An Empirical Bernstein Inequality for Dependent Data in Hilbert Spaces and Applications

Erfan Mirzaei[1,2] **Andreas Maurer**[1] **Vladimir R. Kostic**[1,3] **Massimiliano Pontil**[1,4]
[1]Italian Institute of Technology, [2]University of Genoa, [3]University of Novi Sad, [4]University College London

Abstract

Learning from non-independent and non-identically distributed data poses a persistent challenge in statistical learning. In this study, we introduce data-dependent Bernstein inequalities tailored for vector-valued processes in Hilbert space. Our inequalities apply to both stationary and non-stationary processes and exploit the potential rapid decay of correlations between temporally separated variables to improve estimation. We demonstrate the utility of these bounds by applying them to covariance operator estimation in the Hilbert-Schmidt norm and to operator learning in dynamical systems, achieving novel risk bounds. Finally, we perform numerical experiments to illustrate the practical implications of these bounds in both contexts.

1 INTRODUCTION

Learning from non-independent and identically distributed (non-i.i.d.) data presents significant challenges in machine learning, both from theoretical and practical perspectives. Most real-world data do not follow the neat, predictable patterns of i.i.d. scenarios, creating a demand for statistical learning techniques to handle more complex random processes, thereby broadening the applicability of learning algorithms.

In this paper, we present data-dependent Empirical Bernstein Inequalities (EBIs), which apply to vector-valued random processes in Hilbert space. A driving motivation for this work is recent studies on learning operators associated with stochastic dynamical systems (Kostic et al., 2022, 2024a).

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

Stochastic dynamical systems are essential for modeling complex phenomena across diverse fields, from finance, where they describe asset price fluctuations and stochastic volatility (Tankov, 2003), to neuroscience, where they capture neural variability and synaptic noise (Rusakov et al., 2020; Schug et al., 2021), and climate science, where they model turbulent atmospheric and oceanic dynamics (Majda and Harlim, 2012). A key example is Langevin dynamics, which describes molecular motion in a thermal environment by incorporating both deterministic forces and random fluctuations, making it fundamental for simulating biomolecular systems, modeling Brownian motion, and analyzing stochastic processes in physics, chemistry, and electrical engineering (Coffey and Kalmykov, 2012).

Many such processes are slowly exploring the state space, and one has to wait a long time before two points along the process can be considered independent. Such phenomena are formalized via the notion of mixing. Unfortunately, for largely unknown processes the quantification of mixing is unfeasible, and it may be hard even when dealing with data generated from simulations based on mathematical models which ensure that the dynamical system is mixing. This motivates the empirical concentration inequalities for dependent random variables in Hilbert spaces since their primary objective is to minimize the inequality's dependence on mixing coefficients that are often unknown.

In this respect, EBIs contrast the more classical combinations of Bernstein inequalities and mixing assumptions. Current estimation bounds for covariance operators exhibit a reduced effective sample size, which roughly requires dividing the length of the trajectory by the mixing time. We show that our EBIs allow us to derive estimation bounds in which these large mixing times mainly impact the fast $O(1/n)$ term in the bound, while the slow $O(1/\sqrt{n})$ term involves only the time-lag correlation/variance of within-block average of the process, which may be very small even for slowly mixing processes. Notably, when the variables observed along a trajectory decorrelate much more rapidly than they achieve approximate independence, Bernstein's

inequality, with its smaller variance term, facilitates $O(1/n)$ convergence rates, a significant improvement over standard ones.

This utility is enhanced by combining the inequality with precise estimates of its variance term, further differentiating our empirical bounds from others and highlighting the novelty of our approach compared to related works. Finally, we highlight that our bound only requires prior knowledge of the mixing coefficient of the process, with all other quantities being data-dependent.

Previous Work The idea of combining Bernstein's inequality with estimates of the variance term is not new. It was first applied to reinforcement learning and general learning theory ((Audibert et al., 2007), (Maurer and Pontil, 2009), (Audibert et al., 2009)) and has been extended to improve estimates on Martingales (Peel et al., 2013; Waudby-Smith and Ramdas, 2024), U-statistics (Peel et al., 2010), PAC-Bayesian bounds (Tolstikhin and Seldin, 2013) and certain Banach space-valued random variables (Martinez-Taboada and Ramdas, 2024). To our knowledge, the present paper gives the first application to weakly dependent variables in a Hilbert space. Other relevant works on learning with non-i.i.d. data include (Hang and Steinwart, 2014; Steinwart and Christmann, 2009; Steinwart et al., 2009; Smale and Zhou, 2009; Hang and Steinwart, 2017; Modha and Masry, 1996; Blanchard and Zadorozhnyi, 2019; Liu and Austern, 2023; Alquier et al., 2019; Abeles et al., 2024; Abélès et al., 2024; Chatterjee et al., 2024). These studies do not consider data-dependent bounds for random variables in Hilbert spaces or focus on operator learning, so they cannot be directly compared with ours.

Within the context of operator learning, it's notable that (Kostic et al., 2022) utilizes the block method from (Yu, 1994) to derive estimation bounds for the covariance operator in the domain of transfer operator learning. The concept of a mixing assumption in learning theory is likely first introduced by (Yu, 1994), alongside a technique for method of interlacing block sequences to derive empirical bounds. The authors (Modha and Masry, 1996) soon applied these ideas to provide a version of Bernstein's inequality, a scalar version of Theorem 1. Since then, mixing and the method of blocks have been used by numerous authors ((Meir, 2000), (Mohri and Rostamizadeh, 2008), (Steinwart and Christmann, 2009), (Agarwal and Duchi, 2012), (Shalizi and Kontorovich, 2013), and others). We follow a similar approach, with the distinction that our data consist of a sequence of vectors $X_1, X_2, ...$ in a Hilbert space, and we combine Bernstein's inequality with empirical estimates of its variance term.

Contributions In summary our main contributions are: **1)** We present novel empirical Bernstein inequalities for a sequence of vectors in a Hilbert space; **2)** We apply these inequalities to derive estimation bounds for the covariance and cross-covariance matrices of the process, showing improvement over recent bounds in the context of stochastic dynamical system and Koopman operator regression (KOR); **3)** We use our EBI to prove risk bounds for learning stochastic processes, which due to its empirical nature avoids the need for (typically unverifiable) regularity assumptions; **4)** We present experiments illustrating our theory, and, notably, show that our bounds help understanding generalization in moderate sample-size regimes, and can serve as practical model selection tool in learning dynamical systems.

2 THEORETICAL RESULTS

The objective of this section is to bound the error incurred when estimating the mean of a random vector by its average on an observed trajectory. We study the norm of the random variable

$$\frac{1}{n}\sum_{t=1}^{n}\left(X_t - \mathbb{E}\left[X_t\right]\right), \tag{1}$$

where $\mathbf{X} = (X_1, ..., X_n)$ is a vector of random variables in a separable Hilbert space $\mathcal{H}$, representing the observations along the trajectory. The principal difficulty is the mutual dependence of the X_t. To explain our assumptions to replace independence, we briefly define the β-mixing coefficients for stochastic processes and describe the method of blocks to approximate a sum of dependent random vectors by a sum of independent block-averages. After that, we state our main results and conclude this section with a sketch of their proofs.

2.1 Backgrounds on $\beta-$mixing Coefficients and the Method of Blocks

The key idea for handling dependent data observed from a random, temporal process is that, while subsequent observations may be strongly dependent, they often become approximately independent when separated by sufficiently large time intervals. Such processes tend to forget their distant past.

To estimate the mean of a random vector in Hilbert space from the observation of a single trajectory $\mathbf{X}=(X_1, ..., X_n) \sim \mu$ of a largely unknown random process, for $\tau \in \mathbb{N}$, we define the mixing coefficients (Bradley, 2005),

$$\beta_\mu(\tau) = \sup_{j\in\mathbb{N}} \sup_{B\in\Sigma([1,j]\cup[j+\tau,\infty))} \left|\mu_{[1,j]\cup[j+\tau,\infty)}(B) - \mu_{[1,j]} \times \mu_{[j+\tau,\infty)}(B)\right|.$$

Here $\Sigma([1,j] \cup [j+\tau, \infty))$ is the set of events depending on the X_t with $t \in [1,j]$ and $t \in [j+\tau, \infty)$, the measure $\mu_{[1,j] \cup [j+\tau,\infty)}$ is the joint distribution of these variables, and $\mu_{[1,j]} \times \mu_{[j+\tau,\infty)}$ is the product measure, where events in $\Sigma([1,j])$ and $\Sigma([j+\tau,\infty))$ are independent (we write μ_I for the joint distribution of $X_i : i \in I$). By definition this independence assumption incurs a penalty of $\beta_\mu(\tau)$. For m events, mutually separated by τ time increments, the penalty of assuming them to be independent then increases to $(m-1)\beta_\mu(\tau)$. The mixing coefficients are necessarily non-increasing. Typical assumptions are algebraic mixing, $\beta_\mu(\tau) \approx \tau^{-p}$, or exponential mixing $\beta_\mu(\tau) \approx \exp(-p\tau)$ for $p > 0$.

Let us assume that n is an even multiple of some τ, that is, $n = 2m\tau$. Following the seminal work of (Yu, 1994) we divide the entire time interval into two sequences of blocks of length τ, where the blocks in each sequence are mutually separated by τ. Thus one sequence is $I_1, I_2, ..., I_m$, where each I_k has τ points and the distance between different I_k and I_l is at least τ. The other sequence $I'_1, I'_2, ..., I'_m$ has the same properties and fills the gaps left by the first sequence. In other word for $\tau \in \mathbb{N}$ and $k \in [m]$ these index sets are $I_k = \{2(k-1)\tau+1, ..., (2k-1)\tau\}$ and $I'_k = \{(2k-1)\tau+1, ..., 2k\tau\}$.

If we drop the normalizing factor $1/n$, which can always be re-inserted in our bounds, the random variable in (1) can then be written as

$$\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) = \sum_{k=1}^{m}(Y_k - \mathbb{E}[Y_k]) + \sum_{k=1}^{m}(Y'_k - \mathbb{E}[Y'_k]), \quad (2)$$

where the $Y_1, ..., Y_m$ and $Y'_1, ..., Y'_m$ are the "block sums"

$$Y_k = \sum_{i \in I_k} X_i \text{ and } Y'_k = \sum_{i \in I'_k} X_i. \quad (3)$$

Clearly the Y_i are mutually separated by τ time increments, so they may be assumed mutually independent at a penalty of $(m-1)\beta_\mu(\tau)$, and the same holds for the Y'_i. To express these independencies we write $\Pr_I$ for the probability measure $\mu_{I_1} \times ... \times \mu_{I_m}$ on $\Sigma(I_1 \cup ... \cup I_m)$ and $\Pr_{I'}$ for $\mu_{I'_1} \times ... \times \mu_{I'_m}$ on $\Sigma(I'_1 \cup ... \cup I'_m)$. Then the above ideas are summarized by the following lemma, with detailed proof in Appendix A.3.

Lemma 1. *Let X_i have values in a normed space $(\mathcal{X}, \|\cdot\|)$ and let $F, F' : \mathcal{X}^n \to \mathbb{R}$, where F is $\Sigma(I_1 \cup ... \cup I_m)$-measurable, and F' is $\Sigma(I'_1 \cup ... \cup I'_m)$-measurable. Then*

$$\Pr\left\{\left\|\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right\| > F(\mathbf{X}) + F'(\mathbf{X})\right\}$$
$$\leq \Pr_I\left\{\left\|\sum_{k=1}^{m}(Y_k - \mathbb{E}[Y_k])\right\| > F(\mathbf{X})\right\}$$
$$+ \Pr_{I'}\left\{\left\|\sum_{k=1}^{m}(Y'_k - \mathbb{E}[Y'_k])\right\| > F'(\mathbf{X})\right\} + 2(m-1)\beta_\mu(\tau),$$

where the Y_k and Y'_k are given by (3).

