http://www.diva-portal.org

This is the published version of a paper presented at *35th Conference on Learning Theory, COLT 2022, London, United Kingdom of Great Britain and Northern Ireland, Jul 2 2022 - Jul 5 2022*.

Citation for the original published paper:

Ziemann, I., Sandberg, H., Matni, N. (2022)
Single Trajectory Nonparametric Learning of Nonlinear Dynamics
In: *Proceedings of 35th Conference on Learning Theory, COLT 2022* (pp. 3333-3364).
ML Research Press

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-334444

Single Trajectory Nonparametric Learning of Nonlinear Dynamics

Ingvar Ziemann ZIEMANN@KTH.SE
KTH Royal Institute of Technology

Henrik Sandberg HSAN@KTH.SE
KTH Royal Institute of Technology

Nikolai Matni NMATNI@SEAS.UPENN.EDU
University of Pennsylvania

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

Given a single trajectory of a dynamical system, we analyze the performance of the nonparametric least squares estimator (LSE). More precisely, we give nonasymptotic expected l^2-distance bounds between the LSE and the true regression function, where expectation is evaluated on a fresh, counterfactual, trajectory. We leverage recently developed information-theoretic methods to establish the optimality of the LSE for nonparametric hypotheses classes in terms of supremum norm metric entropy and a subgaussian parameter. Next, we relate this subgaussian parameter to the stability of the underlying process using notions from dynamical systems theory. When combined, these developments lead to rate-optimal error bounds that scale as $T^{-1/(2+q)}$ for suitably stable processes and hypothesis classes with metric entropy growth of order δ^{-q}. Here, T is the length of the observed trajectory, $\delta \in \mathbb{R}_+$ is the packing granularity and $q \in (0, 2)$ is a complexity term. Finally, we specialize our results to a number of scenarios of practical interest, such as Lipschitz dynamics, generalized linear models, and dynamics described by functions in certain classes of Reproducing Kernel Hilbert Spaces (RKHS).

1. Introduction

Consider a time-series model of the form

$$y_t = f_\star(x_t) + w_t, \quad t = 0, \dots, T-1. \tag{1}$$

where $f_\star$ is unknown but belongs to some known function class $\mathbb{F}$. Suppose a learner is given access to samples $(x_0, \dots, x_{T-1}, y_0, \dots, y_{T-1})$, corrupted by noise $(w_0, \dots, w_{T-1})$, from a single trajectory generated by model (1). In this work, we show that the nonparametric least squares estimator (LSE) converges to the ground truth $f_\star$ at the minimax optimal rate. In the i.i.d. setting, in which each observation from the model (1) is drawn independently at random, the optimal rate is $T^{-1/2}$ for parametric models. In the nonparametric setting, this rate degrades gracefully to $T^{-1/(2+q)}$ for models with metric entropy scaling as δ^{-q}, $q \in (0, 2)$ (Tsybakov, 2009). In this paper, we show that these rates can be matched for a class of more general time-series models of the form (1). We note in particular that by setting $y_t = x_{t+1}$ in model (1), we recover the nonlinear stochastic dynamical system

$$x_{t+1} = f_\star(x_t) + w_t. \tag{2}$$

© 2022 I. Ziemann, H. Sandberg & N. Matni.

Estimation of models (1) and (2) remains relatively poorly understood when the data is not i.i.d., with existing results being limited to when the dynamics are known to belong to certain parametric classes. In terms of parameter recovery, the LSE converges at a rate of $T^{-1/2}$ for stable linear autoregressive systems $f_\star(x_t) = A_\star x_t$ (Simchowitz et al., 2018; Sarkar and Rakhlin, 2019; Jedra and Proutiere, 2020). The same rate can also be achieved for linear systems with more general input-output behavior (Oymak and Ozay, 2019; Tsiamis and Pappas, 2019). Moving to nonlinear models, recursive and gradient type algorithms can be shown to converge at a rate of $T^{-1/2}$ for the generalized linear model $f_\star(x_t) = \phi(A_\star x_t)$, where ϕ is a known Lipschitz link function (Foster et al., 2020; Sattar and Oymak, 2020; Jain et al., 2021). In Ziemann and Sandberg (2022), we significantly generalize these results and provide rate-optimal error bounds for nonparametric function classes in terms of their metric entropies. Our approach leverages recently developed information-theoretic tools (Russo and Zou, 2019; Xu and Raginsky, 2017) and the notion of offset complexity (Rakhlin and Sridharan, 2014; Liang et al., 2015).

Problem Formulation The dynamics (1) evolve on two subsets of Euclidean space: $\mathsf{X} \subset \mathbb{R}^{d_x}$ with $x_t \in \mathsf{X}$ and $\mathsf{Y} \subset \mathbb{R}^{d_y}$ with $y_t \in \mathsf{Y}$. We assume that there exists an increasing sequence of σ-fields $\{\mathcal{F}_t\}_{t\in\mathbb{Z}_{\geq -1}}$ such that each x_t is $\mathcal{F}_{t-1}$-measurable, each w_t is $\mathcal{F}_t$-measurable and $\mathbf{E}[w_t|\mathcal{F}_{t-1}] = 0$. In other words, $\{w_t\}_{t\in\mathbb{Z}_{\geq 0}}$ is a martingale difference sequence with respect to the filtration $\{\mathcal{F}_t\}_{t\in\mathbb{Z}_{\geq 0}}$ and $\{x_t\}_{t\in\mathbb{Z}_{\geq 0}}$ is adapted to $\{\mathcal{F}_{t-1}\}_{t\in\mathbb{Z}_{\geq -1}}$. Furthermore, we assume that each w_t is conditionally σ_w^2-subgaussian given $\mathcal{F}_{t-1}$. We denote by $P_\star$ the joint distribution of $Z = (x_0, \ldots, x_{T-1}, y_0, \ldots, y_{T-1})$, with x_t and y_t as in system (1). We assume that $f_\star$ is unknown but that it belongs to a known metric space $(\mathbb{F}, \rho)$ with $\rho(f, g) = \sup_{x\in\mathsf{X}} \|f(x) - g(x)\|_2$. See Section 1.3 for further preliminaries.

Given this, the learning task is to produce an estimate $\hat{f}$ of the model $f_\star$, which is evaluated in terms of the expected Euclidean 2-norm error:

$$\mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2. \tag{3}$$

The expectation (3) is computed with respect to the randomness in the algorithm and the random variable $\xi \sim \nu$ which is independent of all other randomness; here ν is the uniform mixture over $(x_0, \ldots, x_{T-1})$. That is, ξ has the same distribution as random variable x_τ, where the index τ is drawn uniformly at random over $\{0, \ldots, T-1\}$ and is independent of $(x_0, \ldots, x_{T-1})$. If the process (1) is stationary with invariant measure ν, this reduces to the assumption that ξ is drawn from the invariant measure of the process.

In the sequel, we analyze the nonparametric least squares estimator (LSE) $\hat{f}$ of $f_\star$, given below. Namely, we assume that the learner can compute

$$\hat{f} \in \operatorname{argmin}_{f\in\mathbb{F}} \left\{ \frac{1}{T} \sum_{t=0}^{T-1} \|y_t - f(x_t)\|_2^2 \right\}. \tag{4}$$

1.1. Contributions

We derive error bounds for learning nonlinear dynamics (2) and the more general time-series model (1). Theorem 1 provides bounds on $\mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2$ for the LSE (4). These bounds depend only on the metric entropy of the hypothesis class, $\log \mathcal{N}(\mathbb{F}, \|\cdot\|_\infty, \delta)$, the noise level σ_w^2, and a further variance proxy $\sigma_T^2(\mathbb{F}, P_\star)$, measuring the spatiotemporal spread of the covariates relative to

the function class $\mathbb{F}$. Informally, our main result, Theorem 1, states that

$$\mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2 \lesssim \left(\frac{\text{dimensional factors} \times \text{measurement noise}}{\text{trajectory length}}\right)^{1/(2+\text{complexity term})} + \left(\frac{\text{dimensional factors} \times \text{covariates spread}}{\text{trajectory length}}\right)^{1/(2+\text{complexity term})} \tag{5}$$

where the dimensional factors and the complexity term depends on the scaling of the metric entropy $\log \mathcal{N}(\mathbb{F}, \|\cdot\|_\infty, \delta)$ for small $\delta > 0$. Our bounds exhibit optimal scaling in terms of interaction between trajectory length and function class complexity in that they agree with known minimax optimal rates for the i.i.d. setting (Tsybakov, 2009). To the best of our knowledge, these are the first such bounds that are applicable to large nonparametric function classes in the temporally correlated (non-i.i.d.) setting.

Arriving at bounds of the form (5) for temporally correlated data is challenging, as the symmetrization technique typically used to analyze both generalization and training error for regression cannot be applied. Instead, we leverage information-theoretic decoupling arguments introduced by Russo and Zou (2019) and Xu and Raginsky (2017) to reduce the analysis of the error (3) to bounding an in-sample prediction error (training error) and a term measuring the dependence between the sample and the algorithm (generalization error). We analyze the first term using the offset basic inequality (Rakhlin and Sridharan, 2014; Liang et al., 2015). Crucially, this leads to a simplified localization argument that is amenable to modification for correlated data and allows us to obtain fast rates. The second term of the decoupling estimate is bounded by the mutual information between the algorithm and the sample, which we control via metric entropy and discretization. Moreover, the scale of this term is controlled by the spatiotemporal variance proxy, $\sigma_T^2(\mathbb{F}, P_\star)$.

In the information-theoretic generalization bounds literature (Xu and Raginsky, 2017), the term $\sigma_T^2(\mathbb{F}, P_\star)$ is referred to as the subgaussian parameter of the loss function. Here, given the form (1) of the data generating process and our choice of performance metric, the variance proxy $\sigma_T^2(\mathbb{F}, P_\star)$ admits a more direct interpretation in terms of the stability of the process (2). Namely, by an explicit stability argument we show that $\sigma_T^2(\mathbb{F}, P_\star) \lesssim 1/((1-L_\star)^2 T)$ whenever the autoregressive system (2) is $L_\star$-contractive (Proposition 4). We also show that $\sigma_T^2(\mathbb{F}, P_\star) \lesssim 1/T$ holds more generally whenever the covariates of the process (1) form a Markov chain with finite mixing time (Proposition 5). Given the recent line of work emphasizing the role of control-theoretic stability for learning in dynamical systems (Foster et al., 2020; Boffi et al., 2021; Tu et al., 2021), we believe that this is an attractive construction that may have wider applicability[1].

1.2. Further Related Work

Estimation of models of the form (1) and (2) has a rich history in statistics and system identification (Ljung, 1999). Preceding the recent body of work mentioned in the introduction, asymptotically optimal rates for linear stochastic models have been available for some time (Mann and Wald, 1943; Lai and Wei, 1982). Similarly, there is a well-established theory of rate-optimal identification for nonlinear parametric models under various identifiability-type conditions, both in the i.i.d. setting (Van der Vaart, 2000) and under more general assumptions (Le Cam, 2012).

1. See Appendix C.1 for a discussion on how our results apply to generalization bounds for Lipschitz losses.

Perhaps the main motivator for the recent line of work emphasizing nonasymptotic estimation bounds is that these bounds are applicable downstream in control and reinforcement learning pipelines. Regression estimates for linear stochastic systems have been key to understanding both online and offline reinforcement learning in the linear quadratic regulator (Dean et al., 2020; Mania et al., 2019) and can be shown to lead to optimal regret rates (Simchowitz and Foster, 2020; Ziemann and Sandberg, 2022). Extending our understanding of the interaction between learning and control beyond linear-in-the-parameters models (Kakade et al., 2020; Boffi et al., 2021; Lale et al., 2021) inevitably requires new analyses of learning in dynamical systems. This also motivates the present work in that we provide nonasymptotic and counterfactual control of the LSE's estimation error for more general nonlinear and nonparametric models.

Another related field is that of general statistical learning for dependent data, see Agarwal and Duchi (2012), Kuznetsov and Mohri (2017) and the references therein. These works provide generalization bounds for general loss functions and β-mixing processes (x_t, y_t). The assumption of β-mixing processes has also previously been exploited in parametric identification by Vidyasagar and Karandikar (2006). By contrast, our emphasis on regression over general learning is motivated by downstream applications in learning-enabled control, where one first learns a model used to design a controller. Necessarily then, this work builds on a rich line of work in nonparametric regression for the i.i.d. setting, see chapters 13 and 14 of Wainwright (2019) and the references therein. Although there has been some work on the dependent setting, the error in this line of work is typically computed with respect to the design points which is not suitable for the counterfactual reasoning that is key to control (Baraud et al., 2001).

We also draw inspiration from the recent line of work on information-theoretic generalization bounds (Xu and Raginsky, 2017; Russo and Zou, 2019; Bu et al., 2020). There are interesting refinements and variations of this theory using for instance conditional mutual information (Steinke and Zakynthinou, 2020), or Wasserstein distance (Gálvez et al., 2021). However, these more recent bounds rely more explicitly on the tenzorization properties of information measures under i.i.d. data than the earlier work of Russo and Zou (2019) and Xu and Raginsky (2017), and so are not directly amenable to the single trajectory setting. We also note that information-theoretic generalization bounds have previously found other applications, such as in the analysis of stochastic gradient descent (Neu et al., 2021).

