind of processes characterize mixing through the measure $\Phi(\mathcal{X}_{0:i}, \mathcal{X}_{j:T-1}) \doteq \sup_{A\in\mathcal{X}_{0:i}} \sup_{B\in\mathcal{X}_{j:T-1}} |\mathbb{P}(B|A) - \mathbb{P}(B)|$; additionally, let $\Phi_k(i,j) \doteq \sup_{i,j=0,\ldots T-1}\{\Phi(\mathcal{X}_{0:i}, \mathcal{X}_{j:T-1}),\ j-i \geq k\}$. A process is ϕ-mixing if $\Phi_k(i,j) \to 0$ as $k\to\infty$ for all i,j: in other words, the measure $\Phi(\cdot,\cdot)$ quantifies how "independent" two non-overlapping blocks of variables in the trajectory X become as their distance increases. Examples of ϕ-mixing processes are uniformly ergodic Markov processes (Diananda & Bartlett, 1953) and chains of infinite order, which are non-Markovian processes where the transition probability is influenced only slightly by the remote past, that have been deployed to model psychology experiments (Harris, 1955; Bush & Mosteller, 1953; Lamperti & Suppes, 1959).
Noting how Φ enters the definition of the dependence matrix (A.2), $\|\Gamma(\mathbb{P}_X)\|_2$ can be characterized by $\Phi_k \doteq \sup_{i,j}\Phi_k(i,j)$: if Φ_k exhibits an exponential decay, or it holds that $\sum_{k=1}^{\infty}\sqrt{\Phi_k} < \infty$, then we have that $\|\Gamma(\mathbb{P}_X)\|_2$ is a constant (Samson, 2000, p.425).

A.3 Useful properties of sub-Gaussian vectors

We now present two lemmas that will be useful in the proofs of the martingale offset complexity bounds in Section G. First, we recall that, according to Assumption 3, the noise sequence is a martingale difference sequence with sub-Gaussian noise. Recalling the definition, we then have, for every $\xi \in \mathbb{R}$ and every u in the unit sphere in $(\mathbb{R}^{d_Y}, \|\cdot\|_2)$,

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

We now present the lemma on the second moment of a sub-Gaussian random vector.

Lemma A.4. *Let W be a sub-Gaussian random vector according to Assumption 3 and let $\mathfrak{A}$ be a positive constant. Then*

$$\mathbb{E}_W\left[\exp\left\{\xi^2\|W\|_2^2\right\}\right] \leq \exp\left\{\mathfrak{A}^2\xi^2\right\} \quad \textit{for } |\xi| < \frac{1}{\mathfrak{A}}.$$

Proof. The proof for the scalar case can be found in (Vershynin, 2024, Proposition 2.5.2): i.e., when $d_Y = 1$, we obtain that

$$\mathbb{E}_W\left[\exp\left\{\xi^2 W^2\right\}\right] \leq \exp\left\{\mathfrak{A}^2\xi^2\right\} \quad \text{for } |\xi| < \frac{1}{\mathfrak{A}}.$$

To extend the claim to the case $d_Y \geq 1$, we follow the argument in (Ziemann, 2022, Lemma 6.3.4) and consider an auxiliary random vector V that is drawn uniformly over the canonical Euclidean basis of $\mathbb{R}^{d_Y}$, which we denote by $\{e_1,\cdots,e_{d_Y}\}$. With such a construction, we have then that $\mathbb{E}_V\left[VV^\top\right] = \frac{1}{d_Y}\mathbb{I}_{d_Y}$, where $\mathbb{I}_{d_Y}$ is the identity matrix in $\mathbb{R}^{d_Y}$. Then, we can consider

$$\begin{aligned}
\mathbb{E}_W\left[\exp\left\{\xi^2\|W\|_2^2\right\}\right] &= \mathbb{E}_W\left[\exp\left\{\xi^2 W^\top \mathbb{I}_{d_Y} W\right\}\right] \\
&= \mathbb{E}_W\left[\exp\left\{\xi^2 d_Y W^\top \mathbb{E}_V\left[VV^\top\right] W\right\}\right] \\
&\leq \mathbb{E}_{W,V}\left[\exp\left\{\xi^2 d_Y (V^\top W)^2\right\}\right] \quad \text{by Jensen's inequality,} \\
&\leq \exp\left\{\mathfrak{A}^2 d_Y^2 \xi^2\right\} \quad \text{for } |\xi| < \frac{1}{\mathfrak{A} d_Y},
\end{aligned}$$

by applying the result of (Vershynin, 2024, Proposition 2.5.2) on the scalar sub-Gaussian random variable $V^\top W$ conditioned on V. □

Next, we derive a similar result for the 2-norm of a sub-Gaussian random vector.

Lemma A.5. *Let W be a sub-Gaussian random vector according to Assumption 3. Then*

$$\mathbb{E}_W \left[\|W\|_2\right] \leq 3\sqrt{d_Y}\sigma_W.$$

Proof. Again, the claim for the scalar case $d_Y = 1$ can be found in (Vershynin, 2024, Proposition 2.5.2(ii)), where we have that $\mathbb{E}_W \left[|W|\right] \leq 3\sigma_W$: the value for the exact constants can be retrieved from the proof using $K_5^2 = \sigma_W^2/2$, $K_1^2 = 2\sigma_W^2$ and $K_2 = (3p)^{1/p}\sqrt{pK_1^2/2}$.

To obtain the result for the general case $d_Y \geq 1$ we proceed similarly to Theorem A.4 and work with the auxiliary random vector V that is uniformly drawn from the canonical basis of $\mathbb{R}^{d_Y}$, resulting in

$$\mathbb{E}_W\left[\sqrt{W^\top W}\right] = \mathbb{E}_W\left[\sqrt{d_Y W^\top \mathbb{E}_V\left[VV^\top\right] W}\right] = \sqrt{d_Y}\mathbb{E}_{W,V}\left[|V^\top W|\right] \leq 3\sqrt{d_Y}\sigma_W.$$

□

B Sobolev spaces

We now provide the essential information on Sobolev spaces needed in this paper. For a deeper treatment on the subject, we defer to the monograph Adams & Fournier (2003), as well as (Evans, 2010, Chapter 5), Brezis (2011), (Grisvard, 2011, Chapter 1), (Renardy & Rogers, 2004, Chapter 7), (Taylor, 2023, Chapter 4). References focused on weighted Sobolev spaces are Kufner (1985); Kilpelainen (1994); Chua (1996); Gol'dshtein & Ukhlov (2009). Finally, results of vector-valued Banach spaces given by direct sums of scalar spaces can be found in (Conway, 2007, Chapter I.6), (Farenick, 2016, Proposition 5.81), (Bowers & Kalton, 2014, Section 3.7).

B.1 Imbedding properties

We start with the well-known Sobolev imbedding theorem.

Preliminary definitions. A normed space $(\mathcal{H}_a, \|\cdot\|_{\mathcal{H}_a})$ is *imbedded* in another normed space $(\mathcal{H}_b, \|\cdot\|_{\mathcal{H}_b})$ if $\mathcal{H}_a \subseteq \mathcal{H}_b$, and for any $u \in \mathcal{H}_a$ there exists a constant $\mathfrak{C}_b$ such that $\|u\|_{\mathcal{H}_b} \leq \mathfrak{C}_b\|u\|_{\mathcal{H}_a}$. We denote such an imbedding by $\mathcal{H}_a \hookrightarrow \mathcal{H}_b$.

We now define the Banach space of differentiable functions that have continuous partial derivatives up to some order $j \in \mathbb{Z}_{\geq 0}$ for our vector-valued set-up. We will consider the space $\mathscr{C}^j\left(\Omega^T;\mathbb{R}^{d_Y}\right)$ describing functions $f\colon \Omega \to \mathbb{R}^{d_Y}$ evaluated along each component of the input trajectory X. We endow such a space with the norm

$$\|f\|_{\mathscr{C}^j\left(\Omega^T;\mathbb{R}^{d_Y}\right)} \doteq \sum_{i=1}^{d_Y}\sum_{|\alpha|\leq j}\frac{1}{T}\sum_{t=0}^{T-1}\sup_{X_t\in\Omega}|D^\alpha f_i(X_t)| = \sum_{i=1}^{d_Y}\sum_{|\alpha|\leq j}\sup_{X_t\in\Omega}|D^\alpha f_i(X_t)|.$$

Main result. We are now ready to prove the Sobolev imbedding theorem for the vector-valued Sobolev spaces introduced in Section 2.2.

Theorem B.1 (Sobolev imbedding)**.** *Let Assumption 1 and Assumption 2 hold, and let j be a non-negative integer such that Assumption 2 can be written as $s = \lceil \frac{d_X}{2} \rceil + j$. Then*

(a) $\mathscr{H}^s(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y}) \hookrightarrow \mathscr{C}^j(\Omega^T;\mathbb{R}^{d_Y})$;

(b) $\mathscr{H}^s(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y}) \hookrightarrow \mathscr{L}^q(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})$ *for $q \geq 2$.*

Proof. Let us first define the order j. We can write Assumption 2 as $s = 2d_X + j'$ for some $j' \in \mathbb{Z}_{\geq 0}$. Then it has to hold that $j = 2d_X - \lceil \frac{d_X}{2} \rceil + j' \geq 2d_X - \lceil \frac{d_X}{2} \rceil$.

After noting that (i) with our definition of $\mathscr{L}^2$, we can consider each component of the input trajectory separately; (ii) Assumption 1 ensures that the weighted Sobolev spaces are equivalent to the standard ones (Kufner, 1985); and (iii) the structure of the multi-output function spaces is given in terms of direct sums of scalar ones, we can then consider the claim for $\mathscr{H}^s(\Omega, \mu_t; \mathbb{R}^{d_Y})$ and apply (Adams & Fournier, 2003, Theorem 4.12), to which we defer for a complete proof.

The imbedding theorem in Adams & Fournier (2003) is stated under the assumption that the input domain satisfies the "cone condition" (Adams & Fournier, 2003, Definition 4.6). We now argue that our set-up satisfies it. Since our input domain Ω is bounded with locally Lipschitz boundary, we have that it satisfies the *strong local Lipschitz condition* (Adams & Fournier, 2003, Definition 4.9). Then, from (Adams & Fournier, 2003, Paragraph 4.11) we have that such a property implies the *uniform cone condition* (Adams & Fournier, 2003, Definition 4.8) (see also (Grisvard, 2011, Theorem 1.2.2.2)), which in turn implies that the cone condition holds. □

As a corollary of the imbedding theorem, one can obtain the following result:

Corollary B.2 (Berlinet & Thomas-Agnan (2004), Theorem 121)**.** *Under Assumption 1 and Assumption 2, $\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})$ is a Reproducing Kernel Hilbert space.*

B.2 PERIODIC SOBOLEV SPACE $\mathscr{H}^s_{\text{PER}}$

We will now characterize another kind of Sobolev space that will be useful for further analysis. Let us denote $Q_L \doteq [-L, L]^{d_X}$ the hyper-cube in $\mathbb{R}^{d_X}$ that contains the input domain Ω; similarly, we will use $Q_{2L} \doteq [-2L, 2L]^{d_X}$ and $Q_{4L} \doteq [-4L, 4L]^{d_X}$. In the following, we will define and present the main properties of the periodic Sobolev space $\mathscr{H}^s_{\text{per}}\left(Q^T_{2L}, \mu^T_\lambda; \mathbb{R}^{d_Y}\right)$, i.e., considering the Lebesgue measure on Ω instead of the marginals μ_t for each $t = 0, \cdots, T-1$, to simplify the presentation. After that, we will leverage Assumption 1 to handle the general probability measure $\mathbb{P}_X$ and derive the results for $\mathscr{H}^s_{\text{per}}\left(Q^T_{2L}, \mathbb{P}_X; \mathbb{R}^{d_Y}\right)$.
The material in this subsection presents a generalization of the material of (Doumèche et al., 2024, Appendix A). Additional references on periodic Sobolev are Temam (1995); Iorio & Iorio (2001); Berlinet & Thomas-Agnan (2004); Penteker (2015); Gwinner & Stephan (2018); Bogachev & Smolyanov (2020).

Preliminary definitions. Given a point $x = (x_1, \cdots, x_{d_X})$ and a function $f \colon Q_{2L} \to \mathbb{R}^{d_Y}$, its *periodic extension* $E_{\text{per}}(f)(x)$ is the operator mapping $\mathscr{L}^2(Q_{2L}, \mu^T_\lambda; \mathbb{R}^{d_Y})$ to $\mathscr{L}^2(Q_{4L}, \mu^T_\lambda; \mathbb{R}^{d_Y})$; considering $f = (f_1, \cdots, f_{d_Y})$, the periodic extension acts component-wise on f as $[E_{\text{per}}(f)(x)]_i \doteq f_i\left(x_1 - 4L\lfloor \frac{x_1}{4L} \rfloor, \cdots, x_{d_X}) - 4L\lfloor \frac{x_{d_X}}{4L} \rfloor\right)$. Such an operator allows us to define the *periodic Sobolev space* $\mathscr{H}^s_{\text{per}}\left(Q^T_{2L}, \mu^T_\lambda; \mathbb{R}^{d_Y}\right)$ as the space of functions such that $E_{\text{per}}(f)(\cdot)$ belongs to $\mathscr{H}^s(Q^T_{4L}, \mu^T_\lambda; \mathbb{R}^{d_Y})$. Therefore, $\mathscr{H}^s_{\text{per}}\left(Q^T_{2L}, \mu^T_\lambda; \mathbb{R}^{d_Y}\right)$ is a subspace of $\mathscr{H}^s\left(Q^T_{2L}, \mu^T_\lambda; \mathbb{R}^{d_Y}\right)$ consisting of functions whose $4L$-periodic extension is still s-times differentiable.

Fourier characterization of $\mathscr{H}^s_{\text{per}}\left(Q^T_{2L}, \mu^T_\lambda; \mathbb{R}^{d_Y}\right)$. A more practical representation of $\mathscr{H}^s_{\text{per}}\left(Q^T_{2L}, \mu^T_\lambda; \mathbb{R}^{d_Y}\right)$ is given by means of Fourier basis: indeed, for each component $f_i(\cdot)$ in $f(\cdot) = [f_1(\cdot), \cdots, f_{d_Y}(\cdot)]^\top$ there exists a unique (infinite-dimensional) vector z_i indexed by $\mathbb{Z}^{d_X}$ with components in $\mathbb{C}$ (thus, $z \in \mathbb{C}^{\mathbb{Z}^{d_X}}$) such that we have

$$f_i(x) = \sum_{k \in \mathbb{Z}^{d_X}} z_{i,k} \exp\left\{\iota \frac{\pi}{2L} \langle k, x \rangle_2\right\}, \text{ and} \tag{B.1a}$$

$$D^\alpha f_i(x) = \left(\iota \frac{\pi}{2L}\right)^{|\alpha|} \sum_{k \in \mathbb{Z}^{d_X}} z_{i,k} \exp\left\{\iota \frac{\pi}{2L} \langle k, x \rangle_2\right\} \prod_{j=1}^{d_X} k_j^{\alpha_j} \tag{B.1b}$$

for any multi-index $\alpha = (\alpha_1, \cdots, \alpha_{d_X}) \in \mathbb{Z}^{d_X}_{\geq 0}$ such that $|\alpha| \leq s$ (Doumèche et al., 2024, Proposition A.5)). Therefore, we obtain the following result.

Proposition B.3. *The periodic Sobolev space* $\mathscr{H}^s\left(Q_{2L}^T, \mu_\lambda^T; \mathbb{R}^{d_Y}\right)$ *can be characterized as*

$$\mathscr{H}^s_{per}\left(Q_{2L}^T, \mu_\lambda^T; \mathbb{R}^{d_Y}\right) = \left\{ z \in \bigoplus_{i=1}^{d_Y} \mathbb{C}^{\mathbb{Z}^{d_X}} \;\middle|\; \sum_{i=1}^{d_Y} \sum_{k \in \mathbb{Z}^{d_X}} |z_{i,k}|^2 \, \|k\|_2^{2s} < \infty, \; z_{i,-k} = z_{i,k}^* \right\}.$$

Proof. We specify the expression for the Sobolev norm: the rest of the claim follows directly by (Doumèche et al., 2024, Proposition A.5).
By simply considering the norm in $\mathscr{H}^s(Q_{2L}, \mu_\lambda^T; \mathbb{R}^{d_Y})$, we obtain by using (B.1) that

$$\|f\|^2_{\mathscr{H}^s(Q_{2L}, \mu_\lambda^T; \mathbb{R}^{d_Y})} = \sum_{i=1}^{d_Y} \sum_{k \in \mathbb{Z}^{d_X}} |z_{i,k}|^2 \sum_{|\alpha| \le s} \left(\frac{\pi}{2L}\right)^{2|\alpha|} \prod_{j=1}^{d_X} k_j^{2\alpha_j}.$$

We now show that such a norm is equivalent to $\sum_{i=1}^{d_Y} \sum_{k \in \mathbb{Z}^{d_X}} |z_{i,k}|^2 \, \|k\|_2^{2s}$. First, neglecting the sums over i and k, we can find a pair of constants $0 < \underline{\mathfrak{a}} < \overline{\mathfrak{a}}$ such that $\underline{\mathfrak{a}} \sum_{|\alpha| \le s} \prod_{j=1}^{d_X} k_j^{2\alpha_j} \le \sum_{|\alpha| \le s} \left(\frac{\pi}{2L}\right)^{2|\alpha|} \prod_{j=1}^{d_X} k_j^{2\alpha_j} \le \overline{\mathfrak{a}} \sum_{|\alpha| \le s} \prod_{j=1}^{d_X} k_j^{2\alpha_j}$, so we can focus on the term $\sum_{|\alpha| \le s} \prod_{j=1}^{d_X} k_j^{2\alpha_j}$. By an application of the multinomial theorem (see also Royer (2020)), we can find constants $0 < \underline{\mathfrak{b}} < \overline{\mathfrak{b}}$ such that $\underline{\mathfrak{b}} \sum_{|\alpha| \le s} \prod_{j=1}^{d_X} k_j^{2\alpha_j} \le (1 + \|k\|_2^2)^s \le \overline{\mathfrak{b}} \sum_{|\alpha| \le s} \prod_{j=1}^{d_X} k_j^{2\alpha_j}$. Finally, we can find another pair of non-negative constants $\underline{\mathfrak{c}} < \overline{\mathfrak{c}}$, again bounded away from zero, such that $\underline{\mathfrak{c}} \, \|k\|_2^{2s} \le (1 + \|k\|_2^2)^s \le \overline{\mathfrak{c}} \, \|k\|_2^{2s}$ (for instance, $\underline{\mathfrak{c}} = 1$ will do, and $\overline{\mathfrak{c}}$ can be found by upper-bounding the formula of the binomial theorem) – see also (Temam, 1995, Chapter 2.1). □

Finally, we also point out that the characterization of Theorem B.3 can be also more conveniently rewritten using a re-indexing of z: indeed, by (Doumèche et al., 2024, Proposition A.7), there exists a one-to-one mapping $\ell \in \mathbb{N} \mapsto k(\ell) \in \mathbb{Z}^{d_X}$ such that we can write $\phi_\ell(x) \doteq \exp\{\iota\pi \left\langle k(\ell), x \right\rangle_2 / (2L)\}$. With this, we can express each component $f_i(x)$, for $i = 1, \cdots, d_Y$, as $f_i(x) = \sum_{\ell \in \mathbb{N}} z_{i,\ell} \phi_\ell(x)$. Thus, the space is characterized as follows:

$$\mathscr{H}^s_{\text{per}}\left(Q_{2L}^T, \mu_\lambda^T; \mathbb{R}^{d_Y}\right) = \left\{ z \in \bigoplus_{i=1}^{d_Y} \mathbb{C}^{\mathbb{Z}^{d_X}} \;\middle|\; \sum_{i=1}^{d_Y} \sum_{\ell \in \mathbb{N}} |z_{i,\ell}|^2 \ell^{2s/d_X} < \infty \right\}. \tag{B.2}$$

Extension results for $\mathscr{H}^s(\Omega^T, \mu_\lambda^T; \mathbb{R}^{d_Y})$**.** Along the lines of the analysis carried out in (Doumèche et al., 2024, Proposition A.6), it is possible to show that the characterization given in Theorem B.3 holds also for $\mathscr{H}^s(\Omega^T, \mu_\lambda^T; \mathbb{R}^{d_Y})$. Indeed, one can use the Sobolev extension Theorem (Stein, 1970, Chapter VI) to linearly extend every function f in $\mathscr{H}^s(\Omega^T, \mu_\lambda^T; \mathbb{R}^{d_Y})$ to the function $\tilde{E}(f)(x)$, whose i-th component is given by $[\tilde{E}(f)(x)]_i = \sum_{k \in \mathbb{Z}^{d_X}} z_{i,k} \exp\left\{\iota \frac{\pi}{2L} \left\langle k, x \right\rangle_2\right\}$. Thus, $\tilde{E}(f)(x)$ belongs to $\mathscr{H}^s_{\text{per}}(Q_{2L}^T, \mu_\lambda^T; \mathbb{R}^{d_Y})$ and is such that $||\tilde{E}(f)(\cdot)||^2_{\mathscr{H}^s_{\text{per}}(Q_{2L}^T, \mu_\lambda^T; R^{d_Y})} \le \mathfrak{C}_{s,\Omega} \, \|f\|^2_{\mathscr{H}^s(\Omega^T, \mu_\lambda^T; R^{d_Y})}$ for some constant $\mathfrak{C}_{s,\Omega}$, yielding the imbedding $\mathscr{H}^s(\Omega^T, \mu_\lambda^T; R^{d_Y}) \hookrightarrow \mathscr{H}^s_{\text{per}}(Q_{2L}^T, \mu_\lambda^T; R^{d_Y})$. This allows us to leverage the structure of the periodic Sobolev space to derive a key property of $\mathscr{H}^s(\Omega^T, \mu_\lambda^T; R^{d_Y})$, namely the trajectory $(C, 2)$-hypercontractivity (see Section D.2).