So the problem of bounding the norm of the dependent sum is reduced to bounding the norms of two independent sums, albeit with an effective sample size reduced by a factor of 2τ. The interlaced sequences of blocks are a standard method to port bounds from the independent to the dependent case. The novelty here is the introduction of F and F', needed for our empirical bounds.

For an unknown process, explored only by observation of a single trajectory, the coefficients $\beta_\mu(\tau)$ are fixed largely based on plausibility, making τ very uncertain. Any bound on the estimation error should therefore depend as little as possible on τ, which determines the effective sample size $m = n/(2\tau)$.

2.2 Bernstein Inequalities for Vector-Valued Processes

Bernstein-type concentration inequalities for functions of m independent variables bound an estimation error by the sum of two terms, a rapidly decreasing term of order $\frac{1}{m}$ and another term, which decreases as $\sqrt{V/m}$, where V is related to the variance of the variables. For the blocking technique, this variance becomes the average of variances of within-block averages, $\frac{1}{\tau}Y_k$s and $\frac{1}{\tau}Y'_k$s. Bernstein's inequality can exploit the fact that the correlation of temporally separated variables often decreases orders of magnitude faster than they attain approximate independence. The mixing time τ then enters mainly in the rapidly decreasing term of order τ/n. This fact has also been pointed out by (Ziemann et al., 2024).

For a specified τ define the set $S_\tau \subseteq [n] \times [n]$ by $S_\tau = \bigcup_{k=1}^{m}(I_k \times I_k) \cup (I'_k \times I'_k)$, where $|S_\tau| = 2m\tau^2$, and

$$V_\tau(\mathbf{X}) = \frac{1}{|S_\tau|}\sum_{(t,s)\in S_\tau}\mathbb{E}[\langle X_t, X_s\rangle] - \langle \mathbb{E}[X_t], \mathbb{E}[X_s]\rangle \quad (4)$$

The following is our first result.

Theorem 1. *Let $m, \tau \in \mathbb{N}$, $n = 2m\tau$, and let $\mathbf{X} = (X_1, ..., X_n)$ be a vector of random variables in a separable Hilbert-space $\mathcal{H}$, satisfying $\|X_t\| \leq c$ for all t.*

Let $\delta(\tau) = \delta - 2(\frac{n}{2\tau}-1)\beta_\mu(\tau) > 0$. *Then with probability at least* $1-\delta$ *we have*

$$\left\| \frac{1}{n}\sum_{t=1}^{n} (X_t - \mathbb{E}[X_t]) \right\| \leq \sqrt{\frac{2\tau V_\tau(\mathbf{X})}{n}\left(1+2\ln\frac{2}{\delta(\tau)}\right)} + \frac{8\tau c}{3n}\ln\frac{2}{\delta(\tau)}.$$

If the failure probability δ is fixed, τ must be sufficiently large to satisfy $\delta > 2(\frac{n}{2\tau}-1)\beta_\mu(\tau)$ and result from the specific assumptions on the decay of mixing coefficients. The variance surrogate, $V_\tau(\mathbf{X})$, can be bounded by c^2, so the entire term is at worst of order $\sqrt{\tau/n}$, but it may become arbitrarily small depending on the joint distribution of the X_t.

If the law of the process is unknown, then the previous result is not satisfactory, since in addition to the uncertain, but unavoidable, mixing assumptions, we now also need assumptions on the behaviour of correlations, unless we want to return to the worst-case bound. Fortunately $V_\tau(\mathbf{X})$ can be estimated from the same trajectory and the estimates can be combined with Theorem 1 to give empirical Bernstein-type inequalities, an idea which has been successfully applied to a variety of problems (Audibert et al., 2007; Maurer and Pontil, 2009; Audibert et al., 2009; Burgess et al., 2020; Jin et al., 2022; Tolstikhin and Seldin, 2013; Shivaswamy and Jebara, 2010; Peel et al., 2010, 2013).

We give two versions, using a biased variance estimate for general processes and an unbiased estimate for stationary processes. The symmetric structure of S_τ implies, that $\sum_{(t,s)\in S_\tau}\langle \mathbb{E}[X_t], \mathbb{E}[X_s]\rangle \geq 0$, so that the centered correlations in $V_\tau(\mathbf{X})$ can be bounded by uncentered ones. Using this biased estimator

$$\widetilde{V}_\tau(\mathbf{X}) = \frac{1}{|S_\tau|}\sum_{(t,s)\in S_\tau}\langle X_t, X_s\rangle, \tag{5}$$

where $|S_\tau| = 2m\tau^2$, then leads to the following.

Theorem 2. *Under the conditions of Theorem 1 we have for* $\delta \in (0,1)$, *and* $\delta(\tau) = \delta - 2(\frac{n}{2\tau}-1)\beta_\mu(\tau) > 0$ *with probability at least* $1-\delta$ *that*

$$\left\| \frac{1}{n}\sum_{i} (X_i - \mathbb{E}[X_i]) \right\| \leq \sqrt{\frac{2\tau \widetilde{V}_\tau(\mathbf{X})}{n}\left(1+2\ln(\frac{4}{\delta(\tau)})\right)} + \frac{32\tau c}{3n}\ln\frac{4}{\delta(\tau)}.$$

If the observed, uncentered correlations decay quickly relative to τ the first term can be nearly as small as $\sqrt{1/n}$, but no smaller. Because we have at least n elements in the diagonal of the whole matrix so $\widetilde{V}_\tau(\mathbf{X})$ can not be smaller than $\min_t \|X_t\|_{\mathcal{H}}^2/\tau$.

We give an improved estimate for stationary processes. A process is called stationary if the joint distributions satisfy $\mu_I = \mu_{I+t}$ for any $t \in \mathbb{N}$. In this case, we can estimate the unbiased variance using a u-statistic as follows:

$$\widehat{V}_\tau(\mathbf{X}) = \frac{1}{|S_\tau|}\left(\sum_{(t,s)\in S_\tau}\langle X_t, X_s\rangle - \frac{1}{m-1}\sum_{(t,s)\in \tilde{S}_\tau}\langle X_t, X_s\rangle\right) \tag{6}$$

where $|S_\tau| = 2m\tau^2$, and $|\tilde{S}_\tau| = 2m(m-1)\tau^2$. Here

$$\tilde{S}_\tau = \bigcup_{k\neq l : k,l\in[m]} (I_k \times I_l) \cup (I'_k \times I'_l).$$

Theorem 3. *Under the conditions of Theorem 1, if the process is also stationary, we have for* $\delta \in (0, 2/e)$ *and* $\delta(\tau) = \delta - 2(\frac{n}{2\tau}-1)\beta_\mu(\tau) > 0$ *with probability at least* $1-\delta$ *that*

$$\left\| \frac{1}{n}\sum_{i} (X_i - \mathbb{E}[X_i]) \right\| \leq \sqrt{\frac{2\tau \widehat{V}_\tau(\mathbf{X})}{n}\left(1+2\ln(\frac{4}{\delta(\tau)})\right)} + \frac{22\tau c}{n}\ln(\frac{4}{\delta(\tau)}).$$

Think of $[n]\times[n]$ as the surface of a chessboard with fields of side length τ. Then $m = 4$, and S_τ is the union of the squares on the white diagonal, and $\tilde{S}_\tau$ is the union of all other white squares (see Figure 1). If $(s,t) \in S_\tau$ then s and t are no more than $\tau-1$ apart, while for $(s,t) \in \tilde{S}_\tau$ they are at least τ apart. Notice

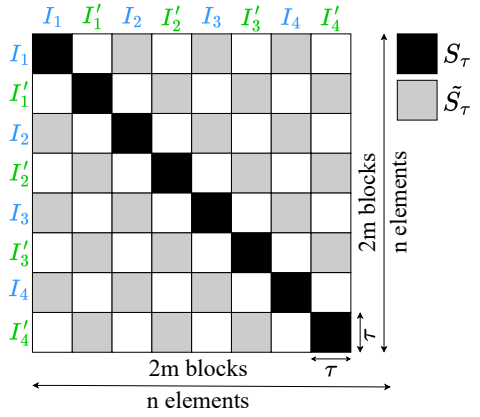

Figure 1: The blocks belonging to S_τ and $\tilde{S}_\tau$.

that the passage to empirical bounds incurs additional constant factors only in the bound for the last term.

2.3 Proofs

Sketch of proof (detailed in the supplementary file). As explained in Section 2.1 the method of blocks provides a tool to port bounds from the independent to

the dependent case, so we first establish inequalities for independent data.

A concentration inequality of (McDiarmid, 1998, Theorem 3.8) quite easily leads to the following Bernstein-type inequality for independent, centered vectors Y_i in Hilbert space, satisfying $\|Y_i - \mathbb{E}Y_i\| \leq c$.

$$\Pr\Bigg\{\left\|\sum_k Y_k - \mathbb{E}\left[Y_k\right]\right\| > \sqrt{\sum_k \mathbb{E}\left\|Y_k - \mathbb{E}\left[Y_k\right]\right\|^2} \left(1+\sqrt{2\ln\left(1/\delta\right)}\right) + \frac{4c}{3}\ln\left(1/\delta\right)\Bigg\} < \delta. \tag{7}$$

The additional 1 in $1+\sqrt{2\ln\left(\frac{1}{\delta}\right)}$ arises from a bound of the expexted norm by the root of the variance. To obtain empirical bounds we want to combine this with estimates of the variance term. Without additional assumptions, we obtain from a concentration inequality for real-valued functions that

$$\sqrt{\sum_k \mathbb{E}\left\|Y_k - \mathbb{E}\left[Y_k\right]\right\|^2} \leq \sqrt{\sum_k \left\|Y_k\right\|^2} + c\sqrt{2\ln\left(1/\delta\right)}.$$

Combining this with (7) in a union bound gives, with probability at least $1-\delta$

$$\Pr\Bigg\{\left\|\sum_k Y_k - \mathbb{E}\left[Y_k\right]\right\| > \sqrt{\sum_k \left\|Y_i\right\|^2}\left(1+\sqrt{2\ln\left(2/\delta\right)}\right) + \frac{16c}{3}\ln\left(2/\delta\right)\Bigg\} < \delta.$$

This is our independent template for Theorem 2.

To obtain Theorem 3, we can use stationarity of the process, which means that the Y_k (or the Y_k') can now be assumed to have identical distribution. In this case, a slightly more involved argument gives the estimate

$$\sqrt{\sum_k \mathbb{E}\left\|Y_k - \mathbb{E}\left[Y_k\right]\right\|^2} \leq \sqrt{\frac{1}{2\left(m-1\right)}\sum_{k,l:k\neq l}\left\|Y_k - Y_l\right\|^2} + 4c\sqrt{2\ln\left(1/\delta\right)}.$$

Again a union bound with (7) gives the independent template for Theorem 3.

Now we port these inequalities to the dependent case using Lemma 1. For Theorem 1, we define

$$F\left(\mathbf{X}\right) = \sqrt{\sum_{k=1}^m \mathbb{E}\left\|Y_k - \mathbb{E}\left[Y_k\right]\right\|^2}\left(1+\sqrt{2\ln\left(2/\delta\right)}\right) + \frac{4\tau c}{3}\ln\left(2/\delta\right)$$

with $F'\left(\mathbf{X}\right)$ defined analogously, replacing $\overline{Y}_k$ by Y_k'. Substitution of these definitions in the body of Lemma 1 , using (7) to bound the two independent probabilities, some algebraic simplifications, and reinsertion of the overall factor of $1/n = 1/\left(2m\tau\right)$ give Theorem 1.