1.3. Preliminaries and Notation

All logarithms used in this paper are base e. For two probability measures $\mathbf{P}$ and $\mathbf{Q}$ we denote by $D(\mathbf{P}\|\mathbf{Q}) = \int \log \frac{d\mathbf{P}}{d\mathbf{Q}} d\mathbf{P}$ their Kullback-Leibler divergence and their total variation distance by $d_{TV}(\mathbf{P}, \mathbf{Q}) = \frac{1}{2}\int |d\mathbf{P} - d\mathbf{Q}|$. For a random variable X we denote its law by P_X, that is $X \sim P_X$. For two random variables X and Y, $(X, Y) \sim P_{X,Y}$, we denote by $I(X;Y) = D(P_{X,Y}\|P_{X\otimes Y})$ their mutual information, where $P_{X\otimes Y} = P_X \otimes P_Y$ denotes the product measure of the marginal distributions of X and Y. If X is a random variable over a finite alphabet $\{1, \dots, M\}$, we denote by $H(X)$ its Shannon entropy which is given by $H(X) = -\sum_{j=1}^{M} [\log P_X(j)] P_X(j)$. Generic expectation (integration with respect to all randomness) is denoted by $\mathbf{E}$. A random variable X taking values in $\mathbb{R}^d$ is said to be σ^2-subgaussian if $\mathbf{E} \exp \lambda \langle X - \mathbf{E}X, s\rangle \leq \exp \lambda^2\sigma^2/2$ for all $s \in \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ is the unit sphere and $\langle \cdot, \cdot \rangle$ is the standard Euclidean inner product. This extends to conditional subgaussianity via conditional expectation, $\mathbf{E}[\cdot|\mathcal{F}]$, with respect to a σ-field $\mathcal{F}$, if the same holds with expectation $\mathbf{E}$ exchanged for $\mathbf{E}[\cdot|\mathcal{F}]$.

Let $(\mathbb{F}, \rho)$ be a metric space. We define its δ-covering number $\mathcal{N}(\mathbb{F}, \rho, \delta)$ as the cardinality of the smallest δ-cover of $\mathbb{F}$ in the metric ρ. In this case we say that $\mathbb{F}$ has metric entropy $\log \mathcal{N}(\mathbb{F}, \rho, \delta)$. If no such covering exists we write $\log \mathcal{N}(\mathbb{F}, \rho, \delta) = \infty$. Recall that $P_\star$ denotes the distribution of $Z = (x_0, \dots, x_{T-1}, y_0, \dots y_{T-1})$ under the dynamics (1). If

$$\sigma_T^2(\mathbb{F}, P_\star) \triangleq \inf \left\{ \sigma^2 : \frac{1}{T} \sum_{t=0}^{T-1} \|f(x_t) - g(x_t)\|_2 \text{ is } \sigma^2\text{-subgaussian under } P_\star \text{ for all } f, g \in \mathbb{F} \right\}$$

is finite, we say that the space $\mathbb{F}$ is $\sigma_T^2(\mathbb{F}, P_\star)$-subgaussian with respect to the dynamics (1). We shall make the assumption that $\mathbb{F}$ is $\sigma_T^2(\mathbb{F}, P_\star)$-subgaussian. Importantly, this implies that all the centered functions $f - f_\star$ are subgaussian. Observe that this a property defined both in terms of the space $\mathbb{F}$ and the system (1). As the variable $f - f_\star$ will appear frequently throughout the text, it will be convenient to define $\mathbb{F}_\star \triangleq \mathbb{F} - f_\star$. Further, it will be useful for purposes of analysis to quantize the estimate $\hat{f}$ given by the LSE (4). For $\mathbb{F}_\delta = \{f_1, \dots, f_M\}$ an optimal δ-covering of $\mathbb{F}$, we define $\hat{f}_\delta$ to be a quantization of the LSE as follows

$$\hat{f}_\delta \in \operatorname{argmin}_{f \in \mathbb{F}_\delta} \rho(f, \hat{f}). \tag{6}$$

The following shorthand notation will also be used to ease the exposition: we write $a_t \lesssim b_t$ if there exists a universal constant C such that $a_t \leq C b_t$ for every $t \geq t_0$ and some $t_0 \in \mathbb{N}$. If $a_t \lesssim b_t$ and $b_t \lesssim a_t$ we write $a_t \asymp b_t$. The same convention applies for functions of δ, the parameter of metric entropy, instead of t, but in the small δ regime (typically δ will be in inverse proportion to some increasing function of t).

2. Results

Our main result is an error bound that controls the distance between the estimate $\hat{f}$, defined by the nonparametric LSE (4), and the ground truth $f_\star$ in terms of a fresh sample ξ drawn independently of the algorithm from the mixture distribution over the samples $(x_0, \dots, x_{T-1})$.

Theorem 1 *Fix a metric space $(\mathbb{F}, \rho)$ with $\rho(f, g) = \sup_{x \in \mathsf{X}} \|f(x) - g(x)\|_2$ and assume that $f_\star \in \mathbb{F}$. Then for any $\delta > 0, \gamma > 0$ and $\alpha \in [0, \gamma]$, the LSE (4), satisfies*

$$\begin{aligned} \mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2 \leq \sqrt{\frac{8\sigma_w^2 \log \mathcal{N}(\mathbb{F}, \rho, \gamma)}{T} + 128\alpha\sigma_w\sqrt{d_y} + 64\frac{\sigma_w}{\sqrt{T}} \int_\alpha^\gamma \sqrt{\log \mathcal{N}(\mathbb{F}, \rho, s)} ds} \\ + 3\delta + \sqrt{2\sigma_T^2(\mathbb{F}, P_\star) \log \mathcal{N}(\mathbb{F}, \rho, \delta)}. \end{aligned} \tag{7}$$

To establish inequality (7), we rely on an information-theoretic decoupling argument given in Proposition 6. Informally, Proposition 6 allows us to decompose the error as

$$\mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2 \lesssim \text{training error}(\gamma, \alpha, T) + \text{generalization error}(\delta, \sigma_T^2(\mathbb{F}, P_\star)).$$

The training error term above is controlled by the offset basic inequality (Lemma 7). Discretizing and proceeding through chaining yields a maximal inequality with α and γ as trade-off parameters.

The generalization error term above is defined in terms of the discretization parameter δ, which controls the discrepancy between the quantized model $\hat{f}_\delta$ (defined in (6)) and the LSE $\hat{f}$. The quantized model $\hat{f}_\delta$, used solely in the proof, "generalizes well" since it belongs to a finite hypothesis class by construction.

While we typically set $\alpha = 0$ in Theorem 1, the optimal choices of δ and γ depend on a critical balance: to arrive at an optimal bound we must balance the complexity of the hypothesis class, $\mathbb{F}$, through its metric entropy, with statistical properties of the model (1), such as the sampling length T, the noise amplitude σ_w^2 and the variance proxy $\sigma_T^2(\mathbb{F}, P_\star)$. The full proof of Theorem 1 can be found in Appendix A and a more detailed outline is given in Section 3.

To make the consequences of Theorem 1 more explicit, we consider two different complexity regimes for $\mathbb{F}$. If there exist $p, q \in \mathbb{R}_+$ such that

$$\log \mathcal{N}(\mathbb{F}, \|\cdot\|_\infty, \delta) \lesssim p \left(\frac{1}{\delta}\right)^q \tag{8}$$

we are in the nonparametric regime. If instead there exist $p, c \in \mathbb{R}_+$ such that

$$\log \mathcal{N}(\mathbb{F}, \|\cdot\|_\infty, \delta) \lesssim p \log\left(1 + \frac{c}{\delta}\right) \tag{9}$$

we are in the parametric regime. Concrete examples of processes satisfying conditions (8) or (9) are given in Section 4. Under the hypothesis (8) we may solve for the critical radii $\gamma \asymp \sqrt{\frac{\log \mathcal{N}(\mathbb{F}, \rho, \gamma)}{T}}$ and $\delta \asymp \sqrt{\frac{\log \mathcal{N}(\mathbb{F}, \rho, \delta)}{\sigma_T^2(\mathbb{F}, P_\star)}}$. This leads to the following result.

Theorem 2 (Nonparametric Rates) *Under the hypotheses of Theorem 1 and if further inequality (8) holds for $p, q \in \mathbb{R}_+$ with $q < 2$, the least squares estimator (4) satisfies*

$$\mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2 \lesssim \sqrt{\frac{1}{2-q}} \left(\frac{p\sigma_w^2}{T}\right)^{\frac{1}{2+q}} + \left(p\sigma_T^2(\mathbb{F}, P_\star)\right)^{\frac{1}{2+q}}. \tag{10}$$

A similar statement holds for $q \geq 2$ and can be found following the proof of Theorem 2 in Appendix A.1. We also have a version of the above theorem, proven in Appendix A.2, applicable to the parametric entropy growth regime. Due to the appearance of the logarithm in inequality (9), γ and δ can be chosen almost arbitrarily (but sub-exponentially) small. This leads to following result.

Theorem 3 (Parametric Rates) *Under the hypotheses of Theorem 1 and if further inequality (9) holds for $p, c \in \mathbb{R}_+$, the least squares estimator (4) satisfies*

$$\mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2 \lesssim \sqrt{\frac{\sigma_w^2 p \log(1 + c\sqrt{d_y}\sigma_w T^2)}{T}} + \sqrt{\sigma_T^2(\mathbb{F}, P_\star) p \log(1 + cT)} + \frac{1}{T}.$$

The error estimates given in Theorems 2 and 3 are rate-optimal[2] in terms of T whenever the generalization term satisfies $\sigma_T^2(\mathbb{F}, P_\star) \lesssim 1/T$. We show that this rate of decay of the subgaussian parameter $\sigma_T^2(\mathbb{F}, P_\star)$ holds for instance when the autoregressive dynamics (2) are contracting or more generally when the process (1) is mixing. In some sense, $\sigma_T^2(\mathbb{F}, P_\star)$ is a measure of the magnitude of the process (1) and its correlation length. We develop this idea next in Section 2.1.

2. Modulo a logarithmic term for the parametric regime.

2.1. Sufficient Conditions for Generalization: Stability and Mixing

As noted above, we crucially need conditions for which the spatiotemporal variance proxy satisifes $\sigma_T^2(\mathbb{F}, P_\star) \lesssim 1/T$. While this scaling typically holds for i.i.d. data, we show that it also holds for temporally correlated data arising from a single trajectory under suitable stability or mixing assumptions on the dynamics (1).

Contracting systems When working with Lipschitz systems of the form (2) one can relate the parameter $\sigma_T^2(\mathbb{F}, P_\star)$ to the stability of the map $f_\star$. Fix a norm $\|\cdot\|_{\mathsf{X}}$ on X. We say that $f_\star$ is $(L_\star, \|\cdot\|_{\mathsf{X}})$-contractive if for some $L_\star < 1$, we have that $\|f(x) - f(z)\|_{\mathsf{X}} \leq L_\star \|x - y\|_{\mathsf{X}}$ for all $x, z \in \mathsf{X}$.

Proposition 4 (Contraction Implies Generalization) *Suppose that we are in the autoregressive setting (2), that $f_\star$ is $(L_\star, \|\cdot\|_{\mathsf{X}})$-contractive, and that all functions $f \in \mathbb{F}$ are L-Lipschitz with respect to $\|\cdot\|_2$. If further the noise process $\{w_t\}_{t \in \mathbb{Z}_{\geq 0}}$ is independent and $\sup_{x \in \mathsf{X}} \|x\|_2 \leq B$, it holds that*

$$\sigma_T^2(\mathbb{F}, P_\star) \leq 64 \cdot \frac{M^2 B^2 L^2}{m^2 (1 - L_\star)^2 T}$$

where

$$M = \sup_{x \neq x'} \frac{\|x - x'\|_2}{\|x - x'\|_{\mathsf{X}}} \qquad \text{and} \qquad m = \sup_{x \neq x'} \frac{\|x - x'\|_{\mathsf{X}}}{\|x - x'\|_2}.$$

The proof of Prop. 4 combines an Azuma-McDiarmid-Hoeffding like argument with a stability argument, and can be found in Appendix B. Proposition 4 allows us to further simplify the bound (10) whenever $f_\star$ is contractive. Namely, when $\mathbb{F}$ is a bounded subset of L-Lipschitz functions with metric entropy scaling as $p\delta^{-q}$ and $q < 2$, the bound in Theorem 2 for the autoregressive system (2) becomes

$$\mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2 \lesssim \sqrt{\frac{1}{2-q}} \left(\frac{\sigma_w^2 p}{T}\right)^{\frac{1}{2+q}} + \left(\frac{M^2 B^2 L^2}{m^2 (1 - L_\star)^2 T}\right)^{\frac{1}{2+q}}.$$

This bound shows that more stable systems, as captured by the Lipschitz constant $L_\star$, have smaller generalization error. This interpretation is in line with the recent trend of using stability bounds to study learning algorithms applied to data generated by a dynamical system, see for example Boffi et al. (2021) and Tu et al. (2021). Finally we note that although we restricted our analysis to contracting systems, our results are easily extended to a more general notion of nonlinear stability. In particular, we extend Proposition 4 in Appendix B.2 to systems satisfying a notion of exponential incremental input-to-state stability, a standard notion from robust nonlinear control theory (Angeli, 2002).

We also point out that in the autoregressive setting $f_\star$ has two roles: it is 1) the unknown function we wish to identify but it also 2) affects the mixing behavior of the covariates $\{x_t\}_{t \in \mathbb{Z}_{\geq 0}}$. The second point imposes the constraint that the unknown function $f_\star$ be contractive $L_\star < 1$. Such an assumption is not necessary in related work (Mohri and Rostamizadeh, 2010), nor is it in the general setting (1) if one by some other means is able to verify the stochastic stability of $\{x_t\}_{t \in \mathbb{Z}_{\geq 0}}$. Alternatively then, one may prefer to work with a stochastic notion of stability. This option is discussed next.