From Lebesgue measure to $\mathbb{P}_X$**.** In the preceding paragraphs we focused on the Lebesgue measure defined on Ω. In such a set-up, it holds that $\|f\|^2_{\mathscr{L}^2(\Omega^T, \mu_\lambda^T; \mathbb{R}^{d_Y})} = \|f\|^2_{\mathscr{L}^2(\Omega, \mu_\lambda^T; \mathbb{R}^{d_Y})}$, and in such a space the functions $\exp\{\iota\pi \left\langle k, x \right\rangle_2 / (2L)\}$ are orthonormal: this enables the application of Parseval's Theorem, leading to the characterization in Theorem B.3. Such a property gets lost as soon we consider the distribution $\mathbb{P}_X$; nevertheless, thanks to Assumption 1, we obtain that $\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y}) \hookrightarrow \mathscr{H}^s(\Omega^T, \mu_\lambda^T; \mathbb{R}^{d_Y})$; therefore, when needed, analysis can be carried out in $\mathscr{H}^s(\Omega^T, \mu_\lambda^T; \mathbb{R}^{d_Y})$ and the results carry over to the space of interest $\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})$.

C On covering numbers

This section focuses on the complexity measure for the hypothesis space, which plays a crucial role in the excess risk bounds derived in this paper. After recalling the definition of covering number and metric entropy, we present how classical results, stated for scalar function spaces, extend to our vector-valued set-up. This section culminates with the derivation of the covering number of the effective hypothesis space $\mathscr{F}^\rho$.

C.1 From scalar to vector-valued hypothesis spaces

To quantify the complexity of a certain hypothesis space $\mathcal{H}$, we will resort to its *covering number*, which is defined as follows:

Definition C.1. Let $\mathcal{S}$ be a subset of a metric space $\mathcal{H}_\mathfrak{a}$ with distance function induced by its norm $\|\cdot\|_{\mathcal{H}_\mathfrak{a}}$. The *$\varepsilon$-cover* of $\mathcal{S}$ is a set $\{f^1, \cdots, f^N\} \subset \mathcal{S}$ such that, for each $f \in \mathcal{S}$, there exists some $\ell = 1, \dots, N$ such that $\left\|f - f^\ell\right\|_{\mathcal{H}_\mathfrak{a}} \leq \varepsilon$. The *$\varepsilon$-covering number*, denoted by $\mathcal{N}_\mathfrak{a}(\mathcal{S}, \varepsilon)$, is the cardinality of the smallest ε-cover.

Additionally, $\log \mathcal{N}_\mathfrak{a}(\mathcal{S}, \varepsilon)$ is called *metric entropy* of $\mathcal{S}$ at resolution ε.

There is a vast literature on bounds for covering numbers: see, e.g., (Cucker & Zhou, 2007, Chapter 5), (Wainwright, 2019, Chapter 5), as well as Zhou (2002); Guo et al. (2002); Zhou (2003); Wang et al. (2009). However, results are typically presented for *scalar* function spaces. In this section, we will adapt covering number estimates to our set-up involving multi-output function spaces.

Proposition C.1. *Let us consider the metric $\|\cdot\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})}$, and let $\overline{\mathcal{H}} = \bigoplus_{i=1}^{d_Y} \mathcal{H}$ be a subset of the Sobolev space $\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})$. Assume there exists a ε/d_Y-cover of $\mathcal{H}$ in the direct sum with the metric of $\mathscr{L}^\infty(\Omega^T; \mathbb{R})$, and let $\mathcal{N}_\infty(\mathcal{H}, \varepsilon/d_Y)$ be its covering number: then it holds that*

$$\mathcal{N}_\infty\left(\overline{\mathcal{H}}, \varepsilon\right) \leq \left(\mathcal{N}_\infty\left(\mathcal{H}, \varepsilon/d_Y\right)\right)^{d_Y}.$$

Proof. For it to be a ε-cover of $\overline{\mathcal{H}}$, it has to hold that, for every $f \in \overline{\mathcal{H}}$ there exists an element f' in the cover such that $\|f - f'\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} \leq \varepsilon$. Similarly, assume there is a $\breve{\varepsilon}$-cover of $\mathcal{H}$ such that, for any arbitrary element $h \in \mathcal{H}$, we have a function h' in the cover such that $\|h - h'\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R})} \leq \breve{\varepsilon}$. Now, by the definition of the vector-valued version of the infinity norm (see Section 2.2), we have

$$\|f - f'\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} = \sup_{x\in\Omega} \|f(x) - f'(x)\|_2 \leq \sum_{i=1}^{d_Y} \sup_{x\in\Omega} |f_i(x) - f_i'(x)| \leq d_Y \breve{\varepsilon}.$$

Thus, letting $\breve{\varepsilon} = \varepsilon/d_Y$, we obtain an ε-approximation for the covering of $\overline{\mathcal{H}}$. The claim follows by observing that, by construction of the direct sum of scalar hypothesis spaces, the cover of $\overline{\mathcal{H}}$ is given by the Cartesian product of the covers of $\mathcal{H}$. □

C.2 Covering numbers for vector-valued hypothesis spaces

We conclude this section by adapting standard covering number bounds for scalar function spaces to vector-valued ones. We start from (Cucker & Zhou, 2007, Theorem 5.3).

Proposition C.2. *Let $\overline{\mathcal{H}} = \bigoplus_{i=1}^{d_Y} \mathcal{H}$ be a finite-dimensional Banach space of dimension E, and let $\mathcal{B}_R$ be the set such that $\mathcal{B}_R = \{f \in \overline{\mathcal{H}} \mid \|f\|_{\overline{\mathcal{H}}} \leq R\}$. Then it holds that*

$$\mathcal{N}_{\overline{\mathcal{H}}}(\mathcal{B}_R, \varepsilon) \leq \left(\frac{2Rd_Y}{\varepsilon} + 1\right)^{E\cdot d_Y}.$$

Proof. This result follows by taking (Cucker & Zhou, 2007, Theorem 5.3) and extending it according to the construction in Theorem C.1. □

We now proceed by characterizing the covering number for a ball in the Sobolev space $\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})$.

Lemma C.3. *Let $\mathcal{B}_R$ be a ball of radius R in the Sobolev space $\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})$ satisfying Assumption 2, where Ω is a domain with locally-Lipschitz boundary. Then the metric entropy of $\mathcal{B}_R$ satisfies*

$$\log \mathcal{N}_\infty\left(\mathcal{B}_R, \varepsilon\right) \leq C_c' d_Y^{\frac{2s+d_X}{2s}} \left(\frac{R}{\varepsilon}\right)^{\frac{d_X}{s}}.$$

Proof. Let us start from the *scalar-valued* Sobolev space $\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R})$ and let $\textsc{b}_R$ be its ball of radius R. If the input domain Ω is smooth, then (Cucker & Smale, 2002, Chapter I, Section 6, Proposition 6) claims that, for some positive constant c, we have

$$\log \mathcal{N}_\infty\left(\textsc{b}_R, \varepsilon\right) \leq \left(\frac{cR}{\varepsilon}\right)^{\frac{d_X}{s}} + 1 \leq C_c' \left(\frac{R}{\varepsilon}\right)^{\frac{d_X}{s}} \tag{C.1}$$

for some other constant C_c' that is big enough to absorb also the contribution of the "+1" in (C.1).

Such a result relies on a bound on the entropy number of the embedding $\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}) \hookrightarrow \mathscr{L}^\infty(\Omega^T; \mathbb{R})$ given by (Edmunds & Triebel, 1996, Section 3.3) (using their notation, we are looking at $F^s_{2,q_1} \hookrightarrow F^0_{\infty,q_2}$). However, in (Edmunds & Triebel, 1996, Section 3.5), such a result is extended to non-smooth domains — specifically, to the *minimally regular* ones (Edmunds & Triebel, 1996, Section 2.5, Definition 2), which include domains with locally-Lipschitz boundary as special cases. By the way, the theorem in (Edmunds & Triebel, 1996, Section 3.5) is stated for the function spaces B^s_{pq}, but the results holds also for the spaces F^s_{pq}: see the argument presented in the proof of the Theorem in (Edmunds & Triebel, 1996, Section 3.3.2).

Thus, overall, (C.1) holds also for our choice of Ω. The proof is concluded by invoking Theorem C.1 to extend (C.1) to the vector-valued case. Specifically, we have that $\mathcal{N}_\infty\left(\mathcal{B}_R, \varepsilon\right) \leq \left(\mathcal{N}_\infty\left(\textsc{b}_R, \varepsilon/d_Y\right)\right)^{d_Y}$, and taking logarithms we obtain $\log \mathcal{N}_\infty\left(\mathcal{B}_R, \varepsilon\right) \leq d_Y \log \mathcal{N}_\infty\left(\textsc{b}_R, \varepsilon/d_Y\right)$, and substituting (C.1) leads to the final claim. □

We conclude this section by deriving a bound for the metric entropy of the effective hypothesis space $\mathscr{F}^\rho$ presented in (3.7). The idea is to leverage the structure of the differential operator (Assumption 4) and find a ball in the Sobolev space that approximates $\mathscr{F}^\rho$, and then invoke Theorem C.3 to bound its metric entropy.

Proposition C.4. *Let Assumptions 2 and 4 hold and consider the effective hypothesis space*

$$\mathscr{F}^\rho = \left\{ f \in \mathscr{F}_\star \mid \|\mathscr{D}(f)\|^2_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} \leq \rho \right\}. \tag{3.7}$$

Then, for some positive constant C_c not depending on ε and ρ, we have that

$$\log \mathcal{N}_\infty\left(\mathscr{F}^\rho, \varepsilon\right) \leq C_c d_Y^{\frac{2s+d_X}{2s}} \left(\frac{\sqrt{\rho}}{\varepsilon}\right)^{\frac{d_X}{s}}.$$

Proof. Similarly to Theorem C.3, we prove the result for scalar-valued function spaces and then invoke Theorem C.1 to obtain the claim for the vector-valued case.

We start by recalling an important property of elliptic operators derived from (Evans, 2010, Chapter 6.3, Theorem 5) that will allow us to find the Sobolev ball centered at $f_\star$ containing $\mathscr{F}^\rho$. Specifically, for some positive constant C_e and $f \in \mathscr{F}_\star$, we have that

$$\|f\|_{\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} \leq C_e \left(\|\mathscr{D}(f)\|_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} + \|f\|_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} \right). \tag{C.2}$$

Let $\mathcal{H}$ be the scalar-valued version of $\mathscr{F}^\rho$. We decompose $\mathcal{H}$ into the null-space of $\mathscr{D}$ and its orthogonal complement, obtaining $\mathcal{H} = \ker(\mathscr{D}) \oplus \ker(\mathscr{D})^\perp$. Accordingly, any $f \in \mathcal{H}$ can be written as $f = g + h$, with $g \in \ker(\mathscr{D})$ and $h \in \ker(\mathscr{D})^\perp$, and the constraint $\|\mathscr{D}(f)\|^2_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R})} \leq \rho$ reduces to $\|\mathscr{D}(h)\|^2_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R})} \leq \rho$.

We first focus on the subspace $\ker(\mathscr{D})^\perp$ and prove a preliminary result that allows us to rewrite (C.2). Specifically, we want to show that, for some positive constant C_l and any $h \in \ker(\mathscr{D}(f))^\perp$,

$$\|h\|_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R})} \leq C_l \|\mathscr{D}(h)\|_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R})}. \tag{C.3}$$

We show this by contradiction. If the claim were false, then we could find a sequence $\{h_k\}_k$ in $\ker(\mathscr{D}(f))^\perp$ such that $\|h_k\|_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R})} = 1$ and $\|\mathscr{D}(h_k)\|_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R})} = 1/k$. By (C.2), the sequence $\{h_k\}_k$ is uniformly bounded, which implies that there exists a subsequence $\{h_{k_\ell}\}_\ell \subset \{h_k\}_k$ that converges weakly to some $h \in \ker(\mathscr{D})^\perp$ (Evans, 2010, Appendix D, Theorem 3) (see also (Adams & Fournier, 2003, Theorem 3.6)). However, this implies that $\mathscr{D}(h_{k_\ell}) \to \mathscr{D}(h) = 0$, and $h \neq 0$ because $\|h_k\|_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R})} = 1$ by construction. Hence, we would obtain that $h \in \ker(\mathscr{D})$, which is absurd.

Thanks to (C.3), we can re-write (C.2) for $h \in \ker(\mathscr{D})^\perp$ as $\|h\|_{\mathscr{H}^s(\Omega^T,\mathbb{P}_X;\mathbb{R})} \le C_e(C_l+1)\sqrt{\rho}$, showing that any $h \in \ker(\mathscr{D})^\perp$ is contained in a ball of the Sobolev space of radius proportional to $\sqrt{\rho}$. Finally, we can determine the covering number for $\ker(\mathscr{D})^\perp$ by invoking (C.1).

We now proceed by focusing on $\ker(\mathscr{D})$. Ellipticity stated in Assumption 4 implies that such a subspace is finite-dimensional. Additionally, by Theorem B.1 (Sobolev imbedding), there exists a uniform positive constant $\tilde{B}$ such that $\|f\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R})} \le \tilde{B}$ for every function $f \in \mathcal{H}$, which is a closed and convex subset of the Sobolev space $\mathscr{H}^s(\Omega^T,\mathbb{P}_X;\mathbb{R})$ — thus, also $\|g\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R})} \le \tilde{B}$. In light of these considerations, $\ker(\mathscr{D})$ belongs to a ball of radius $\tilde{B}$ of a finite-dimensional Euclidean space with dimension $\dim\ker(\mathscr{D})$, and its covering number can be calculated according to (Cucker & Zhou, 2007, Theorem 5.3).

At this point, since the decomposition of $\mathcal{H}$ into $\ker(\mathscr{D})$ and $\ker(\mathscr{D})^\perp$ is orthogonal, the covering number of $\mathcal{H}$ is given by the product of the covering numbers of the two subspaces. Taking logarithms, we obtain, for sufficiently large constants $\mathfrak{c}$ and C_c,

$$\log \mathcal{N}_\infty(\mathcal{H},\varepsilon) \le \mathfrak{c}\left(\left(\frac{\sqrt{\rho}}{\varepsilon}\right)^{\frac{d_X}{s}} + \dim\ker(\mathscr{D})\log\left(\frac{\tilde{B}}{\varepsilon}\right)\right) \le C_c\left(\frac{\sqrt{\rho}}{\varepsilon}\right)^{\frac{d_X}{s}}, \tag{C.4}$$

where the contribution of $\ker(\mathscr{D})$ is incorporated in C_c as the logarithmic term is negligible for small ε.

To conclude the proof, we proceed along the lines of the proof for Theorem C.3 and obtain the final claim by invoking Theorem C.1. □

D Properties of the hypothesis spaces

We now demonstrate some useful properties of our hypothesis spaces $\mathscr{F}$ and $\mathscr{F}^\rho$. We will start by focusing on convexity of the spaces and on the boundedness of the functions that belong to them; next, we proceed by showing that our effective hypothesis space $\mathscr{F}^\rho$ satisfies the small-ball condition introduced in Mendelson (2014) at least on a subset of interest for the following proofs.

D.1 Convexity and B-boundedness

Lemma D.1. *The hypothesis space $\mathscr{F}$ presented in* (3.6) *is convex – i.e., for any $0 \le \xi \le 1$ and any $f, h \in \mathscr{F}$, we have that $\xi f + (1-\xi)h$ still belongs to $\mathscr{F}$. Additionally, there exists a constant $B > 0$ such that every $f \in \mathscr{F}$ is B-bounded, i.e., $\|f\|_\infty \le B$ for all $f \in \mathscr{F}$.*

Proof. Given an arbitrary Hilbert space $\mathcal{H}$ with norm induced by the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, the parallelogram law yields

$$\|x-y\|_{\mathcal{H}} + \|x+y\|_{\mathcal{H}} = 2\|x\|_{\mathcal{H}} + 2\|y\|_{\mathcal{H}} \Rightarrow \|x-y\|_{\mathcal{H}} \le 2\|x\|_{\mathcal{H}} + 2\|y\|_{\mathcal{H}} \tag{D.1}$$

This leads to showing that Hilbert spaces (then, also the Sobolev space $\mathscr{H}^s(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})$) are uniformly convex (Adams & Fournier, 2003, Definition 1.20 and Theorem 3.5): therefore, the first claim follows by noting that $\mathscr{F}$ is a convex subset of the Sobolev space. The second claim is a consequence of the imbedding presented in Theorem B.1(b): indeed, there exist a constant $\mathfrak{C}_\infty$ such that $\|f\|_{\mathscr{L}^\infty(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} \le \mathfrak{C}_\infty\|f\|_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} \le \mathfrak{C}_\infty\rho_f \doteq B$ by definition of $\mathscr{F}$ in (3.6). □

D.2 $(C,2)$-hypercontractivity

We start by providing the general definition.

Definition D.1. For a given hypothesis space $\mathcal{H}$ of vector-valued functions and uniform constants $C > 0$ and $\alpha \in [1, 2]$, the tuple $(\mathcal{H}, \mathbb{P}_\Omega)$ is (C, α)-*hypercontractive* if it holds that, for every $f \in \mathcal{H}$,

$$\mathbb{E}_{\mathbb{P}_X}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^4\right] \leq C\left(\mathbb{E}_{\mathbb{P}_X}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right]\right)^\alpha. \tag{D.2}$$

In this paper, we will be focusing on the corner case $\alpha = 2$, which implies that the *small-ball condition* (Mendelson, 2014) holds:

Lemma D.2. *Let $(\mathcal{H}, \mathbb{P}_X)$ be $(C, 2)$-hypercontractive for a suitable positive constant C. Then it holds that, for any f, h in $\mathcal{H}$, there exist ε and ξ such that*

$$\mathbb{P}_X\left(\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t) - h(X_t)\|_2 \geq \xi\,\|f - h\|_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})}\right) \geq \varepsilon.$$

Proof. First, as pointed out in the discussion in Section 2.2, note that Definition D.1 can be written as $\|f\|^4_{\mathscr{L}^4(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} \leq C\,\|f\|^{2\cdot 2}_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})}$. Next, let $c_{2,4} \doteq \|f - h\|_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} / \|f - h\|_{\mathscr{L}^4(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})}$. By the Paley-Zygmund inequality (De La Peña & Giné, 1999, Corollary 3.3.2), we have that

$$\mathbb{P}_X\left(\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t) - h(X_t)\|_2 \geq u\,\|f - h\|_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})}\right) \geq [(1-u^2)c_{2,4}^2]^2 \geq (1-u^2)^2/C,$$

where the last step follows by Definition D.1. Conclusion follows as soon as we let $u \leftrightarrow \xi$ and $(1-u^2)^2/C \leftrightarrow \varepsilon$. □

Drawing inspiration from (Ziemann, 2022, Proposition 3.4.4), we now prove that the hypothesis space $\mathscr{F}$ satisfies this particular kind of hypercontractivity on a subset of interest for Theorem F.2, which will allow us to quantify the probability of the lower isometry event.

Theorem D.3. *Let Assumptions 1, 2, 5 and 6 hold. Given some $r > 0$, consider the set $\partial B(r) = \left\{f \in \mathscr{F} \,|\, \|f - f_\star\|^2_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} = r^2\right\}$ and let $\mathscr{F}_\epsilon$ be its cover with balls of radius ϵ. Furthermore, let $\widetilde{\rho_f} \propto \rho_f/\overline{\kappa}$, where $\overline{\kappa}$ is as per Assumption 1. Then, we have that the covering number of $\partial B(r)$ (in other words, the cardinality of $\mathscr{F}_\epsilon$) satisfies*

$$\mathcal{N}_\infty\left(\partial B(r), \epsilon\right) \leq \left(\frac{8\widetilde{\rho_f}m_\epsilon^{s/d_X}d_Y}{\epsilon} + 1\right)^{m_\epsilon d_Y} \tag{D.3}$$

with m_ϵ being the smallest integer solution of

$$m \geq \left(\frac{16\widetilde{\rho_f}d_X}{(2s - d_X)\epsilon^2}\right)^{d_X/(2s-d_X)}.$$

Additionally, as long as $\epsilon \leq \inf_{f \in \partial B(r/\sqrt{(\overline{\kappa})})}\|f\|_{\mathscr{L}^2(\Omega^T, \mu_\lambda^T; \mathbb{R}^{d_Y})}$, the tuple $(\mathscr{F}_\epsilon, \mathbb{P}_X)$ is $(C(\epsilon), 2)$-hypercontractive, with

$$C(\epsilon) \propto \left(\frac{\mu_\lambda(\Omega)}{32} + 8\mu_\lambda(\Omega)m_\epsilon^2\left(\frac{\mu_\lambda(\Omega)}{8} + 2\right)^2\right). \tag{D.4}$$

Proof. We will make use of the construction presented in Section B.2 and focus on the Fourier characterization of $\mathscr{H}^s(\Omega^T, \mu_\lambda^T; \mathbb{R}^{d_Y})$: the result carries over to $\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})$ by the imbedding $\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y}) \hookrightarrow \mathscr{H}^s(\Omega^T, \mu_\lambda^T; \mathbb{R}^{d_Y})$. As such, we will consider the space $\mathscr{F}_{\mu_\lambda} \doteq \left\{f \in \mathscr{H}^s(\Omega^T, \mu_\lambda^T; \mathbb{R}^{d_Y}) \mid \|f\|^2_{\mathscr{H}^s(\Omega^T, \mu_\lambda^T; \mathbb{R}^{d_Y})} \leq \rho_f^2/\overline{\kappa}\right\}$ such that $\mathscr{F} \subseteq \mathscr{F}_{\mu_\lambda}$: indeed, the condition on the norm reads as $\overline{\kappa}\,\|f\|^2_{\mathscr{H}^s(\Omega^T, \mu_\lambda^T; \mathbb{R}^{d_Y})} \geq \|f\|^2_{\mathscr{H}^s(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})}$ by Assumption 1. According to the construction in Section B.2 and the Fourier decomposition in terms of the basis functions

$\phi_\ell(x) = \exp\{\iota\pi \left\langle k(\ell),\, x\right\rangle_2 /(2L)\}$, we can leverage (B.2) to write

$$\mathscr{F}_{\mu_\lambda} = \left\{ f(x) = \sum_{\ell\in\mathbb{N}} z_\ell\phi_\ell(x),\ z_\ell \in \mathbb{R}^{d_Y}\ \forall\ell\in\mathbb{N} \;\middle|\; \sum_{\ell\in\mathbb{N}} \frac{\|z_\ell\|_2^2}{\ell^{-2s/d_X}} \le \widetilde{\rho_f}^2, \right\}, \tag{D.5}$$

where $\widetilde{\rho_f}^2$ is equal to $\rho_f^2/\overline{\kappa}$ times a multiplicative constant that can be retrieved along the lines of the proof of Theorem B.3, but we do not specify it because it does not affect the main message of our results. Additionally, we will consider $\widetilde{r}^2 \doteq r^2/\overline{\kappa}$ such that $B(\widetilde{r}) \supset B(r)$. Note that, due to such an inclusion, $\mathcal{N}_\infty\left(\partial B(r), \epsilon\right) \le \mathcal{N}_\infty\left(\partial B(\widetilde{r}), \epsilon\right)$.