Above F and F' were simply constant functions. This is different for the empirical bounds, where they contain the variance estimates. To obtain Theorem 2 we let

$$F\left(\mathbf{X}\right) = \sqrt{\sum_{k=1}^m \left\|Y_k\right\|^2}\left(1+\sqrt{2\ln\left(4/\delta\right)}\right) + \frac{16c\tau}{3}\ln\left(4/\delta\right),$$

which is $\Sigma\left(I_1 \cup \ldots \cup I_m\right)$-measurable, and replace Y_k by Y_k' for the analogous definition of $F'\left(\mathbf{X}\right)$. Then, using Lemma 1 and (2.3), unraveling the definitions and some simplifications, give Theorem 2. Similarly

$$F\left(\mathbf{X}\right) = \sqrt{\frac{1}{2\left(m-1\right)}\sum_{k,l:k\neq l}\left\|Y_k - Y_l\right\|^2}\left(1+\sqrt{2\ln\left(4/\delta\right)}\right) + 11c\tau\ln\left(4/\delta\right),$$

and the corresponding $F'\left(\mathbf{X}\right)$ give Theorem 3.

3 APPLICATIONS

In this section, we address two important fields of study where our bound can lead to new advances. First, we note that the concentration inequality is naturally linked to covariance estimation, which plays an important role in machine learning and statistics (Markowitz, 1952; von Storch and Navarra, 1999; Schäfer and Strimmer, 2005; Mollenhauer et al., 2022). The second application concerns data-driven dynamical systems and in particular transfer operators, which are also widely used in science and engineering (see e.g. Brunton et al., 2022; Tuckerman, 2023, and references therein).

3.1 Covariance Estimation

We apply our results to the estimation of covariance operators on the Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ with an associated kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Letting $\phi : \mathcal{X} \to \mathcal{H}$ be a *feature map* such that $k(x, x') = \langle \phi(x), \phi(x') \rangle$ for all $x, x' \in \mathcal{X}$, we aim to bound the Hilbert-Schmidt norm estimation error. Hence, in this setting, the observed vectors are operators, and to apply our bounds one needs to replace X_t by the rank-one operator $Y_t = \phi(X_t) \otimes \phi(X_t)$. Then, $\widetilde{V}_\tau(\mathbf{Y})$ and $\widehat{V}_\tau(\mathbf{Y})$ can be easily computed by using entries of the kernel matrix. If the process is not stationary, we can only estimate the ergodic average of covariance operators, that is, $(1/n)\sum_{t=1}^n \mathbb{E}\left[\phi(X_t) \otimes \phi(X_t)\right]$, which becomes $C = \mathbb{E}\left[\phi(X_1) \otimes \phi(X_1)\right]$ for stationary processes. The transcription of the previous results fortunately is affected by simply replacing all

$T = (1/n)\sum_{t=1}^{n} \mathbb{E}\left[\phi(X_t) \otimes \phi(X_{t+1})\right]$ because the inner products change; the block variables are then only separated by $\tau - 1$ instead of τ, and we also need to observe one more point on the trajectory.

To demonstrate the application of our empirical Bernstein's inequalities to estimate the covariance and the improvements it introduces, we first adapt Pinelis and Sakhanenko's concentration inequality for random variables in a separable Hilbert space (see (Caponnetto and De Vito, 2007, Proposition 2)) to trajectory data, using the method of blocks and β-mixing, as it was suggested in (Kostic et al., 2023), see Appendix C.1

Assuming that the data is sampled from the stationary distribution, our bounds in Theorems 2 and 3 apply. Compared to worst-case bounds, the improvement lies in the correlation factors $\widehat{V}_\tau(\mathbf{X})$ and $\widetilde{V}_\tau(\mathbf{X})$ in (6) and (5), respectively, affecting the slow term in the bound. Before we illustrate this improvement empirically, we discuss a related application.

3.2 Learning Dynamical Systems with Transfer Operators

Recent advances in the statistical theory for Koopman operator learning have highlighted the significant impact of covariance estimation on the ability to generalize when forecasting and interpreting dynamical systems from data-driven models (Kostic et al., 2022; Philipp et al., 2024). While statistical learning theory has been thoroughly developed for kernel methods (Kostic et al., 2023, 2024a) under β-mixing assumptions, an important gap remains between the learning rates and generalization bounds and the practical performance of the methods. This is particularly interesting when comparing recent deep learning advances with kernel methods (Kostic et al., 2024c). Here, we briefly review the transfer operator learning and then present novel contributions to this field based on our EBI.

For a *time-homogeneous* Markov chain with an invariant (stationary) distribution π the (stochastic) *Koopman operator* $A_\pi \colon L^2_\pi(\mathcal{X}) \to L^2_\pi(\mathcal{X})$ is given by

$$[A_\pi f](x) := \int_{\mathcal{X}} p(x, dy) f(y) = \mathbb{E}\left[f(X_{t+1}) | X_t = x\right],$$

where $f \in L^2_\pi(\mathcal{X}), x \in \mathcal{X}$, and p is the transition kernel.

In many practical cases, A_π is unknown, but data from one or multiple system trajectories are available. A framework for operator regression learning was introduced in (Kostic et al., 2022) to estimate the Koopman operator on $L^2_\pi(\mathcal{X})$ within a RKHS using an associated feature map $\phi : \mathcal{X} \to \mathcal{H}$. In this vector-valued regression, the risk functional is defined as $\mathcal{R}(G) = \mathbb{E}_{X\sim\pi, X^+\sim p(\cdot|X))} \|\phi(Y) - G^*\phi(X)\|^2_{\mathcal{H}}$, and the task is to learn A_π by minimizing the risk over some class of operators $G \colon \mathcal{H} \to \mathcal{H}$ using a dataset of consecutive states $\mathcal{D}_n := (x_i, x_i^+)_{i=1}^n$. Typical scenario is to obtain the states from a single trajectory of the process after reaching the equilibrium distribution, that is $X_0 \sim \pi$, $X_i^+ \equiv X_{i+1} \sim p(\cdot \,|\, X_i)$, $i = 2, \ldots, n$, and the popular estimator in this setting is the Reduced Rank Regression (RRR) one $\widehat{G}_\lambda$ obtained by minimizing regularized empirical risk $\widehat{\mathcal{R}}_\lambda(G) := \frac{1}{n}\sum_{i\in[n]} \|\phi(x_i^+) - G^*\phi(x_i)\|^2_{\mathcal{H}} + \lambda\|G\|^2_{\text{HS}}$, over operators G of rank at most r, that is $\widehat{G}^{\text{RRR}}_{r,\lambda} = \widehat{C}_\lambda^{-1/2} [\![\widehat{C}_\lambda^{-1/2} \widehat{T}]\!]_r$ is computed via r-truncated SVD $[\![\cdot]\!]_r$ and *input* and *cross* empirical covariances $\widehat{C} = \frac{1}{n}\sum_{i\in[n]} \phi(x_i) \otimes \phi(x_i)$ and $\widehat{T} = \frac{1}{n}\sum_{i\in[n]} \phi(x_i) \otimes \phi(x_i^+)$, respectively, while $\widehat{C}_\lambda = \widehat{C} + \lambda I_{\mathcal{H}}$, c.f. (Kostic et al., 2022, 2023).

The recent works (Li et al., 2022; Kostic et al., 2023) on the mini-max optimal learning rate for KOR in i.i.d. settings crucially rely on Pinelis and Sakhanenko's inequality. As observed above, applying the method of blocks and β-mixing extends the i.i.d. analysis of KOR to realistic scenarios of learning from data trajectories of a stationary process. In the following, we present a novel risk bound for the reduced-rank Tikhonov estimator that circumvents the need for regularity assumptions, in (Li et al., 2022; Kostic et al., 2023, 2024a,b).

Theorem 4. *Let $\mathbf{X} = (X_t)_{t=1}^n$ be a stationary Markov chain with distribution μ, and the risk definitions as above noting that $\pi = \mu_1$. Denote $\widehat{G}_{r,\lambda}$ be a minimizer of reduced-rank Tikhonov regularized empirical risk and $\mathbf{Y} = (\phi(X_t) \otimes \phi(X_t))_{t=1}^n$, $\mathbf{Z} = (\phi(X_t) \otimes \phi(X_{t+1}))_{t=1}^n$, and $\mathbf{W} = (\|\phi(X_t)\|^2)_{t=1}^n$. Assume $n = 2m\tau$, and exists $c_{\mathcal{H}} > 0$ such that $\|\phi(X_t)\|^2 \leq c_{\mathcal{H}}$ a.s. for all t. Let $\delta \geq 0$ and assume $\widehat{\delta}_\mu(\tau, \lambda) := 0.5\,\delta / \|\widehat{G}_{r,\lambda}\| - 2(\frac{n}{2\tau} - 1)\beta_\mu(\tau) > 0$. Then, with probability at least $1 - \delta$ we have for every $\widehat{G}_{r,\lambda}$ such that [1] $\|\widehat{G}_{r,\lambda}\|_{HS} \geq 1$*

$$\begin{aligned}
|\mathcal{R}(\widehat{G}) - \widehat{\mathcal{R}}(\widehat{G})| \leq & \frac{128 c_{\mathcal{H}} \tau \|\widehat{G}_{r,\lambda}\| (\sqrt{r} + \|\widehat{G}_{r,\lambda}\|)}{3n} \ln \frac{12}{\widehat{\delta}_\mu(\tau, \lambda)} \\
& + \frac{14 c_{\mathcal{H}} \tau^2}{3n - 2\tau} \ln \frac{12}{\widehat{\delta}_\mu(\tau, \lambda)} \\
& + \sqrt{\frac{32 \|\widehat{G}_{r,\lambda}\|^4 \widetilde{V}_\tau(\mathbf{Y}) \tau}{n} \left(1 + 2 \ln \frac{12}{\widehat{\delta}_\mu(\tau, \lambda)}\right)} \\
& + \sqrt{\frac{8r \|\widehat{G}_{r,\lambda}\|^2 \widetilde{V}_\tau(\mathbf{Z}) \tau}{n} \left(1 + 2 \ln \frac{12}{\widehat{\delta}_\mu(\tau, \lambda)}\right)} \\
& + \sqrt{\frac{2 \overline{V}_\tau(\mathbf{W}) \tau}{n} \ln \frac{12}{\widehat{\delta}_\mu(\tau, \lambda)}},
\end{aligned}$$

where $\overline{V}_\tau(\mathbf{W}) = \frac{1}{m(m-1)\tau^2} \sum_{1 \leq i < j \leq m} (\overline{W}_i - \overline{W}_j)^2 +$

[1] This is an artifact of the conversion from Ivanov to Tikhonov-type bounds. In practice, we expect the Hilbert-Schmidt norm to be much larger than 1.

$(\overline{W'}_i - \overline{W'}_j)^2$, $\overline{W}_j = \sum_{i\in I_j} W_i$ *and* $\overline{W'}_j = \sum_{i\in I'_j} W_i$. $\widetilde{V}_\tau(\mathbf{Y})$ *and* $\widetilde{V}_\tau(\mathbf{Z})$ *were defined before.*

We remark that the only non-computable part in the bound is the mixing coefficient in the log terms, while the powers of the estimator's norm are weighted by the correlation coefficients. This is particularly interesting because the bound holds in the non-asymptotic regime, and even with moderate sample sizes, it reveals the impact of hyperparameters, namely the ridge parameter $\lambda > 0$ and rank parameter r, on the risk concentration. We illustrate this feature in the practical problem of model selection when learning molecular dynamics. Further, note that our approach can easily relax the restrictive stationarity assumptions at the cost of an additive term quantifying the distance of the initial distribution from the equilibrium one, see Appendix B for a discussion. This is crucial for scenarios where data is collected out of equilibrium, a challenge not easily addressed by approaches like (Kostic et al., 2022, 2024a).

4 EXPERIMENTS

In this section, we showcase the improvements of our empirical Bernstein inequalities for covariance estimation and for learning dynamics systems with moderate sample sizes. To facilitate the reproducibility of our main experimental results, we have made the code publicly available in a GitHub repository, which can be accessed via the link `https://github.com/erfunmirzaei/EBI4LDS`.