Mixing systems We now demonstrate that our approach is equally applicable to mixing systems. To this end, we recall the definition of a mixing time: if $\{x_t\}_{t\in\mathbb{Z}_{\geq 0}}$ is a Markov chain with transition kernel $P(x,\cdot)$ and invariant measure ν_∞, its mixing time t_{mix} is given by

$$t_{\mathsf{mix}} \triangleq \min\{t \in \mathbb{N} : \sup_{x\in X} d_{TV}(P^t(x,\cdot),\nu_\infty) \leq 1/4\}.$$

Equipped with this notion, the following result can be inferred from Paulin (2015) (see Definition 1.3 and Corollary 2.10 therein).

Proposition 5 (Mixing Implies Generalization) *Assume that* $\sup_{y\in\mathsf{Y}} \|y\|_2 \leq B$. *If the sequence* $\{x_t\}_{t\in\mathbb{Z}_{\geq 0}}$ *in system (1) is a* t_{mix}*-mixing Markov chain, then the class* $\mathbb{F}$ *is* $\sigma_T^2(\mathbb{F},P_\star)$*-subgaussian with*

$$\sigma_T^2(\mathbb{F},P_\star) \lesssim \frac{B^2 t_{\mathsf{mix}}}{T}.$$

Comparing Proposition 5 with Proposition 4 we see that we obtain similar bounds on the subgaussian parameter $\sigma_T^2(\mathbb{F},P_\star)$. While Proposition 4 is easier to prove and has a direct interpretation in terms of the model (2), Proposition 5 has the advantage of being equally applicable to both the more general time series model (1) and the dynamical system (2).

2.2. Summary

Our results show that a large class of nonlinear systems can be learned at the minimax optimal nonparametric rate $T^{-1/(2+q)}$ using the LSE (4). This significantly extends our current understanding of nonasymptotic learning of dynamical systems from single trajectory data. By contrast, previous work assumes i.i.d. data or focuses on either linear models (Simchowitz et al., 2018; Tsiamis and Pappas, 2019; Jedra and Proutiere, 2020) or parametric models with known nonlinearities (Foster et al., 2020; Sattar and Oymak, 2020; Mania et al., 2020; Jain et al., 2021).

3. Proof Strategy for Theorem 1

Our analysis of the least squares estimator (4) begins with the following information-theoretic decoupling estimate inspired by Russo and Zou (2019) and Xu and Raginsky (2017).

Proposition 6 *Let* f *and* g *be random variables (functions) taking values in* $\mathbb{F}$. *If* $\mathbb{F}$ *is* $\sigma_T^2(\mathbb{F},P_\star)$*-subgaussian with respect to dynamics (1), we have that*

$$\mathbf{E}\|f(\xi)-g(\xi)\|_2 \leq \sqrt{\frac{1}{T}\mathbf{E}\sum_{t=0}^{T-1}\|f(x_t)-g(x_t)\|_2^2} + \sqrt{2\sigma_T^2(\mathbb{F},P_\star)I((f,g);Z)}, \tag{11}$$

where $Z = (x_0,\ldots,x_{T-1},y_0,\ldots,y_{T-1})$ *and where* ξ *has uniform mixture distribution over the covariates* $(x_0,\ldots,x_{T-1})$ *of system (1) and is independent of all other randomness.*

The proof of the estimate (11) relies on the Donsker-Varadhan variational representation of relative entropy and is given in Appendix C.

To arrive at Theorem 1 we set $f = \hat{f}_\delta$, the discretized least squares estimator (6), and $g = f_\star$ in inequality (11). Now, the discretized estimator $\hat{f}_\delta$ behaves similarly to $\hat{f}$ since the covering $\mathbb{F}_\delta$ to which $\hat{f}_\delta$ belongs is with respect to the uniform metric $\|\cdot\|_\infty$; that is $\|\hat{f}_\delta - \hat{f}\|_\infty \leq \delta$. Exploiting this similarity in behavior between $\hat{f}$ and $\hat{f}_\delta$ yields a bound of the form

$$\mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2 \lesssim \sqrt{\frac{1}{T}\mathbf{E}\sum_{0=1}^{T-1}\|\hat{f}(x_t) - f_\star(x_t)\|_2^2} + \delta + \sqrt{2\sigma_T^2(\mathbb{F}, P_\star)I(\hat{f}_\delta; Z)}. \tag{12}$$

The advantage of inequality (12) over directly choosing $f = \hat{f}$ in inequality (11) is that the mutual information term in (12) is with respect to $\hat{f}_\delta$ instead of $\hat{f}$. By finiteness of $\mathbb{F}_\delta$ this mutual information term is readily controlled by the metric entropy: $I(\hat{f}_\delta; Z) \leq \log \mathcal{N}(\mathbb{F}, \|\cdot\|_\infty, \delta)$. This yields the second term appearing in inequality (7) of Theorem 1 which controls the generalization performance of the discretized estimator $\hat{f}_\delta$. It remains to control the first term appearing on right of inequality (12).

3.1. Offset Basic Inequality Analysis

We now describe our analysis of the in-sample prediction (or training) error, namely the first term appearing on the right hand side of inequality (12). We start with an inequality due to Liang et al. (2015), which is a variant of the basic inequality of least squares and that is crucial to analyzing the in-sample prediction error for correlated data.

Lemma 7 *For the system (1) the LSE (4) satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\hat{f}(x_t) - f_\star(x_t)\|_2^2 \leq \frac{1}{T}\sup_{f\in\mathbb{F}_\star}\sum_{t=0}^{T-1} 4\langle w_t, f(x_t)\rangle - \|f(x_t)\|_2^2. \tag{13}$$

Inequality (13) implies that it suffices to control the family of tilted random walks with increments $4\langle w_t, f(x_t)\rangle - \|f(x_t)\|_2^2$. The right hand side of equation (13) is the supremum of a stochastic process over $\mathbb{F}_\star$. This becomes more clear if we define

$$M_T(f) \triangleq \sum_{t=0}^{T-1} 4\langle w_t, f(x_t)\rangle - \|f(x_t)\|_2^2, \tag{14}$$

which for each fixed T is a real-valued process over $\mathbb{F}_\star$. The supremum of the process $M_T(f)$ in (14) can be viewed as a self-normalized version of the (subgaussian) complexity of $\mathbb{F}_\star$; as the next lemma shows, regardless of the member f and the time-horizon T, its scale is always unity.

Lemma 8 *For any function space $\mathbb{F}$, any $f \in \mathbb{F}_\star$ and $\lambda \in [0, 1/2\sigma_w^2]$ we have that*

$$\mathbf{E}\exp\left(\lambda M_T(f)\right) \leq 1.$$

This leads to the maximal inequality (15) of Lemma 9 below.

Lemma 9 *Let S be a finite subset of the shifted metric space $\mathbb{F}_\star$. Then*

$$\mathbf{E}\sup_{f\in S} M_T(f) \leq 2\sigma_w^2 \log|S|. \tag{15}$$

As observed by Liang et al. (2015), if we had not included the offset term $-\|f(x_t)\|^2$, a naive bound would have yielded $\mathbf{E}\sup_{f\in S} M_T(f) \lesssim \sqrt{T\ln|S|}$, penalizing us by a factor $\sqrt{T}$ for the scale of $\sum_{t=0}^{T-1}\langle w_t, f(x_t)\rangle$.

If we are given a finite class $|\mathbb{F}| < \infty$, we can take $\delta = 0$ in (12) and Lemma 9 in combination with Proposition 6 directly give generalization bounds with optimal dependency on $\log|\mathbb{F}|$, see Appendix A.4 for details. For large spaces with metric structure, we combine this analysis with discretization and chaining to arrive at Theorem 1, see Appendix A.

4. Applications of Theorem 1

Having outlined the proof of Theorem 1, this section is devoted to various examples, demonstrating that we obtain tight rates. As a first example, consider the learning Lipschitz dynamics on $\mathbb{R}$. Combining Theorem 2 with Proposition 4, and that 1-dimensional 1-Lipschitz functions exhibit entropy growth $\log\mathcal{N}(\mathbb{F},\rho,\delta) \lesssim \frac{1}{\delta}$ the following is immediate.

Example 1 *Suppose that we are in the autoregressive setting (2). If $\mathbb{F} = \{f : \mathbb{R} \to [0,1]; f(0) = 0$ and each f is 1-Lipsc with $f_\star \in \mathbb{F}$ being $L_\star$-Lipschitz, $L_\star < 1$, then the least squares estimator (4) achieves*

$$\mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2 \lesssim \left[\frac{1}{(1-L_\star)^{2/3}} + \sigma_w^{2/3}\right] T^{-1/3}.$$

The factor $T^{-1/3}$ is optimal, and appears even in the i.i.d. case, see for example (Wainwright, 2019, Ch. 13).

More generally, growth rates of the form $\log\mathcal{N}(\mathbb{F}_\star,\rho,\delta) \lesssim p\left(\frac{1}{\delta}\right)^q$, as considered in Theorem 2, are typical for spaces comprised of smooth functions as computed in (Kolmogorov and Tikhomirov, 1961, Sec. 5). For instance, it can be shown that the space $C^\beta(\mathsf{X}\to\mathsf{Y})$ of β-times differentiable functions between connected compact subsets of Euclidean space X and Y can be bounded as[3]

$$\log\mathcal{N}(C^\beta(\mathsf{X}\to\mathsf{Y}), \|\cdot\|_\infty, \delta) \lesssim d_y^{1+\frac{d_x}{2\beta}}\left(\frac{1}{\delta}\right)^{\frac{d_x}{\beta}}.$$

In the following example, we consider function spaces $\mathbb{F}$ specified by a Reproducing Kernel Hilbert Space (RKHS) satisfying an eigenvalue decay condition.

Example 2 *Consider model (1) with $d_y = 1$, and suppose that $f_\star$ is known to belong to some RKHS $\mathbb{H}$. More precisely, we fix a probability measure $\mathbf{P}$ on a compact set X and let $K : \mathsf{X}\times\mathsf{X}\to\mathbb{R}$ be a differentiable positive semidefinite kernel function. Assume K has eigenexpansion $K(x,z) = \sum_{i=1}^\infty \lambda_i\phi_i(x)\phi_i(z)$ where $\{\phi_i\}_{i=1}^\infty$ is an orthonormal basis of $L^2(\mathbf{P})$, and where $\{\lambda_i\}_{i=1}^\infty$ is a sequence of nonnegative real numbers. Recall that the RKHS associated to K is then given by*

$$\mathbb{H} = \left\{ f = \sum_{i=1}^\infty b_i\phi_i \;\middle|\; \{b_i\}\subset l^2(\mathbb{N}), \sum_{i=1}^\infty \frac{b_i^2}{\lambda_i} < \infty \right\}.$$

3. This is a consequence of Theorem XIII in Kolmogorov and Tikhomirov (1961).

Denote also by $\mathbb{B}_{\mathbb{H}}$ the unit ball in $\mathbb{H}$ induced by the inner product $\langle f, g\rangle_{\mathbb{H}} = \sum_{i=1}^{\infty} \frac{\langle f, \phi_i\rangle\langle g, \phi_i\rangle}{\lambda_i}$, where $\langle\cdot,\cdot\rangle$ is the standard inner product in $L^2(\mathbf{P})$. See Chapter 12 of Wainwright (2019) for further background.

Suppose the kernel K satisfies the regularity conditions $\sup_{x\in\mathsf{X}} |\phi_j(x)| \le A$ and $\lambda_j \lesssim j^{-2\alpha}$ for some $\alpha > 1/2$ and all $j \in \mathbb{N}$. Assume further that $\sup_{y\in\mathsf{Y}} |y| \le B$ and that $(x_0, \dots, x_{T_1})$ is a Markov chain with finite mixing time t_{mix}. Then if $f_\star \in \mathbb{B}_{\mathbb{H}}$, the least squares estimator $\hat{f}$ defined in equation (4) with $\mathbb{F} = \mathbb{B}_{\mathbb{H}}$ satisfies

$$\mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2 \lesssim \left(A^{1/(2\alpha+1)} \sqrt{\frac{\alpha}{2\alpha-1}} \left(\frac{\sigma_w^2}{T}\right)^{\frac{\alpha}{2\alpha+1}} + A^{1/(2\alpha+1)} \left(\frac{B^2 t_{\mathsf{mix}}}{T}\right)^{\frac{\alpha}{2\alpha+1}} \right) \operatorname{polylog}(1 + \lambda_1 A T).$$

The result relies on a metric entropy calculation of $\mathbb{B}_{\mathbb{H}}$ in the supremum metric, which can be found in Appendix D.1. Notice that a faster eigenvalue decay corresponds to a faster rate of convergence. In particular, as $\alpha \to \infty$ the rate of convergence approaches the parametric rate $T^{-1/2}$, modulo a polylogarithmic factor. Again, the rate $T^{-\alpha/(2\alpha+1)}$ is optimal even in the i.i.d. setting ($t_{\mathsf{mix}} = 1$). A supporting experiment can be found in Appendix D.2.

The next example revisits the generalized linear models using Theorem 3. This model has recently been analyzed using recursive methods (Foster et al., 2020; Sattar and Oymak, 2020; Jain et al., 2021).