Covering. We start by proving the result on the covering of $\partial B(\widetilde{r})$. The idea is to find a suitable finite-dimensional approximation of the space $\mathscr{F}_{\mu_\lambda}$, compute its covering number using Theorem C.2, and then express $\mathcal{N}_\infty\left(\partial B(\widetilde{r}), \epsilon\right)$ in terms of such a cover.

We first seek an approximation of $\mathscr{F}_{\mu_\lambda}$ at resolution $\epsilon/4$ by considering, for some $m \in \mathbb{Z}_{\ge 0}$, the finite-dimensional space

$$\mathscr{F}^m_{\mu_\lambda} = \left\{ f(x) = \sum_{\ell=1}^{m} z_\ell\phi_\ell(x),\ z_\ell \in \mathbb{R}^{d_Y} \;\middle|\; \sum_{\ell\in\mathbb{N}} \frac{\|z_\ell\|_2^2}{\ell^{-2s/d}} \le \widetilde{\rho_f}^2 \right\}.$$

Now, fix $f \in \mathscr{F}_{\mu_\lambda}$ with coordinates $\{z_\ell\}_{\ell\in\mathbb{N}}$ and let f' be its projection onto $\mathscr{F}^m_{\mu_\lambda}$. Then the following holds:

$$\begin{aligned}
\|f - f'\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} &= \left\| \sum_{\ell=m+1}^{\infty} z_\ell\phi_\ell \right\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} \\
&\le \left\| \sqrt{\sum_{\ell=m+1}^{\infty} \frac{\|z_\ell\|_2^2}{\ell^{-2s/d_X}}} \sqrt{\sum_{\ell=m+1}^{\infty} \ell^{-2s/d_X}|\phi_\ell|^2} \right\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} \\
&\overset{\text{(D.5)}}{\le} \widetilde{\rho_f} \left\| \sqrt{\sum_{\ell=m+1}^{\infty} \ell^{-2s/d_X}|\phi_\ell|^2} \right\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} \\
&\le \widetilde{\rho_f} \sqrt{\sum_{\ell=m+1}^{\infty} \ell^{-2s/d_X}\, \||\phi_\ell|^2\|_{\mathscr{L}^\infty(\Omega;\mathbb{R})}} \le \widetilde{\rho_f} \sqrt{\sum_{\ell=m+1}^{\infty} \ell^{-2s/d_X}},
\end{aligned} \tag{D.6}$$

where the first inequality is given by the Cauchy-Schwarz one, and the last one follows by definition of the basis $\phi_\ell(\cdot)$. We now use the integral test (Rudin, 1976, Chapter 6, Exercise 8) to upper-bound the last expression. Let $\mathfrak{p} \doteq 2s/d_X$ to simplify notation, and note that $\mathfrak{p} > 1$ by Assumption 2. Then we have that

$$\sum_{\ell=m+1}^{\infty} \left(\frac{1}{\ell}\right)^{\mathfrak{p}} \le \sum_{\ell=m+1}^{\infty} \int_{\ell-1}^{\ell} \left(\frac{1}{\mathrm{x}}\right)^{\mathfrak{p}} d\mathrm{x} = \int_{m}^{\infty} \left(\frac{1}{\mathrm{x}}\right)^{\mathfrak{p}} d\mathrm{x} = \frac{1}{\mathfrak{p}-1}\frac{1}{m^{\mathfrak{p}-1}}.$$

Plugging such a bound in (D.6), to ensure that $\|f - f'\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} \le \epsilon/4$ for any $f \in \mathscr{F}_{\mu_\lambda}$ we then require that

$$\widetilde{\rho_f}\sqrt{\frac{1}{\mathfrak{p}-1}\frac{1}{m^{p-1}}} \le \epsilon/4 \iff m \ge \left(\frac{16\widetilde{\rho_f}^2 d_X}{(2s-d_X)\epsilon^2}\right)^{\frac{d_X}{2s-d_X}}. \tag{D.7}$$

Thus, we take m_ϵ the smallest integer m satisfying (D.7) to have the approximation of $\mathscr{F}_{\mu_\lambda}$ at resolution $\epsilon/4$.

We can now proceed by constructing the covering for $\mathscr{F}^{m_\epsilon}_{\mu_\lambda}$ at resolution $\epsilon/4$. We start by characterizing such a space in terms of the coefficients $\{z_\ell\}_{\ell=1}^{m_\epsilon}$ by considering

$$\mathcal{Z}^{m_\epsilon} \doteq \left\{ z_\ell \in \mathbb{R}^{d_Y},\ \ell = 1, \cdots, m_\epsilon \;\middle|\; \sum_{\ell=1}^{m_\epsilon} \frac{\|z_\ell\|_2^2}{\ell^{-2s/d_X}} \le \widetilde{\rho_f}^2. \right\}.$$

This is a ball of radius $\widetilde{\rho_f}$ in a finite-dimensional Banach space with a weighted 2-norm, which we denote by $\|\cdot\|_w$. With such a norm, by Theorem C.2, its covering number with balls of radius $\bar{\epsilon}$ satisfies

$$\mathcal{N}_w\left(\mathcal{Z}^{m_\epsilon}, \bar{\epsilon}\right) \leq (2\widetilde{\rho_f} d_Y/\bar{\epsilon} + 1)^{m_\epsilon d_Y} \doteq N. \tag{D.8}$$

Now, denote the elements of the optimal covering of $\mathcal{Z}^{m_\epsilon}$ as $\{z_\ell^1, \cdots, z_\ell^N\}_{\ell=1,\cdots,m_\epsilon}$. These identify a further approximation $\mathscr{F}_{\mu_\lambda}^{m_\epsilon, N} \subset \mathscr{F}_{\mu_\lambda}^{m_\epsilon}$ defined as

$$\mathscr{F}_{\mu_\lambda}^{m_\epsilon, N} \doteq \left\{ \sum_{\ell=1}^{m_\epsilon} z_\ell^1 \phi_\ell, \cdots, \sum_{\ell=1}^{m_\epsilon} z_\ell^N \phi_\ell \right\}.$$

We can use this construction to characterize the radius $\bar{\epsilon}$. Let $f'(\cdot) = \sum_{\ell=1}^{m_\epsilon} z'_\ell \phi_\ell(\cdot)$ be an arbitrary function in $\mathscr{F}_{\mu_\lambda}^{m_\epsilon}$. Then we have that

$$\begin{aligned}
\min_{n=1,\cdots,N} \left\| f' - \sum_{\ell=1}^{m_\epsilon} z_\ell^n \phi_\ell \right\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} &= \min_{n=1,\cdots,N} \left\| \sum_{\ell=1}^{m_\epsilon} (z'_\ell - z_\ell^n)\phi_\ell \right\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} \\
&\leq \min_{n=1,\cdots,N} \left\| \sqrt{\sum_{\ell=1}^{m_\epsilon} \frac{\|z'_\ell - z_\ell^n\|_2^2}{\ell^{-2s/d_X}}} \sqrt{\sum_{\ell=1}^{m_\epsilon} \ell^{-2s/d_X} |\phi_\ell|^2} \right\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} \\
&\overset{(D.8)}{\leq} \bar{\epsilon} \sqrt{\sum_{\ell=1}^{m_\epsilon} \left(\frac{1}{\ell}\right)^{\mathfrak{p}}} \leq \bar{\epsilon} m_\epsilon^{\mathfrak{p}/2}.
\end{aligned}$$

Therefore, if we take $\bar{\epsilon} \leq \epsilon/(4m_\epsilon^{\mathfrak{p}/2})$ we obtain the $\epsilon/4$-cover of $\mathscr{F}_{\mu_\lambda}^{m_\epsilon}$ we seek. This leads to the claim that the covering number of $\mathscr{F}_{\mu_\lambda}^{m_\epsilon}$ satisfies

$$\mathcal{N}_\infty\left(\mathscr{F}_{\mu_\lambda}^{m_\epsilon}, \frac{\epsilon}{4}\right) \leq \left(\frac{8\widetilde{\rho_f} m_\epsilon^{\mathfrak{p}/2} d_Y}{\epsilon} + 1\right)^{d_Y m_\epsilon}. \tag{D.9}$$

This part of the proof is concluded by converting the covering $\mathscr{F}_{\mu_\lambda}^{m_\epsilon, N}$ into an *exterior cover* of the set $\partial B(\widetilde{r}) \subset \mathscr{F}$ — that is, its elements cover $\partial B(\widetilde{r})$ but they are required to belong to $\mathscr{F}$ and not necessarily to $\partial B(\widetilde{r})$. Indeed, with the construction carried out so far, for each $f \in \partial B(\widetilde{r})$ we can identify $f' \in \mathscr{F}_{\mu_\lambda}^{m_\epsilon}$ such that $\|f - f'\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} \leq \epsilon/4$, and a $f'' \in \mathscr{F}_{\mu_\lambda}^{m_\epsilon, N}$ such that $\|f' - f''\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} \leq \epsilon/4$: thus, by the triangle inequality, $\|f - f''\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} \leq \epsilon/2$, implying that $\mathscr{F}_{\mu_\lambda}^{m_\epsilon, N}$ is an exterior cover of $\partial B(\widetilde{r})$ of resolution $\epsilon/2$. The final claim follows by applying (Vershynin, 2024, Exercise 4.2.9).

Hypercontractivity. We now prove $(C(\epsilon), 2)$-hypercontractivity of the tuple $(\mathscr{F}_\epsilon, \mu_\lambda^T)$ — the same claim will hold also for $(\mathscr{F}_\epsilon, \mathbb{P}_X)$ by multiplying C_ϵ by a constant not depending on ϵ and is thus not relevant to our analysis.
We start by showing that $\mathscr{F}_{\mu_\lambda}^{m_\epsilon}$ satisfies the hypercontractivity condition (Definition D.1) with $\alpha = 2$. Letting $f = \sum_{\ell=1}^{m_\epsilon} z_\ell \phi_\ell \in \mathscr{F}_{\mu_\lambda}^{m_\epsilon}$, we first observe that

$$\left\| \sum_{\ell=1}^{m_\epsilon} z_\ell \phi_\ell \right\|_{\mathscr{L}^2(\Omega^T, \mu_\lambda^T;\mathbb{R}^{d_Y})}^2 = \left\| \sum_{\ell=1}^{m_\epsilon} z_\ell \phi_\ell \right\|_{\mathscr{L}^2(\Omega, \mu_\lambda;\mathbb{R}^{d_Y})}^2 = \sum_{\ell=1}^{m_\epsilon} \|z_\ell\|_2^2. \tag{D.10}$$

Next, looking at the fourth moment,

$$\begin{aligned}
\left\| \sum_{\ell=1}^{m_\epsilon} z_\ell \phi_\ell \right\|_{\mathscr{L}^4(\Omega^T, \mu_\lambda^T;\mathbb{R}^{d_Y})}^4 &= \left\| \sum_{\ell=1}^{m_\epsilon} z_\ell \phi_\ell \right\|_{\mathscr{L}^4(\Omega, \mu_\lambda;\mathbb{R}^{d_Y})}^4 \\
&= \int_\Omega \left\| \sum_{\ell=1}^{m_\epsilon} z_\ell \phi_\ell(\mathrm{x}) \right\|_2^4 d\mathrm{x}
\end{aligned}$$

$$
\begin{aligned}
&\leq \int_{\Omega} \left(\sum_{\ell=1}^{m_\epsilon} \|z_\ell\|_2 \, |\phi_\ell(x)| \right)^4 d\mathrm{x} \\
&\leq \mu_\lambda(\Omega) \left(\sum_{\ell=1}^{m_\epsilon} \|z_\ell\|_2 \right)^4 \\
&\leq \mu_\lambda(\Omega) \left(\sqrt{\sum_{\ell=1}^{m_\epsilon} \|z_\ell\|_2^2} \right)^4 (\sqrt{m_\epsilon})^4 \\
&\overset{\text{(D.10)}}{=} \mu_\lambda(\Omega) m_\epsilon^2 \left(\left\| \sum_{\ell=1}^{m_\epsilon} z_\ell \phi_\ell \right\|^2_{\mathscr{L}^2(\Omega^T,\mu_\lambda^T;\mathbb{R}^{d_Y})} \right)^2, \qquad \text{(D.11)}
\end{aligned}
$$

which shows the $(\mu_\lambda(\Omega) m_\epsilon^2, 2)$-hypercontractivity of $\mathscr{F}_{\mu_\lambda}^{m_\epsilon}$.

Before showing hypercontractivity of $(\mathscr{F}_\epsilon, \mu_\lambda^T)$, we first state some additional useful relations. Recalling that $\epsilon \leq \inf_{f\in\partial B(\widetilde{r})} \|f\|_{\mathscr{L}^2(\Omega^T,\mu_\lambda^T;\mathbb{R}^{d_Y})}$, letting f be an arbitrary element in the cover $\mathscr{F}_\epsilon$ and f' be a function in $\mathscr{F}_{\mu_\lambda}^{m_\epsilon}$ such that $\|f - f'\|_{\mathscr{L}^\infty(\Omega,\mathbb{R}^{d_Y})} \leq \epsilon/4$, we have:

$$
\|f(x)\|_2^4 \leq 8 \left(\|f(x) - f'(x)\|_2^4 + \|f'(x)\|_2^4 \right) \leq \frac{\epsilon^4}{32} + 8 \|f'(x)\|_2^4 \qquad \text{(D.12a)}
$$

$$
\|f'(x)\|_2^2 \leq 2 \left(\|f(x) - f'(x)\|_2^2 + \|f(x)\|_2^2 \right) \leq \frac{\epsilon^2}{8} + 2 \|f(x)\|_2^2 \qquad \text{(D.12b)}
$$

$$
\epsilon^2 \leq \|f\|^2_{\mathscr{L}^2(\Omega^T,\mu_\lambda^T;\mathbb{R}^{d_Y})}, \qquad \epsilon^4 \leq \left(\|f\|^2_{\mathscr{L}^2(\Omega^T,\mu_\lambda^T;\mathbb{R}^{d_Y})} \right)^2. \qquad \text{(D.12c)}
$$

We can now get to the final claim and show hypercontractivity for $\mathscr{F}_\epsilon$. Letting again f be an arbitrary element in the cover $\mathscr{F}_\epsilon$, it holds that

$$
\begin{aligned}
\|f\|_{\mathscr{L}^4(\Omega^T,\mu_\lambda^T;\mathbb{R}^{d_Y})} &= \int_\Omega \|f(\mathrm{x})\|_2^4 \, d\mathrm{x} \\
&\overset{\text{(D.12a)}}{\leq} \mu_\lambda(\Omega) \frac{\epsilon^4}{32} + 8 \int_\Omega \|f'(\mathrm{x})\|_2^4 \, d\mathrm{x} \\
&= \mu_\lambda(\Omega) \frac{\epsilon^4}{32} + 8 \int_\Omega \left\| \sum_{\ell=1}^{m_\epsilon} z'_\ell \phi_\ell(\mathrm{x}) \right\|_2^4 d\mathrm{x} \\
&\overset{\text{(D.11)}}{\leq} \mu_\lambda(\Omega) \frac{\epsilon^4}{32} + 8\mu_\lambda(\Omega) m_\epsilon^2 \left(\int_\Omega \|f'(\mathrm{x})\|_2^2 \, d\mathrm{x} \right)^2 \\
&\overset{\text{(D.12b)}}{\leq} \mu_\lambda(\Omega) \frac{\epsilon^4}{32} + 8\mu_\lambda(\Omega) m_\epsilon^2 \left(\int_\Omega \frac{\epsilon^2}{8} + 2 \|f(\mathrm{x})\|_2^2 \, d\mathrm{x} \right)^2 \\
&\overset{\text{(D.12c)}}{\leq} \left(\frac{\mu_\lambda(\Omega)}{32} + 8\mu_\lambda(\Omega) m_\epsilon^2 \left(\frac{\mu_\lambda(\Omega)}{8} + 2 \right)^2 \right) \left(\|f\|_{\mathscr{L}^2(\Omega^T,\mu_\lambda;\mathbb{R}^{d_Y})} \right)^2,
\end{aligned}
$$

which concludes the proof. □

E ON THE REGULARIZER $\Psi(f)$

We now present the main properties of the regularizer $\Psi(f) \colon \mathscr{F} \to \mathbb{R}_{\geq 0}$ introduced in the learning problem (3.3), where we recall that $\Psi(f) = ||\mathscr{D}(f)||^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;R^{d_Y})}$.

E.1 PHYSICS-INFORMED REGULARIZATION IS 2-PROPER

We start from the definition of η-regularizer (Lecué & Mendelson, 2017):

Definition E.1. An η-proper regularizer defined for a hypothesis space $\mathcal{H}$ is a function $\Psi(\cdot)\colon \mathcal{H} \to \mathbb{R}$ satisfying the following properties:

(a) it is non-negative, even, convex, and such that $\Psi(0) = 0$;

(b) for $\eta \geq 1$, it holds for every f, h in $\mathscr{F}$ that $\Psi(f + h) \leq \eta\,(\Psi(f) + \Psi(h))$;

(c) for every $0 \leq a \leq 1$, $\Psi(af) \leq a\Psi(f)$. Additionally, if $\eta = 2$, it holds that $\Psi(af) \leq a^2\Psi(f)$.

In particular, any square-norm-based regularizer is 2-proper. Therefore, by construction, the physics-informed regularizer considered in this paper (see (3.5)) is 2-proper.

E.2 Useful inequality from Lecué & Mendelson (2017)

We now report an inequality that will be used in the proof of Theorem 4.1 proved in Section H.1.

Lemma E.1 (Lecué & Mendelson (2017), Inequality 2.3)**.** *Denote with $\hat{f}$ the solution of the regularized empirical risk minimization over $\mathscr{F}^\rho$. Write $\hat{f} = f_\star + R(h - f_\star)$, where $R \geq 1$ and $\Psi(h - f_\star) = \rho$, with $\rho \geq 5\eta\Psi(f_\star)$. Then it holds that*

$$\Psi(\hat{f}) - \Psi(f_\star) \geq \frac{R}{2\eta^2}(\Psi(h) - \Psi(f_\star)). \tag{E.1}$$

Proof. By the triangle inequality and the fact that the regularizer is an even function (both reported in Definition E.1), it holds that

$$\Psi(\hat{f}) = \Psi(f_\star + R(h - f_\star)) \geq \frac{1}{\eta}\Psi(R(h - f_\star)) - \Psi(f_\star) \geq \frac{R}{\eta}\Psi(h - f_\star) - \Psi(f_\star)$$

recalling that $R \geq 1$. Adding $\Psi(f_\star)$ on both sides,

$$\Psi(\hat{f}) - \Psi(f_\star) \geq \frac{R}{\eta}\Psi(h - f_\star) - 2\Psi(f_\star). \tag{E.2}$$

As an intermediate step, we find a lower bound on the term $\frac{R}{\eta}\Psi(h - f_\star)$ in (E.2): specifically, it holds that

$$\begin{aligned}
\frac{R}{\eta}\Psi(h - f_\star) &= \frac{R}{2\eta}\Psi(h - f_\star) + \frac{R}{2\eta}\Psi(h - f_\star) \\
&\geq \frac{R}{2\eta}\Psi(h - f_\star) + \frac{5R}{2}\Psi(f_\star) \quad \text{because } \rho = \Psi(h - f_\star) \geq 5\eta\Psi(f_\star)) \\
&\geq \frac{R}{2\eta}\Psi(h - f_\star) + \frac{R}{2}\Psi(f_\star) + 2\Psi(f_\star) \quad \text{because } R \geq 1, \\
&\geq \frac{R}{2\eta}\left(\Psi(h - f_\star) + \Psi(f_\star)\right) + 2\Psi(f_\star) \quad \text{as } \eta \geq 1 \text{ and } R \geq 1, \\
&\geq \frac{R}{2\eta^2}\Psi(h) + 2\Psi(f_\star)
\end{aligned}$$

again as a consequence of the triangle inequality in Definition E.1(b). Plugging such an inequality back in (E.2), we obtain

$$\Psi(\hat{f}) - \Psi(f_\star) \geq \frac{R}{2\eta^2}\Psi(h) \geq \frac{R}{2\eta^2}\Psi(h) - \frac{R}{2\eta^2}\Psi(f_\star),$$

which yields the claim. □

F Lower isometry bound

We start by presenting in Section F.1 an ancillary result combining (C, α)-hypercontractivity (Definition D.1) and S-persistence (Section A.2). This will then play a key role in Section F.2, where we prove an upper bound for the probability of the lower isometry event, which will be crucial in our main results stated in Sections 4 and 5.