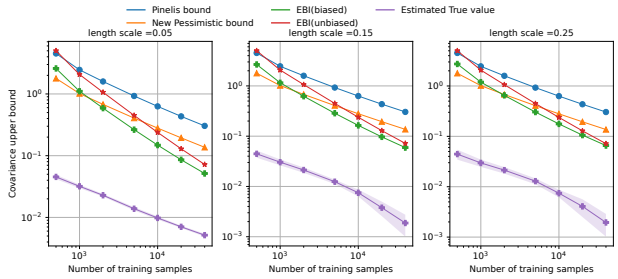

Figure 2: Covariance upper bound as a function of the number of training points for three different length scales of Gaussian kernel in logarithmic scale. The failure probability is assumed to be 0.05, and the plots have averaged over 30 independent simulations.

4.1 Covariance Estimation Using Samples from Ornstein-Uhlenbeck Process

In the first experiment, we illustrate the quantitative improvement of our EBIs in comparison to its pessimistic non-empirical version, noting that a slightly different form has been used in (Kostic et al., 2022). We use the proposed empirical inequalities to determine the concentration of the covariance operator in the Hilbert-Schmidt norm. The usefulness of the new bounds can be particularly exploited when the process decorrelates much faster than it attains independence. To demonstrate this, we use 1D equidistant sampling of the Ornstein-Uhlenbeck process, obtained by integrating $X_t = e^{-1}X_{t-1} + \sqrt{1-e^{-2}}\,\epsilon_t$, where $\{\epsilon_t\}_{t\geq 1}$ are i.i.d. samples from the standard Gaussian distribution. For this process, it is well-known (Pavliotis, 2014) that the invariant distribution, π, coincides $\mathcal{N}(0,1)$. One point that makes the use of this as a toy example beneficial is the fact that this process belongs to exponential mixing processes $\beta_\mu(\tau) \approx \exp(-p\tau)$ where p is the gap between the first and second eigenvalues of its transfer operator is also known $p = 1-1/e$. We apply a Gaussian kernel with different length scales to map our data to the RKHS with the corresponding feature map, $\phi(x_t)$.

Initially, we set a probability threshold for the failure of the inequalities. Subsequently, with this fixed failure probability in mind, we determined the appropriate mixing time τ for a given sample size n, defined as the smallest value satisfying $\delta(\tau)$. We opted for τ due to its optimality since we observed through experimental validation, that there is a consistent monotonic increase in the relationship between the empirical bounds and τ, across various training set sizes and different failure probabilities. For further details, please refer to Appendix C.1.2. In Figure 2, we plotted empirical upper bounds for covariance estimation for different numbers of training points across three different choices of length scales over 50 independent simulations. We compared the new data-dependent upper bounds against the conservative bounds derived from Pinelis and Sakhanenko's inequality, as well as the concentration inequality presented in Theorem 4. The details of computing the true error value are explained in Appendix C.1.3.

Figure 2 shows that one can significantly overestimate when using classical Bernstein-type inequalities pessimistically, especially as the number of training points grows. In other words, the slow term in the classical Bernstein inequalities may not be too slow, especially for processes where elements decorrelate rapidly. Importantly, notice that the slopes of the EBI bounds show a faster rate $\approx 1/n$ for the moderate sample size regime than the rate $1/\sqrt{n}$ that is asymptotically optimal.

As noted above, the sample covariance operator matches the square of the kernel matrix. For a better understanding, we can examine the extreme cases of the two variance proxy estimates in Appendix C.1.4.

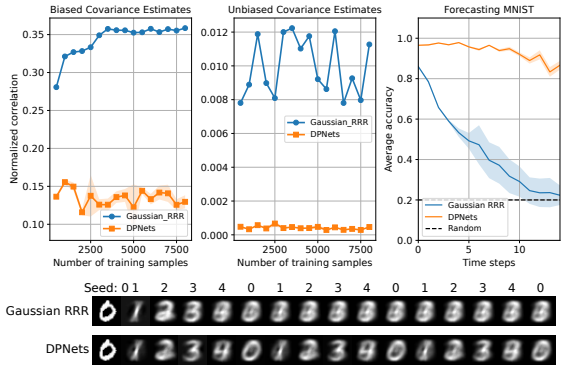

Figure 3: Performance evaluation of rank-5 RRR estimators using Gaussian and DPNet kernels on MNIST with $\eta = 0.1$: normalized correlations and forecast accuracy.

4.2 Noisy Ordered MNIST

Now we design an experiment involving a stochastic process with K states, $S_1, S_2, \ldots, S_K$, corresponding to each integer class of the MNIST dataset. At each time step of a trajectory, we sample an MNIST image from that class, with the classes appearing in an ordered sequence, subject to a small perturbation probability. More precisely, the probability of transitioning from S_i to $S_{(i+1) \bmod K}$ is $1 - \eta$, while the transition to any other state occurs uniformly at random with probability η. Notice the invariant distribution π is uniform in the above setting. Moreover, the mixing time is $\tau \geq \eta^{-1} \ln(1/\epsilon)$ for a distance ϵ from the invariant distribution, π (Levin and Peres, 2017). This provides an estimate of the mixing coefficients $\beta_\mu(\tau) \approx \exp(-\eta\tau)$.

To show the importance of selecting an appropriate representation for learning the dynamics, we selected the first 5 classes of the MNIST dataset as training data points, and we compared the Reduced Rank Regression (RRR) estimator in (Kostic et al., 2022) with rank 5, using two different kernels. The first representation is the Gaussian kernel which is known to be a universal kernel. The second kernel is linear in the space parametrized by a neural network $\phi_{\boldsymbol{\theta}} \in \mathbb{R}^5$ trained according to Deep Projection Neural Network (DPNet) (Kostic et al., 2024c), which is designed to minimize the representation error for the operator regression task by minimizing the empirical risk. While for the Gaussian kernel, we performed hyperparameter tuning on validation data points, resulting in a length scale of 784 and a regularization parameter of 10^{-7}, for DPNet we use a CNN architecture; see Appendix C.2 for more information.

We trained the two transfer operator estimators on different samples. Figure 3 demonstrates that the normalized correlations, $\widehat{V}_\tau(\mathbf{X})$, and $\widehat{V}_\tau(\mathbf{X})$, are effective estimators of the generalization performance of the learned models. We applied min-max normalization to the kernel matrices. The rightmost figure in the first row shows the average accuracy of all forecasts of test points for each model, varying the number of forecasting steps. In this process, the model predicts an image, and we use an Oracle CNN to predict the label. We then match this predicted label with the true label and compute the accuracy accordingly. In the second row, we illustrated how this forecasting works for a single sample of the test set. As expected, only the forecasts using the DPNet kernel maintained a sharp (and correct) shape for the predicted digits over a long horizon. In contrast, the Gaussian kernels were less effective in capturing visual structures, and their predictions quickly lost resemblance to digits.

The results indicate that our bounds provide valuable insights into the generalization performance of operator regression methods for dynamical systems, which depend crucially on the concentration of covariances and time-delayed cross-covariances, c.f. (Kostic et al., 2023, 2024a). We repeated the procedure with two different values of $\eta = 0.2$ and 0.05. The results can be found in Appendix C.2.

4.3 EBI-based Model Selection

In this experiment, we demonstrate that minimizing the bound Theorem 4 can serve as an effective criterion for selecting Koopman models. We applied this approach to a realistic simulation of the small molecule Alanine Dipeptide, c.f. (Wehmeyer and Noé, 2018). Sixteen RRR estimators, each associated with a different kernel's length scale and/or Ridge regularization parameter, were trained, and their forecasting RMSE was assessed using 2000 initial conditions from a test dataset. In this procedure, for the β-mixing coefficient, we used an estimate via the second eigenvalue of the Koopman operator based on the independent studies (Bonati et al., 2021). Figure 4 presents these errors, highlighting the model with the lowest empirical risk bound (as defined in Theorem 4), which is one of the best estimators.

5 CONCLUSION

Motivated by recent progress on stochastic dynamical systems and the fact that, in real-world situations, data are neither i.i.d. nor in the stationary regime, we derived empirical Bernstein inequalities for a general class of stochastic processes in Hilbert space. Bernstein-type inequalities have been shown to be a key component of learning dynamical systems in previous studies (Kostic

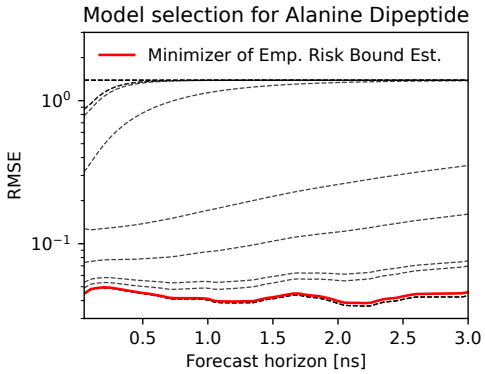

Figure 4: Forecasting RMSE on the Alanine Dipeptide dataset for 16 different RRR estimators, each corresponding to a different kernel's length scale and/or regularization parameter, which show how the best model, according to the empirical risk bound metric, also attains one of the best forecasting performances.

et al., 2022, 2024a). In this paper, we showed that one can retain the guarantees in these papers while relaxing the restrictive assumptions on the data sequence. Notably, our inequalities expose the effect of correlations in improving the bounds. Further, we developed an alternative approach to risk bounds of reduced rank regression that is less restrictive and more practical than the existing ones, as also illustrated numerically. A limitation of our work is that our bounds in their current form are still not adapted to exploit the effective dimension of the RKHS in deriving the minimax rate and theoretically proving the impact of the correlation coefficients on the excess risk. Future work could address this question as well as investigate the application of our bound to the vector-valued problem in the context of partial differential equations.

6 Acknowledgments

We acknowledge financial support from NextGenerationEU and MUR PNRR project PE0000013 CUP J53C22003010006 "Future Artificial Intelligence Research (FAIR)", and EU Project ELIAS (GA no 101120237). We would like to thank the CSML-IIT lab members for their valuable discussions and support during this research. In particular, we are especially grateful to Matia Bojovic for his substantial help.

References

Abeles, B., Clerico, E., and Neu, G. (2024). Generalization bounds for mixing processes via delayed online-to-pac conversions. *arXiv preprint arXiv:2406.12600*.

Abélès, B., Clerico, E., and Neu, G. (2024). Online-to-pac generalization bounds under graph-mixing dependencies. *arXiv preprint arXiv:2410.08977*.

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

Alquier, P., Doukhan, P., and Fan, X. (2019). Exponential inequalities for nonstationary markov chains. *Dependence Modeling*, 7(1):150–168.

Anthony, M. M. and Bartlett, P. (1999). *Learning in neural networks: theoretical foundations*. Cambridge University Press.

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2007). Variance estimates and exploration function in multi-armed bandit. In *CERTIS Research Report 07–31*. Citeseer.

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902.

Blanchard, G. and Zadorozhnyi, O. (2019). Concentration of weakly dependent banach-valued sums and applications to statistical learning methods. *Bernoulli*, 25(4B):3421—3458.

Bonati, L., Piccini, G., and Parrinello, M. (2021). Deep learning the slow modes for rare events sampling. *Proceedings of the National Academy of Sciences*, 118(44):e2113533118.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions.

Brunton, S. L., Budišić, M., Kaiser, E., and Kutz, J. N. (2022). Modern Koopman Theory for Dynamical Systems. *SIAM Review*, 64(2):229–340.

Burgess, M. A., Chapman, A. C., and Scott, P. (2020). An engineered empirical bernstein bound. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part III*, pages 86–102. Springer.

Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368.

Chatterjee, S., Mukherjee, M., and Sethi, A. (2024). Generalization bounds for dependent data using online-to-batch conversion. *arXiv preprint arXiv:2405.13666*.

Coffey, W. and Kalmykov, Y. P. (2012). *The Langevin equation: with applications to stochastic problems in physics, chemistry and electrical engineering*, volume 27. World Scientific.