Example 3 *Consider a system of the form*

$$x_{t+1} = \phi(A_\star x_t) + w_t \tag{16}$$

and define

$$\mathbb{F}^\phi = \{f \in C(X \to X); f(\cdot) = \phi(A\,\cdot) \text{ with } A \in \mathbb{R}^{d_x \times d_x} \text{ and } \|A\|_F \le C\}$$

for some $C > 0$ and where the Frobenius norm of A is given by $\|A\|_F = \sqrt{\operatorname{tr} A^\top A}$. This setting is a special case of the autoregressive system (2) with $f_\star(\cdot) = \phi(A_\star\,\cdot)$.

If ϕ is 1-Lipschitz with respect to $\|\cdot\|_2$, $\sigma_{\max}(A_\star) \le L_\star < 1$ so that $\phi(A_\star\cdot)$ is $L_\star$-contractive[4]. Suppose further $\sup_{x\in\mathsf{X}} \|x\|_2 \le B$, then the least squares estimator $\hat{f}$ (4) for the system (16) with hypothesis class $\mathbb{F} = \mathbb{F}^\phi$ satisfies the error bound

$$\mathbf{E}\|\hat{f}(\xi) - \phi(A_\star\xi)\|_2 \lesssim \sqrt{\frac{\sigma_w^2}{T} d_x^2 \log(1 + 2\sigma_w\sqrt{d_x}BCT^2)} + \sqrt{\frac{B^2}{(1-L_\star)^2 T} d_x^2 \log(1 + 2BCT)}.$$

A proof of this claim can be found in Appendix D.3. The dependency on T, d_x and $L_\star$ matches Theorem 2 of Foster et al. (2020), which is optimal in T and d_x.

4. If the link function ϕ is chosen to be the identity, the system (16) becomes a linear dynamical system. In this case, the assumption $\sigma_{\max}(A_\star) < 1$ can be replaced by the constraint that the spectral radius of $A_\star$ is less than unity.

5. Discussion

We have leveraged recently developed information-theoretic tools (Russo and Zou, 2019; Xu and Raginsky, 2017) to analyze the nonparametric LSE (4) for learning dynamical systems. Our analysis yields, to the best of our knowledge, the first rate-optimal bounds for nonparametric estimation of stable or otherwise mixing nonlinear systems from a single trajectory. In addition, our results are able to capture, as a special case, existing parametric rates in the literature (Foster et al., 2020; Sattar and Oymak, 2020).

While our bounds are in expectation, similar tools applied via exponential stochastic inequalities have recently been used to provide high probability generalization bounds for statistical learning (Hellström and Durisi, 2020; Grünwald et al., 2021). Combining our results with these methods could potentially also yield control of $\|\hat{f}(\xi) - f_\star(\xi)\|_2$ with high probability, and is an exciting direction for future work. To arrive at our bounds, we leveraged the decoupling technique of Russo and Zou (2019) and Xu and Raginsky (2017). To apply these techniques to the system (1), we had to control the subgaussian parameter $\sigma_T^2(\mathbb{F}, P_\star)$, which captures the spatiotemporal spread of the covariates. We showed that this term can be controlled using either control-theoretic stability notions, or more general mixing properties.

While this paper develops tools to estimate $f_\star$ in system (1), we believe that the general technique developed is more broadly applicable, and of independent interest. For example, since the variance proxy $\sigma_T^2(\mathbb{F}, P_\star)$ captures the subgaussian parameter of the loss function in statistical learning (Xu and Raginsky, 2017), we prove in Appendix C.1 a generalization bound for single-trajectory learning for Lipschitz loss functions.

Finally, an open problem is to determine for which learning problems the system (1) is required to mix in the single trajectory setting. A recent lower bound (Nagaraj et al., 2020) indicates that in the worst case agnostic setting a corresponding degradation in the rate of convergence is necessary. However, in the realizable setting (1) this question remains unanswered. Most previous works on learning in nonlinear dynamical systems rely on similar mixing time or stability arguments. The cost of this is typically a multiplicative factor in the final bound that degrades as stability is lost (Foster et al., 2020; Sattar and Oymak, 2020; Boffi et al., 2021). In contrast, it is well-known that this dependency can be avoided for learning in linear systems (Lai and Wei, 1982; Simchowitz et al., 2018; Tu et al., 2022). Recently Jain et al. (2021) showed under a strong invertibility condition that dependency on the mixing time can also be avoided for the generalized linear model (16). This leaves open the question whether learning without mixing is possible in situations beyond the generalized linear model.

Acknowledgments

Ingvar Ziemann and Henrik Sandberg are supported by the Swedish Research Council (grant 2016-00861). Nikolai Matni is supported in part by NSF awards CPS-2038873 and CAREER award ECCS-2045834, and a Google Research Scholar award.

References

Alekh Agarwal and John C Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2012.

David Angeli. A Lyapunov approach to incremental stability properties. *IEEE Transactions on Automatic Control*, 47(3):410–421, 2002.

Yannick Baraud, Fabienne Comte, and Gabrielle Viennet. Adaptive estimation in autoregression or-mixing regression via model selection. *The Annals of Statistics*, 29(3):839–875, 2001.

Nicholas M Boffi, Stephen Tu, and Jean-Jacques E Slotine. Regret bounds for adaptive nonlinear control. In *Learning for Dynamics and Control*, pages 471–483. PMLR, 2021.

Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 121–130, 2020.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4): 633–679, 2020.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

Dylan Foster, Tuhin Sarkar, and Alexander Rakhlin. Learning nonlinear dynamical systems from a single trajectory. In *Learning for Dynamics and Control*, pages 851–861. PMLR, 2020.

Borja Rodríguez Gálvez, Germán Bassi, Ragnar Thobaben, and Mikael Skoglund. Tighter expected generalization error bounds via wasserstein distance. In *Advances in Neural Information Processing Systems*, 2021.

Peter Grünwald, Thomas Steinke, and Lydia Zakynthinou. Pac-bayes, mac-bayes and conditional mutual information: Fast rate bounds that handle general vc classes. In *Conference on Learning Theory*, pages 2217–2247. PMLR, 2021.

Fredrik Hellström and Giuseppe Durisi. Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory*, 1(3):824–839, 2020.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Prateek Jain, Suhas S Kowshik, Dheeraj Nagaraj, and Praneeth Netrapalli. Near-optimal offline and streaming algorithms for learning non-linear dynamical systems. *arXiv preprint arXiv:2105.11558*, 2021.

Yassir Jedra and Alexandre Proutiere. Finite-time identification of stable linear systems optimality of the least-squares estimator. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 996–1001. IEEE, 2020.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 33:15312–15325, 2020.

Andrei N Kolmogorov and Vladimir M Tikhomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *Amer. Math. Soc. Transl.(Ser. 2)*, 17:277–364, 1961.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017.

Tze Leung Lai and Ching Zong Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1): 154–166, 1982.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Model learning predictive control in nonlinear dynamical systems. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 757–762. IEEE, 2021.

Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285. PMLR, 2015.

Lennart Ljung. System identification: theory for the user. *PTR Prentice Hall, Upper Saddle River, NJ*, 28, 1999.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. *arXiv preprint arXiv:1902.07826*, 2019.

Horia Mania, Michael I Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *arXiv preprint arXiv:2006.10277*, 2020.

Henry B Mann and Abraham Wald. On the statistical treatment of linear stochastic difference equations. *Econometrica, Journal of the Econometric Society*, pages 173–220, 1943.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary ϕ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11(26):789–814, 2010. URL http://jmlr.org/papers/v11/mohri10a.html.

Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least squares regression with markovian data: Fundamental limits and algorithms. *Advances in Neural Information Processing Systems*, 33:16666–16676, 2020.

Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M. Roy. Information-theoretic generalization bounds for stochastic gradient descent. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3526–3545. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/neu21a.html.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. In *2019 American control conference (ACC)*, pages 5655–5661. IEEE, 2019.

Daniel Paulin. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20:1–32, 2015.

Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264. PMLR, 2014.

Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pages 5610–5618. PMLR, 2019.

Yahya Sattar and Samet Oymak. Non-asymptotic and accurate learning of nonlinear dynamical systems. *arXiv preprint arXiv:2002.08538*, 2020.

Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR, 2020.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pages 3437–3452. PMLR, 2020.

Anastasios Tsiamis and George J Pappas. Finite sample analysis of stochastic system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3648–3654. IEEE, 2019.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Stephen Tu, Alexander Robey, Tingnan Zhang, and Nikolai Matni. On the sample complexity of stability constrained imitation learning. *arXiv preprint arXiv:2102.09161*, 2021.

Stephen Tu, Roy Frostig, and Mahdi Soltanolkotabi. Learning from many trajectories. *arXiv preprint arXiv:2203.17193*, 2022.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Mathukumalli Vidyasagar and Rajeeva L Karandikar. A learning theory approach to system identification and stochastic adaptive control. In *Probabilistic and randomized methods for design under uncertainty*, pages 265–302. Springer, 2006.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ingvar Ziemann and Henrik Sandberg. Regret lower bounds for learning linear quadratic gaussian systems. *arXiv preprint arXiv:2201.01680*, 2022.

Contents

1 Introduction **1**
- 1.1 Contributions . . . 2
- 1.2 Further Related Work . . . 3
- 1.3 Preliminaries and Notation . . . 4

2 Results **5**
- 2.1 Sufficient Conditions for Generalization: Stability and Mixing . . . 7
- 2.2 Summary . . . 8

3 Proof Strategy for Theorem 1 **8**
- 3.1 Offset Basic Inequality Analysis . . . 9

4 Applications of Theorem 1 **10**

5 Discussion **12**

A Proof of Theorem 1 and its Corollaries **18**
- A.1 Proof of Theorem 2 . . . 19
- A.2 Proof of Theorem 3 . . . 20
- A.3 Proof of Auxilliary Results . . . 21
- A.4 Finite Classes . . . 22

B Proofs Related to Stability and Learning **22**
- B.1 Proof of Proposition 4 . . . 23
- B.2 Extension to Exponential Incremental Input-to-State Stability . . . 24
- B.3 Proof of Proposition 5 . . . 26

C Proof of the Decoupling Estimate, Proposition 6 **26**
- C.1 Extension: Generalization Bounds for Dynamical Systems . . . 28

D Supporting Material for the Examples in Section 4 **29**
- D.1 Metric Entropy Calculations for Reproducing Kernel Hilbert Spaces . . . 29
- D.2 An Experiment Supporting Example 2 . . . 30
- D.3 Proof of the claim in Example 3 . . . 32

Appendix A. Proof of Theorem 1 and its Corollaries

We now turn to the proof of Theorem 1. First, we begin by applying Proposition 6 to the discretized estimator $\hat{f}$. Let us begin by bounding the generalization error of the quantized estimator $\hat{f}_\delta$ as defined by (6). We may write

$$\begin{aligned}\mathbf{E}\|\hat{f}_\delta(\xi) - f_\star(\xi)\|_2 &\leq \sqrt{\int \frac{1}{T}\sum_{t=1}^{T} \|\hat{f}_\delta(x_t) - f_\star(x_t)\|_2^2 dP_{x^T,\hat{f}_\delta}} + \sqrt{2\sigma_T^2(\mathbb{F}, P_\star) I(\hat{f}_\delta; x^T)} \\ &\leq \sqrt{\int \frac{1}{T}\sum_{t=1}^{T} \|\hat{f}_\delta(x_t) - f_\star(x_t)\|_2^2 dP_{x^T,\hat{f}_\delta}} + \sqrt{2\sigma_T^2(\mathbb{F}, P_\star) \log \mathcal{N}(\mathbb{F}, \rho, \delta)}.\end{aligned} \tag{17}$$

The first inequality is just Proposition 6 applied to $\hat{f}_\delta$, whereas the second inequality follows from the fact that $I(\hat{f}_\delta; x^T) \leq \log \mathcal{N}(\mathbb{F}, \rho, \delta)$, since the random variable $\hat{f}_\delta$ can take on at most $\mathcal{N}(\mathbb{F}, \rho, \delta)$ different values and the bound $I(\hat{f}_\delta; x^T) \leq H(\hat{f}_\delta) \leq \log \mathcal{N}(\mathbb{F}, \rho, \delta)$.

It remains to bound the in-sample-prediction error. We have

$$\int \frac{1}{T}\sum_{t=1}^{T} \|\hat{f}_\delta(x_t) - f_\star(x_t)\|_2^2 dP_{x^T,\hat{f}_\delta} \leq \int \frac{2}{T}\sum_{t=1}^{T} \|\hat{f}(x_t) - f_\star(x_t)\|_2^2 dP_{x^T,\hat{f}} + 2\delta^2 \tag{18}$$

by construction of $\hat{f}_\delta$, the parallelogram law, and the fact that $\mathbb{F}$ is metrized by the supremum norm.