F.1 Combining (C, α)-hypercontractivity and S-persistence

We now present an ancillary result obtained by generalizing (Ziemann, 2022, Lemma 3.1.1.).

Lemma F.1. *Consider $g \colon \Omega \to \mathbb{R}_{\geq 0}$ satisfying (C, α)-hypercontractivity (see Definition D.1) and S-persistence (see Assumption 6). Then, for $\theta \geq 8$, it holds that*

$$\mathbb{P}_X \left(\sum_{t=0}^{T-1} g(X_t) \leq \frac{4}{\theta} \sum_{t=0}^{T-1} \mathbb{E}_{\mathbb{P}_X} \left[g(X_t) \right] \right) \leq \exp \left(- \frac{8T}{CS\theta^2} \left(\sum_{t=0}^{T-1} \mathbb{E}_{\mathbb{P}_X} \left[g(X_t) \right] \right)^{2-\alpha} \right) \tag{F.1}$$

Proof. We start by generalizing the S-persistence bound in Assumption 6. Specifically, introducing $\varepsilon > 0$, it holds that

$$\mathbb{E} \left[\exp \left(-\xi \sum_{t=0}^{T-1} g(X_t) \right) \right] \leq \exp \left(- \frac{8\xi}{\theta} \sum_{t=0}^{T-1} \mathbb{E}[g(X_t)] + \frac{\xi^2 S \varepsilon}{\theta} \sum_{t=0}^{T-1} \mathbb{E}[g^2(X_t)] \right), \tag{F.2}$$

where it is required that $\theta \geq 8$ and $\varepsilon / \theta \geq 1/2$.
Now we consider the left-hand side of (F.1) and apply a Chernoff bound to obtain

$$\begin{aligned} \mathbb{P} \left(\sum_{t=0}^{T-1} g(X_t) \leq \frac{4}{\theta} \sum_{t=0}^{T-1} \mathbb{E}[g(X_t)] \right) &\leq \inf_{\xi > 0} \mathbb{E} \left[\exp \left(\frac{4\xi}{\theta} \sum_{t=0}^{T-1} \mathbb{E}[g(X_t)] - \xi \sum_{t=0}^{T-1} g(X_t) \right) \right] \\ &\overset{\text{(F.2)}}{\leq} \inf_{\xi > 0} \mathbb{E} \left[\exp \left(- \frac{4\xi}{\theta} \sum_{t=0}^{T-1} \mathbb{E}[g(X_t)] + \frac{\xi^2 S \varepsilon}{\theta} \sum_{t=0}^{T-1} \mathbb{E}[g^2(X_t)] \right) \right]. \end{aligned}$$

We find the optimal ξ, which reads as

$$\xi = \frac{2 \sum_{t=0}^{T-1} \mathbb{E}[g(X_t)]}{S\varepsilon \sum_{t=0}^{T-1} \mathbb{E}[g^2(X_t)]},$$

and plugging it in we obtain

$$\begin{aligned} \mathbb{P} \left(\sum_{t=0}^{T-1} g(X_t) \leq \frac{4}{\theta} \sum_{t=0}^{T-1} \mathbb{E}[g(X_t)] \right) &\leq \exp \left(- \frac{4}{\theta S \varepsilon} \frac{\left(\sum_{t=0}^{T-1} \mathbb{E}[g(X_t)] \right)^2}{\sum_{t=0}^{T-1} \mathbb{E}[g^2(X_t)]} \right) \\ &\leq \exp \left(- \frac{4T}{\theta C S \varepsilon} \left(\sum_{t=0}^{T-1} \mathbb{E}[g(X_t)] \right)^{2-\alpha} \right) \end{aligned}$$

by (C, α)-hypercontractivity given in Definition D.1. Conclusion follows by minimizing the bound over $\varepsilon \geq \theta/2$. □

F.2 The main bound on lower isometry

We are now ready to prove the key bound for the lower isometry event by generalizing (Ziemann, 2022, Theorem 3.1.2.).

Theorem F.2. *Assume that the tuple $(\mathscr{F}^\rho, \mathbb{P}_X)$ is S-persistent (Assumption 6). For a given $r > 0$, define $B(r) \doteq \left\{ f \in \mathscr{F} \mid \|f\|_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} \leq r^2 \right\}$ and let $\partial B(r)$ be its boundary. Additionally, assume that the hypothesis space satisfies the (C, α)-hypercontractivity condition (Definition D.1) on $\partial B(r)$. For a fixed $\theta > 8$, define $\mathscr{F}_r$ the $r/\sqrt{\theta}$-cover in the $\|\cdot\|_{\mathscr{L}^\infty(\Omega^T; \mathbb{R}^{d_Y})}$ of $\partial B(r)$, and denote by $\mathcal{N}_\infty \left(\partial B(r), \frac{r}{\sqrt{\theta}} \right)$ the corresponding covering number. Define the lower-isometry event*

$$\mathcal{A}_r \doteq \sup_{f \in \mathscr{F}^\rho_\star \setminus B(r)} \left\{ \frac{1}{T} \sum_{t=0}^{T-1} \|f(X_t)\|_2^2 - \frac{1}{\theta} \|f\|^2_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} \leq 0 \right\}. \tag{F.3}$$

Then the following lower-isometry estimate holds:

$$\mathbb{P}_X\left(\mathcal{A}_r\right) \le \mathcal{N}_\infty\left(\partial B(r), \frac{r}{\sqrt{\theta}}\right)\exp\left\{-\frac{8Tr^{4-2\alpha}}{\theta^2 CS}\right\}.$$

Proof. We first show a preliminary result that allows us to focus just on the boundary $\partial B(r)$ instead of the full $\mathscr{F}_\star \setminus B(r)$. Specifically, we show that, if $f \in \mathscr{F}_\star^\rho$ and $0 \le \xi \le 1$, then ξf still belongs to $\mathscr{F}_\star^\rho$: that is, we show that $\mathscr{F}_\star^\rho$ is *star-shaped* around 0 (see (Mendelson, 2014, Definition 5.1)). To prove that $\xi(f - f_\star)$ belongs to $\mathscr{F}_\star^\rho$ for $f, f_\star \in \mathscr{F}^\rho$ we deploy Theorem D.1: specifically, we have

$$\xi f - \xi f_\star \pm f_\star = \underbrace{\xi f + (1-\xi) f_\star}_{w} - f_\star,$$

and $w \in \mathscr{F}^\rho$ by convexity, thus proving the claim.

Thanks to the result above obtained, we can focus on $\partial B(r)$ and then obtain the final claim by rescaling: if $f' \in \mathscr{F}_\star^\rho \setminus B(r)$, then $\|f'\|_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} > r$ by construction, which implies $\frac{r}{\|f'\|_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}} < 1$; thus, if we consider $f = f' \frac{r}{\|f'\|_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}}$, we are on $\partial B(r)$, and $f \in \mathscr{F}_\star^\rho$ by it being star-shaped around 0.

Define the event

$$\mathcal{E} \doteq \bigcup_{f^i \in \mathscr{F}_r} \left\{\frac{1}{T}\sum_{t=0}^{T-1} \left\|f^i(X_t)\right\|_2^2 \le \frac{4}{\theta}\|f^i\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}\right\}.$$

By Theorem F.1 with $g(x) = \left\|f^i(x)\right\|_2^2$, the union bound yields

$$\mathbb{P}_X\left(\mathcal{E}\right) \le \mathcal{N}_\infty\left(\partial B(r), \frac{r}{\sqrt{\theta}}\right)\exp\left\{-\frac{8Tr^{4-2\alpha}}{\theta^2 CS}\right\}.$$

Now, fixing an arbitrary $f \in \partial B(r)$:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2 &\ge \frac{1}{2T}\sum_{t=0}^{T-1}\left\|f^i(X_t)\right\|_2^2 - \frac{r^2}{\theta} && \text{Equation (D.1),}\\
&\ge \frac{2}{\theta}\left\|f^i\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} - \frac{r^2}{\theta} && \text{on } \mathcal{E}^{\complement},\\
&= \frac{2r^2}{\theta} - \frac{r^2}{\theta} = \frac{r^2}{\theta} && \text{by definition of } \mathscr{F}_r.
\end{aligned}$$

Since f was arbitrary, we obtain

$$\mathbb{P}\left(\sup_{f\in\partial B(r)}\left\{\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2 - \frac{r^2}{\theta} \le 0\right\}\right) \le \mathcal{N}_\infty\left(\partial B(r), \frac{r}{\sqrt{\theta}}\right)\exp\left\{-\frac{8Tr^{4-2\alpha}}{\theta^2 CS}\right\}.$$

The claim is finally obtained by rescaling. □

We can now provide a special case of Theorem F.2 that will be useful in the derivations of the paper.

Corollary F.3. *Under the assumptions of Theorem F.2, assume that the hypothesis space satisfies the $(C(r), 2)-$hypercontractivity condition according to Theorem D.3. Then, the following lower-isometry estimate holds:*

$$\mathbb{P}_X(\mathcal{A}_r) \le \left(C_L\left(\frac{1}{r}\right)^{\frac{4s-d_X}{2s-d_X}} + 1\right)^{d_Y C_m\left(\frac{1}{r}\right)^{\frac{2d_X}{2s-d_X}}} \exp\left\{-\frac{8Tr^{\frac{4d_X}{2s-d_X}}}{\theta^2 C_h S}\right\},$$

where C_L, C_m and C_h are constants that depend on $\rho_f, \overline{\kappa}, d_Y, \theta, s, d_X$ and Ω.

Proof. The corollary is obtained by leveraging Theorem D.3, which provides expressions for both the covering number of $\mathscr{F}_r$ and for the hypercontractivity parameter $C(r)$, setting the covering of $\partial B(r)$ to have resolution equal to $r/\sqrt{\theta}$.

Covering number. Using the notation of Theorem D.3, we have that

$$\mathcal{N}_\infty\left(\partial B(r), \frac{r}{\sqrt{\theta}}\right) \overset{\text{(D.3)}}{\leq} \left(\frac{8\widetilde{\rho_f} d_Y \sqrt{\theta} m_{\frac{r}{\sqrt{\theta}}}^{\frac{s}{d_X}}}{r} + 1\right)^{d_Y m_{\frac{r}{\sqrt{\theta}}}},$$

$$\text{where } m_{\frac{r}{\sqrt{\theta}}} \geq \left(\frac{16\widetilde{\rho_f} d_X \theta}{2s - d_X} \cdot \left(\frac{1}{r^2}\right)\right)^{\frac{d_X}{2s-d_X}} \longrightarrow m_{\frac{r}{\sqrt{\theta}}} = C_m \left(\frac{1}{r}\right)^{\frac{2d_X}{2s-dx}}. \tag{F.4}$$

With such a value for $m_{\frac{r}{\sqrt{\theta}}}$, we obtain that the covering number admits the following upper bound:

$$\mathcal{N}_\infty\left(\partial B(r), \frac{r}{\sqrt{\theta}}\right) \leq \left(\underbrace{8\widetilde{\rho_f} d_Y \sqrt{\theta} C_m^{\frac{s}{d_X}}}_{\doteq C_L} \left(\frac{1}{r}\right)^{\frac{4s-d_X}{2s-d_X}} + 1\right)^{d_Y C_m \left(\frac{1}{r}\right)^{\frac{2d_X}{2s-d_X}}}. \tag{F.5}$$

Hypercontractivity parameter. Again, using the result in Theorem D.3 and using the expression for $m_{\frac{r}{\sqrt{\theta}}}$ in (F.4), we obtain that

$$C(r) \overset{\text{(D.4)}}{\propto} \left(\frac{\mu_\lambda(\Omega)}{32} + 8\mu_\lambda(\Omega)\left(\frac{\mu_\lambda(\Omega)}{8} + 2\right)^2 C_m^2 \left(\frac{1}{r}\right)^{\frac{4d_X}{2s-d_X}}\right)$$

$$\longrightarrow C(r) = C_h \left(\frac{1}{r}\right)^{\frac{4d_X}{2s-d_X}} \quad \text{for some sufficiently large constant } C_h. \tag{F.6}$$

The lower-isometry probability bound is then obtained by plugging (F.5) and (F.6) in the claim of Theorem F.2. □

G MARTINGALE OFFSET COMPLEXITY BOUNDS

In this section we focus on some useful results concerning the martingale offset complexity presented in (4.2) and that will play a prominent role in the main results of Sections 4 and 5, being an upper-bound on the empirical excess risk. We start by proving the inequality leading to the definition of the martingale offset complexity, building upon Liang et al. (2015). Next, we report its bounds in probability and in expectation obtained by the chaining arguments of (Ziemann, 2022, Theorem 4.2.2, Theorem 3.2.1). The proofs of the latter, given in Section G.2 and Section G.3 respectively, are given in full detail to keep track of all of the constants involved.

G.1 BEHIND THE SCENES OF THE DEFINITION

The first result is an ancillary inequality derived by extending Liang et al. (2015) to the regularized case (see also (Ziemann et al., 2022, Lemma 1)). The following lemma is the basis yielding the definition of martingale offset complexity.

Lemma G.1. *Let $\hat{f}$ be the solution of the regularized empirical risk minimization problem* (3.3). *Then, it holds that*

$$\frac{1}{T}\sum_{t=0}^{T-1} \left\|\hat{f}(X_t) - f_\star(X_t)\right\|_2^2 \leq \frac{1}{T}\sum_{t=0}^{T-1} 4\left\langle W_t,\, \hat{f}(X_t) - f_\star(X_t)\right\rangle_2 - \left\|\hat{f}(X_t) - f_\star(X_t)\right\|_2^2.$$

Proof. We start by showing the following facts:

- *Fact 1:* for any f and the measurement model $Y_t = f_\star(X_t) + W_t$ in (3.1), it holds that

  $$\|f(X_t) - f_\star(X_t)\|_2^2 = \|Y_t - f(X_t)\|_2^2 - \|Y_t - f_\star(X_t)\|_2^2 + 2\left\langle W_t,\, f(X_t) - f_\star(X_t)\right\rangle_2;$$

- *Fact 2:* from the construction of (3.3), we have that

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|Y_t-\hat{f}(X_t)\right\|_2^2+\lambda_T\Psi(\hat{f})\leq\frac{1}{T}\sum_{t=0}^{T-1}\|Y_t-f_\star(X_t)\|_2^2+\lambda_T\Psi(f_\star). \tag{G.1}$$

  Additionally, since $\Psi(f_\star)\leq\Psi(\hat{f})$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|Y_t-\hat{f}(X_t)\right\|_2^2-\|Y_t-f_\star(X_t)\|_2^2\leq 0. \tag{G.2}$$

Fact 2 follows immediately from optimality of $\hat{f}$. To see why *Fact 1* holds:

$$\|f(X_t)-f_\star(X_t)\pm Y_t\|_2^2=\|Y_t-f(X_t)\|_2^2+\|Y_t-f_\star(X_t)\|_2^2-2\left\langle Y_t-f(X_t),\,Y_t-f_\star(X_t)\right\rangle_2.$$

Considering the last addendum on the right-hand side, adding and subtracting $f_\star(X_t)$ in $(Y_t-f(X_t))$ and using the definition of W_t for the other term, we obtain

$$\begin{aligned}\|f(X_t)-f_\star(X_t)\|_2^2=&\|Y_t-f(X_t)\|_2^2+\|Y_t-f_\star(X_t)\|_2^2\\&-2\|Y_t-f_\star(X_t)\|_2^2-2\left\langle W_t,\,f_\star(X_t)-f(X_t)\right\rangle_2,\end{aligned}$$

thus proving the claim in *Fact 1*.

We are now ready to prove Theorem G.1. We start by applying *Fact 1* to the estimate $\hat{f}$ of (3.3) and multiplying everything by 2. Rearranging the terms, we then obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\left\|\hat{f}(X_t)-f_\star(X_t)\right\|_2^2=\frac{1}{T}\sum_{t=0}^{T-1}2&\left[\overbrace{\left\|Y_t-\hat{f}(X_t)\right\|_2^2-\|Y_t-f_\star(X_t)\|_2^2}^{(\natural)}\right]\\&+4\left\langle W_t,\,\hat{f}(X_t)-f_\star(X_t)\right\rangle_2-\left\|\hat{f}(X_t)-f_\star(X_t)\right\|_2^2.\end{aligned}$$

The conclusion follows by applying *Fact 2* to $(\natural)$. In particular, the claim is obtained by deploying (G.2). If on the other hand one would use (G.1), we would obtain

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\left\|\hat{f}(X_t)-f_\star(X_t)\right\|_2^2\\\leq&\frac{1}{T}\sum_{t=0}^{T-1}4\left\langle W_t,\,\hat{f}(X_t)-f_\star(X_t)\right\rangle_2-\left\|\hat{f}(X_t)-f_\star(X_t)\right\|_2^2+2\lambda_T\left(\Psi(f_\star)-\Psi(\hat{f})\right) &\text{(G.3)}\\\leq&\frac{1}{T}\sum_{t=0}^{T-1}4\left\langle W_t,\,\hat{f}(X_t)-f_\star(X_t)\right\rangle_2-\left\|\hat{f}(X_t)-f_\star(X_t)\right\|_2^2+2\lambda_T\Psi(f_\star). &\text{(G.4)}\end{aligned}$$

□

G.2 Bound in probability

We now provide the high-probability bound for the martingale offset complexity of a given hypothesis space $\mathcal{H}$. This will then be deployed in the high-probability bound for the excess risk in Theorem 4.1, and its analysis will be key to determine the desired rate.

Theorem G.2 (Ziemann (2022), Theorem 4.2.2)**.** *Let Assumptions 1 to 3 and 6 hold, and let $\mathcal{H}\subset\mathscr{H}^s(\Omega^T;\mathbb{P}_X;\mathbb{R}^{d_Y})$ be a convex hypothesis space satisfying Assumption 5. Let $u,v,w\geq 0$. Then,* *with probability* $1 - 4\exp\left\{-u^2/2\right\} - \exp\left\{-v/2\right\}$ *the martingale offset complexity satisfies*

$$\mathbf{M}_T\left[\mathcal{H}\right] \le \inf_{\gamma>0}\left\{8\gamma(u+1)\sqrt{\frac{\sigma_W^2}{T}} + 8\int_0^\gamma \sqrt{\frac{\sigma_W^2 \log \mathcal{N}_\infty\left(\mathcal{H},\varepsilon\right)}{T}}d\varepsilon \right.$$
$$\left. +32\sigma_W^2\frac{\left(v+\log\mathcal{N}_\infty\left(\mathcal{H},\gamma\right)\right)}{T} + 4\gamma^2\right\}.$$

Proof. The core of the proof consists in a *chaining* argument (Talagrand, 2005), i.e., in finding a suitable finite cover of $\mathcal{H}$ and deploying it to derive the desired bounds. We start by defining the terms of the chaining.

Chaining set-up. Let F_k denote the cover of $\mathcal{H}$ with radius $\epsilon_k = \frac{1}{2^k}$; given two arbitrary positive scalars $\delta < \gamma$, the values of k belong to an interval of integers $[\underline{K}, \overline{K}]$ such that

$$\frac{1}{2^{\overline{K}+1}} \le \delta \le \frac{1}{2^{\overline{K}}} \le \frac{1}{2^{\underline{K}+1}} \le \gamma. \tag{G.5}$$

For an arbitrary $f \in \mathcal{H}$, denote with $\pi_k(f)$ the center of the ball in the cover F_k that contains f – i.e., the function such that $\|f - \pi_k(f)\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} \le \epsilon_k$.

Let us write the martingale offset complexity $\mathbf{M}_T\left[\mathcal{H}\right]$ using the following notation:

$$\mathbf{M}_T\left[\mathcal{H}\right] = \sup_{f\in\mathcal{H}}\frac{1}{T}\underbrace{\left[\underbrace{\sum_{t=0}^{T-1} 4\left\langle W_t,\, f(X_t)\right\rangle_2}_{\doteq M_T(f)} - \underbrace{\sum_{t=0}^{T-1}\|f(X_t)\|_2^2}_{\doteq S_T(f)}\right]}_{\doteq N_T(f)}. \tag{G.6}$$

Now, let us exploit the sequence of coverings to write $M_T(f)$ as a telescopic sum:

$$\begin{aligned} M_T(f) &= M_T(f) \pm M_T(\pi_{\overline{K}}(f)) \pm \cdots \pm M_T(\pi_{\underline{K}}(f)) \\ &= \left[M_T(f) - M_T(\pi_{\overline{K}}(f))\right] + D_T(f) + M_T(\pi_{\underline{K}}(f)), \end{aligned}$$

where we set $D_T(f) \doteq \sum_{k=\underline{K}+1}^{\overline{K}} M_T(\pi_k(f)) - M_T(\pi_{k-1}(f))$. We can now focus on using such a rewriting of $M_T(f)$ in the expression of $N_T(f)$ of Equation (G.6). With adding and subtracting $S_T(\pi_{\underline{K}}(f))/2$, it reads as

$$\begin{aligned} N_T(f) &= M_T(f) - S_T(f) \\ &= \left[M_T(f) - M_T(\pi_{\overline{K}}(f))\right] + D_T(f) + \left[M_T(\pi_{\underline{K}}(f)) - \frac{S_T(\pi_{\underline{K}}(f))}{2}\right] \\ &\quad + \left[\frac{S_T(\pi_{\underline{K}}(f))}{2} - S_T(f)\right] \\ &\le \left[M_T(f) - M_T(\pi_{\overline{K}}(f))\right] + D_T(f) + \left[M_T(\pi_{\underline{K}}(f)) - \frac{S_T(\pi_{\underline{K}}(f))}{2}\right] + T\epsilon_{\underline{K}}^2, \end{aligned}$$

where the inequality acting on the last term is obtained through (D.1). Thus, overall, to find a bound for $\mathbf{M}_T\left[\mathcal{H}\right] = \sup_{f\in\mathcal{H}} N_T(f)/T$, we are interested in

$$\begin{aligned} \sup_{f\in\mathcal{H}} N_T(f) \le &\underbrace{\sup_{\substack{f,g\in\mathcal{H} \\ \|f-g\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})}\le 2^{-\overline{K}}}} \left[M_T(f) - M_T(g)\right]}_{\doteq\texttt{(N.A)}} + \underbrace{\sup_{f\in\mathcal{H}} D_T(f)}_{\doteq\texttt{(N.B)}} \\ &+ \underbrace{\sup_{f\in F_{\underline{K}}}\left[M_T(f) - \frac{S_T(f)}{2}\right]}_{\doteq\texttt{(N.C)}} + T\underbrace{\left(\frac{1}{2^{\underline{K}}}\right)^2}_{\overset{\text{(G.5)}}{\le} 4\gamma^2}. \end{aligned} \tag{G.7}$$

The proof proceeds with the following steps. For each **(`N.x`)** with $\mathtt{x} = \mathtt{A,B,C}$, we derive high-probability bounds of the form

$$\boxed{\mathbb{P}\left(\frac{1}{T}(\mathtt{N.x}) > \frac{1}{T}\boxed{\text{value}(\mathtt{x})}\right) \leq \boxed{\text{probability bound}(\mathtt{x})};}$$

then, by deploying the union bound, we combine those results and obtain an upper-bound for $\mathbf{M}_T\left[\mathcal{H}\right]$ that depends on the parameters δ and γ introduced in the chaining set-up (G.5). Finally, by leveraging the condition on the Sobolev order in Assumption 2, we show that we can let $\delta \to 0$ and $w \to +\infty$ to obtain the final claim.