Hang, H. and Steinwart, I. (2014). Fast learning from α-mixing observations. *Journal of Multivariate Analysis*, 127:184–199.

Hang, H. and Steinwart, I. (2017). A bernstein-type inequality for some mixing processes and dynamical systems with an application to learning.

Jin, Y., Ren, Z., Yang, Z., and Wang, Z. (2022). Policy learning" without"overlap: Pessimism and generalized empirical bernstein's inequality. *arXiv preprint arXiv:2212.09900*.

Kostic, V., Inzerili, P., Lounici, K., Novelli, P., and Pontil, M. (2024a). Consistent long-term forecasting of ergodic dynamical systems. In *2024 International Conference on Machine Learning*.

Kostic, V., Lounici, K., Novelli, P., and Pontil, M. (2023). Sharp spectral rates for koopman operator learning. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 32328–32339. Curran Associates, Inc.

Kostic, V., Novelli, P., Maurer, A., Ciliberto, C., Rosasco, L., and Pontil, M. (2022). Learning dynamical systems via Koopman operator regression in reproducing kernel hilbert spaces. In *Advances in Neural Information Processing Systems*.

Kostic, V. R., Lounici, K., Halconruy, H., Devergne, T., and Pontil, M. (2024b). Learning the infinitesimal generator of stochastic diffusion processes. *arXiv preprint arXiv:2405.12940*.

Kostic, V. R., Novelli, P., Grazzi, R., Lounici, K., and Pontil, M. (2024c). Learning invariant representations of time-homogeneous stochastic dynamical systems. In *ICLR 2024*.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Li, Z., Meunier, D., Mollenhauer, M., and Gretton, A. (2022). Optimal rates for regularized conditional mean embedding learning. In *Advances in Neural Information Processing Systems*.

Liu, T. and Austern, M. (2023). Wasserstein-p bounds in the central limit theorem under local dependence. *Electronic Journal of Probability*, 28:1–47.

Luise, G., Stamos, D., Pontil, M., and Ciliberto, C. (2019). Leveraging low-rank relations between surrogate tasks in structured prediction. In *International Conference on Machine Learning*, pages 4193–4202. PMLR.

Majda, A. J. and Harlim, J. (2012). *Filtering complex turbulent systems*. Cambridge University Press.

Markowitz, H. (1952). *Portfolio selection*, volume 7. Wiley Online Library.

Martinez-Taboada, D. and Ramdas, A. (2024). Empirical bernstein in smooth banach spaces. *arXiv preprint arXiv:2409.06060*.

Maurer, A. (2012). Thermodynamics and concentration. *Bernoulli*, 18(2):434–454.

Maurer, A. and Pontil, M. (2009). Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*.

Maurer, A. and Pontil, M. (2018). Empirical bounds for functions with weak interactions. In *Conference On Learning Theory*, pages 987–1010. PMLR.

McDiarmid, C. (1998). Concentration. In *Probabilistic Methods of Algorithmic Discrete Mathematics*, pages 195–248, Berlin. Springer.

Meir, R. (2000). Nonparametric time series prediction through adaptive model selection. *Machine learning*, 39:5–34.

Modha, D. S. and Masry, E. (1996). Minimum complexity regression estimation with weakly dependent observations. *IEEE Transactions on Information Theory*, 42(6):2133–2145.

Mohri, M. and Rostamizadeh, A. (2008). Rademacher complexity bounds for non-iid processes. *Advances in Neural Information Processing Systems*, 21.

Mollenhauer, M., Klus, S., Schütte, C., and Koltai, P. (2022). Kernel autocovariance operators of stationary processes: Estimation and convergence. *Journal of Machine Learning Research*, 23(327):1–34.

Oneto, L., Ridella, S., and Anguita, D. (2016). Tikhonov, ivanov and morozov regularization for support vector machine learning. *Machine Learning*, 103:103–136.

Pavliotis, G. A. (2014). *Stochastic Processes and Applications*. Springer New York.

Peel, T., Anthoine, S., and Ralaivola, L. (2010). Empirical bernstein inequalities for u-statistics. In *Neural Information Processing Systems (NIPS)*, number 23, pages 1903–1911.

Peel, T., Anthoine, S., and Ralaivola, L. (2013). Empirical bernstein inequality for martingales: Application to online learning.

Philipp, F. M., Schaller, M., Worthmann, K., Peitz, S., and Nüske, F. (2024). Error bounds for kernel-based approximations of the koopman operator. *Applied and Computational Harmonic Analysis*, 71:101657.

Rusakov, D. A., Savtchenko, L. P., and Latham, P. E. (2020). Noisy synaptic conductance: bug or a feature? *Trends in Neurosciences*, 43(6):363–372.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation

and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article 32.

Schug, S., Benzing, F., and Steger, A. (2021). Presynaptic stochasticity improves energy efficiency and helps alleviate the stability-plasticity dilemma. *Elife*, 10:e69884.

Shalizi, C. and Kontorovich, A. (2013). Predictive pac learning and process decompositions. *Advances in neural information processing systems*, 26.

Shivaswamy, P. and Jebara, T. (2010). Empirical bernstein boosting. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 733–740. JMLR Workshop and Conference Proceedings.

Smale, S. and Zhou, D.-X. (2009). Online learning with markov sampling. *Analysis and Applications*, 7(01):87–113.

Steinwart, I. and Christmann, A. (2009). Fast learning from non-iid observations. *Advances in neural information processing systems*, 22.

Steinwart, I., Hush, D., and Scovel, C. (2009). Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175–194.

Tankov, P. (2003). *Financial modelling with jump processes*. Chapman and Hall/CRC.

Tolstikhin, I. O. and Seldin, Y. (2013). Pac-bayes-empirical-bernstein inequality. *Advances in Neural Information Processing Systems*, 26.

Tuckerman, M. E. (2023). *Statistical Mechanics: Theory and Molecular Simulation*. Oxford university press.

von Storch, H. and Navarra, A. (1999). Principal oscillation patterns: A review. *Journal of Climate*, 12(12):3519–3535.

Waudby-Smith, I. and Ramdas, A. (2024). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27.

Wehmeyer, C. and Noé, F. (2018). Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *The Journal of chemical physics*, 148(24).

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116.

Ziemann, I., Tu, S., Pappas, G. J., and Matni, N. (2024). The noise level in linear regression with dependent data. *Advances in Neural Information Processing Systems*, 36.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] Paper is theoretical with clear mathematical formalism
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] It concerns experiments, and info is provided in Appendix C.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] The implementation of illustrative experiments will be made public

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes] Proof sketches are in the main body, full proofs in the Appendix
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] See Appendix C.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes] For the experiment on molecular dynamics, see Appendix C.3.
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Material

A Hilbert space-valued concentration for dependent, non-stationary sequences.

First, we give a version of Bernstein's inequality for independent vector-valued random variables. Then we specialize in the dependent case under β-mixing assumptions.

A.1 Vector valued concentration for independent variables

Theorem 5. *Suppose that the X_i are m independent mean zero random variables with values in a Hilbert-space H, satisfying $\|X_i\| \leq c$. Then for $\delta > 0$*

$$\Pr\left\{\left\|\sum_i X_i\right\| > \sqrt{\sum_i \mathbb{E}\left\|X_i\right\|^2}\left(1+\sqrt{2\ln\left(1/\delta\right)}\right)+\frac{4c}{3}\ln\left(1/\delta\right)\right\} < \delta.$$

A well-known bound of Pinelis and Sakhanenko (Caponnetto and De Vito, 2007, Proposition 2), when specialized to bounded vectors, would be

$$2\sqrt{\sum_i \mathbb{E}\left\|X_i\right\|^2}\ln\left(\frac{2}{\delta}\right)+4c\ln\left(\frac{2}{\delta}\right),$$

which is slightly worse, not only with respect to constants but also in its dependence on the confidence parameter δ.

To prove Theorem 5 we use the version of Bernstein's inequality is below. For $\mathbf{z} = (z_1, ..., z_m) \in \mathcal{X}^m$, $k \in [m]$ and $y \in \mathcal{X}$ define the substitution $S_y^k \mathbf{z} = (z_1, ..., z_{k-1}, y, z_{k+1}, ..., z_m)$.

Theorem 6. *(see (McDiarmid, 1998), Theorem 3.8 or (Maurer, 2012), Theorem 11) Let $\mathbf{X} = (X_1, ..., X_m)$ be a vector of independent random variables, with values in $\mathcal{X}$ and $f : \mathcal{X}^m \to \mathbb{R}$. Assume that $\forall k \in [m]$, $\mathbf{x} \in \mathcal{X}^m$, $f(\mathbf{x}) - \mathbb{E}\left[f\left(S_{X_k}^k \mathbf{x}\right)\right] \leq b$. Let denote*

$$V = \frac{1}{2}\sup_{\mathbf{x}\in\mathcal{X}^m}\sum_{k=1}^m \mathbb{E}\left[\left(f\left(S_{X_k}^k\mathbf{x}\right)-f\left(S_{X_k'}^k\mathbf{x}\right)\right)^2\right].$$

Then for $t > 0$

$$\Pr\left\{f(\mathbf{X}) - \mathbb{E}\left[f(\mathbf{X})\right] > t\right\} \leq \exp\left(\frac{-t^2}{2V+2bt/3}\right).$$

The quantity V is the "maximal sum of conditional variances" ((McDiarmid, 1998)): we compute the variance of f in the k-th variable (here X_k' is an independent copy of X_k), holding all the other variables of $\mathbf{x}$ fixed, then we sum over k and finally take the supremum in $\mathbf{x}$.

Proof of Theorem 5. We will apply Theorem 6 with $\mathcal{X} = \{x \in \mathcal{H} : \|x\| \leq c\}$. Define $f(\mathbf{x}) = \|\sum_i x_i\|$ and note that for $y, y' \in \mathcal{H}$, $f\left(S_y^k\mathbf{x}\right) - f\left(S_{y'}^k\mathbf{x}\right) \leq \|y - y'\|$. This implies that $f(\mathbf{x}) - \mathbb{E}\left[f\left(S_{X_k}^k\mathbf{x}\right)\right] \leq 2c$ and also

$$V = \frac{1}{2}\sup_{\mathbf{x}\in\mathcal{X}^m}\sum_{k=1}^m \mathbb{E}\left[\left(f\left(S_{X_k}^k\mathbf{x}\right)-f\left(S_{X_k'}^k\mathbf{x}\right)\right)^2\right] \leq \frac{1}{2}\sum_{k=1}^m \mathbb{E}\left[\|X_k - X_k'\|^2\right] = \sum_{i=1}^m \mathbb{E}\left\|X_i\right\|^2.$$

By Bernstein's inequality, Theorem 6, for $t > 0$,

$$\Pr\left\{\left\|\sum_i X_i\right\| - \mathbb{E}\left[\left\|\sum_i X_i\right\|\right] > t\right\} \leq \exp\left(\frac{-t^2}{2\sum_i \mathbb{E}\left\|X_i\right\|^2+4ct/3}\right).$$

Solving for t gives for $\delta > 0$

$$\Pr\left\{\left\|\sum_i X_i\right\| \leq \mathbb{E}\left[\left\|\sum_i X_i\right\|\right] + \sqrt{2\sum_i \mathbb{E}\left\|X_i\right\|^2\ln\left(1/\delta\right)}+\frac{4c}{3}\ln\left(1/\delta\right)\right\} \leq \delta.$$

Using Jensen's inequality, independence, and the mean-zero assumption to bound $\mathbb{E}\left[\left\|\sum_i X_i\right\|\right] \leq \sqrt{\sum_i \mathbb{E}\left\|X_i\right\|^2}$ complete the proof. □

A.2 Empirical bounds

If we drop the mean-zero assumption, then the inequality in Theorem 5 reads

$$\left\|\sum_i \left(X_i - \mathbb{E}\left[X_i\right]\right)\right\| \leq \sqrt{\sum_i \mathbb{E}\left\|X_i - \mathbb{E}\left[X_i\right]\right\|^2}\left(1+\sqrt{2\ln\left(1/\delta\right)}\right)+\frac{4c}{3}\ln\left(1/\delta\right).$$

In many applications, we observe the X_i, but we do not know the expectations and wish to use the inequality to estimate $\sum_i \mathbb{E}\left[X_i\right]$. In this case, it is useful to have an empirical estimate of the variance term on the right-hand side. We give two such estimates, a simple biased one and an unbiased one for i.i.d data. Both use the following concentration inequality, which can be found in (Maurer and Pontil, 2018, Corollary 10).