The main technincal chaining step is given in Lemma 10. Namely, by appealing to Lemma 10 and combining with (17) and (18) we find

$$\begin{aligned}&\mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2 \\ &\quad\leq \sqrt{\frac{8\sigma_w^2 \log \mathcal{N}(\mathbb{F}, \rho, \gamma)}{T} + 128\alpha\sigma_w\sqrt{d_y} + 64\frac{\sigma_w}{\sqrt{T}}\int_\alpha^\gamma \sqrt{\log \mathcal{N}(\mathbb{F}, \rho, s)} ds + 2\delta^2} \\ &\qquad\qquad + \delta + \sqrt{2\sigma_T^2(\mathbb{F}, P_\star) \log \mathcal{N}(\mathbb{F}, \rho, \delta)}.\end{aligned} \tag{19}$$

since $\mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2 \leq \mathbf{E}\|\hat{f}_\delta(\xi) - f_\star(\xi)\|_2 + \delta$ by the triangle inequality. The result follows after pulling the factor $2\delta^2$ out of the square root sign using the triangle inequality. ■

For large spaces and fine grained coverings, the metric entropy starts to dominate the scale free process $M_T(f)$ appearing in Lemma 9. The analysis of $M_T(f)$ in Lemma 10 below essentially follows that in Liang et al. (2015) (compare with their Lemma 6) with certain slight simplifications due to the added structure the uniform topology on $C(\mathsf{X} \to \mathsf{Y})$ affords us. We begin with an analogue of Lemma 9 which takes the scale of the functions considered into account.

The proof of Theorem 1 requires both Lemma 11 and Lemma 9 to accomplish succesful chaining for a variety of metric entropy scalings. While these results are quite similar, for typically metric entropy scalings ($q < 2$) Lemma 9 performs better due to self-normalization. However, once $q \geq 2$ Dudley's entropy integral becomes singular near 0, wherefore we also require the cruder discretization provided by Lemma 11. In other words, there is a critical radius $r \asymp \sqrt{\frac{\log \mathcal{N}(\mathbb{F}_\star, \rho, r)}{T}}$ below which self-normalization becomes insignificant and the supremum norm bound $\|f\|_\infty \leq r$ starts to become increasingly important to control the supremum of $M_T(f)$.

Lemma 10 *Fix a metric space $(\mathbb{F}, \rho)$ with $\rho(f,g) = \sup_{x\in X} \|f(x) - g(x)\|_2$. Then with $M_T(f)$ defined by equation (14), we have that*

$$\mathbf{E} \sup_{f\in\mathbb{F}_\star} \frac{1}{T} M_T(f)$$
$$\leq \inf_{\gamma>0,\alpha\in[0,\gamma]} \left\{ \frac{8\sigma_w^2 \log \mathcal{N}(\mathbb{F},\rho,\gamma)}{T} + 128\alpha\sigma_w\sqrt{d_y} + 64\frac{\sigma_w}{\sqrt{T}} \int_\alpha^\gamma \sqrt{\log \mathcal{N}(\mathbb{F},\rho,s)} ds \right\}.$$

As noted in Liang et al. (2015), the optimal value for γ in Lemma 10 is of the same nature as when obtained by other methods, see for example Chapter 13 of Wainwright (2019) for a more standard approach.

The idea below is to decompose the supremum over $\mathbb{F}_\star$ in inequality (13) by $\mathbb{F}_\star = \{\mathbb{F}_\star \cap \alpha\mathbb{B}_\star\} \cup \{\mathbb{F}_\star \setminus \alpha\mathbb{B}_\star\}$ where $\mathbb{B}_\star$ is the unit ball in the space of bounded functions, centered at $f_\star$. On $\mathbb{F}_\star \cap \alpha\mathbb{B}_\star$, $M_T(f)$, the process (14), is small since $\|f\|_\infty \leq \alpha$. The role of chaining is to show that it suffices to approximate $M_T(f)$ on $\mathbb{F}_\star \setminus \alpha\mathbb{B}_\star$ at low resolution and thus rely on Lemma 9 with small $|S|$.

Proof Observe first that for any fixed $\alpha > 0$ we have, simply by discarding the negative second order term, Cauchy-Schwarz, and a standard subgaussian concentration inequality for $\mathbf{E}\|w_t\|_2$:

$$\mathbf{E} \sup_{f\in\mathbb{F}_\star\cap\alpha\mathbb{B}_\star} \frac{1}{T} M_T(f) \leq \mathbf{E} \sup_{f\in\mathbb{F}_\star\cap\alpha\mathbb{B}} \frac{4}{T} \sum_{t=0}^{T-1} \langle w_t, f(x_t)\rangle \leq 16\alpha\sigma_w\sqrt{d_y}. \tag{20}$$

A standard one-step discretization bound (c.f. the proof of Proposition 5.17 in Wainwright (2019)) combined with the finite class maximal inequality of Lemma 9 yields for fixed $\gamma > 0$:

$$\mathbf{E} \sup_{f\in\mathbb{F}_\star} \frac{1}{T} M_T(f) \leq \frac{8\sigma_w^2 \log \mathcal{N}(\mathbb{F}_\star,\rho,\gamma)}{T} + 2\mathbf{E} \sup_{f\in\mathbb{F}_\star\cap\gamma\mathbb{B}_\star} \frac{1}{T} M_T(f)$$

Having extracted the fast rate term for scales larger than γ, we proceed with a chaining bound on the second term above. Since $M_T(f)$ satisfies the maximal inequality (23) with $r = \gamma$, chaining (as in Theorem 5.22 of Wainwright (2019)) yields

$$2\mathbf{E} \sup_{f\in\mathbb{F}_\star\cap\gamma\mathbb{B}_\star} \frac{1}{T} M_T(f) \leq 32\alpha\sigma_w\sqrt{d_y} + 64\frac{\sigma_w}{\sqrt{T}} \int_{\alpha/4}^\gamma \sqrt{\log \mathcal{N}(\mathbb{F}_\star,\rho,s)} ds.$$

Note that $\mathcal{N}(\mathbb{F}_\star,\rho,s) = \mathcal{N}(\mathbb{F},\rho,s)$ by translation invariance of the metric ρ. The results now follows by terminating the chaining at scale α, using (20) to bound that which remains and rescaling $\alpha \leftrightarrow 4\alpha$. ∎

A.1. Proof of Theorem 2

Under the hypothesis (8) we may use Theorem 1 with $\alpha = 0$ to write

$$\begin{aligned}
&\mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2 \\
&\lesssim \sqrt{\frac{8\sigma_w^2 p \left(\frac{1}{\gamma}\right)^q}{T} + 64\frac{\sigma_w}{\sqrt{T}} \int_0^\gamma \sqrt{p\left(\frac{1}{s}\right)^q} ds} + 3\delta + \sqrt{2\sigma_T^2(\mathbb{F}, P_\star) p \left(\frac{1}{\delta}\right)^q} \\
&= \sqrt{\frac{8\sigma_w^2 p \left(\frac{1}{\gamma}\right)^q}{T} + 64\frac{\sigma_w}{\sqrt{T}}\sqrt{p}\frac{2}{2-q}\left(\frac{1}{\gamma}\right)^{1-q/2}} + 3\delta + \sqrt{2\sigma_T^2(\mathbb{F}, P_\star) p \left(\frac{1}{\delta}\right)^q}.
\end{aligned} \tag{21}$$

Choosing γ and δ to satisfy the optimal balance: $\gamma \asymp \left(\frac{\sigma_w^2 p}{T}\right)^{\frac{1}{2+q}}$ and $\delta \asymp \left(p\sigma_T^2(\mathbb{F}, P_\star)\right)^{\frac{1}{2+q}}$ (21) becomes

$$\begin{aligned}
&\mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2 \\
&\lesssim \sqrt{\frac{2}{2-q}\frac{\sigma_w^2 p\left(\frac{\sigma_w^2 p}{T}\right)^{\frac{-q}{2+q}}}{T}} + \left(p\sigma_T^2(\mathbb{F}, P_\star)\right)^{\frac{1}{2+q}} \\
&\asymp \sqrt{\frac{2}{2-q}}\left(\frac{\sigma_w^2 p}{T}\right)^{\frac{1}{2+q}} + \left(p\sigma_T^2(\mathbb{F}, P_\star)\right)^{\frac{1}{2+q}}.
\end{aligned}$$

This verifies the claim. ■

Remark: If instead $q > 2$, we have

$$\mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2 \lesssim \sigma_w\sqrt{d_y}\left(\frac{1}{q-2}\right)^{1/q}\left(\frac{p}{d_y T}\right)^{\frac{1}{2q}} + \left(p\sigma_T^2(\mathbb{F}, P_\star)\right)^{\frac{1}{2+q}}$$

and similarly but with an extra logarithmic factor at $q = 2$. To show this, we again use Theorem 1 but with γ chosen sufficiently large such that $\log \mathcal{N}(\mathbb{F}_\star, \rho, \gamma) \asymp 1$. In this case

$$\begin{aligned}
&\mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2 \\
&\lesssim \sqrt{\alpha\sigma_w\sqrt{d_y} + \frac{\sigma_w\sqrt{p}}{\sqrt{T}}\left(\frac{2}{q-2}\right)\alpha^{1-q/2}} + \delta + \sqrt{2\sigma_T^2(\mathbb{F}, P_\star)\log\mathcal{N}(\mathbb{F}, \rho, \delta)}.
\end{aligned}$$

The claim follows by solving for the optimal balance

$$\alpha \asymp \left(\frac{2}{q-2}\right)^{1/q}\left(\frac{p}{d_y T}\right)^{\frac{1}{2q}}$$

and $\delta \asymp \left(p\sigma_T^2(\mathbb{F}, P_\star)\right)^{\frac{1}{2+q}}$.

A.2. Proof of Theorem 3

By virtue of Theorem 1 and by selecting $\gamma = \alpha$ we have

$$\begin{aligned}
\mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2 \leq &\sqrt{\frac{8\sigma_w^2\log\mathcal{N}(\mathbb{F}_\star, \rho, \alpha)}{T} + 128\alpha\sigma_w\sqrt{d_y}} \\
&+ 3\delta + \sqrt{2\sigma_T^2(\mathbb{F}, P_\star)\log\mathcal{N}(\mathbb{F}, \rho, \delta)}.
\end{aligned} \tag{22}$$

Let now $\alpha = 1/\sqrt{d_y}\sigma_w T^2$ and $\delta = 1/T$. Then (22) by using the hypothesis (9) becomes

$$\begin{aligned}
\mathbf{E}\|\hat{f}(\xi) - f_\star(\xi)\|_2 \lesssim &\sqrt{\frac{\sigma_w^2 p\log(1 + c\sqrt{d_y}\sigma_w T^2)}{T} + \frac{1}{T^2}} \\
&+ \frac{1}{T} + \sqrt{2\sigma_T^2(\mathbb{F}, P_\star)p\log(1 + cT)}.
\end{aligned}$$

The result follows. ■

A.3. Proof of Auxilliary Results

Proof of Lemma 7 By optimality of $\hat{f}$ to the prediction error objective we have that

$$\sum_{t=0}^{T-1} \|\hat{f}(x_t) - y_t\|_2^2 \leq \sum_{t=0}^{T-1} \|f_\star(x_t) - y_t\|_2^2.$$

Rearranging and expanding the square gives the basic inequality

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\hat{f}(x_t) - f_\star(x_t)\|_2^2 \leq \frac{2}{T}\sum_{t=0}^{T-1} \langle w_t, \hat{f}(x_t) - f_\star(x_t)\rangle$$

which after multiplying both sides by 2 can be rearranged again to give

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\hat{f}(x_t) - f_\star(x_t)\|_2^2 \leq \frac{1}{T}\sum_{t=0}^{T-1} 4\langle w_t, \hat{f}(x_t) - f_\star(x_t)\rangle - \|\hat{f}(x_t) - f_\star(x_t)\|_2^2$$

so that the result follows by taking the supremum over the variable $\hat{f} - f_\star \in \mathbb{F}_\star$. ■

Proof of Lemma 9 The proof is a straight-forward modification of the standard proof for bounding the expected supremum of subgaussian maxima. By Jensen's inequality and monotonicity of the exponential it follows that

$$\begin{aligned}
\exp\left(\lambda \mathbf{E} \sup_{f\in S} M_T(f)\right) &\leq \mathbf{E}\exp\left(\lambda \max_{f\in S} M_T(f)\right) \\
&\leq \mathbf{E}\max_{f\in S}\exp\left(\lambda M_T(f)\right) \\
&\leq \sum_{f\in S}\mathbf{E}\exp\left(\lambda M_T(f)\right).
\end{aligned}$$

Choosing $\lambda = 1/2\sigma_w^2$, application of Lemma 8 yields $\exp\left(\frac{1}{2\sigma^2}\mathbf{E}\sup_{f\in S} M_T(f)\right) \leq |S|$ which is equivalent to the result. ■

Proof of Lemma 8 Write by the tower property

$$\begin{aligned}
&\mathbf{E}\exp\left(\lambda\sum_{t=1}^{T-1} 4\langle w_t, f(x_t)\rangle - \|f(x_t)\|_2^2\right) \\
&= \mathbf{E}\exp\left(\lambda\sum_{t=1}^{T-2} 4\langle w_t, f(x_t)\rangle - \|f(x_t)\|_2^2\right)\exp\left(-\lambda\|f(x_{T-1})\|_2^2\right)\mathbf{E}_{T-2}\exp\left(\lambda 4\langle w_{T-1}, f(x_{T-1})\rangle\right) \\
&\leq \mathbf{E}\exp\left(\lambda\sum_{t=1}^{T-2} 4\langle w_t, f(x_t)\rangle - \|f(x_t)\|_2^2\right)\exp\left(\left[2\lambda^2\sigma_w^2 - \lambda\right]\|f(x_{T-1})\|_2^2\right) \\
&\leq \mathbf{E}\exp\left(\lambda\sum_{t=1}^{T-2} 4\langle w_t, f(x_t)\rangle - \|f(x_t)\|_2^2\right) \\
&\leq \cdots \leq 1
\end{aligned}$$

as per requirement. ■

Lemma 11 *Let S be a finite subset of the shifted metric space $\mathbb{F}_\star$ with $\|f\|_\infty \leq r$ for all $f \in S$. Then*

$$\mathbf{E} \sup_{f \in S} M_T(f) \leq \sqrt{2T\sigma_w^2 r^2 \log |S|}. \tag{23}$$