Bound for (`N.A`). Defining $\mathscr{F}_{\overline{K}} \doteq \{f = f^{\mathfrak{a}} - f^{\mathfrak{b}};\ f^{\mathfrak{a}}, f^{\mathfrak{b}} \in \mathcal{H} \mid \|f\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} \leq 2^{-\overline{K}}\}$ and by linearity of $M_T(\cdot)$, we are interested in

$$\mathbb{P}\left(\sup_{f\in\mathscr{F}_{\overline{K}}} M_T(f) > w\right) \leq e^{-\xi w}\mathbb{E}\left[\exp\left\{\xi \sup_{f\in\mathscr{F}_{\overline{K}}} \sum_{t=0}^{T-1} 4\left\langle W_t,\, f(X_t)\right\rangle_2\right\}\right], \tag{G.8}$$

where the inequality follows by applying a Chernoff bound with $\xi > 0$. Now, by deploying monotonicity of the exponential function, we can work on finding upper-bounds for the term in curly brackets in (G.8). Specifically, we consider

$$\begin{aligned}
\xi \sup_{f\in\mathscr{F}_{\overline{K}}} \sum_{t=0}^{T-1} 4\left\langle W_t,\, f(X_t)\right\rangle_2 &\leq \xi \sup_{f\in\mathscr{F}_{\overline{K}}} 4\sqrt{\sum_{t=0}^{T-1}\|W_t\|_2^2}\sqrt{\sum_{t=0}^{T-1}\|f(X_t)\|_2^2} \\
&\leq \xi\frac{4\sqrt{T}}{2^{\overline{K}}}\sqrt{\sum_{t=0}^{T-1}\|W_t\|_2^2} \\
&\leq \frac{1}{2}\left(\frac{4\sqrt{T}\xi}{2^{\overline{K}}}\right)^2\sum_{t=0}^{T-1}\|W_t\|_2^2 + \frac{1}{2} \qquad \text{by Young's inequality.}
\end{aligned} \tag{G.9}$$

Now, by plugging (G.9) into (G.8), we obtain

$$\begin{aligned}
\mathbb{P}\left(\sup_{f\in\mathscr{F}_{\overline{K}}} M_T(f) > w\right) &\leq e^{-\xi w + 1/2}\mathbb{E}\left[\exp\left\{\left(\frac{4\sqrt{T}\xi}{\sqrt{2}\cdot 2^{\overline{K}}}\right)^2\sum_{t=0}^{T-1}\|W_t\|_2^2\right\}\right] \\
&\overset{(Lemma\ A.4)}{\leq} \exp\left\{-\xi w + \frac{1}{2} + \xi^2\left(\frac{4Td_Y\sigma_W^2}{\sqrt{2}\cdot 2^{\overline{K}}}\right)^2\right\},
\end{aligned} \tag{G.10}$$

provided that $\xi < \frac{\sqrt{2}\cdot 2^{\overline{K}}}{4Td_Y\sigma_W^2}$ and using the law of total expectation on the sum of $\|W_t\|_2^2$. In view of obtaining a bound in terms of w and not w^2, we can choose at our convenience ξ such that the quadratic term in (G.10) becomes equal to 1/2. To this aim, setting $\xi = \frac{2^{\overline{K}}}{4Td_Y\sigma_W^2}$ (note that it satisfies the constraint of Theorem A.4) and deploying the definition of δ in (G.5), we obtain

$$\mathbb{P}\left(\sup_{f\in\mathscr{F}_{\overline{K}}} M_T(f) > w\right) \leq \exp\left\{1 - \frac{w}{4\delta Td_Y\sigma_W^2}\right\}.$$

By substituting $w \leftrightarrow w \cdot 4\delta Td_Y\sigma_W^2$ and dividing by T, we finally obtain

$$\boxed{\mathbb{P}\left(\sup_{f\in\mathscr{F}_{\overline{K}}} \frac{M_T(f)}{T} > 4w\delta d_Y\sigma_W^2\right) \leq \exp\{-w+1\}.} \tag{G.11}$$

Bound for (`N.B`). We start by introducing the short-hand notation for the function space $\widetilde{F}_k \doteq \left\{ f = f^{\mathfrak{a}} - f^{\mathfrak{b}};\ f^{\mathfrak{a}} \in F_k, f^{\mathfrak{b}} \in F_{k-1} \mid \|f\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} \leq 2^{-k} \right\}$ for all $k = \underline{K}+1, \ldots, \overline{K}$. Additionally, by linearity of $M_T(\cdot)$, we also have that

$$\sup_{f\in\mathcal{H}} D_T(f) \leq \sum_{k=\underline{K}+1}^{\overline{K}} \sup_{f\in\mathcal{H}} M_T\left(\pi_k(f) - \pi_{k-1}(f)\right) = \sum_{k=\underline{K}+1}^{\overline{K}} \max_{f\in\widetilde{F}_k} M_T(f).$$

We proceed by first studying the single addendum $\max_{f\in\widetilde{F}_k} M_T(f)$, and then apply a union bound to reach the desired claim for (`N.B`).

Letting $u_k > 0$, by deploying a Chernoff bound, we get

$$\begin{aligned}\mathbb{P}\left(\max_{f\in\widetilde{F}_k} M_T(f) > u_k\right) &\leq \min_{\xi} e^{-\xi u_k}\mathbb{E}\left[\exp\left\{\xi \max_{f\in\widetilde{F}_k} M_T(f)\right\}\right] \\ &\leq \min_{\xi} e^{-\xi u_k}\mathbb{E}\left[\sum_{f\in\widetilde{F}_k} \exp\left\{\xi M_T(f)\right\}\right]. \qquad \text{(G.12)}\end{aligned}$$

We now upper-bound (G.12) by iteratively applying the law of total expectation: specifically, we have that

$$\begin{aligned}&\mathbb{E}\left[\sum_{f\in\widetilde{F}_k} \exp\left\{\xi M_T(f)\right\}\right] = \sum_{f\in\widetilde{F}_k} \mathbb{E}\left[\mathbb{E}\left[\exp\left\{\xi \sum_{t=0}^{T-1} 4\left\langle W_t,\, f(X_t)\right\rangle_2\right\} \,\middle|\, \mathcal{X}_{T-2}\right]\right] \\ &= \sum_{f\in\widetilde{F}_k} \mathbb{E}\left[\exp\left\{\xi \sum_{t=0}^{T-2} 4\left\langle W_t,\, f(X_t)\right\rangle_2\right\}\right] \mathbb{E}\left[\exp\left\{\xi 4\left\langle W_{T-1},\, f(X_{T-1})\right\rangle_2\right\} \,\middle|\, \mathcal{X}_{T-2}\right] \\ &\overset{(3.2)}{\leq} \sum_{f\in\widetilde{F}_k} \mathbb{E}\left[\exp\left\{\xi \sum_{t=0}^{T-2} 4\left\langle W_t,\, f(X_t)\right\rangle_2\right\}\right] \exp\left\{\frac{8\xi^2\sigma_W^2}{2^{2k}}\right\} \\ &\leq \vdots \quad \text{(i.e., repeating the argument with the next filtrations } \mathcal{X}_{T-3}, ..., \mathcal{X}_0\text{)} \\ &\leq |\widetilde{F}_k| \exp\left\{\frac{8T\xi^2\sigma_W^2}{2^{2k}}\right\}, \qquad \text{(G.13)}\end{aligned}$$

where $|\widetilde{F}_k|$ is the cardinality of $\widetilde{F}_k = F_k \times F_{k-1}$. Now, after noting that $|\widetilde{F}_k| \leq \left(\mathcal{N}_\infty\left(\mathcal{H}, 2^{-k}\right)\right)^2$, we can plug the bound of (G.13) into (G.12) and obtain, by minimizing over ξ (yielding $\xi = {}^{2^{2k}u_k}\!/\!_{(16T\sigma_W^2)}$),

$$\mathbb{P}\left(\max_{f\in\widetilde{F}_k} M_T(f) > u_k\right) \leq \left(\mathcal{N}_\infty\left(\mathcal{H}, \frac{1}{2^k}\right)\right)^2 \exp\left\{-\frac{2^{2k}u_k^2}{32T\sigma_W^2}\right\}$$

Additionally, by substituting $u_k \leftrightarrow u_k + 2^{-k}\sqrt{64T\sigma_W^2 \log \mathcal{N}_\infty\left(\mathcal{H}, \frac{1}{2^k}\right)} > 0$, we can remove the dependence on the covering number from the probability bound. Applying the union bound over all $k = \underline{K}+1, \cdots, \overline{K}$, we obtain that the bound for (`N.B`) can be written as

$$\begin{aligned}&\mathbb{P}\left(\sup_{f\in\mathcal{H}} D_T(f) > \sum_{k=\underline{K}+1}^{\overline{K}} u_k + 2^{-k}\sqrt{64T\sigma_W^2 \log \mathcal{N}_\infty\left(\mathcal{H}, \frac{1}{2^k}\right)}\right) \\ &\leq \sum_{k=\underline{K}+1}^{\overline{K}} \exp\left\{-\frac{2^{2k}u_k^2}{32\sigma_W^2 T}\right\}. \qquad \text{(G.14)}\end{aligned}$$

We now want the right-hand side of (G.14) to depend on a single $u \in \mathbb{R}$, in order to obtain an upper-bound that reads, informally, as $\boxed{\textit{constant}} \times \exp\left\{-u^2/2\right\}$. To do so, we operate on the terms u_k,

$k = \underline{K}+1, \cdots, \overline{K}$ and set them to $u_k = 2^{2-k}\sqrt{\sigma_W^2 T}\sqrt{u^2 - \log 2^{-k+\underline{K}+1}}$: thanks to this choice, the right-hand side of (G.14) becomes

$$\begin{aligned}
&\sum_{k=\underline{K}+1}^{\overline{K}} \exp\left\{-\frac{2^{2k}}{32T\sigma_W^2}\cdot\left(2^{2-k}\sqrt{\sigma_W^2 T}\sqrt{u^2-\log 2^{-k+\underline{K}+1}}\right)^2\right\} \\
&= \sum_{k=\underline{K}+1}^{\overline{K}} \exp\left\{\frac{-u^2}{2}\right\}(\sqrt{2})^{-k+\underline{K}+1} \\
&\leq \exp\left\{\frac{-u^2}{2}\right\}\sum_{k=0}^{\infty}\left(\frac{1}{\sqrt{2}}\right)^k = (2+\sqrt{2})\exp\left\{\frac{-u^2}{2}\right\} \qquad \text{(G.15)}
\end{aligned}$$

as desired. Now we can analyze such a choice for u_k in the left-hand side of (G.14), which becomes

$$\begin{aligned}
\mathbb{P}\Bigg(\sup_{f\in\mathcal{H}} D_T(f) > &\overbrace{\sum_{k=\underline{K}+1}^{\overline{K}} 2^{2-k}\sqrt{\sigma_W^2 T(u^2-\log 2^{-k+\underline{K}+1})}}^{\doteq s_1} \\
&+\underbrace{\sum_{k=\underline{K}+1}^{\overline{K}} 2^{-k}\sqrt{64T\sigma_W^2\log\mathcal{N}_\infty\left(\mathcal{H},2^{-k}\right)}}_{\doteq s_2}\Bigg). \qquad \text{(G.16)}
\end{aligned}$$

We now want to find upper bounds for s_1 and s_2 and remove the sum over k. Regarding s_1, we have

$$\begin{aligned}
s_1 &\leq \sum_{k=\underline{K}+1}^{\overline{K}} 2^{2-k}\sqrt{\sigma_W^2 T u^2} + \sum_{k=\underline{K}+1}^{\overline{K}} 2^{2-k}\sqrt{\sigma_W^2 T}\sqrt{\log\left(2^{k-\underline{K}-1}\right)} \\
&\leq 4\cdot 2^{-\underline{K}-1}\sqrt{\sigma_W^2 T u^2}\sum_{k=0}^{\infty}2^{-k} + 4\cdot 2^{-\underline{K}-1}\sqrt{\sigma_W^2 T\log 2}\underbrace{\sum_{k=0}^{\infty}2^{-k}\sqrt{k}}_{=\mathrm{Li}_{-1/2}(1/2)} \\
&\overset{\text{(G.5)}}{\leq} 8\gamma\sqrt{\sigma_W^2 T}(u+1), \qquad \text{(G.17)}
\end{aligned}$$

because the polylogarithmic function satisfies $\mathrm{Li}_{1/2}(1/2) \approx 1.35$. Now, going to s_2, noting that $2^{-k} = (2^{-k+1} - 2^{-k})$ and by deploying a truncated Dudley's entropy integral (Wainwright, 2019, Theorem 5.22), we have

$$\begin{aligned}
s_2 &\leq \sum_{k=\underline{K}+1}^{\overline{K}}\left(\frac{1}{2^{k-1}}-\frac{1}{2^k}\right)\sqrt{64T\sigma_W^2\log\mathcal{N}_\infty\left(\mathcal{H},2^{-k}\right)} \\
&\leq \int_{2^{-\overline{K}}}^{2^{-\underline{K}-1}}\sqrt{64T\sigma_W^2\log\mathcal{N}_\infty\left(\mathcal{H},\varepsilon\right)}d\varepsilon \\
&\overset{\text{(G.5)}}{\leq} \int_{\delta}^{\gamma}\sqrt{64T\sigma_W^2\log\mathcal{N}_\infty\left(\mathcal{H},\varepsilon\right)}d\varepsilon. \qquad \text{(G.18)}
\end{aligned}$$

Thus, plugging (G.17,G.18) in (G.16), dividing by T and using the bound in (G.15), we obtain the desired bound for **(N.B)**, which reads as

$$\boxed{\mathbb{P}\left(\sup_{f\in\mathcal{H}}\frac{D_T(f)}{T} > \frac{8(u+1)\gamma\sqrt{\sigma_W^2}}{\sqrt{T}} + \frac{8}{\sqrt{T}}\int_{\delta}^{\gamma}\sqrt{\sigma_W^2\log\mathcal{N}_\infty\left(\mathcal{H},\varepsilon\right)}d\varepsilon\right) \leq 4\exp\left\{-\frac{u^2}{2}\right\}.} \qquad \text{(G.19)}$$

Bound for (N.C). Applying again a Chernoff bound along the lines of the manipulations for **(N.B)** in (G.12), we consider

$$
\begin{aligned}
&\mathbb{P}\left(\max_{f\in F_{\underline{K}}}\sum_{t=0}^{T-1} 4\left\langle W_t,\, f(X_t)\right\rangle_2 - \frac{1}{2}\left\|f(X_t)\right\|_2^2 > v\right)\\
&\leq \min_{\xi} e^{-\xi v}\sum_{f\in F_{\underline{K}}}\mathbb{E}\left[\exp\left\{\xi\left(\sum_{t=0}^{T-1} 4\left\langle W_t,\, f(X_t)\right\rangle_2 - \frac{1}{2}\left\|f(X_t)\right\|_2^2\right)\right\}\right]\\
&\leq \min_{\xi} e^{-\xi v}\sum_{f\in F_{\underline{K}}}\mathbb{E}\left[\mathbb{E}\left[\exp\left\{\xi\left(\sum_{t=0}^{T-1} 4\left\langle W_t,\, f(X_t)\right\rangle_2 - \frac{1}{2}\left\|f(X_t)\right\|_2^2\right)\right\}\middle|\mathcal{X}_{T-2}\right]\right]\\
&\leq \min_{\xi} e^{-\xi v}\left(\xi\sum_{f\in F_{\underline{K}}}\mathbb{E}\left[\exp\left(\left\{\sum_{t=0}^{T-2} 4\left\langle W_t,\, f(X_t)\right\rangle_2 - \frac{1}{2}\left\|f(X_t)\right\|_2^2\right)\right\}\right]\right)\\
&\qquad\qquad\times\mathbb{E}\left[\exp\left\{4\xi\left\langle W_{T-1},\, f(X_{T-1})\right\rangle_2 - \xi\frac{1}{2}\left\|f(X_{T-1})\right\|_2^2\right\}|\mathcal{X}_{T-2}\right]. \qquad \text{(G.20)}
\end{aligned}
$$

We now focus on last expected value in (G.20) and discuss its upper bound. Specifically, we have

$$
\begin{aligned}
&\exp\left\{-\xi\frac{1}{2}\left\|f(X_{T-1})\right\|_2^2\right\}\mathbb{E}\left[\exp\left\{4\xi\left\langle W_{T-1},\, f(X_{T-1})\right\rangle_2\right\}\right]\\
&\overset{(3.2)}{\leq}\exp\left\{\left\|f(X_{T-1})\right\|_2^2\left(-\frac{\xi}{2}+8\xi^2\sigma_W^2\right)\right\}\leq 1
\end{aligned}
$$

by setting $\xi = (32\sigma_W^2)^{-1}$. By this choice of ξ, applying the law of total expectation iteratively over t in (G.20) and using the definition of γ in (G.5), we obtain

$$
\mathbb{P}\left(\max_{f\in F_{\underline{K}}}\sum_{t=0}^{T-1} 4\left\langle W_t,\, f(X_t)\right\rangle_2 - \frac{1}{2}\left\|f(X_t)\right\|_2^2 > v\right)\leq\mathcal{N}_\infty\left(\mathcal{H},\gamma\right)\exp\left\{-\frac{v}{32\sigma_W^2}\right\};
$$

finally, substituting $v \leftrightarrow 32\sigma_W^2(v/2+\log\mathcal{N}_\infty\left(\mathcal{H},\gamma\right))$ and dividing by T, we obtain the bound for **(N.C)**:

$$
\boxed{\mathbb{P}\left(\sup_{f\in F_{\underline{K}}}\frac{M_T(f)}{T}-\frac{1}{2T}S_T(f)>\frac{32\sigma_W^2}{T}\left(\frac{v}{2}+\log\mathcal{N}_\infty\left(\mathcal{H},\gamma\right)\right)\right)\leq\exp\left\{-\frac{v}{2}\right\}.} \qquad \text{(G.21)}
$$

Obtaining the final bound. We can now combine these results and derive the high-probability bound for the martingale offset complexity $\mathbf{M}_T\left[\mathcal{H}\right] = \sup_{f\in\mathcal{H}} N_T(f)/T$. Leveraging the decomposition of $N_T(f)$ according to (G.7), we combine the bounds on **(N.A)**, **(N.B)** and **(N.C)** in (G.11), (G.19) and (G.21) using the union bound and obtain that

$$
\begin{aligned}
\mathbb{P}\Bigg(\mathbf{M}_T\left[\mathcal{H}\right] > 4w\delta d_Y\sigma_W^2 &+ 8\sqrt{\frac{\sigma_W^2}{T}}\left((u+1)\gamma+\int_\delta^\gamma\sqrt{\log\mathcal{N}_\infty\left(\mathcal{H},\varepsilon\right)}d\varepsilon\right)\\
&+\frac{32\sigma_W^2}{T}\left(\frac{v}{2}+\log\mathcal{N}_\infty\left(\mathcal{H},\gamma\right)\right)+4\gamma^2\Bigg)\\
\leq\exp\left\{-w+1\right\}&+4\exp\left\{-\frac{u^2}{2}\right\}+\exp\left\{-\frac{v}{2}\right\},
\end{aligned}
$$

and the expression for the lower bound of the martingale offset complexity is to be maximized with respect to γ and $\delta < \gamma$.

We now claim that we can set $\delta = 0$ and simplify the bound. Setting $\delta = 0$ is possible only if the integral in the term **(N.B)** converges. Under our assumptions, we have that (see Section C.2)

$$
\int_\delta^\gamma\sqrt{\log\mathcal{N}_\infty\left(\mathcal{H},\varepsilon\right)}d\varepsilon\propto\int_\delta^\gamma\left(\frac{1}{\varepsilon}\right)^{\frac{d_X}{2s}}d\varepsilon=\frac{\varepsilon^{1-d_X/2s}}{1-d_X/2s}, \qquad \text{(G.22)}
$$

and the value is finite for $\delta \to 0$ if and only if ${}^{d_X}/_{2s} < 1$, which is guaranteed by Assumption 2. Therefore, we can set $\delta = 0$, and since the term associated to **(N.A)** becomes 0, we can also let $w \to \infty$ and increase the final probability level in the claim of the theorem.