Theorem 7. *Define an operator D^2 acting on measurable functions $f : \mathcal{X}^m \to \mathbb{R}$ by*

$$D^2 f\left(\mathbf{x}\right) = \sum_k \left(f\left(\mathbf{x}\right) - \inf_{y\in\mathcal{X}} f\left(S_y^k \mathbf{x}\right)\right)^2,$$

and let $\mathbf{X} = (X_1, ..., X_m)$ be a vector of independent variables with values in $\mathcal{X}$. Now suppose $f : \mathcal{X}^m \to \mathbb{R}$ satisfies $f\left(\mathbf{x}\right) - \inf_{y\in\mathcal{X}} f\left(S_y^k \mathbf{x}\right) \leq b$ for all $k \in \{1, ..., m\}$ and all $\mathbf{x} \in \mathcal{X}^m$, and for some $a > 0$

$$D^2 f\left(\mathbf{x}\right) \leq a f\left(\mathbf{x}\right), \forall \mathbf{x} \in \mathcal{X}^m. \tag{8}$$

Then for all $\delta > 0$ with probability at least $1 - \delta$

$$\sqrt{\mathbb{E}\left[f\left(\mathbf{X}\right)\right]} \leq \sqrt{f\left(\mathbf{X}\right)}+\sqrt{2\max\left\{a, b\right\}\ln\left(1/\delta\right)}.$$

For the simple biased estimate, we use the elementary inequality

$$\sqrt{\sum_i \mathbb{E}\left\|X_i - \mathbb{E}\left[X_i\right]\right\|^2} = \sqrt{\sum_i \left(\mathbb{E}\left\|X_i\right\|^2 - \left\|\mathbb{E}\left[X_i\right]\right\|^2\right)} \leq \sqrt{\sum_i \mathbb{E}\left\|X_i\right\|^2}. \tag{9}$$

Let $f\left(\mathbf{x}\right) = \sum_i \left\|x_i\right\|^2$. The minimizer in y of $f\left(S_y^k \mathbf{x}\right)$ is clearly $y = 0$. Thus $f\left(\mathbf{x}\right) - \inf_{y\in\mathcal{X}} f\left(S_y^k \mathbf{x}\right) \leq \left\|x_k\right\|^2 \leq c^2$ and

$$D^2 f\left(\mathbf{x}\right) \leq \sum_k \left\|x_k\right\|^4 \leq c^2 \sum_k \left\|x_k\right\|^2 = c^2 f\left(\mathbf{x}\right).$$

The theorem above, with $a = b = c^2$ gives with probability at least $1 - \delta$ that

$$\sqrt{\sum_i \mathbb{E}\left\|X_i\right\|^2} \leq \sqrt{\sum_i \left\|X_i\right\|^2}+c\sqrt{2\ln\left(1/\delta\right)}.$$

A union bound with Theorem 5 gives

Proposition 1. *Under the conditions of Theorem 5, but with uncentered X_i, we have for $0 < \delta < 1$ with probability at least $1 - \delta$ in $\mathbf{X}$*

$$\left\|\sum_i \left(X_i - \mathbb{E}\left[X_i\right]\right)\right\| \leq \sqrt{\sum_i \left\|X_i\right\|^2}\left(1+\sqrt{2\ln\left(2/\delta\right)}\right)+\frac{16c}{3}\ln\left(2/\delta\right).$$

For identically distributed X_i we can do better, since then

$$\sqrt{\sum_i \mathbb{E}\left\|X_i - \mathbb{E}\left[X_i\right]\right\|^2} = \sqrt{\frac{1}{2\left(m-1\right)}\mathbb{E}\sum_{i,j:i\neq j}\left\|X_i - X_j\right\|^2}, \tag{10}$$

and we can apply Theorem 7 to

$$f(\mathbf{x}) = \sum_{i,j:i\neq j} \|x_i - x_j\|^2 .$$

Then for any $k \in [m]$ and $y \in \mathcal{X}$ we have

$$f(\mathbf{x}) - f\left(S_y^k \mathbf{x}\right) = 2 \sum_{i:i\neq k} \left(\|x_i - x_k\|^2 - \|x_i - y\|^2 \right).$$

It is easy to see that the minimizer in y of $f\left(S_y^k \mathbf{x}\right)$ is the average $y_k = 1/(m-1) \sum_{j:j\neq k} x_j$. A computation gives

$$\sum_{i:i\neq k} \|x_i - y_k\|^2 = \sum_{i:i\neq k} \|x_i - x_k\|^2 - \frac{1}{m-1} \left\| \sum_{i:i\neq k} (x_i - x_k) \right\|^2$$

Thus

$$f(\mathbf{x}) - \inf_y f\left(S_y^k \mathbf{x}\right) = \frac{2}{m-1} \left\| \sum_{i:i\neq k} (x_i - x_k) \right\|^2 \leq 8(m-1)c^2,$$

Thus, $b = 8(m-1)c^2$, and

$$\begin{aligned} D^2 f(\mathbf{x}) &= \frac{4}{(m-1)^2} \sum_k \left\| \sum_{i:i\neq k} (x_i - x_k) \right\|^4 \\ &= \frac{4(m-1)^4}{(m-1)^2} \sum_k \left(\left\| \frac{1}{m-1} \sum_{i:i\neq k} (x_i - x_k) \right\|^2 \right)^2 \\ &\leq \frac{4(m-1)^4}{(m-1)^2} \sum_k \left(\frac{1}{m-1} \sum_{i:i\neq k} \|x_i - x_k\|^2 \right)^2 \\ &\leq 16c^2 (m-1) \sum_k \sum_{i:i\neq k} \|x_i - x_k\|^2 . \end{aligned}$$

The first inequality follows from Jensen's inequality; the second is obtained by bounding $1/(m-1) \sum_{i:i\neq k} \|x_i - x_k\|^2 \leq 4c^2$. Thus $D^2 f(\mathbf{x}) \leq 16(m-1)c^2 f(\mathbf{x})$ where $a = 16(m-1)c^2$.

It follows from Theorem 7 that with probability at least $1-\delta$

$$\sqrt{\sum_i \mathbb{E} \|X_i - \mathbb{E}[X_i]\|^2} \leq \sqrt{\frac{1}{2(m-1)} \sum_{i,j:i\neq j} \|X_i - X_j\|^2} + 4c\sqrt{\ln(1/\delta)}$$

and the union bound with Theorem 5 gives (with some rather crude estimates) for $\delta < 2/e$ with probability at least $1-\delta$

$$\begin{aligned} \left\| \sum_i (X_i - \mathbb{E}[X_i]) \right\| > \sqrt{\frac{1}{2(m-1)} \sum_{i,j:i\neq j} \|X_i - X_j\|^2} \left(1 + \sqrt{2\ln(2/\delta)}\right) \\ + (4\sqrt{2} + 4 + \frac{4}{3}) c \ln(2/\delta) . \end{aligned}$$

Using $4\sqrt{2} + 4 + \frac{4}{3} = 10.99 < 11$ we obtain

Proposition 2. *Let $\mathbf{X} = (X_1, ..., X_m)$ be a vector of i.i.d. random variables with values in a Hilbert space, satisfying $\|X_1\| \leq c$ almost surely. Then for $0 < \delta < 2/e$ with probability at least $1-\delta$ in $\mathbf{X}$*

$$\left\| \sum_i (X_i - \mathbb{E}[X_i]) \right\| \leq \sqrt{\frac{1}{2(m-1)} \sum_{i,j:i\neq j} \|X_i - X_j\|^2} \left(1 + \sqrt{2\ln(2/\delta)}\right) + 11c \ln(2/\delta) .$$

A.3 Mixing and its consequences

Let $\mathbf{X} := \{X_t\}_{t\in\mathbb{N}}$ be a sequence of random variables with values in some measurable space $\mathcal{X}$. For a set $I \subseteq \mathbb{N}$ use $\Sigma(I)$ for the σ-field generated by $\{X_i\}_{i\in I}$ and μ_I for the joint distribution of $\{X_i\}_{i\in I}$. Then the definition of the mixing coefficients $\beta_\mu(\tau)$ for $\tau \in \mathbb{N}$ reads (Bradley, 2005)

$$\beta_\mu(\tau) := \sup_{j\in\mathbb{N}} \sup_{B\in\Sigma([1,j]\cup[j+\tau,\infty))} \left|\mu_{[1,j]\cup[j+\tau,\infty)}(B) - \mu_{[1,j]} \times \mu_{[j+\tau,\infty)}(B)\right|.$$

In the sequel, we let $T, m, \tau \in \mathbb{N}$ with $T = 2m\tau$ the length of the trajectory observed. We call τ the "mixing time" with the intuitive understanding that events separated by more than τ can be regarded as independent, with a penalty in probability of order $\beta_\mu(\tau)$.

For $j \in [m]$ define the index sets $I_j = \{2(j-1)\tau+1, ..., (2j-1)\tau\}$ and $I'_j = \{(2j-1)\tau+1, ..., 2j\tau\}$. Then $[T] = \bigcup_{k=1}^m I_k \cup I'_k$. We also set

Lemma 2. *For $m \in \mathbb{N}$ and $B \in \Sigma(I_1 \cup ... \cup I_m)$*

$$|\mu_{I_1\cup...\cup I_m}(B) - \mu_{I_1} \times ... \times \mu_{I_m}(B)| \leq (m-1)\beta_\mu(\tau),$$

and for $B' \in \Sigma(I'_1 \cup ... \cup I'_m)$

$$\left|\mu_{I'_1\cup...\cup I'_m}(B') - \mu_{I'_1} \times ... \times \mu_{I'_m}(B')\right| \leq (m-1)\beta_\mu(\tau),$$

Proof. By Fubini's Theorem and the definition of the mixing coefficients, we have for $1 \leq k < m$, that

$$\left|\mu_{I_1} \times ... \times \mu_{I_k\cup...\cup I_m}(B) - \mu_{I_1} \times ... \times \mu_{I_k} \times \mu_{I_{k+1}\cup...\cup I_m}(B)\right| \leq \beta_\mu(\tau).$$

Then with a telescopic expansion,

$$\begin{aligned}
&|\mu_{I_1\cup...\cup I_m}(B) - \mu_{I_1} \times ... \times \mu_{I_m}(B)| \\
&= \left|\sum_{k=1}^{m-1} \mu_{I_1} \times ... \times \mu_{I_k\cup...\cup I_m}(B) - \mu_{I_1} \times ... \times \mu_{I_k} \times \mu_{I_{k+1}\cup...\cup I_m}(B)\right| \\
&\leq \sum_{k=1}^{m-1} \left|\mu_{I_1} \times ... \times \mu_{I_k\cup...\cup I_m}(B) - \mu_{I_1} \times ... \times \mu_{I_k} \times \mu_{I_{k+1}\cup...\cup I_m}(B)\right| \\
&\leq (m-1)\beta_\mu(\tau).
\end{aligned}$$

The argument for B' is analogous. □

In the sequel, we write $\Pr_I$ for the probability measure $\mu_{I_1} \times ... \times \mu_{I_m}$ on $\Sigma(I_1 \cup ... \cup I_m)$ and $\Pr_{I'}$ for $\mu_{I'_1} \times ... \times \mu_{I'_m}$ on $\Sigma(I'_1 \cup ... \cup I'_m)$. The previous lemma can then be stated more concisely as $|\Pr_I(B) - \Pr(B)| \leq (m-1)\beta_\mu(\tau)$ for $B \in \Sigma(I_1 \cup ... \cup I_m)$ and $|\Pr_{I'}(B) - \Pr(B)| \leq (m-1)\beta_\mu(\tau)$ for $B \in \Sigma(I'_1 \cup ... \cup I'_m)$.