Proof Fix $\lambda > 0$. By Jensen's inequality and monotonicity of the exponential it follows that

$$\begin{aligned} \exp\left(\lambda \mathbf{E} \sup_{f \in S} M_T(f)\right) &\leq \mathbf{E} \exp\left(\lambda \max_{f \in S} M_T(f)\right) \\ &= \mathbf{E} \max_{f \in S} \exp\left(\lambda M_T(f)\right) \\ &\leq \sum_{f \in S} \mathbf{E} \exp\left(\lambda M_T(f)\right). \end{aligned} \tag{24}$$

Using $\|f\|_\infty \leq r$ and the tower property let us now estimate

$$\begin{aligned} \mathbf{E} \exp\left(\lambda M_T(f)\right) &\leq \mathbf{E} \exp\left(\lambda \sum_{t=1}^{T-1} 4\langle w_t, f(x_t)\rangle\right) \\ &\leq \mathbf{E} \exp\left(\lambda \sum_{t=1}^{T-2} 4\langle w_t, f(x_t)\rangle\right) \exp\left(2\lambda^2 r^2 \sigma_w^2\right) \\ &\leq \dots \\ &\leq \exp\left(2T\lambda^2 r^2 \sigma_w^2\right). \end{aligned}$$

Hence after applying logarithms to both sides of equation (24), we find

$$\mathbf{E} \sup_{f \in S} M_T(f) \leq \frac{1}{\lambda}\left(\log |S|\right) + \lambda\left(2Tr^2\sigma_w^2\right)$$

which yields the result after optimizing over $\lambda > 0$. ∎

A.4. Finite Classes

The generalization bound of Proposition 6 in combination with the control of the subgaussian parameter $\sigma_T^2(\mathbb{F}, P_\star)$ Proposition 4 affords us, together with Lemmas 7 and 9 immediately yields Theorem 12 below.

Theorem 12 *Assume that $\mathbb{F}$ has finite cardinality. Then under the assumptions of Proposition 4 it holds that*

$$\mathbf{E}\|\hat{f}(\xi) - f(\xi)\|_2 \lesssim \sqrt{\frac{\sigma_w^2}{T}\log(|\mathbb{F}|)} + \sqrt{\frac{M^2B^2L^2}{m^2(1-L_\star)^2T}\log(|\mathbb{F}|)}. \tag{25}$$

Appendix B. Proofs Related to Stability and Learning

Let us now prove that contraction in an arbitrary norm $\|\cdot\|_{\mathsf{X}}$ implies that the subgaussian parameter $\sigma_T^2(\mathbb{F}, P_\star)$ decays gracefully with time, T. We remind the reader that the idea is to combine a Azuma-McDiarmid-Hoeffding style of analysis with a stability argument. We now procede with this program.

B.1. Proof of Proposition 4

Fix two functions $f, g \in \mathbb{F}$ and denote $h = f - g$. Define also the function $F : \mathbb{R}^{d_x T} \to \mathbb{R}$ by

$$F(\zeta_0, \dots, \zeta_{T-1}) \triangleq \frac{1}{T} \sum_{t=0}^{T-1} \|h(\zeta_t)\|_2$$

where the dummy variables ζ_t are elements of $\mathbb{R}^{d_x}$. Let also $\mathbf{E}_t$ denote conditional expectation with respect to $\mathcal{F}_t$ and define the Doob martingale difference sequence

$$\Delta_s \triangleq \mathbf{E}_s F(x_0, \dots, x_{T-1}) - \mathbf{E}_{s-1} F(x_0, \dots, x_{T-1})$$

with the convention $\mathbf{E}_{-1} = \mathbf{E}$. Note now that $F(x_0, \dots, x_{T-1}) - \mathbf{E}F(x_0, \dots, x_{T-1}) = \sum_{t=0}^{T-1} \Delta_t$ so to arrive at the desired conclusion we need to prove that the Δ_s are uniformly bounded.

To this end, for a fixed s, we define two couplings of $(x_s, \dots, x_{T-1})$ via

$$\begin{aligned} z_{t+1} &= f_\star(z_t), + w_t & z_s &= z, & t &= s, \dots, T-1, \\ z'_{t+1} &= f_\star(z'_t), + w_t & z'_s &= z', & t &= s, \dots, T-1. \end{aligned}$$

which vary only in their initial condition z, z' but are constructed with the same sequence $(w_s, \dots, w_{T-1})$. We now compute

$$\begin{aligned} \Delta_s &= \mathbf{E}_s F(x_0, \dots, x_{T-1}) - \mathbf{E}_{s-1} F(x_0, \dots, x_{T-1}) \\ &= \mathbf{E}_s \frac{1}{T} \sum_{t=0}^{T-1} \|h(x_t)\|_2 - \mathbf{E}_{s-1} \frac{1}{T} \sum_{t=0}^{T-1} \|h(x_t)\|_2 \\ &= \frac{1}{T} \mathbf{E}_s \sum_{t=s}^{T-1} \|h(x_t)\|_2 - \frac{1}{T} \mathbf{E}_{s-1} \sum_{t=s}^{T-1} \|h(x_t)\|_2 \\ &\le \sup_{z,z'} \frac{1}{T} \sum_{t=s}^{T-1} \|h(z_t)\|_2 - \|h(z'_t)\|_2 \\ &\le \sup_{z,z'} \frac{1}{T} \sum_{t=s}^{T-1} \big| \|h(z_t)\|_2 - \|h(z'_t)\|_2 \big| \\ &\le \sup_{z,z'} \frac{1}{T} \sum_{t=s}^{T-1} \|h(z_t) - h(z'_t)\|_2 \\ &\le \sup_{z,z'} \frac{2L}{T} \sum_{t=s}^{T-1} \|z_t - z'_t\|_2 \end{aligned} \tag{26}$$

where the first inequality uses the Markov property to realize the conditional expectations as functions of x_s and x_{s-1} respectively. The other inequalities follow by application of the triangle inequality and the $2L$-Lipschitzness of h.

Let us now bound the $\|\cdot\|_{\mathsf{X}}$-distance between z_t and z'_t:

$$\begin{aligned} \|z_t - z'_t\|_{\mathsf{X}} &= \|f_\star(z_{t-1}) + w_{t-1} - f_\star(z'_{t-1}) + w_{t-1}\|_{\mathsf{X}} \\ &\le L_\star \|z_{t-1} - z'_{t-1}\|_{\mathsf{X}} \\ &\dots \\ &\le L_\star^{t-s-1} \|z' - z\|_{\mathsf{X}}. \end{aligned} \tag{27}$$

Combining equations (26) and (27), and noting that a symmetric argument applies to $-\Delta_s$ it follows that

$$|\Delta_s| \leq \frac{4MLB}{m(1-L_\star)T}. \tag{28}$$

Expressing F as a telescoping sum over Δ_s, we can compute its moment generating function in combination with the tower property:

$$\begin{aligned}
&\mathbf{E}\exp\left(\lambda\left[F(w_0,\dots,w_{T-1}) - \mathbf{E}F(w_0,\dots,w_{T-1})\right]\right) \\
&= \mathbf{E}\exp\left(\lambda\sum_{j=0}^{T-1}\Delta_j\right) \\
&= \mathbf{E}\exp\left(\lambda\sum_{j=0}^{T-2}\Delta_j\right)\mathbf{E}_{T-2}\exp\left(\lambda\Delta_{T-1}\right) \\
&\leq \mathbf{E}\exp\left(\lambda\sum_{j=0}^{T-2}\Delta_j\right)\exp\left(2\lambda^2\left(\frac{4MLB}{m(1-L_\star)T}\right)^2\right) \\
&\leq \cdots \leq \exp\left(\frac{32\lambda^2M^2L^2B^2}{m^2(1-L_\star)^2T}\right)
\end{aligned}$$

using Hoeffding's inequality to bound the conditional moment generating functions of the bounded random variables Δ_j using the inequality (28) (see Hoeffding (1963) or Example 2.4. in Wainwright (2019)). ■

B.2. Extension to Exponential Incremental Input-to-State Stability

In the main text we described how contraction properties of $f_\star$ lead to bounds on the subgaussian parameter $\sigma_T^2(\mathbb{F}, P_\star)$. We now show that another control-theoretic notion of stability, known as Exponential Incremental Input-to-State Stability (E-δISS) is also amenable to this analysis. The E-δISS framework was introduced by Angeli (2002). Let us fix two metric spaces $(\mathsf{X}, \rho_{\mathsf{X}})$ and $(\mathsf{W}, \rho_{\mathsf{W}})$. A family of functions $\{G_t\}_{t\in\mathbb{Z}_{\geq 0}}$, $G_t : \mathsf{X}\times\mathsf{W}\to\mathsf{X}$ is (a,b,r)-E-δISS if for each $T\in\mathbb{N}$, every pair of sequences $\{\eta_t\}$ and $\{\zeta_t\}$, and system of equations satisfying

$$\begin{aligned}
\phi_{t+1} &= G_t(\phi_t,\eta_t) && t=0,\dots,T-1 \\
\psi_{t+1} &= G_t(\psi_t,\zeta_t) && t=0,\dots,T-1
\end{aligned}$$

with $\phi_t,\psi_t\in\mathsf{X}, \eta_t,\zeta_t\in\mathsf{W}$ it holds for all $t\in[T]$ that

$$\rho_{\mathsf{X}}(\phi_t,\psi_t) \leq ar^t\rho_{\mathsf{X}}(\phi_0,\psi_0) + b\sum_{k=0}^{t-1} r^{t-k-1}\rho_{\mathsf{W}}(\eta_k,\zeta_k). \tag{29}$$

Proposition 13 *Fix a sequence $\{\eta_t\}$ of i.i.d. random variables, and assume that W is bounded, $\sup_{w,w'\in\mathsf{W}}\rho_{\mathsf{W}}(w,w')\leq B$. Suppose that $\{G_t\}_{t\in\mathbb{Z}_{\geq 0}}$ is (a,b,r)-E-δISS and consider the process*

$$x_{t+1} = G_t(x_t,\eta_t) \qquad t=0,\dots,T-1. \tag{30}$$

Then for every function $l : \mathsf{X} \to \mathbb{R}_{\geq 0}$, L-Lipschitz with respect to ρ_{X}:

$$|l(x) - l(x')| \leq L\rho_{\mathsf{X}}(x, x'), \forall x, x' \in \mathsf{X}$$

we have

$$\mathbf{E} \exp\left(\lambda\left[\frac{1}{T}\sum_{t=0}^{T-1} l(x_t) - \mathbf{E}\frac{1}{T}\sum_{t=0}^{T-1} l(x_t)\right]\right) \leq \exp\left(\frac{8\lambda^2 L^2 B^2 b^2}{(1-r)^2 T}\right).$$

By choosing $l = \|f - g\|_2$ every space $\mathbb{F}$ of L-Lipschitz functions is $\sigma_T^2(\mathbb{F}, P_\star)$-subgaussian with respect to the process (30) with $\sigma_T^2(\mathbb{F}, P_\star) = \frac{64L^2B^2b^2}{(1-r)^2T}$. Note that the constant M^2/m^2 appearing in Proposition 4 can be subsumed into the constants a, b and L by appropriate rescaling of ρ_{X} and ρ_{W}.

Proof As before, the idea is to lean on an Azuma-McDiarmid-Hoeffding style of analysis, but now combined with the bound (29). Fix two functions $f, g \in \mathbb{F}$ and denote $h = f - g$. Consider the function $F(\eta_0, \ldots, \eta_{T-1}) = \frac{1}{T}\sum_{t=0}^{T-1} l(x_t)$, which becomes a function of the random sequence $\{\eta_t\}$ via (30). We shall show that this function is $\frac{4MLBb}{m(1-r)T}$-Lipschitz with respect to the Hamming metric. To this end, introduce a coupling of x_t by defining the system $z_t = G(z_{t-1}, \zeta_t)$ (and with the same initial condition) and observe that

$$\begin{aligned}\left|\frac{1}{T}\sum_{t=0}^{T-1} l(x_t) - \frac{1}{T}\sum_{t=0}^{T-1} l(z_t)\right| &\leq \frac{1}{T}\sum_{t=0}^{T-1} |l(x_t) - l(z_t)| \\ &\leq \frac{L}{T}\sum_{t=0}^{T-1} \rho_{\mathsf{X}}(x_t, z_t)\end{aligned} \tag{31}$$

by repeated application of the triangle inequality and since l is L-Lipschitz.

Let us now bound the ρ_{X}-distance between x_t and z_t under the hypothesis that $\eta_t = \zeta_t, \forall t \neq j$. Then we have using the E-δ-ISS bound in equation (29) that

$$\rho_{\mathsf{X}}(x_t, z_t) \leq br^{t-j-1}\rho_{\mathsf{W}}(\eta_j, \zeta_j). \tag{32}$$

Thus, combining (31) with (32) gives

$$\left|\frac{1}{T}\sum_{t=0}^{T-1} l(x_t) - \frac{1}{T}\sum_{t=0}^{T-1} l(z_t)\right| \leq \frac{L}{T(1-r)}\rho_{\mathsf{W}}(\eta_j, \zeta_j) \leq \frac{LBb}{(1-r)T} \tag{33}$$

by boundedness of W.