□

G.3 BOUND IN EXPECTATION

Along the lines of the result in probability of the previous subsection, we now present the bound in expectation for the martingale offset complexity.

Theorem G.3 (Ziemann (2022), Theorem 3.2.1)**.** *Let Assumptions 1 to 3 and 6 hold, and let $\mathcal{H} \subset \mathscr{H}^s(\Omega^T; \mathbb{P}_X; \mathbb{R}^{d_Y})$ be a convex hypothesis space satisfying Assumption 5. Then, the martingale offset complexity satisfies*

$$\mathbb{E}\left[\mathbf{M}_T\left[\mathcal{H}\right]\right] \leq \inf_{\gamma > 0} \left\{ 8 \int_0^\gamma \sqrt{\frac{\sigma_W^2 \log \mathcal{N}_\infty\left(\mathcal{H}, \varepsilon\right)}{T}} d\varepsilon + \frac{32\sigma_W^2 \log \mathcal{N}_\infty\left(\mathcal{H}, \gamma\right)}{T} + 4\gamma^2 \right\}.$$

Proof. The result is again obtained by chaining, using the construction leading to (G.7). Using the definitions of $N_T(f)$, $M_T(f)$ and $S_T(f)$ in (G.6), as well as the ones for the chaining resolutions δ, γ in (G.5), we are looking at

$$\begin{aligned} \mathbb{E}\left[\sup_{f \in \mathcal{H}} \frac{1}{T} N_T(f)\right] \leq & \frac{1}{T} \mathbb{E}\left[\underbrace{\sup_{\substack{f, g \in \mathcal{H} \\ \|f-g\|_{\mathscr{L}^\infty(\Omega^T; \mathbb{R}^{d_Y})} \leq 2^{-\overline{K}}}} M_T(f) - M_T(g)}_{\textbf{(N.A)}}\right] + \frac{1}{T} \mathbb{E}\left[\underbrace{\sup_{f \in \mathcal{H}} D_T(f)}_{\textbf{(N.B)}}\right] \\ & + \frac{1}{T} \mathbb{E}\left[\underbrace{\sup_{f \in F_{\underline{K}}} M_T(f) - \frac{S_T(f)}{2}}_{\textbf{(N.C)}}\right] + \underbrace{\left(\frac{1}{2^{\underline{K}}}\right)^2}_{\overset{\text{(G.5)}}{\leq} 4\gamma^2}. \end{aligned} \tag{G.23}$$

We now proceed with deriving the bounds for the expected values of the terms **(N.A)**, **(N.B)** and **(N.C)**; the final claim is obtained by summing all of the contributions together. Finally, we will discuss the fact that we are allowed to let $\delta = 0$ and simplify the bound.

Bound for (N.A). Define $\mathscr{F}_{\overline{K}} \doteq \{f = f^{\mathfrak{a}} - f^{\mathfrak{b}};\ f^{\mathfrak{a}}, f^{\mathfrak{b}} \in \mathcal{H} \mid \|f\|_{\mathscr{L}^\infty(\Omega^T; \mathbb{R}^{d_Y})} \leq 2^{-\overline{K}}\}$. By linearity of $M_T(\cdot)$, we are looking at

$$\begin{aligned} & \mathbb{E}\left[\sup_{f \in \mathscr{F}_{\overline{K}}} \sum_{t=0}^{T-1} 4 \left\langle W_t,\, f(X_t) \right\rangle_2\right] \leq \mathbb{E}\left[\sup_{f \in \mathscr{F}_{\overline{K}}} \sum_{t=0}^{T-1} 4 \left\|W_t\right\|_2 \left\|f(X_t)\right\|_2\right] \\ \leq & \frac{4}{2^{\overline{K}}} \sum_{t=0}^{T-1} \mathbb{E}\left[\|W_t\|_2\right] \overset{\textit{Lemma A.5}}{\leq} \frac{12T}{2^{\overline{K}}} \sqrt{d_Y \sigma_W^2} \overset{\text{(G.5)}}{\leq} 24T\delta\sqrt{d_Y \sigma_W^2}. \end{aligned}$$

Dividing by T, we obtain the first term in the bound (G.23).

Bound for (N.B). We start by defining the auxiliary search space $\widetilde{F}_k \doteq \left\{f = f^{\mathfrak{a}} - f^{\mathfrak{b}};\ f^{\mathfrak{a}} \in F_k, f^{\mathfrak{b}} \in F_{k-1} \mid \|f\|_{\mathscr{L}^\infty(\Omega^T; \mathbb{R}^{d_Y})} \leq 2^{-k}\right\}$ for all $k = \underline{K} + 1, \ldots, \overline{K}$. Next, using the definition of $D_T(f)$ and exploiting its linearity, we consider

$$\mathbb{E}\left[\sup_{f \in \mathcal{H}} \sum_{k=\underline{K}+1}^{\overline{K}} M_T\left(\pi_k(f) - \pi_{k-1}(f)\right)\right] \leq \sum_{k=\underline{K}+1}^{\overline{K}} \mathbb{E}\left[\sup_{f \in \widetilde{F}_k} M_T(f)\right]. \tag{G.24}$$

To upper-bound the right-hand side of (G.24), we focus on its addenda and proceed with the following argument. Noting that $\widetilde{F}_k$ is a finite-dimensional class, let us consider, for some $\xi > 0$,

$$
\begin{aligned}
\exp\left\{\xi \mathbb{E}\left[\max_{f\in\widetilde{F}_k} M_T(f)\right]\right\} &\leq \mathbb{E}\left[\exp\left\{\xi \max_{f\in\widetilde{F}_k} M_T(f)\right\}\right] \quad \text{by Jensen's inequality,}\\
&= \mathbb{E}\left[\max_{f\in\widetilde{F}_k} \exp\left\{\xi M_T(f)\right\}\right] \quad \text{by monotonicity,}\\
&\leq \sum_{f\in\widetilde{F}_k} \mathbb{E}\left[\exp\left\{\xi M_T(f)\right\}\right]\\
&= \sum_{f\in\widetilde{F}_k} \mathbb{E}\left[\mathbb{E}\left[\exp\left\{\xi \sum_{t=0}^{T-1} 4\left\langle W_t,\, f(X_t)\right\rangle_2\right\}\middle|\mathcal{X}_{T-2}\right]\right]\\
&\overset{(3.2)}{\leq} \mathbb{E}\left[\exp\left\{\xi \sum_{t=0}^{T-2} 4\left\langle W_t,\, f(X_t)\right\rangle_2\right\}\right]\exp\left\{\frac{8\xi^2\sigma_W^2}{2^{2k}}\right\}\\
&\leq \vdots \quad \text{(i.e., repeating the argument with the subsequent filtrations)}\\
&\leq \left(\mathcal{N}_\infty\left(\mathcal{H}, 2^{-k}\right)\right)^2 \exp\left\{\frac{8T\xi^2\sigma_W^2}{2^{2k}}\right\},
\end{aligned}
\tag{G.25}
$$

noting that the cardinality of $\widetilde{F}_k = F_k \times F_{k-1}$ is upper-bounded by $\left(\mathcal{N}_\infty\left(\mathcal{H}, 2^{-k}\right)\right)^2$. Now, taking logarithms of both sides of the whole inequality (G.25), we obtain

$$
\begin{aligned}
\mathbb{E}\left[\max_{f\in\widetilde{F}_k} M_T(f)\right] &\leq \frac{2}{\xi}\log\mathcal{N}_\infty\left(\mathcal{H}, 2^{-k}\right) + \frac{8T\xi\sigma_W^2}{2^{2k}}\\
\rightarrow \mathbb{E}\left[\max_{f\in\widetilde{F}_k} M_T(f)\right] &\leq 2^{-k}\sqrt{64T\sigma_W^2 \log\mathcal{N}_\infty\left(\mathcal{H}, 2^{-k}\right)}
\end{aligned}
\tag{G.26}
$$

after minimizing with respect to ξ.

We can now go back to (G.24). Plugging (G.26), we obtain

$$
\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{H}} \sum_{k=\underline{K}+1}^{\overline{K}} M_T\left(\pi_k(f) - \pi_{k-1}(f)\right)\right] &\leq \sum_{k=\underline{K}+1}^{\overline{K}} \frac{1}{2^k}\sqrt{64T\sigma_W^2 \log\mathcal{N}_\infty\left(\mathcal{H}, 2^{-k}\right)}\\
&= \sum_{k=\underline{K}+1}^{\overline{K}} \left(\frac{1}{2^{k-1}} - \frac{1}{2^k}\right)\sqrt{64T\sigma_W^2 \log\mathcal{N}_\infty\left(\mathcal{H}, 2^{-k}\right)}\\
&\leq \sum_{k=\underline{K}+1}^{\overline{K}} \int_{2^{-\overline{K}}}^{2^{-\underline{K}-1}} \sqrt{64T\sigma_W^2 \log\mathcal{N}_\infty\left(\mathcal{H}, \varepsilon\right)} d\varepsilon\\
&\overset{(G.5)}{\leq} 8\int_{\delta}^{\gamma} \sqrt{T\sigma_W^2 \log\mathcal{N}_\infty\left(\mathcal{H}, \varepsilon\right)} d\varepsilon
\end{aligned}
$$

having used in the second inequality a truncated Dudley entropy integral (Wainwright, 2019, Theorem 5.22). Finally, the second term in (G.11) is obtained by dividing the last inequality by T.

Bound for (`N.C`). We are now working to find the upper bound for the expected value of the martingale offset complexity of a finite class of functions, $\mathbb{E}\left[\mathbf{M}_T\left[F_{\underline{K}}\right]\right]$, where $F_{\underline{K}}$ is the $2^{-\underline{K}}$-cover of the hypothesis space $\mathcal{H}$. Similarly to what has been done for (**`N.B`**), we start by noticing that, for any $\xi > 0$,

$$
\exp\left\{\xi \mathbb{E}\left[\max_{f\in F_{\underline{K}}} M_T(f) - \frac{1}{2}S_T(f)\right]\right\}
$$

$$
\begin{aligned}
&\le \mathbb{E}\left[\max_{f\in F_{\underline{K}}} \exp\left\{\xi\left(M_T(f) - \frac{1}{2}S_T(f)\right)\right\}\right] \qquad \text{(Jensen's inequality)}\\
&\le \sum_{f\in F_{\underline{K}}} \mathbb{E}\left[\exp\left\{\xi\left(M_T(f) - \frac{1}{2}S_T(f)\right)\right\}\right]\\
&= \sum_{f\in F_{\underline{K}}} \mathbb{E}\left[\mathbb{E}\left[\exp\left\{\xi\left(\sum_{t=0}^{T-1} 4\xi \left\langle W_t,\, f(X_t)\right\rangle_2 - \frac{\xi}{2}\left\|f(X_t)\right\|_2^2\right)\right\}\Big|\mathcal{X}_{T-2}\right]\right] \qquad \text{(total expectation)}\\
&= \sum_{f\in F_{\underline{K}}} \mathbb{E}\left[\exp\left\{\sum_{t=0}^{T-1} 4\left\langle W_t,\, f(X_t)\right\rangle_2 - \frac{1}{2}\left\|f(X_t)\right\|_2^2\right\}\right]\\
&\qquad\quad \cdot\, \mathbb{E}\left[4\left\langle W_{T-1},\, f(X_{T-1})\right\rangle_2 - \frac{1}{2}\left\|f(X_{T-1})\right\|_2^2\,\Big|\mathcal{X}_{T-2}\right]\\
&\overset{(3.2)}{\le} \sum_{f\in F_{\underline{K}}} \mathbb{E}\left[\exp\left\{\sum_{t=0}^{T-1} 4\left\langle W_t,\, f(X_t)\right\rangle_2 - \frac{1}{2}\left\|f(X_t)\right\|_2^2\right\}\right] \underbrace{\exp\left\{\left\|f(X_{T-1})\right\|_2^2\left(-\frac{\xi}{2} + 8\xi^2\sigma_W^2\right)\right\}}_{\le 1 \text{ by letting } \xi = (32\sigma_W^2)^{-1}}\\
&\le \;\vdots\; \quad \text{(iterating over the subsequent filtrations)}\\
&\le \mathcal{N}_\infty\left(\mathcal{H},\, \frac{1}{2^{\underline{K}}}\right). \qquad\qquad \text{(G.27)}
\end{aligned}
$$

Now, by taking the logarithm on both sides of (G.27) and using the value $\xi = (32\sigma_W^2)^{-1}$ found above, we obtain

$$
\mathbb{E}\left[\max_{f\in F_{\underline{K}}} M_T(f) - \frac{1}{2}S_T(f)\right] \le 32\sigma_W^2 \log\mathcal{N}_\infty\left(\mathcal{H},\, \frac{1}{2^{\underline{K}}}\right) \overset{\text{(G.5)}}{\le} 32\sigma_W^2 \log\mathcal{N}_\infty\left(\mathcal{H}, \gamma\right),
$$

and the bound for the third term in (G.23) is obtained by dividing the terms above by T.

Wrapping up. Putting the bounds for all the terms **(N.A)**, **(N.B)** and **(N.C)** together, we obtain

$$
\boxed{\mathbb{E}\left[\mathbf{M}_T\left[\mathcal{H}\right]\right] \le 24\delta\sqrt{d_Y\sigma_W^2} + \int_\delta^\gamma \sqrt{\frac{64\sigma_W^2\log\mathcal{N}_\infty\left(\mathcal{H},\varepsilon\right)}{T}}d\varepsilon + \frac{32\sigma_W^2\log\mathcal{N}_\infty\left(\mathcal{H},\varepsilon\right)}{T} + 4\gamma^2.}
$$

Following the reasoning at the end of the proof for Theorem G.2, we have that in the scenarios of our interest the integral does not diverge at $\delta = 0$. For this reason, in the final claim we will make use of $\delta = 0$ and simplify the bound. □

H Proofs of the excess risk bounds in Section 4

This section provides the proof of Theorems 4.1 and 4.2. As discussed in Section 4.1, the results are derived leveraging the lower isometry event (F.3) and the bound on its probability presented in Section F. Moreover, we make use of the regularizer's properties elucidated in Section E, and of $(C(r), 2)$-hypercontractivity proved in Section D.2. Ultimately, we obtain that our *complexity-dependent* bounds on the excess risk feature three main ingredients: the complexity of the hypothesis class, captured by the martingale offset complexity; the critical radius r identifying the set $B(r)$ and determining its size (thus, the covering number of its boundary, see Theorem F.2); and the ground-truth regularizer $\Psi(f_\star)$. These results bring together the small-ball method with learning with dependent data, and are the starting point for the derivation of our convergence rate results presented in Section 5 and proved in Section I.

H.1 Proof of Theorem 4.1 (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

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

Proof. We start by noting that $\mathcal{A}_r^{\complement}$ can happen in the following situations:

(i) $\left\| \hat{f} - f_\star \right\|^2_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} \leq r^2$;

(ii) $\hat{f} - f_\star$ is in $\mathscr{F}^\rho \setminus B(r)$, but it happens that $\|\hat{f} - f_\star\|^2_{\mathscr{L}^2} \leq \frac{\theta}{T} \sum_{t=0}^{T-1} \left\| \hat{f}(X_t) - f_\star(X_t) \right\|_2^2$ (see (4.1));

(iii) $\hat{f}$ is outside $\mathscr{F}^\rho$.

The key idea of this Theorem is to prove that scenario (iii) cannot occur with our choice of the regularization parameter λ_T. We will now analyze each situation separately.

Case (i). This is the simple situation in which we are already in the $\mathscr{L}^2$-ball with radius r, $B(r)$, leading to $\|f - f_\star\|^2_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} \leq r^2$.

Case (ii). On this event, we have

$$\begin{aligned}
\left\| \hat{f} - f_\star \right\|^2_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} &\leq \frac{\theta}{T} \sum_{t=0}^{T-1} \left\| \hat{f}(X_T) - f_\star(X_t) \right\|_2^2 \\
&\overset{\text{(G.4)}}{\leq} \frac{\theta}{T} \sup_{g \in \mathscr{F}_\star^\rho} \sum_{t=0}^{T-1} \left[4 \left\langle W_t,\, g(X_t) \right\rangle_2 - \|g(X_t)\|_2^2 \right] + 2\lambda_T \Psi(f_\star) \\
&\overset{\text{(4.2)}}{\leq} \theta \mathbf{M}_T \left[\mathscr{F}^\rho\right] + 2\lambda_T \Psi(f_\star).
\end{aligned}$$

Case (iii). By Theorem D.1, the hypothesis space is convex, and the regularizer $\Psi(\cdot)$ is continuous: therefore, there exists $R > 1$ and $h \in \partial \mathscr{F}^\rho$ such that $\hat{f} = f_\star + R(h - f_\star)$. Additionally, by Definition E.1(b), we have that

$$\Psi(h) \geq \frac{1}{2}\Psi(h - f_\star) - \Psi(f_\star) \Rightarrow \Psi(h) - \Psi(f_\star) \geq \frac{3\rho}{10} \tag{H.1}$$

by virtue of our choice $\Psi(f_\star - h) = \rho$ and by the assumption $\Psi(f_\star) \leq \frac{\rho}{10}$. We can use this in our construction and consider

$$\begin{aligned}
&\frac{1}{T} \sum_{t=0}^{T-1} \left\| \hat{f}(X_t) - f_\star(X_t) \right\|_2^2 \\
&\overset{\text{(G.3)}}{\leq} \frac{1}{T} \sum_{t=0}^{T-1} 4 \left\langle W_t,\, \hat{f}(X_t) - f_\star(X_t) \right\rangle_2 - \left\| \hat{f}(X_t) - f_\star(X_t) \right\|_2^2 + 2\lambda_T \left(\Psi(f_\star) - \Psi(\hat{f}) \right) \\
&\overset{Theorem\ E.1}{\leq} \frac{1}{T} \sum_{t=0}^{T-1} 4R \left\langle W_t,\, h(X_t) - f_\star(X_t) \right\rangle_2 - R^2 \|h(X_t) - f_\star(X_t)\|_2^2 - \frac{R\lambda_T}{4} \left(\Psi(h) - \Psi(f_\star) \right) \\
&\leq R \left[\frac{1}{T} \sum_{t=0}^{T-1} 4 \left\langle W_t,\, h(X_t) - f_\star(X_t) \right\rangle_2 - \|h(X_t) - f_\star(X_t)\|_2^2 - \frac{\lambda_T}{4} \left(\Psi(h) - \Psi(f_\star) \right) \right]
\end{aligned}$$

$$\overset{\text{(H.1)}}{\leq} R\left[\frac{1}{T}\sum_{t=0}^{T-1} 4\left\langle W_t,\, h(X_t) - f_\star(X_t)\right\rangle_2 - \|h(X_t) - f_\star(X_t)\|_2^2 - \frac{3\rho\lambda_T}{40}\right].$$

However, by taking $\lambda > 40\mathbf{M}_T\left[\mathscr{F}^\rho\right]/(3\rho)$, the term in the square brackets becomes negative, leading to an absurd statement.

In light of the analysis for cases (i)-(iii), it results that only cases (i) and (ii) are of interest under the assumptions of Theorem 4.1. Therefore, it holds that

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

as we wanted to prove. □

H.2 Proof of Theorem 4.2 (result in expectation)

Theorem 4.2. Let Assumptions 1 to 3, 5 and 6 hold. Having the set $\partial B(r)$ that is $(C(r), 2)$-hypercontractive, let $\mathscr{F}_r$ be a $r/\sqrt{\theta}$-cover in the infinity norm of $\partial B(r)$, and let $\mathcal{N}_\infty\left(\partial B(r), \frac{r}{\sqrt{\theta}}\right)$ be the associated covering number. 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 $f \in \mathscr{F}$, we have

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

Proof. First, we observe that $\mathscr{F}_r$ is $(C(r), 2)$-hypercontractive as shown in Theorem D.3, and B-boundedness of $\mathscr{F}$ (thus, also of $\mathscr{F}^\rho \subset \mathscr{F}$) follows from Theorem D.1.

We now use the lower isometry event $\mathcal{A}_r$ in (F.3) to decompose the expected value as

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{f} - f_\star\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}\right] &= \mathbb{E}\left[\left\|\left(\hat{f} - f_\star\right)\chi_{\mathcal{A}_r}\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}\right]\\ &\quad + \mathbb{E}\left[\left\|\left(\hat{f} - f_\star\right)\chi_{\mathcal{A}_r^\complement}\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}\right],\end{aligned} \tag{H.2}$$

where $\chi_{\mathfrak{A}}$ is the indicator function of the event $\mathfrak{A}$, such that it is equal to 1 if $\mathfrak{A}$ is true, and 0 otherwise. To obtain the desired bound, we proceed by analyzing the two addenda separately.