Lemma 1. *Let X_i have values in a normed space $(\mathcal{X}, \|\cdot\|)$ and let $F, F' : \mathcal{X}^n \to \mathbb{R}$, where F is $\Sigma(I_1 \cup ... \cup I_m)$-measurable, and F' is $\Sigma(I'_1 \cup ... \cup I'_m)$-measurable. Then*

$$\begin{aligned}
&\Pr\left\{\left\|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right\| > F(\mathbf{X}) + F'(\mathbf{X})\right\} \\
&\leq \Pr_I\left\{\left\|\sum_{k=1}^m (Y_k - \mathbb{E}[Y_k])\right\| > F(\mathbf{X})\right\} \\
&+ \Pr_{I'}\left\{\left\|\sum_{k=1}^m (Y'_k - \mathbb{E}[Y'_k])\right\| > F'(\mathbf{X})\right\} + 2(m-1)\beta_\mu(\tau),
\end{aligned}$$

where the Y_k and Y'_k are given by (3).

Proof. Write $Y_j = \sum_{i\in I_j} X_i$ and $Y'_j = \sum_{i\in I'_j} X_i$. Then

$$\left\|\sum_{i=1}^{n} X_i\right\| = \left\|\sum_{j=1}^{m} Y_j + \sum_{j=1}^{m} Y'_j\right\| \leq \left\|\sum_{j=1}^{m} Y_j\right\| + \left\|\sum_{j=1}^{m} Y'_j\right\|,$$

Thus

$$\begin{aligned}
&\Pr\left\{\left\|\sum_{i=1}^{n} X_i\right\| > F(\mathbf{X}) + F'(\mathbf{X})\right\} \\
&\leq \Pr\left\{\left\|\sum_{j=1}^{m} Y_j\right\| + \left\|\sum_{j=1}^{m} Y'_j\right\| > F(\mathbf{X}) + F'(\mathbf{X})\right\} \\
&\leq \Pr\left\{\left\|\sum_{j=1}^{m} Y_j\right\| > F(\mathbf{X})\right\} + \Pr\left\{\left\|\sum_{j=1}^{m} Y'_j\right\| > F'(\mathbf{X})\right\}.
\end{aligned}$$

The conclusion then follows from applying Lemma 2 to the events $B = \left\{\left\|\sum_{j=1}^{m} Y_j\right\| > F(\mathbf{X})\right\} \in \Sigma(I_1 \cup ... \cup I_m)$ and $B' = \left\{\left\|\sum_{j=1}^{m} Y'_j\right\| > F'(\mathbf{X})\right\} \in \Sigma(I'_1 \cup ... \cup I'_m)$. □

A.4 Concentration for dependent variables

Now let X_i be a possibly dependent sequence of mean zero random vectors in H. Write $Y_j = \sum_{i\in I_j} X_i$ and $Y'_j = \sum_{i\in I'_j} X_i$. Note the bounds $\|Y_k\| \leq \tau c$ and $\|Y'_k\| \leq \tau c$ and that the Y_k are independent under $\Pr_I$, as are the Y'_k under $\Pr_{I'}$. Define

$$\begin{aligned}
F(\mathbf{X}) &= \sqrt{\sum_{k=1}^{m} \mathbb{E}\|Y_k\|^2}\left(1+\sqrt{2\ln(2/\delta)}\right) + \frac{4\tau c}{3}\ln(2/\delta) \\
F'(\mathbf{X}) &= \sqrt{\sum_{k=1}^{m} \mathbb{E}\|Y'_k\|^2}\left(1+\sqrt{2\ln(2/\delta)}\right) + \frac{4\tau c}{3}\ln(2/\delta).
\end{aligned}$$

Lemma 1 gives

$$\begin{aligned}
\Pr\left\{\left\|\sum_{i=1}^{n} X_i\right\| > F(\mathbf{X}) + F'(\mathbf{X})\right\} &\leq \Pr_I\left\{\left\|\sum_{j=1}^{m} Y_j\right\| > F(\mathbf{X})\right\} + \Pr_{I'}\left\{\left\|\sum_{j=1}^{m} Y'_j\right\| > F'(\mathbf{X})\right\} + 2(m-1)\beta_\mu(\tau) \\
&\leq \delta/2 + \delta/2 + 2(m-1)\beta_\mu(\tau),
\end{aligned}$$

where the second inequality follows from Theorem 5. Substitution of the expressions for $F(\mathbf{X})$ and $F'(\mathbf{X})$ and using $\sqrt{a}+\sqrt{b} \leq \sqrt{2}\sqrt{a+b}$, we obtain

$$\begin{aligned}
&\Pr\left\{\left\|\sum_{i=1}^{n} X_i\right\| > \sqrt{\sum_{k=1}^{m}\left(\mathbb{E}\|Y_k\|^2 + \mathbb{E}\|Y'_k\|^2\right)}\left(\sqrt{2}+2\sqrt{\ln(2/\delta)}\right) + \frac{8\tau c}{3}\ln(2/\delta)\right\} \\
&< \delta + 2(m-1)\beta_\mu(\tau).
\end{aligned}$$

Now $\mathbb{E}\|Y_k\|^2 = \sum_{(t,s)\in I_k\times I_k} \mathbb{E}[\langle X_t, X_s\rangle]$ and $\mathbb{E}\|Y'_k\|^2 = \sum_{(t,s)\in I'_k\times I'_k} \mathbb{E}[\langle X_t, X_s\rangle]$. Thus

$$\sum_{k=1}^{m}\left(\mathbb{E}\|Y_k\|^2 + \mathbb{E}\|Y'_k\|^2\right) = \sum_{(t,s)\in S_\tau} \mathbb{E}[\langle X_t, X_s\rangle],$$

where S_τ is the set of index pairs

$$S_\tau = \bigcup_{k=1}^{m} (I_k \times I_k) \cup (I'_k \times I'_k), \tag{11}$$

which is a disjoint union of $\tau \times \tau$-squares along the diagonal of the $T \times T$ matrix. If $(t,s) \in S_\tau$ then the times s and t are never more than τ apart. We have proved

Theorem 8. *Let $m, \tau \in \mathbb{N}$, $T = 2m\tau$, and let $\mathbf{X} = (X_1, ..., X_T)$ be a vector of mean zero random variables in a separable Hilbert-space H, satisfying $\|X_t\| \leq c$. Let $\delta > 0$. Then with a probability of at least $1 - \delta - 2(m-1)\beta_\mu(\tau)$ we have*

$$\left\| \sum_{t=1}^{T} X_t \right\| \leq \sqrt{\sum_{(t,s)\in S_\tau} \mathbb{E}\left[\langle X_t, X_s \rangle\right]} \left(\sqrt{2} + 2\sqrt{\ln(2/\delta)}\right) + \frac{8\tau c}{3} \ln(2/\delta),$$

where $S_\tau \subseteq [T] \times [T]$ is given by (11).

1. If we drop the mean zero assumption this becomes Theorem 1.
2. To interpret the first term on the right-hand side, note the similarity to the variance of $\sum X_t$, which would be

$$\mathbb{E}\left\| \sum X_t \right\|^2 = \sum_{(t,s)\in[T]\times[T]} \mathbb{E}\left[\langle X_t, X_s \rangle\right].$$

3. If the X_t are independent, we can set $\tau = 1$ and $\beta_{\mathbf{X}}(\tau) = 0$, and we recover Theorem 5 up to a constant factor of $\sqrt{2}$ on the first, and 2 on the second term.

In the proof above the functions $F(\mathbf{X})$ and $F'(\mathbf{X})$ were constants. But let

$$F(\mathbf{X}) = \sqrt{\sum_{k=1}^{m} \left\| \sum_{i\in I_k} X_i \right\|^2} \left(1 + \sqrt{2\ln(4/\delta)}\right) + \frac{16c\tau}{3} \ln(4/\delta),$$

which is $\Sigma(I_1 \cup ... \cup I_m)$-measurable, and replace I_k by I'_k for the analogous definition of $F'(\mathbf{X})$. Then Lemma 1 and Proposition 1 give a proof of Theorem 2. In the same way, defining

$$F(\mathbf{X}) = \sqrt{\frac{1}{2(m-1)} \sum_{k,l:k\neq l} \left\| \sum_{i\in I_k} X_i - \sum_{j\in I_l} X_j \right\|^2} \left(1 + \sqrt{2\ln(4/\delta)}\right) + 11c\tau \ln(4/\delta),$$

together with the corresponding $F'(\mathbf{X})$, yields the proof of Theorem 3.

B Theoretical result for learning dynamical systems

In this section, we begin by briefly reviewing the notations and definitions pertinent to learning dynamical systems. Next, we establish bounds on the deviation of the true risk from the empirical risk, uniformly across a prescribed set of HS operators on $\mathcal{H}$. Finally, we demonstrate how this analysis, originally formulated for Ivanov regularization, can be extended to Tikhonov regularization.

B.1 Notation and background

Let us first review some fundamental concepts related to Markov chains and Koopman operators. Consider $\mathbf{X} := \{X_t : t \in \mathbb{N}\}$, a collection of random variables taking values in a measurable space $(\mathcal{X}, \Sigma_\mathcal{X})$, known as the state space. We define $\mathbf{X}$ as a *Markov chain* if $\mathbb{P}\{X_{t+1} \in B | X_{[t]}\} = \mathbb{P}\{X_{t+1} \in B | X_t\}$. Additionally, $\mathbf{X}$ is termed *time-homogeneous* if there exists a function $p : \mathcal{X} \times \Sigma_\mathcal{X} \to [0,1]$, known as the *transition kernel*, such that for all $(x, B) \in \mathcal{X} \times \Sigma_\mathcal{X}$ and each $t \in \mathbb{N}$,

$$\mathbb{P}\{X_{t+1} \in B | X_t = x\} = p(x, B).$$

A broad class of Markov chains includes those with an *invariant measure* π, which satisfies $\pi(B)=\int_{\mathcal{X}}\pi(dx)p(x,B)$ for $B\in\Sigma_{\mathcal{X}}$. For a *time-homogeneous* Markov chain with an invariant (stationary) distribution π the (stochastic) *Koopman operator* $A_\pi\colon L^2_\pi(\mathcal{X})\to L^2_\pi(\mathcal{X})$ is given by

$$A_\pi f(x) := \int_{\mathcal{X}} p(x,dy)f(y) = \mathbb{E}\left[f(X_{t+1})|X_t=x\right], \quad f\in L^2_\pi(\mathcal{X}), x\in\mathcal{X}. \tag{12}$$