We may proceed with the analysis by defining the martingale difference sequence

$$\Delta_j = \mathbf{E}[F(\eta_0, \ldots, \eta_{T-1})|\eta_0, \ldots, \eta_j] - \mathbf{E}[F(\eta_0, \ldots, \eta_{T-1})|\eta_0 \ldots \eta_{j-1}]$$

which has bounded absolute value by independence of the sequence $\{\eta_t\}$ and (33). Observe that this allows us to express F as a telescoping sum, which we can readily use to compute the moment

generating function in combination with the tower property:

$$\begin{aligned}
&\mathbf{E}\exp\left(\lambda\left[F(\eta_0,\dots,\eta_{T-1}) - \mathbf{E}F(\eta_0,\dots,\eta_{T-1})\right]\right)\\
&= \mathbf{E}\exp\left(\lambda\sum_{j=0}^{T-1}\Delta_j\right)\\
&= \mathbf{E}\exp\left(\lambda\sum_{j=0}^{T-2}\Delta_j\right)\mathbf{E}_{T-2}\exp\left(\lambda\Delta_{T-1}\right)\\
&\leq \mathbf{E}\exp\left(\lambda\sum_{j=0}^{T-2}\Delta_j\right)\exp\left(2\lambda^2\left(\frac{2LBb}{(1-r)T}\right)^2\right)\\
&\leq \cdots \leq \exp\left(\frac{8\lambda^2L^2B^2b^2}{(1-r)^2T}\right)
\end{aligned}$$

using Hoeffding's inequality to bound the conditional moment generating functions of the bounded random variables Δ_j using (33) (see Hoeffding (1963) or Example 2.4. in Wainwright (2019)). ■

B.3. Proof of Proposition 5

Observe that the function

$$F(z_0,\dots,z_{T-1}) = \frac{1}{T}\sum_{t=0}^{T-1}\|f(z_t) - g(z_t)\|_2$$

is $2B/T$ Hamming-Lipschitz for every fixed choice of $f, g \in \mathbb{F}$. Hence we may apply Corollary 2.10 of Paulin (2015) to obtain a high probability tail bound. This is equivalent to the desired statement by Proposition 2.5.2 in Vershynin (2018). ■

Appendix C. Proof of the Decoupling Estimate, Proposition 6

In what follows, we compare probability integrals under different distributions. More precisely, we wish to relate the joint distribution of the least squares estimator (4) and the samples from the system (1) with the product measure of their marginals. The following variational formulation of $D(P\|Q)$, due to Donsker and Varadhan (1975), is key:

Lemma 14 *Fix two probability measures* $\mathbf{P}$ *and* $\mathbf{Q}$ *on a measure space* $(\Omega, \mathcal{F})$. *Then for every* $\mathcal{F}$*-measurable* $F : \Omega \to \mathbb{R}$ *such that* $\int e^F d\mathbf{Q}$ *is finite, it holds that*

$$\int F d\mathbf{P} - \log\int e^F d\mathbf{Q} \leq D(\mathbf{P}\|\mathbf{Q}). \tag{34}$$

Moreover, if $D(\mathbf{P}\|\mathbf{Q}) < \infty$, *then equality in (34) is attained at* $F = \log\frac{d\mathbf{P}}{d\mathbf{Q}}$.

Equipped with Lemma 14, and inspired by the work of Russo and Zou (2019) and Xu and Raginsky (2017), we now turn to the proof of Proposition 6. We remark that the first paragraph of the proof is identical to the proof of Lemma 1 in Xu and Raginsky (2017). As it is central to our argument, we reproduce it below.

Proof of Proposition 6 We begin by observing that by rescaling F in (34) by λ, we obtain

$$\int \lambda F d\mathbf{P} \leq \log \int e^{\lambda F} dQ + D(\mathbf{P}\|\mathbf{Q}). \tag{35}$$

For any F which is σ^2-subgaussian under $\mathbf{Q}$, we have that

$$\log \int e^{\lambda F} d\mathbf{Q} \leq \int \lambda F d\mathbf{Q} + \frac{\lambda^2 \sigma^2}{2}. \tag{36}$$

Combining inequalities (35) and (36), we see that

$$\int \lambda F d\mathbf{P} - \int \lambda F d\mathbf{Q} \leq D(\mathbf{P}\|\mathbf{Q}) + \frac{\lambda^2 \sigma^2}{2}$$

which after choice of $\lambda = -\frac{\sqrt{D(P\|Q)}}{\sqrt{2\sigma^2}}$ and rearranging becomes

$$\int F d\mathbf{Q} \leq \int F d\mathbf{P} + \sqrt{2\sigma^2 D(\mathbf{P}\|\mathbf{Q})}. \tag{37}$$

We now specialize this known result to our setting. Let us now choose

$$\mathbf{P} = P_{Z,(f,g)}, \qquad \mathbf{Q} = P_Z \otimes P_{(f,g)},$$
$$F = \frac{1}{T}\sum_{t=0}^{T-1} \|f(x_t) - g(x_t)\|_2.$$

Observe that for $\mathbf{P}, \mathbf{Q}$ as above, $D(\mathbf{P}\|\mathbf{Q}) = I((f,g); Z)$. Let further $(x'_0, \ldots, x'_{T-1})$ be equal in distribution to $(x_0, \ldots, x_{T-1})$ but independent from f and g. In other words $(x_0, \ldots, x_{T-1}, f, g)$ is drawn from $\mathbf{P}$ and $(x'_0, \ldots, x'_{T-1}, f, g)$ is drawn from $\mathbf{Q}$. Let also τ be uniformly distributed over $\{0, \ldots, T-1\}$ and independent of all other randomness so that we may take $\xi = x'_\tau$. Hence, for these choices, inequality (37) combined with Jensen's inequality yield

$$\begin{aligned}
\mathbf{E}\|f(\xi) - g(\xi)\|_2 &= \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(x'_t) - g(x'_t)\|_2 \\
&\overset{\text{(i)}}{\leq} \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(x_t) - g(x_t)\|_2 + \sqrt{2\sigma_T^2(\mathbb{F}, P_\star) I(f,g;Z)} \\
&= \mathbf{E}\|f(x_\tau) - g(x_\tau)\|_2 + \sqrt{2\sigma_T^2(\mathbb{F}, P_\star) I(f,g;Z)} \\
&= \sqrt{(\mathbf{E}\|f(x_\tau) - g(x_\tau)\|_2)^2} + \sqrt{2\sigma_T^2(\mathbb{F}, P_\star) I(f,g;Z)} \\
&\overset{\text{(ii)}}{\leq} \sqrt{\mathbf{E}\|f(x_\tau) - g(x_\tau)\|_2^2} + \sqrt{2\sigma_T^2(\mathbb{F}, P_\star) I(f,g;Z)} \\
&= \sqrt{\frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(x_t) - g(x_t)\|_2^2} + \sqrt{2\sigma_T^2(\mathbb{F}, P_\star) I(f,g;Z)}
\end{aligned}$$

by linearity of expectation and reformulating the mixture component. Inequality (i) follows from inequality (37) and inequality (ii) from Jensen's inequality. ■

C.1. Extension: Generalization Bounds for Dynamical Systems

It has previously been observed in the context of the generalized linear model (16) that system-theoretic notions are useful to provide learning guarantees, see Section 4 of Foster et al. (2020) for an interesting discussion. Here, we show that the bounds on $\sigma_T^2(\mathbb{F}, P_\star)$ in Propositions 4 and 13 yield generalization bounds for more general statistical learning. Consider a loss function $l : \mathsf{X} \times \mathsf{Y} \times \mathsf{H} \to \mathbb{R}_{\geq 0}$ and assume that the sequences $(x_0, \dots, x_{T-1})$ and $(y_0, \dots, y_{T-1})$ are generated by E-δISS systems (30), $\{G_t^x\}$ and $\{G_t^y\}$ respectively. Assume that these are driven by the same i.i.d. noise sequence $(w_0, \dots, w_{T-1})$. If not, we can always define such a sequence on a space of the form $\mathsf{W} = \mathsf{W}_\mathsf{X} \times \mathsf{W}_\mathsf{Y}$.

The problem of statistical learning is to find a hypothesis $h \in \mathsf{H}$ that minimizes

$$\mathbf{E}_Z L(Z, h) = \mathbf{E}_Z \frac{1}{T} \sum_{t=0}^{T-1} l(x_{t+1}, x_t, h)$$

with $Z = (x_0, \dots, x_{T-1}, y_0, \dots, y_{T-1})$ and where $\mathbf{E}_Z$ denotes integration over the randomness in Z. Let H be a randomized learning algorithm (a random, data-dependent element of H). We define its generalization error by

$$\mathsf{gen}(H) = \mathbf{E}[\mathbf{E}_{\bar{Z}} L(\bar{Z}, H) - L(Z, H)]$$

where $\bar{Z}$ is equal to Z in distribution but independent of H. By combining Lemma 1 of Xu and Raginsky (2017) with Proposition 13 we arrive at the following inequality.

Proposition 15 *Suppose that l is L-lipschitz in its first two arguments:*

$$l|(x, y, h) - l(x', y', h)| \leq L\rho_\mathsf{X}(x, x') + L\rho_\mathsf{Y}(y, y'), \forall x, x' \in \mathsf{X}, y, y' \in \mathsf{Y}, h \in \mathbb{H},$$

that $(x_0, \dots, x_{T-1})$ is (a, b, r) E-δISS and that $(y_0, \dots, y_{T-1})$ is (a', b', r') E-δISS. Then

$$|\mathsf{gen}(H)| \leq \sqrt{\frac{64L^2B^2b^2}{(1 - r_{\max})^2 T} I(Z; H)}$$

where $r_{\max} = \max(r, r')$ and $B = \sup_{w,w' \in \mathsf{W}} \rho_\mathsf{W}(w, w')$.

In principle a direct proof using the methods from Appendix B is possible. For brevity, we instead show how the result can be reduced to the statement of Proposition 13.

Proof Let $\mathsf{X}' = \mathsf{X} \times \mathsf{Y}$ and define the extended dynamics $\phi_t^1 = x_t, \phi_t^2 = y_t$. Then

$$\begin{bmatrix} \phi_{t+1}^1 \\ \phi_{t+1}^2 \end{bmatrix} = \begin{bmatrix} G_t^x(\phi_t^1, w_t) \\ G_t^y(\phi_t^2, w_t) \end{bmatrix}, \tag{38}$$

or $\phi_{t+1} = G_t(\phi_t, w_t)$ in brief. Since G^x and G^y are both E-δISS as system from $(\mathsf{W}, \rho_\mathsf{W})$ to $(\mathsf{X}, \rho_\mathsf{X})$ and $(\mathsf{Y}, \rho_\mathsf{Y})$ respectively, it follows that G is $(\max(a, a'), 2b, \max(r, r'))$ E-δISS from $(\mathsf{W}, \rho_\mathsf{W})$ to $(\mathsf{X}', \rho_{\mathsf{X}'})$ with $\rho_{\mathsf{X}'} = \rho_\mathsf{X} + \rho_\mathsf{Y}$. Hence, we may apply Proposition 13 to conclude that

$$L(Z, h) = \frac{1}{T} \sum_{t=0}^{T-1} l(y_t, x_t, h)$$

is $\frac{32L^2B^2b^2}{(1-r_{\max})^2 T}$-subgaussian for each fixed h where $r_{\max} = \max(r, r')$. The result follows by applying Lemma 1 of Xu and Raginsky (2017). ∎

Appendix D. Supporting Material for the Examples in Section 4

D.1. Metric Entropy Calculations for Reproducing Kernel Hilbert Spaces

In this section we compute the metric entropy of the unit ball of a Reproducing Kernel Hilbert Space (RKHS) of real-valued functions $f : \mathsf{X} \to \mathbb{R}$ subject to an eigenvalue decay condition.

Let us recall some facts about Reproducing Kernel Hilbert Spaces and their embeddings into $L^2(\mathbf{P})$. Assume that X is a compact subset of $\mathbb{R}^{d_x}$ and let $K : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$ be a continuous positive semidefinite kernel function. Suppose further that the Hilbert-Schmidt norm of K with respect to the probability measure $\mathbf{P}$ on X is finite: $\int\int K^2(x,z)d\mathbf{P}(x)d\mathbf{P}(z) < \infty$. A consequence of Mercer's Theorem (Theorem 12.20 and Corollary 12.26 of Wainwright (2019)) is that there exists an orthonormal basis of $\{\phi_i\}_{i=1}^{\infty}$ of $L^2(\mathbf{P})$ and a sequence of nonnegative real numbers $\{\lambda_i\}_{i=1}^{\infty}$ such that $K(x,z) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(z)$. Moreover, the RKHS associated to K is given by

$$\mathbb{H} = \left\{ f = \sum_{i=1}^{\infty} b_i \phi_i \middle| \{b_i\} \subset l^2(\mathbb{N}), \sum_{i=1}^{\infty} \frac{b_i^2}{\lambda_i} < \infty \right\}$$

with the inner product $\langle f, g \rangle_{\mathbb{H}} = \sum_{i=1}^{\infty} \frac{\langle f, \phi_i \rangle \langle g, \phi_i \rangle}{\lambda_i}$, where $\langle \cdot, \cdot \rangle$ is the standard inner product in $L^2(\mathbf{P})$. The unit ball in $\mathbb{H}$ is therefore given by

$$\mathbb{B}_{\mathbb{H}} = \left\{ f = \sum_{i=1}^{\infty} b_i \phi_i \middle| \{b_i\} \subset l^2(\mathbb{N}), \sum_{i=1}^{\infty} \frac{b_i^2}{\lambda_i} \leq 1 \right\}.$$

With this background established, we are now poised to compute the metric entropy of $\mathbb{B}$.