First scenario ($\mathcal{A}_r$ is true). In the lower isometry event, we can write

$$\mathbb{E}\left[\left\|\left(\hat{f} - f_\star\right)\chi_{\mathcal{A}_r}\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}\right] \leq \left\|\hat{f} - f_\star\right\|^2_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})}\mathbb{P}_X(\mathcal{A}_r).$$

Then, we can bound the norm on the right-hand side of such an expression by $(2B)^2$, being $\sup_x\left\|\hat{f}(x) - f_\star(x)\right\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} \leq \sup_x\left(\left\|\hat{f}(x)\right\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})} + \|f_\star(x)\|_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})}\right) \leq 2B$, and the bound for $\mathbb{P}_X(\mathcal{A}_r)$ follows by Theorem F.2. Ultimately, we obtain the bound for the first addendum in (H.2) as

$$\mathbb{E}\left[\left\|\left(\hat{f} - f_\star\right)\chi_{\mathcal{A}_r}\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}\right] \leq 4B^2\mathcal{N}_\infty\left(\partial B(r), \frac{r}{\sqrt{\theta}}\right)\exp\left\{-\frac{8T}{\theta^2 C_r S}\right\}. \tag{H.3}$$

Second scenario ($\mathcal{A}_r$ is false). This case is treated as in the high-probability bound of Theorem 4.1. Again, we express the cases leading to the realization of $\mathcal{A}_r^\complement$ as (i) $\hat{f} \in B(r)$; (ii) $\hat{f} \in \mathscr{F}^\rho \setminus B(r)$, but it happens that $\|\hat{f} - f_\star\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} \leq \theta\frac{1}{T}\sum_{t=0}^{T-1}||\hat{f}(X_t) - f_\star(X_t)||_2^2$; (iii) $\hat{f} \in \mathscr{F} \setminus \mathscr{F}^\rho$. Along the lines of Theorem 4.1, we find an upper bound for the second addendum in (H.2) by showing that, with our choice of λ_T, case (iii) does not happen.

CASE (I) When $\hat{f} \in B(r)$, by definition we have that $\mathbb{E}\left[\left\|\hat{f} - f_\star\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}\right] \leq r^2$.

CASE (II) Following the steps in the proof of Theorem 4.1, in this high-probability scenario we have that

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

CASE (III) The argument in the corresponding part of the proof of Theorem 4.1 carries out also when considering the expected value, leading to an absurd conclusion as soon as $\lambda_T \geq \frac{40\mathbb{E}[\mathbf{M}_T[\mathscr{F}^\rho]]}{3\rho}$.

Therefore, overall, the term for the case in which $\mathcal{A}_r$ is false (i.e., the second addendum in (H.2)) is upper-bounded by

$$\mathbb{E}\left[\left\|\left(\hat{f} - f_\star\right)\chi_{\mathcal{A}_r^\complement}\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}\right] \leq r^2 + \theta\mathbb{E}\left[\mathbf{M}_T\left[\mathscr{F}^\rho\right]\right] + 2\lambda_T\Psi(f_\star); \tag{H.4}$$

thus, the claim follows by upper-bounding (H.2) by the sum of (H.3) and (H.4). □

I PROOFS OF CONVERGENCE RATE RESULTS IN SECTION 5

We now present the proofs of Theorems 5.1 and 5.2. These results build upon Theorems 4.1 and 4.2 and rely on specifying the martingale offset complexity bounds (Section G) and the covering number of the boundary of $B(r)$ (Section D.2). By setting the squared critical radius r^2 be dominated by the martingale offset complexity term, we obtain the desired complexity-dependent bounds for the excess risk.

I.1 PROOF OF THEOREM 5.1 (RESULT 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 $d = {}^{2s}\!/_{2s+d_X}$, $d' = {}^{2d_X}\!/_{2s+d_X}$, and 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:

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

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

Proof. The starting point is the bound in probability on the excess risk of Theorem 4.1 given in Equation (4.3). As one of its main ingredients is the bound on the martingale offset complexity, we start the proof by characterizing such a bound reported in Theorem G.2 for the effective hypothesis space $\mathscr{F}^\rho$. A key role is also played by the covering number of $\mathscr{F}^\rho$, which is derived in Theorem C.4. Next, we choose the parameters ρ, λ_T and r^2 according to the requirements of Theorem 4.1, and this leads to the desired excess risk bound. The proof is concluded by characterizing the lower isometry event probability, which leads to the specification of the burn-in time stated in the claim.

Martingale offset complexity bound. We start by determining the bound for $\mathbf{M}_T\left[\mathscr{F}^\rho\right]$ entering (4.3) using the general result of Theorem G.2. By setting $u = \sqrt{2\log(1/\delta)}$ and $v = 2\log(1/\delta)$, we have that, with probability at least $1 - 5\delta$,

$$\mathbf{M}_T\left[\mathscr{F}^\rho\right] \le \inf_{\gamma>0}\left\{8\gamma\sqrt{\frac{\sigma_W^2}{T}}(1+\sqrt{2\log(1/\delta)}) + 8\sqrt{\frac{\sigma_W^2}{T}}\int_0^\gamma \sqrt{\mathcal{N}_\infty\left(\mathscr{F}^\rho,\varepsilon\right)}d\varepsilon \right.$$
$$\left. + \frac{64\sigma_W^2\log(1/\delta)}{T} + \frac{32\sigma_W^2}{T}\mathcal{N}_\infty\left(\mathscr{F}^\rho,\gamma\right) + 4\gamma^2\right\}.$$

By using the covering number result in Theorem C.4 and noting that, according to (G.22),

$$\int_0^\gamma \left(\frac{1}{\varepsilon}\right)^{\frac{d_X}{2s}} d\varepsilon = \frac{\gamma^{1-{d_X}/{2s}}}{1-{d_X}/{2s}}, \tag{I.1}$$

the bound on the martingale offset complexity can be re-written as

$$\mathbf{M}_T\left[\mathscr{F}^\rho\right] \le \inf_{\gamma>0}\left\{8\gamma\sqrt{\frac{\sigma_W^2}{T}}(1+\sqrt{2\log(1/\delta)}) + 8\sqrt{\frac{\sigma_W^2}{T}}\frac{\sqrt{C_c}d_Y^{\frac{2s+d_X}{4s}}}{1-\frac{d_X}{2s}}(\sqrt{\rho})^{\frac{d_X}{2s}}\gamma^{1-d_X/2s} \right.$$
$$\left. + \frac{64\sigma_W^2\log(1/\delta)}{T} + \frac{32\sigma_W^2}{T}C_c d_Y^{\frac{2s+d_X}{2s}}\left(\frac{\sqrt{\rho}}{\gamma}\right)^{\frac{d_X}{s}} + 4\gamma^2\right\}. \tag{I.2}$$

By following the reasoning presented in Liang et al. (2015) (see also Yang & Barron (1999)), minimization over γ is obtained by equating the last two terms in (I.2), which yields

$$\gamma = \left(8C_c d_Y^{\frac{2s+d_X}{2s}}\right)^{\frac{s}{2s+d_X}}\left(\frac{\sigma_W^2}{T}\right)^{\frac{s}{2s+d_X}}(\sqrt{\rho})^{\frac{d_X}{2s+d_X}}.$$

Plugging in such a value for γ in (I.2), and recalling the definitions $d \doteq {2s}/{2s+d_X}$ and $d' = {2d_X}/{2s+d_X}$, we obtain

$$\mathbf{M}_T\left[\mathscr{F}^\rho\right] \le C_I\frac{\sigma_W^{1+d}}{T^{\frac{1+d}{2}}}(\sqrt{\rho})^{\frac{d'}{2}} + C_{II}\frac{(\sigma_W^2)^d}{T^d}(\sqrt{\rho})^{d'} + C_{III}\frac{\sigma_W^2\log(1/\delta)}{T} + C_{IV}\frac{(\sigma_W^2)^d}{T^d}(\sqrt{\rho})^{d'},$$
$$\text{where}\quad\begin{cases} C_I & \doteq 8(1+\sqrt{2\log(1/\delta)})\left(8C_c d_Y^{1/d}\right)^{d/2} \\ C_{II} & \doteq \frac{2s}{2s-d_X}8\sqrt{C_c d_Y^{1/d}}\left(8C_c d_Y^{1/d}\right)^{\frac{2s-d_X}{2(2s+d_X)}} \\ C_{III} & \doteq 64 \\ C_{IV} & \doteq 8\left(8C_c d_Y^{1/d}\right)^d \end{cases} \tag{I.3}$$

Choice of the parameters ρ, λ_T and r. According to Theorem 4.1, we set the radius of the effective hypothesis class $\mathscr{F}^\rho$ to satisfy $\rho = 10\Psi(f_\star)$; similarly, the regularization parameter is chosen as $\lambda_T = \frac{40}{3\rho}\mathbf{M}_T\left[\mathscr{F}^\rho\right]$. Regarding the radius of the $\mathscr{L}^2$-ball $B(r)$, we conveniently set it as $r^2 = \lambda_T\Psi(f_\star) + \frac{\sigma_W^2\log(1/\delta)}{T}$. Thanks to these choices, the excess risk bound in (4.3) reads as

$$\left\|\hat{f} - f_\star\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} \le (\theta+4)\left(C_I\frac{\sigma_W^{1+d}}{T^{\frac{1+d}{2}}}(\sqrt{\rho})^{\frac{d'}{2}} + C_{II}\frac{(\sigma_W^2)^d}{T^d}(\sqrt{\rho})^{d'}\right.$$
$$\left. + C_{III}\frac{\sigma_W^2\log(1/\delta)}{T} + C_{IV}\frac{(\sigma_W^2)^d}{T^d}(\sqrt{\rho})^{d'}\right) + \frac{\sigma_W^2\log(1/\delta)}{T}.$$

Now, noting that $(1+d)/2 > d$, we obtain the desired claim, namely that

$$\left\|\hat{f} - f_\star\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} \le C_{\texttt{slow}}\frac{\max\left\{\Psi(f_\star)^{d'/4},\Psi(f_\star)^{d'/2}\right\}}{T^d} + C_{\texttt{fast}}\frac{\sigma_W^2\log(1/\delta)}{T},$$
$$\text{where}\quad\begin{cases} C_{\texttt{slow}} & \doteq (\theta+4)\left(C_I 10^{d'/4}\sigma_W^{1+d} + C_{II}10^{d'/2}\sigma_W^{2d} + C_{IV}\sigma_W^{2d}10^{d'/2}\right) \\ C_{\texttt{fast}} & \doteq (\theta+4)C_{III} + 1. \end{cases}$$

Characterization of the burn-in time. We conclude the proof by setting the probability of the lower isometry event $\mathcal{A}_r$ equal to δ, so that the overall claim can hold with the desired probability $1 - 5\delta - \delta$.

By Theorem F.3, we have the following bound for the probability of the lower isometry event:

$$\mathbb{P}_X(\mathcal{A}_r) \leq \left(C_L \left(\frac{1}{r}\right)^{\frac{4s-d_X}{2s-d_X}} + 1 \right)^{d_Y C_m \left(\frac{1}{r}\right)^{\frac{2d_X}{2s-d_X}}} \exp\left\{ -\frac{8Tr^{\frac{4d_X}{2s-d_X}}}{\theta^2 C_h S} \right\} \overset{!}{\leq} \delta.$$

Taking logarithms on both sides of the last inequality, letting $C_M \doteq C_m d_Y$, we obtain that T has to satisfy the condition

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

The effective condition is obtained by substituting $r^2 = \lambda_T \Psi(f_\star) + \sigma_W^2/T$. □

I.2 Proof of Theorem 5.2 (result 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 $d = {}^{2s}/_{2s+d_X}$ is the Sobolev minimax rate, $d' = {}^{2d_X}/_{2s+d_X}$, and C_I and C_{II} are constants depending only on s, d_X and d_Y. If the amount of samples 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:

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

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

Proof. Similarly to the proof of Theorem 5.1, we start by characterizing the bound on the expected value of the martingale offset complexity of $\mathscr{F}^\rho$. Next, by choosing the parameters ρ, λ_T and r according to the requirements of Theorem 4.2, we arrive to the desired claim on the bound. Finally, we discuss the burn-in time by characterizing the lower-isometry event probability.

Martingale offset complexity bound. As stated in Theorem G.3, we have that

$$\mathbb{E}\left[\mathbf{M}_T\left[\mathscr{F}^\rho\right]\right] \leq \inf_{\gamma > 0} 8\sqrt{\frac{\sigma_W^2}{T}} \int_0^\gamma \sqrt{\log \mathcal{N}_\infty\left(\mathscr{F}^\rho, \varepsilon\right)} d\varepsilon + \frac{32\sigma_W^2 \log \mathcal{N}_\infty\left(\mathscr{F}^\rho, \gamma\right)}{T} + 4\gamma^2.$$

Leveraging C.4 to characterize the metric entropy of $\mathscr{F}^\rho$ and leveraging (I.1), such a bound can be re-written as

$$\mathbb{E}\left[\mathbf{M}_T\left[\mathscr{F}^\rho\right]\right] \leq \inf_{\gamma > 0} 8\sqrt{\frac{\sigma_W^2}{T}} \frac{\sqrt{C_c d_Y^{\frac{2s+d_X}{2s}}}}{1 - {}^{d_X}/_{2s}} (\sqrt{\rho})^{\frac{d_X}{2s}} \gamma^{1-\frac{d_X}{2s}} + \frac{32\sigma_W^2}{T} (C_c d_Y^{\frac{2s+d_X}{2s}}) \left(\frac{\sqrt{\rho}}{\gamma}\right)^{\frac{d_X}{s}} + 4\gamma^2.$$

As done in Liang et al. (2015); Yang & Barron (1999), we minimize the right-hand side by balancing the last two addenda, which leads to

$$\gamma = \left(8 C_c d_Y^{\frac{2s+d_X}{2s}}\right)^{\frac{s}{2s+d_X}} \left(\frac{\sigma_W^2}{T}\right)^{\frac{s}{2s+d_X}} (\sqrt{\rho})^{\frac{d_X}{2s+d_X}}.$$

By substituting such a value of γ in the martingale offset complexity bound, we obtain that

$$\mathbb{E}\left[\mathbf{M}_T\left[\mathscr{F}^\rho\right]\right] \leq (C_I + C_{II})\left(\frac{\sigma_W^2}{T}\right)^d (\sqrt{\rho})^{d'},$$

where we recall that $d = {}^{2s}\!/_{2s+d_X}$ and $d' = {}^{2d_X}\!/_{2s+d_X}$, and the constants C_I and $C_I I$ are equal to

$$\begin{cases} C_I & \doteq \frac{8\sqrt{C_c d_Y^d}}{1-{}^{d_X}/_{2s}}\left(8C_c d_Y^{\frac{1}{d}}\right)^{\frac{2s-d_X}{2(2s+d_X)}} \\ C_{II} & \doteq 8\left(8C_c d_Y^{\frac{1}{d}}\right)^d \end{cases}$$

Choosing parameters ρ, λ_T and r. We proceed by following the requirements of Theorem 4.2, setting $\rho = 10\Psi(f_\star)$ and $\lambda_T = \frac{40}{3\rho}\mathbb{E}\left[\mathbf{M}_T\left[\mathscr{F}^\rho\right]\right]$. Furthermore, setting $r^2 = 2\lambda_T\Psi(f_\star) + \frac{\sigma_W^2}{T}$, we obtain that the desired bound reads as

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{f} - f_\star\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}\right] \leq & 4B^2\mathcal{N}_\infty\left(\partial B(r), \frac{r}{\sqrt{\theta}}\right)\exp\left\{-\frac{8T}{\theta^2 C(r)S}\right\} \\ & + (\theta+4)10^{\frac{d'}{2}}(C_I + C_{II})\left(\frac{\sigma_W^2}{T}\right)^d \Psi(f_\star)^{\frac{d'}{2}} + \frac{\sigma_W^2}{T}. \qquad \text{(I.4)}\end{aligned}$$

Characterizing the burn-in. We conclude the proof by imposing that the first term on the right-hand side of (I.4) is upper-bounded by $\frac{\sigma_W^2}{T}$, i.e.,

$$4B^2\mathcal{N}_\infty\left(\partial B(r), \frac{r}{\sqrt{\theta}}\right)\exp\left\{-\frac{8T}{\theta^2 C(r)S}\right\} \leq \frac{\sigma_W^2}{T}.$$

Leveraging Theorem F.3, we deploy the values for the covering number and the hypercontractivity parameter and obtain that the number of samples T has to satisfy

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

The effective condition for the burn-in is obtained by letting $r^2 = 2\lambda_T\Psi(f_\star) + \sigma_W^2/T$. Ultimately, with such a choice of T, we obtain that

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

where the constants read as

$$\begin{cases} C_{\texttt{slow}} & \doteq (\theta+4)10^{\frac{d'}{2}}(C_I + C_{II})(\sigma_W^2)^d \\ C_{\texttt{fast}} & \doteq 2. \end{cases}$$

□

J Results for the case without physics-informed regularization

We now derive the bounds in the situation in which there is no physics-informed regularization (i.e., $\lambda_T = 0$ in (3.3)). The obtained bounds will be the benchmark against which we compare Theorems 5.1 and 5.2, showing the impact of knowledge alignment in the excess risk bounds to obtain faster convergence.

We start by recalling that, when physics-informed regularization is absent, the empirical risk minimization problem (3.3) reads as

$$\hat{f}' = \underset{f\in\mathscr{F}}{\arg\min}\,\frac{1}{T}\sum_{t=0}^{T-1}\|Y_t - f(X_t)\|_2^2, \qquad \text{(J.1)}$$

and the lower isometry event takes this expression:

$$\mathcal{A}'_r \doteq \sup_{f\in\mathscr{F}_\star\setminus B(r)}\left\{\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2 - \frac{1}{\theta}\|f\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} \leq 0\right\}. \qquad \text{(J.2)}$$

J.1 Corollary of Theorem 5.1 (result in probability)

Corollary J.1. *Let Assumptions 1 to 3, 5 and 6 hold, and let $\hat{f}'$ be the solution of the estimation problem* (J.1)*. Setting $\theta > 8$, if*

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

where C_h, C_M, C_L are uniform constants, then the excess risk satisfies

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

with probability at least $1-6\delta$, where $C'_{\mathtt{slow}}$ is a constant that depends only on $s, d_X, d_Y, \rho_f, \sigma_W^2, \theta$ and $\sqrt{\log(1/\delta)}$, and $C'_{\mathtt{fast}}$ is a constant that depends only on θ.

Proof. We first adapt Theorem 4.1 to the non-regularized case, and then derive the bounds along the lines of Theorem 5.1.

Expression for the bound. By adapting Theorem 4.1 to the lower-isometry event (J.2), we have that the event $(\mathcal{A}'_r)^{\complement}$ happens in two scenarios:

(i) $\hat{f}' \in B(r) \Longrightarrow \left\|\hat{f}'-f_\star\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} \leq r^2$;

(ii) $\hat{f}-f_\star$ belong to $\mathscr{F}\setminus B(r)$, but it holds that

$$\left\|\hat{f}'-f_\star\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} \leq \frac{\theta}{T}\sum_{t=0}^{T-1}\left\|\hat{f}'(X_t)-f_\star(X_t)\right\|_2^2 \leq \theta\mathbf{M}_T\left[\mathscr{F}\right].$$

Therefore, on $(\mathcal{A}'_r)^{\complement}$, we have that

$$\left\|\hat{f}'-f_\star\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})} \leq \theta\mathbf{M}_T\left[\mathscr{F}\right]+r^2. \tag{J.3}$$

Bounding the martingale offset complexity. We can now rely on Theorem G.2 to bound the martingale offset complexity: setting $u=\sqrt{2\log(1/\delta)}$ and $v=2\log(1/\delta)$, and using Theorem C.3 to characterize the metric entropy of the hypothesis space $\mathscr{F}$, we have that

$$\begin{aligned}\mathbf{M}_T\left[\mathscr{F}\right] \leq \inf_{\gamma>0} 8\sqrt{\frac{\sigma_W^2}{T}}(\sqrt{2\log(1/\delta)}+1)+8\sqrt{\frac{\sigma_W^2}{T}}\frac{\sqrt{C'_c d_Y^{\frac{2s+d_X}{2s}}}}{1-{d_X}/{2s}}\rho_f^{\frac{d_X}{2s}}\gamma^{\frac{2s-d_X}{2s}}\\ +\frac{64\sigma_W^2\log(1/\delta)}{T}+\frac{32\sigma_W^2}{T}C'_c d_Y^{\frac{2s+d_X}{2s}}\left(\frac{\rho_f}{\gamma}\right)^{\frac{d_X}{s}}+4\gamma^2.\end{aligned} \tag{J.4}$$

As done in Theorem 5.1, we balance the last two addenda to get

$$\gamma=\left(8C'_c d_Y^{\frac{2s+d_X}{2s}}\right)^{\frac{s}{2s+d_X}}\rho_f^{\frac{d_X}{2s+d_X}}\left(\frac{\sigma_W^2}{T}\right)^{\frac{s}{2s+d_X}}.$$

Substituting into (J.4) and recalling the definition of the Sobolev minimax rate $d=2s/(2s+d_X)$, we obtain

$$\mathbf{M}_T\left[\mathscr{F}\right] \leq C'_I\left(\frac{\sigma_W^2}{T}\right)^{\frac{d+1}{2}}+C'_{II}\left(\frac{\sigma_W^2}{T}\right)^{d}+C'_{III}\log(1/\delta)\frac{\sigma_W^2}{T}+C'_{IV}\left(\frac{\sigma_W^2}{T}\right)^{d},$$

$$\text{where} \quad \begin{cases} C'_I & \doteq 8\left(8C'_c d_Y^{\frac{2s+d_X}{2s}}\right)^{\frac{s}{2s+d_X}} (1+\sqrt{2\log(1/\delta)})\rho_f^{\frac{d_X}{2s+d_X}} \\ C'_{II} & \doteq 8\frac{\sqrt{C'_c d_Y^{\frac{2s+d_X}{2s}}}}{1-{}^{d_X/2s}}\left(8C'_c d_Y^{\frac{2s+d_X}{2s}}\right)^{\frac{2s-d_X}{2(2s+d_X)}} \rho_f^{\frac{2d_X}{2s+d_X}} \\ C'_{III} & \doteq 64 \\ C'_{IV} & \doteq 8\left(8C'_c d_Y^{\frac{2s+d_X}{2s}}\right)^{\frac{2s}{2s+d_X}} \rho_f^{\frac{2d_X}{2s+d_X}}. \end{cases}$$

Final expression for the bound. Going back to (J.3), noting that $T^{-(d+1)/2} < T^{-d}$ and setting $r^2 \leq \sigma_W^2/T$, we obtain the bound for the excess risk

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

where $C'_{\mathtt{slow}} = \theta(C'_I + C'_{II} + C'_{IV})\max\{\sigma_W^{1+d}, \sigma_W^{2d}\}$ and $C'_{\mathtt{fast}} = 1 + (\theta)C'_{III}$.

Characterization of the burn-in. Similarly to the derivation in Theorem 5.1, the burn-in condition consists in having the amount T satisfying

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

where $r^2 \leq \sigma_W^2/T$. □

J.2 Corollary of Theorem 5.2 (result in expectation)

Corollary J.2. *Let Assumptions 1 to 3, 5 and 6 hold, and let $\hat{f}'$ be the solution of the estimation problem* (J.1)*. Let B the infinity-norm bound of functions in $\mathscr{F}$ and $\theta > 8$. If the number of samples T satisfies*

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

with $r \leq \sqrt{\sigma_W^2/T}$ and uniform constants C_h, C_L, C_M, then we have that

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

Proof. Similarly to the previous Corollary, we first adapt Theorem 4.2 to the non-regularized case and then derive the bounds following Theorem 5.2.

Expression for the bound. Considering the lower-isometry event (J.2), we decompose the expected value as follows:

$$\mathbb{E}\left[\left\|\hat{f}' - f_\star\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}\right] = \overbrace{\mathbb{E}\left[\left\|\left(\hat{f}' - f_\star\right)\chi_{\mathcal{A}'_r}\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}\right]}^{\text{(I)}} + \underbrace{\mathbb{E}\left[\left\|\left(\hat{f}' - f_\star\right)\chi_{(\mathcal{A}'_r)^\complement}\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}\right]}_{\text{(II)}} \tag{J.5}$$

along the lines of the proof of Theorem 4.2. Looking at the two addenda separately:

(I) when $\mathcal{A}'_r$ is true, we have that

$$\mathbb{E}\left[\left\|\left(\hat{f} - f_\star\right)\chi_{\mathcal{A}'_r}\right\|^2_{\mathscr{L}^2(\Omega^T,\mathbb{P}_X;\mathbb{R}^{d_Y})}\right] \leq \left\|\hat{f}' - f_\star\right\|^2_{\mathscr{L}^\infty(\Omega^T;\mathbb{R}^{d_Y})}\mathbb{P}_X(\mathcal{A}'_r)$$

$$\overset{\textit{Theorem F.2}}{\leq} 4B^2\mathcal{N}_\infty\left(\partial B(r), \frac{r}{\sqrt{\theta}}\right)\exp\left\{-\frac{8T}{\theta^2 C(r)S}\right\}$$

by bounding the worst-case distance between $\hat{f}'$ and $f_\star$.

(II) when $\mathcal{A}'_r$ is false, there are two scenarios possible: (i) $\hat{f}' \in B(r)$; or (ii) $\hat{f} \in \mathscr{F} \setminus B(r)$, but it happens that $\|\hat{f}' - f_\star\|^2_{\mathscr{L}^2} \leq \frac{\theta}{T}\sum_{t=0}^{T-1} ||\hat{f}'(X_t) - f_\star(X_t)||_2^2$. Looking at these two cases:

(i) $\|\hat{f}' - f_\star\|^2_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} \leq r^2$ by definition;

(ii) $\|\hat{f}' - f_\star\|^2_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})} \leq \theta \mathbf{M}_T\left[\mathscr{F}\right]$.

Bringing all these terms together, the bound in (J.5) reads as

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{f}' - f_\star\right\|^2_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})}\right] \leq\ & 4B^2\mathcal{N}_\infty\left(\partial B(r), \frac{r}{\sqrt{\theta}}\right)\exp\left\{-\frac{8T}{\theta^2 C(r)S}\right\} \\ & + \theta\mathbb{E}\left[\mathbf{M}_T\left[\mathscr{F}\right]\right] + r^2. \end{aligned} \tag{J.6}$$

Bounding the martingale offset complexity. We can proceed by upper-bounding the martingale offset complexity term by deploying Theorem G.3. Specifically, we have that

$$\mathbb{E}\left[\mathbf{M}_T\left[\mathscr{F}\right]\right] \leq \inf_{\gamma>0} 8\sqrt{\frac{\sigma_W^2}{T}}\int_0^\gamma \sqrt{\log \mathcal{N}_\infty\left(\mathscr{F}, \varepsilon\right)}d\varepsilon + \frac{32\sigma_W^2 \log \mathcal{N}_\infty\left(\mathscr{F}, \gamma\right)}{T} + 4\gamma^2.$$

We deploy Theorem C.3 to provide the expression for the metric entropy of $\mathscr{F}$ (see also the proof of Theorem 5.2) and obtain

$$\mathbb{E}\left[\mathbf{M}_T\left[\mathscr{F}\right]\right] \leq \inf_{\gamma>0} 8\sqrt{\frac{\sigma_W^2}{T}}\frac{\sqrt{C'_c d_Y^{\frac{2s+d_X}{2s}}}}{1 - {}^{d_X}\!/_{2s}}\rho_f^{\frac{d_X}{2s}}\gamma^{\frac{2s-d_X}{2s}} + \frac{32\sigma_W^2}{T}C'_c d_Y^{\frac{2s+d_X}{2s}}\left(\frac{\rho_f}{\gamma}\right)^{\frac{d_X}{s}} + 4\gamma^2$$

Balancing the last two terms of the right-hand side, we obtain

$$\gamma = \left(8C'_c d_Y^{\frac{2s+d_X}{2s}}\rho_f^{\frac{d_X}{s}}\right)^{\frac{s}{2s+d_X}}\left(\frac{\sigma_W^2}{T}\right)^{\frac{s}{2s+d_X}},$$

which we substitute back in the expression of the martingale offset complexity to get (recalling that $d = 2s/(2s + d_X)$ is the Sobolev minimax exponent)

$$\begin{aligned} &\mathbb{E}\left[\mathbf{M}_T\left[\mathscr{F}\right]\right] \leq \frac{C'_I}{T^d} + \frac{C'_{II}}{T^d}, \\ &\text{where} \quad \begin{cases} C'_I & \doteq 8\frac{\sqrt{C'_c d_Y^{\frac{2s+d_X}{2s}}}}{1 - {}^{d_X}\!/_{2s}}\left(8C'_c d_Y^{\frac{2s+d_X}{2s}}\rho_f^{\frac{d_X}{s}}\right)^{\frac{s}{2s+d_X}}\rho_f^{\frac{d_X}{s} + \frac{d_X}{2s+d_X}} \\ C'_{II} & \doteq 8\left(8C'_c d_Y^{\frac{2s+d_X}{2s}}\rho_f^{\frac{d_X}{s}}\right)^{\frac{2s}{2s+d_X}}. \end{cases} \end{aligned} \tag{J.7}$$

Final expression for the bound. We can now go back to the excess risk bound (J.6). We let $r^2 \leq \sigma_W^2/T$, substitute (J.7) in the expected value for the martingale offset complexity, and let the first addendum in (J.6) be upper-bounded by $\frac{\sigma_W^2}{T}$. Ultimately, this yields

$$\begin{aligned} &\mathbb{E}\left[\left\|\hat{f}' - f_\star\right\|^2_{\mathscr{L}^2(\Omega^T, \mathbb{P}_X; \mathbb{R}^{d_Y})}\right] \leq \frac{C'_{\texttt{slow}}}{T^d} + C'_{\texttt{fast}}\frac{\sigma_W^2}{T}, \\ &\text{where} \quad \begin{cases} C'_{\texttt{slow}} & \doteq \theta(C'_I + C'_{II})(\sigma_W^2)^d \\ C'_{\texttt{fast}} & \doteq 2. \end{cases} \end{aligned}$$

Characterization of the burn-in. The bound has been obtained by imposing that the first addendum in (J.6) satisfies

$$4B^2 \mathcal{N}_\infty \left(\partial B(r), \frac{r}{\sqrt{\theta}}\right) \exp\left\{-\frac{8T}{\theta^2 C(r) S}\right\} \leq \frac{\sigma_W^2}{T}.$$

Leveraging Theorem F.3, we deploy the values for the covering number and the hypercontractivity parameter and obtain that the number of samples T has to satisfy

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

The effective condition for the burn-in is obtained by letting $r^2 = \sigma_W^2/T$. □

K Addendum to numerical tests

In the following, we provide in Section K.1 the full details of the numerical test presented in Section 6, also providing more comments on the burn-in, the choice of the regularization parameter λ_T, and the impact of misaligned priors. Additionally, we present in Section K.2 an additional numerical example, in which the prior is given by the Poisson equation.

K.1 Unicycle experiment

Model set-up. We consider a nonlinear dynamical system that describes the dynamics of a unicycle robot. The ground-truth model is given by

$$\begin{aligned}\dot{x}_1(t) &= \nu(t)\cos\vartheta(t),\\ \dot{x}_2(t) &= \nu(t)\sin\vartheta(t)\\ \dot{\vartheta}(t) &= \omega(t),\end{aligned}$$

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.
We simulate the continuous dynamics using a forward Euler discretization with step size $dt = 0.05$, yielding the discrete-time dynamical system

$$\begin{aligned}x_{1,t+1} &= x_{1,t} + \nu_t\cos(\vartheta_t)dt,\\ x_{2,t+1} &= x_{2,t} + \nu_t\sin(\vartheta_t)dt,\\ \vartheta_{t+1} &= \vartheta_t + \omega_t dt.\end{aligned}$$

We generate training pairs $\{(s_t, u_t), s_{t+1}\}$ where $s_t = (x_{1,t}, x_{2,t}, \vartheta_t)$ and $u_t = (\nu_t, \omega_t)$, corrupted by an additive Gaussian noise on s_{t+1} with variance $\sigma_W^2 = 1$.

Physics-informed regularization. The unicycle kinematics enforce that the velocity has no lateral component. This constraint takes the form

$$q(s_t, u_t) = (x_{2,t+1} - x_{2,t})\cos\vartheta_t - (x_{1,t+1} - x_{1,t})\sin\vartheta_t = 0.$$

To promote models consistent with the physics, we penalize the squared $\mathscr{L}^2$-norm of this residual over a compact domain of states and inputs, i.e.,

$$\|q\|^2_{\mathscr{L}^2(\Omega)} \;=\; \int_\Omega q(\xi)^2 d\xi, \text{ with } \xi = (s, u).$$

We approximate the above integral with Monte Carlo sampling from the input domain Ω. For each mini-batch, we evaluate the residuals under the adopted predictor $g_\theta(s_t, u_t)$ (which we specify below) and compute

$$\widehat{\|q\|}^2_{\mathscr{L}^2(\Omega)} = \frac{\mu_\lambda(\Omega)}{N}\sum_{i=1}^{N} q(\xi_i)^2, \text{ with } \xi_i \text{ uniformly sampled from } \Omega.$$

The total loss combines the data mean squared error and the physics-informed penalty according to (3.3). This ensures the learned predictor both fits noisy data and respects the no-slip constraint.

Adopted estimator. We use a feedforward neural network $g_\theta : (s_t, u_t) \in \mathbb{R}^5 \to \hat{s}_{t+1} \in \mathbb{R}^3$ to approximate the discrete-time dynamics. The architecture is a two-hidden-layer multilayer perceptron (MLP) with ReLU activation function and 64 inner nodes.
We train the estimator using the Adam optimizer with learning rate 0.5×10^{-3} and batch size $N = 200$. We vary the effective number of training samples T over the range $T \in [300, 10^6]$, and compute the average learning rate after the burn-in.

Experiment regimes. We compare three learning regimes:

1. **Without knowledge**: $\lambda_T = 0$.
2. **With knowledge**: the correct non-slip operator is enforced,
$$q(s_t, u_t) = (x_{2,t+1} - x_{2,t}) \cos \vartheta_t - (x_{1,t+1} - x_{1,t}) \sin \vartheta_t.$$
3. **Misaligned knowledge**: an incorrect operator with two sources of error is enforced. First, we use a multiplicative distortion of the physics, given by
$$q_{\texttt{mis}}(s_t, u_t) = (x_{2,t+1} - x_{2,t}) \cos \vartheta_t - \beta\, (x_{1,t+1} - x_{1,t}) \sin \vartheta_t.$$
for some $\beta \neq 1$. Second, we use an angle perturbation, where the physics operator is evaluated using a perturbed angle $\vartheta_{\text{mis},t} = \vartheta_t + \delta\vartheta$, with $\delta\vartheta \neq 0$. We obtain $\|\mathscr{D}(f_\star)\|_{\mathscr{L}^2} = 0.15$ with $\delta\vartheta = 0.005$ and $\beta = 1.2$, and $\|\mathscr{D}(f_\star)\|_{\mathscr{L}^2} = 0.50$ with $\delta\vartheta = 0.5$ and $\beta = 0.5$.

Results. The top plot of panel of Figure 3 complements the results displayed in Figure 2 on page 10, presenting the log-log plot of the empirical excess risk (estimation error) as a function of the number of samples T. While Figure 2 displayed models trained with and without physics-informed regularization, we now illustrate the effect of incorrect physical priors on the learning rate. Each curve is obtained again 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. We can observe that, when the degree of misalignment is mild (for instance, when $\|\mathscr{D}(f_\star)\|_{\mathscr{L}^2} = 0.15$), the estimator still substantially benefits from incorporating physics: the decay rate remains faster than in the case without knowledge. In contrast, when the misalignment grows larger (e.g., $\|\mathscr{D}(f_\star)\|_{\mathscr{L}^2} = 0.5$), the advantage of the prior diminishes, and the observed rate approaches the slower Sobolev minimax behavior of the purely data-driven estimator. We summarize the empirical rates for all the tests in Table 2.

The bottom panel of Figure 3 reports the behavior of the regularization parameter λ_T as a function of the sample size T. Consistent with our theory, as the prior becomes increasingly unreliable, λ_T must decrease more aggressively: this prevents the learner from enforcing an inaccurate physical bias too strongly, striking an appropriate balance between data fit and physics regularization.

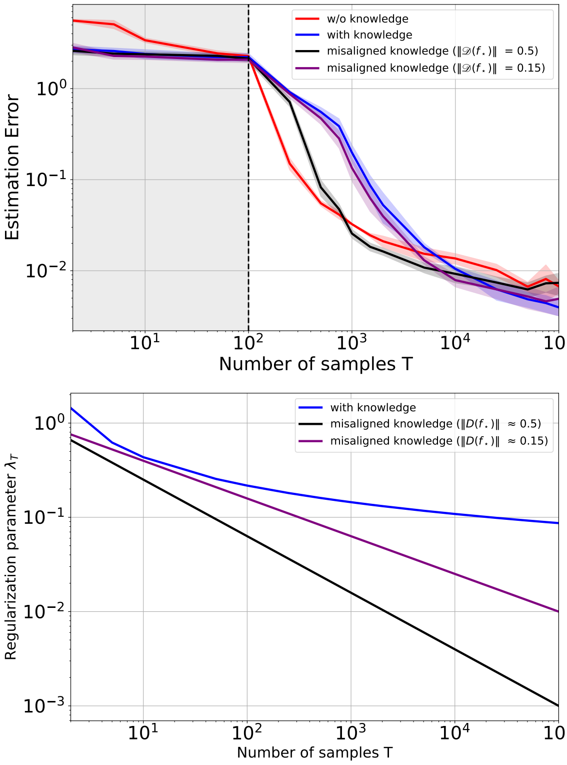

Figure 3: Log–log plots of the empirical excess risk (estimation error) for the unicycle experiment (top panel), shown as a function of the sample size T. The gray shaded region on the left highlights the estimated burn-in period, after which the learning rate becomes apparent. The bottom panel displays the behavior of the regularization parameter λ_T as T increases.

Table 2: Comparison of average learning rates for the unicycle experiment.

Experiment	No-Knowledge	Full-Knowledge $\|\mathscr{D}(f_\star)\|_{\mathscr{L}^2} = 0$	Misaligned Knowledge $\|\mathscr{D}(f_\star)\|_{\mathscr{L}^2} = 0.15$	Misaligned Knowledge $\|\mathscr{D}(f_\star)\|_{\mathscr{L}^2} = 0.5$
Unicycle	$T^{-0.646}$	$T^{-0.993}$	$T^{-0.972}$	$T^{-0.700}$

K.2 Poisson Equation Experiment

We complement the unicycle experiment with an additional numerical study based on a two–dimensional Poisson equation. This example illustrates the same qualitative behavior observed in the unicycle setting: physics-informed regularization accelerates convergence even when the data are dependent.

Model setup. Consider the Poisson equation on the unit square $\Omega = [0,1]^2$:

$$-\Delta f_\star(x,y) = g(x,y), \text{ for } (x,y) \in \Omega,$$

with homogeneous Dirichlet boundary conditions. We choose a ground-truth solution

$$f_\star(x,y) = \sin(\pi x)\sin(\pi y),$$

for which the forcing term is analytically given by

$$g(x,y) = 2\pi^2 \sin(\pi x)\sin(\pi y).$$

The physical operator considered in the regularization term is the Laplace operator $\mathscr{D}(f) = -\Delta f$, which satisfies Assumption 4.

Sequential data generation. To illustrate the dependent-data setting, we generate inputs sequentially. Starting from an initial point (x_0, y_0) drawn uniformly from Ω, we propagate the trajectory according to the stochastic Markovian dynamics

$$(x_{t+1}, y_{t+1}) = (x_t, y_t) + \varepsilon_t, \;\; \varepsilon_t \sim \mathcal{N}(0, \sigma^2_{\texttt{step}} \mathbb{I}_2)$$

with $\sigma^2_{\texttt{step}} = 0.05$ and $\mathbb{I}_2$ is the 2×2 identity matrix. To make the generated data satisfy Assumption 1, we project them onto Ω by clipping each coordinate.
At each time step we observe the noisy evaluation

$$Z_t = f_\star(X_t, Y_t) + W_t,$$

where W_t is an additive Gaussian noise with variance $\sigma_W^2 = 1$.

Adopted estimator. We approximate $f_\star$ with a multilayer perceptron with two hidden layers of width 64 and ReLU activations. We train the estimator using the Adam optimizer with learning rate 0.5×10^{-3}. We vary the number of training samples T over the range $T \in [10, 5 \times 10^4]$. The approximated rate is computed after the burn-in time.

Experiment regimes. We compare three learning regimes:

1. **Without knowledge**: $\lambda_T = 0$.
2. **With knowledge**: use the correct operator $\mathscr{D}(f) = -\Delta f$.
3. **Misaligned knowledge**: enforce the incorrect physics prior

$$-\Delta f = \beta g, \;\; \beta \neq 1,$$

leading to the regularizer

$$\|\mathscr{D}(f_\star)\|_{\mathscr{L}^2} = |1-\beta|\pi^2,$$

where we use $\beta = 0.93$ to obtain $\|\mathscr{D}(f_\star)\|_{\mathscr{L}^2} = 0.7$ and $\beta = 0.8$ for $\|\mathscr{D}(f_\star)\|_{\mathscr{L}^2} = 1.97$.

Results. Figure 4 reports the log–log plot of the empirical excess risk as a function of the number of samples T for the Poisson equation experiment, comparing the learning curves obtained with the three experiment regimes presented above. As in the unicycle experiment, each curve is computed by averaging over 20 independent realizations of the training data, with solid lines showing the empirical mean and shaded regions representing 95% confidence intervals. Consistently with the theoretical predictions in Section 5, the estimator that incorporates the correct PDE structure exhibits a faster decay of the estimation error than the purely data-driven baseline.

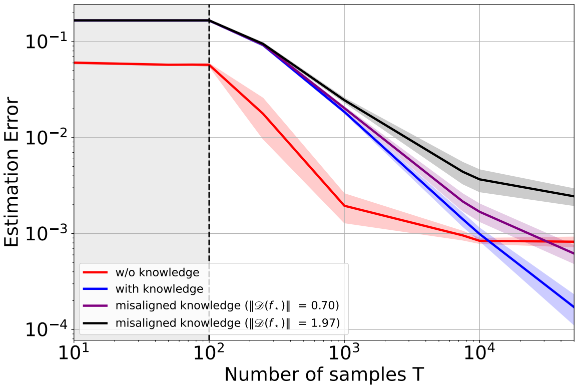

Figure 4: Log–log plot of the empirical excess risk (estimation error) as a function of the sample size T for the Poisson equation experiment. The gray shaded region on the left highlights the estimated burn-in period, after which the learning rate becomes apparent.

Beyond the aligned-knowledge setting, Figure 4 also highlights the effect of misaligned physical priors on the convergence behavior. When the degree of misspecification is moderate, for instance when $\|\mathscr{D}(f_\star)\|_{\mathscr{L}^2} = 0.70$, the learning rate remains significantly improved relative to the regime without knowledge, demonstrating robustness of physics-informed regularization to small prior errors. On the other hand, as the misalignment increases further, e.g., when $\|\mathscr{D}(f_\star)\|_{\mathscr{L}^2} = 1.97$, the benefits of incorporating the PDE prior gradually diminish: the empirical decay rate slows down and approaches the behavior of the estimator trained without physics. This transition mirrors the trends predicted by our theory, where the rate interpolates between the fast i.i.d.-like regime under knowledge alignment and the slower Sobolev minimax rate as the prior becomes increasingly incorrect. For a summary of the obtained empirical average learning rates obtained in the studied scenarios, we defer to Table 3.

Table 3: Comparison of Learning Rates for the Poisson Equation Experiment.

Experiment	No-Knowledge	Full-Knowledge $\|\mathscr{D}(f_\star)\|_{\mathscr{L}^2} = 0$	Misaligned Knowledge $\|\mathscr{D}(f_\star)\|_{\mathscr{L}^2} = 0.70$	Misaligned Knowledge $\|\mathscr{D}(f_\star)\|_{\mathscr{L}^2} = 1.95$
Poisson	$T^{-0.741}$	$T^{-1.105}$	$T^{-0.935}$	$T^{-0.733}$