Proposition 16 *Let X be a compact subset of $\mathbb{R}^{d_x}$ and K be a continuous positive semidefinite kernel function. Assume that $\mathbb{H}$ is a RKHS generated by the kernel K, which further satisfies the eigenvalue decay condition $\lambda_j \lesssim j^{-2\alpha}$, for some $\alpha > 0$. Assume further that the eigenfunctions ϕ_j of K are uniformly bounded; $\sup_{x \in \mathsf{X}} |\phi_j(x)| \leq A$ for all $j = 1, 2, \ldots,$. Then*

$$\log \mathcal{N}(\mathbb{B}_{\mathbb{H}}, \|\cdot\|_{\infty}, \delta) \lesssim A^{\alpha} \varepsilon^{-1/\alpha} \log \left(1 + \lambda_1 \frac{A^{1+1/\alpha}}{\delta^{1+1/2\alpha}} \right).$$

Observe that $\mathbb{B}_{\mathbb{H}}$ is an ellipsoid in L^2, which is essentially ill-conditioned due to the eigenvalue decay condition. We shall show that it suffices to construct a covering for a finite-dimensional section of this ellipsoid corresponding to the large eigenvalues of the kernel K. In other words, at scale δ the ellipsoid $\mathbb{B}_{\mathbb{H}}$ "looks" finite-dimensional. The proof determines this critical dimension for a given $\delta > 0$.

Proof We may assume that the eigenvalues $\{\lambda_i\}$ are ordered as $\lambda_1 \geq \lambda_2 \geq \ldots$. Fix an integer m and define

$$\mathbb{B}_m = \left\{ g = \sum_{i=1}^{m} b_i \phi_i \middle| \sum_{i=1}^{m} \frac{b_i^2}{\lambda_i} \leq 1 \right\}.$$

Observe that for every $f \in \mathbb{B}_{\mathbb{H}}$ there exists $g \in \mathbb{B}_m$ such that $\|f - g\|_{L^{\infty}} \leq A\sqrt{\lambda_{m+1}}$.

Observe now that for every $g \in \mathbb{B}_m$, $g = \sum_{i=1}^m b_i \phi_i(\cdot), b = (b_1, \dots, b_m)$ we have

$$\|g\|_{L^\infty} = \left\| \sum_{i=1}^m b_i \phi_i(\cdot) \right\|_{L^\infty} \leq A \|b\|_{l^1(\mathbb{R}^m)}. \tag{39}$$

Using this, we obtain a covering of $((B)_m, \|\cdot\|_\infty)$ by regarding it as a subset of $\mathbb{R}^m$. Namely, choose $N \in \mathbb{N}$ so that $\{b^1, \dots, b^N\}$ is an optimal (δ/A)-covering of

$$B_m = \left\{ b \in \mathbb{R}^m \middle| \sum_{i=1}^m \frac{b_i^2}{\lambda_i} \leq 1 \right\}$$

in the metric of $l^1(\mathbb{R}^m)$ and extend it to a δ-covering of $\mathbb{B}_m$ in supremum norm by introducing $\{(b^1)^\top \phi(\cdot), \dots, (b^N)^\top \phi(\cdot)\}$ where $\phi(\cdot) = (\phi_1(\cdot), \dots, \phi_m(\cdot))$ and using (39). Now, the finite-dimensional norms are all equivalent and in particular we have that$\|\cdot\|_{l^1(\mathbb{R}^m)} \leq \sqrt{m} \|\cdot\|_{l^2(\mathbb{R}^m)}$. Hence, by rescaling apropriately, we require at most $\left(1 + \frac{2A\lambda_1\sqrt{m}}{\delta}\right)^m$ points to cover B_m in l^1-metric, which we may thus take as an upper bound for N.

By hypothesis that $\lambda_j \lesssim j^{-2\alpha}$ it suffices to take $m \asymp (\delta/A)^{-1/\alpha}$ for the above covering to also cover the entirety of $\mathbb{B}_\mathbb{H}$ in L^∞ since then every point of $\mathbb{B}$ is at most distance δ removed from a point of $\mathbb{B}_m$, which in turn is at most δ removed from the covering. It follows that

$$\begin{aligned} \log \mathcal{N}(\mathbb{B}_\mathbb{H}, \|\cdot\|_\infty, \delta) &\lesssim \log \left(1 + \frac{A\lambda_1\sqrt{m}}{\delta}\right)^m \\ &\lesssim (\delta/A)^{-1/\alpha} \log \left(1 + \lambda_1 \frac{A^{1+1/2\alpha}}{\delta^{1+1/2\alpha}}\right) \end{aligned}$$

which we sought to prove. ■

D.2. An Experiment Supporting Example 2

To empirically verify our claim regarding the rate of convergence of the LSE (4) for Example 2 we simulate data from an autoregressive system (2) with $f_\star$ belonging to the RKHS with radial basis function kernel $K(x,z) = \exp\left(-\frac{1}{2}\|x-z\|_2^2\right)$. More precisely, we generate a random $f_\star$ with order $k = 10000$ and state dimension $d_x = 5$ by first generating $\eta \in \mathbb{R}^{d_x \times k}$ with entries η_{ij}, drawn i.i.d. from a standard normal distribution and $\Theta \in \mathbb{R}^{k \times d_x}$ with Θ_{ij} drawn i.i.d., also from a standard normal distribution. We then set $\bar{K}(\eta, \cdot) = (K(\eta)_i, \cdot)_{i=1}^k$, and $\tilde{\Theta} = \rho\Theta/|\Theta|$, for $\rho > 0$ a parameter used to control the Lipschitz constant of $f_\star$ and where $|\cdot|$ denotes the matrix operator norm. Finally, we choose $f_\star(\cdot) = \tilde{\Theta}^\top \bar{K}(\eta, \cdot)$. Note that since $K(x,z)$ is 1-Lipschitz in either argument, $f_\star$ is guaranteed to be ρ-contractive if $\rho < 1$.

We then use $f_\star$ to generate training trajectories of varying length to be used in the LSE (4), as well as use $f_\star$ to generate 500 i.i.d. draws from the stationary distribution[5]. To implement the LSE (4) we pass by the dual problem, kernel ridge regression, to estimate $\hat{f}$. We then approximate the 2-norm distance $\|\hat{f}(\xi) - f_\star(\xi)\|_2$ by drawing 1000 fresh trajectories of length and averaging over the final sample. We average our results over 10 independent systems (random draws of $f_\star$) and plot

5. Approximated by running the system for a burn in time of $T = 1000$ time-steps before sampling from it.

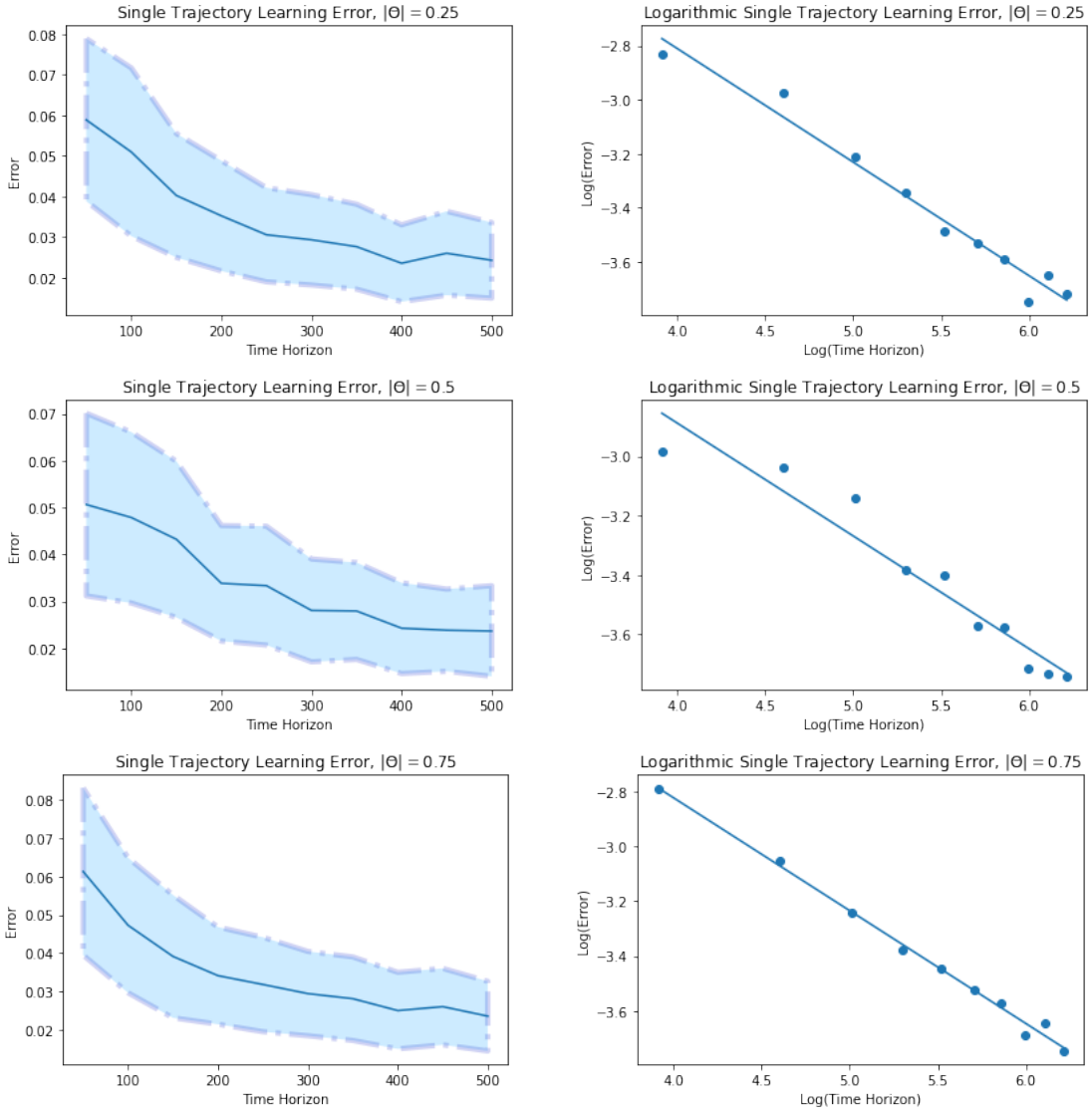

Figure 1: Convergence of the LSE (4) in terms of the error $\|\hat{f}(\xi) - f_\star(\xi)\|_2$ using data from a single trajectory and with a time horizon $T = 500$. The plots on the left illustrate the convergence of the LSE with error bars and the right plots show this on a logarithmic scale. Notice that the slope of the line any of the rightmost plots is slightly less steep than $-1/2$.

our experiment in Figure 1. It is interesting to note that the slope of the logarithmic plot is slightly less steep than $-1/2$. This is consistent with the near parametric rate of convergence suggested by Example 2 and the exponential eigenvalue decay of the kernel $K(x, z) = \exp\left(-\frac{1}{2}\|x - z\|_2^2\right)$, see Wainwright (2019), page 399.

D.3. Proof of the claim in Example 3

To use Theorem 3 we need to bound the covering number of $\mathbb{F}^\phi$. Define

$$\mathbb{M}_{d_x,d_x} = \{A \in \mathbb{R}^{d_x \times d_x}, \|A\|_F \leq C\}.$$

Then it is well-known that $\log \mathcal{N}(\mathbb{M}_{d_x,d_x}, \|\cdot\|_F, \delta) \leq 2d_x \log(1 + 2C/\delta)$. Let now $\{A_1, \dots A_N\}$ be an optimal δ-cover of $\mathbb{M}_{d_x,d_x}$. Then for every $A \in \mathbb{M}_{d_x,d_x}$ we can find $A_i \in \{A_1, \dots A_N\}$ such that

$$\|\phi(A\,\cdot\,) - \phi(A_i\,\cdot\,)\|_\infty \leq \sup_{x \in X} \|(A - A_i)x\|_2 \leq \|A - A_i\|_F \sup_{x \in X} \|x\|_2 \leq \delta B.$$

Hence any δ-covering of $\mathbb{M}_{d_x,d_x}$ induces a $B\delta$-covering of $\mathbb{F}^\phi$ and we have established the upper bound

$$\begin{aligned}\log \mathcal{N}(\mathbb{F}^\phi, \|\cdot\|_\infty, \delta) &\leq \log \mathcal{N}(\mathbb{M}_{d_x,d_x}, \|\cdot\|_F, \delta/BC) \\ &\leq 2d_x^2 \log(1 + 2BC/\delta).\end{aligned}$$

By Theorem 3 we thus have

$$\mathbf{E}\|\hat{f}(\xi) - f(\xi)\|_2 \lesssim \sqrt{\frac{\sigma_w^2}{T} d_x^2 \log(1 + 2\sigma_w \sqrt{d_x} BCT^2)} + \sqrt{\sigma_T^2(\mathbb{F}^\phi) d_x^2 \log(1 + 2BCT)}.$$

so that the result follows by using Proposition 4 to bound $\sigma_T^2(\mathbb{F}^\phi, P_\star)$. ■