In many practical scenarios, the Koopman operator A_π is unknown, but data from one or more trajectories are available. The framework for operator regression learning introduced in (Kostic et al., 2022) estimates the Koopman operator on $L^2_\pi(\mathcal{X})$ within an RKHS, using an associated feature map $\phi:\mathcal{X}\to\mathcal{H}$. In this vector-valued regression, the population risk functional is defined as

$$\mathcal{R}(G) = \mathbb{E}_{X\sim\pi, X^+\sim p(\cdot|X))}[\|\phi(X^+)-G\phi(X)\|^2_{\mathcal{H}}], \tag{13}$$

and the goal is to learn A_π by minimizing the risk over a class of operators $G\colon\mathcal{H}\to\mathcal{H}$, using a dataset of consecutive states $\mathcal{D}_n := (x_i,x_i^+)_{i=1}^n$. A typical setting involves obtaining these states from a single trajectory of the process after it reaches the equilibrium distribution, where $X_0\sim\pi$ and $X_i^+\equiv X_{i+1}\sim p(\cdot,|,X_i)$ for $i=2,\dots,n$. A common estimator in this context is the Reduced Rank Regression (RRR) estimator $\widehat{G}_\lambda$, obtained by minimizing the regularized empirical risk

$$\widehat{\mathcal{R}}_\lambda(G) := \tfrac{1}{n}\sum_{i\in[n]}\|\phi(x_i^+)-G\phi(x_i)\|^2_{\mathcal{H}}+\lambda\|G\|^2_{\mathrm{HS}}, \tag{14}$$

over operators G of rank at most r. The estimator $\widehat{G}^{\mathrm{RRR}}_{r,\lambda}=\widehat{C}_\lambda^{-1/2}[\![\widehat{C}_\lambda^{-1/2}\widehat{T}]\!]_r$ is computed via an r-truncated SVD $[\![\cdot]\!]_r$, where the empirical input and cross covariances are respectively

$$\widehat{C}=\tfrac{1}{n}\textstyle\sum_{i\in[n]}\phi(x_i)\otimes\phi(x_i), \quad \text{and} \quad \widehat{T}=\tfrac{1}{n}\sum_{i\in[n]}\phi(x_i)\otimes\phi(x_i^+).$$

In this section we denote as $\widehat{C}_\lambda=\widehat{C}+\lambda I_{\mathcal{H}}$. Further, similar to the relationship between population and empirical risk, the population covariance and cross-covariance are given respectively by

$$C=\mathbb{E}_{X\sim\pi}[\phi(X)\otimes\phi(X)], \quad \text{and} \quad T=\mathbb{E}_{X\sim\pi,X^+\sim p(\cdot|X))}[\phi(X)\otimes\phi(X^+)].$$

B.2 Uniform bound for Ivanov regularization

In view of the fact that both Tikhonov regularization and Ivanov regularization versions of the problem are equivalent, we will first focus on the Ivanov regularization formulation, which is more convenient to theoretical analysis(Oneto et al., 2016; Luise et al., 2019). We first state that the uniform bound for the Ivanov regularization case. Then reformulate to Tikhonov regularization since it is more convenient from a computational standpoint.

Theorem 9. *Let $\mathbf{X}=(X_t)_{t=1}^n$ be a stationary Markov chain with distribution μ, and the risk definitions as above note that $\pi=\mu_1$. Denote $\mathcal{G}_{r,\gamma}=\{G\in\mathrm{HS}_r(\mathcal{H}):\|G\|_{\mathrm{HS}}\le\gamma\}$ and $\mathbf{Y}=(\phi(X_0)\otimes\phi(X_0),...,\phi(X_{n-1})\otimes\phi(X_{n-1}))$, $\mathbf{Z}=(\phi(X_0)\otimes\phi(X_1),...,\phi(X_{n-1})\otimes\phi(X_n))$, and $\mathbf{W}=\big(\|\phi(X_1)\|^2,...,\|\phi(X_n)\|^2\big)$. Assume $n=2m\tau$, and exists $c_{\mathcal{H}}>0$ such that $\|\phi(X_t)\|^2\le c_{\mathcal{H}}$ a.s. for all t. Let $\delta>0$ and assume $\delta(\tau)=\delta-2(\frac{n}{2\tau}-1)\beta_\mu(\tau)>0$ and $\delta'(\tau)=\delta-2(\frac{n}{2\tau}-1)\beta_\mu(\tau-1)>0$. Then, with probability at least $1-\delta$ we have for every $\widehat{G}\in\mathcal{G}_{r,\gamma}$*

$$\begin{aligned}|\mathcal{R}(\widehat{G})-\widehat{\mathcal{R}}(\widehat{G})| \le{}& \frac{32\gamma^2c_{\mathcal{H}}\tau}{3n}\ln\frac{12}{\delta(\tau)}+\frac{64\sqrt{r}\gamma c_{\mathcal{H}}\tau}{3n}\ln\frac{12}{\delta'(\tau)}+\frac{7c_{\mathcal{H}}\tau}{3(\frac{n}{2\tau}-1)}\ln\frac{12}{\delta(\tau)}\\ &+\sqrt{\frac{2\gamma^4\widetilde{V}_\tau(\mathbf{Y})\tau}{n}\left(1+2\ln\frac{12}{\delta(\tau)}\right)}+\sqrt{\frac{2r\gamma^2\widetilde{V}_\tau(\mathbf{Z})\tau}{n}\left(1+2\ln\frac{12}{\delta'(\tau)}\right)}+\sqrt{\frac{2\overline{V}_\tau(\mathbf{W})\tau}{n}\ln\frac{12}{\delta(\tau)}}\end{aligned} \tag{15}$$

where $\overline{V}_\tau(\mathbf{W})=\frac{1}{m(m-1)\tau^2}\sum_{1\le i<j\le m}(\overline{W}_i-\overline{W}_j)^2+(\overline{W'}_i-\overline{W'}_j)^2$, $\overline{W}_j=\sum_{i\in I_j}W_i$ and $\overline{W'}_j=\sum_{i\in I'_j}W_i$. $\widetilde{V}_\tau(\mathbf{Y})$ and $\widetilde{V}_\tau(\mathbf{Z})$ were defined before.

Proof. We can restate both the true risk and the empirical risk as follows:

$$\begin{aligned}\mathcal{R}(G) &= \mathbb{E}_{(x,y)\sim\rho}\|\phi(y) - G^*\phi(x)\|^2 \\ &= \mathbb{E}_{(x,y)\sim\rho}[\operatorname{tr}(\phi(X_{i+1}) \otimes \phi(X_{i+1})) - 2\langle \phi(X_{i+1}), G^*\phi(x_i)\rangle_{\mathcal{H}} + \operatorname{tr}(GG^*\phi(x_i) \otimes \phi(x_i))] \\ &= \operatorname{tr}(D) + \operatorname{tr}(GG^*C) - 2\operatorname{tr}(G^*T)\end{aligned}$$

Similarly, we have $\widehat{\mathcal{R}}(G) = \operatorname{tr}(\widehat{D}) + \operatorname{tr}\left(GG^*\widehat{C}\right) - 2\operatorname{tr}\left(G^*\widehat{T}\right)$.

Let $\mathcal{G}_{r,\gamma} = \{G \in \mathrm{HS}_r(\mathcal{H}) : \|G\|_{\mathrm{HS}} \leq \gamma\}$. Then for $\forall G \in \mathcal{G}_{r,\gamma}$ we can easily show:

$$\begin{aligned}\mathcal{R}(G) - \widehat{\mathcal{R}}(G) &= \operatorname{tr}(D - \widehat{D}) + \operatorname{tr}\left(GG^*(C - \widehat{C})\right) - 2\operatorname{tr}\left(G^*(T - \widehat{T})\right) \\ &\leq \operatorname{tr}(D - \widehat{D}) + \gamma^2\|C - \widehat{C}\| + 2\sqrt{r}\gamma\|T - \widehat{T}\|\end{aligned}$$

where we have used Hölder inequality to obtain the last two terms.

First, we use an empirical Bernstein's inequality for bounded random variables, $W_i = \operatorname{tr}(\phi(X_{i+1}) \otimes \phi(X_{i+1})) = \|\phi(X_{i+1})\|^2$ and $|W_i| \leq c_{\mathcal{H}}$ since

$$\begin{aligned}\operatorname{tr}(D - \widehat{D}) &= \operatorname{tr}(\frac{1}{n}\sum_{i=1}^{n}\phi(X_{i+1}) \otimes \phi(X_{i+1}) - \mathbb{E}[\phi(X_{i+1}) \otimes \phi(X_{i+1})]) \\ &= \frac{1}{n}\sum_{i=1}^{n}\operatorname{tr}(\phi(X_{i+1}) \otimes \phi(X_{i+1}) - \mathbb{E}[\phi(X_{i+1}) \otimes \phi(X_{i+1}))] \\ &= \frac{1}{n}\sum_{i=1}^{n}\operatorname{tr}(\phi(X_{i+1}) \otimes \phi(X_{i+1})) - \mathbb{E}[\operatorname{tr}(\phi(X_{i+1}) \otimes \phi(X_{i+1}))].\end{aligned}$$

Now, write $\overline{W}_j = \sum_{i\in I_j} W_i$ and $\overline{W'}_j = \sum_{i\in I'_j} W_i$. Note the bounds, $\left\|\overline{W}_k\right\| \leq \tau c_{\mathcal{H}}$ and $\left\|\overline{W}'_k\right\| \leq \tau c_{\mathcal{H}}$ since we know there exists $c_{\mathcal{H}} > 0$ such that $\|\phi(X_t)\|^2 \leq c_{\mathcal{H}}$ a.s. for all t. In addition, we know that the $\overline{W}_k$ are independent under $\Pr_I$, as are the $\overline{W}'_k$ under $\Pr_{I'}$.

Define

$$\begin{aligned}F(\mathbf{W}) &= \frac{7mc_{\mathcal{H}}\tau\ln\frac{2}{\delta'}}{3(m-1)} + \sqrt{2mV_m(\overline{W})\ln\frac{2}{\delta'}}, V_m(\overline{W}) = \frac{1}{m(m-1)}\sum_{1\leq i<j\leq m}(\overline{W}_i - \overline{W}_j)^2 \\ F'(\mathbf{W}) &= \frac{7mc_{\mathcal{H}}\tau\ln\frac{2}{\delta'}}{3(m-1)} + \sqrt{2mV_m(\overline{W'})\ln\frac{2}{\delta'}}, V_m(\overline{W'}) = \frac{1}{m(m-1)}\sum_{1\leq i<j\leq m}(\overline{W'}_i - \overline{W'}_j)^2.\end{aligned}$$

Lemma 1 gives

$$\begin{aligned}&\Pr\left\{\left\|\sum_{i=1}^{n}(W_i - \mathbb{E}[W_i])\right\| > F(\mathbf{W}) + F'(\mathbf{W})\right\} \\ &\leq \Pr_I\left\{\left\|\sum_{k=1}^{m}(\overline{W}_k - \mathbb{E}\left[\overline{W}_k\right])\right\| > F(\mathbf{W})\right\} + \Pr_{I'}\left\{\left\|\sum_{k=1}^{m}(\overline{W}'_k - \mathbb{E}\left[\overline{W}'_k\right])\right\| > F'(\mathbf{W})\right\} + 2(m-1)\beta_{\mathbf{W}}(\tau), \\ &\leq \delta/2 + \delta/2 + 2(m-1)\beta_\mu(\tau),\end{aligned}$$

where the second inequality follows from adapting to arbitrarily bounded random variables of Thm 4 (Maurer and Pontil, 2009). Substitution of the expressions for $F(\mathbf{W})$ and $F'(\mathbf{W})$ and using $\sqrt{a}+\sqrt{b} \leq \sqrt{2}\sqrt{a+b}$, and $\delta = 2\delta'$, we obtain

$$\begin{aligned}&\Pr\left\{\left\|\sum_{i=1}^{n}(W_i - \mathbb{E}[W_i])\right\| > \frac{14mc_{\mathcal{H}}\tau\ln\frac{4}{\delta}}{3(m-1)} + \sqrt{4m(V_m(\overline{W}) + V_m(\overline{W'}))\ln\frac{4}{\delta}}\right\} \\ &< \delta + 2(m-1)\beta_\mu(\tau)\end{aligned}$$

Let define $\overline{V}_\tau(\mathbf{W}) = \frac{1}{\tau^2}(V_m(\overline{W}) + V_m(\overline{W'}))$. Now, if we divide both sides in the probability by n and substitute $\frac{2\tau}{n} = m$ we have

$$\operatorname{tr}(D - \widehat{D}) \leq \frac{7c_{\mathcal{H}}\tau}{3(m-1)} \ln \frac{4}{\delta - 2(m-1)\beta_\mu(\tau)} + \sqrt{\frac{\overline{V}_\tau(\mathbf{W}) \ln \frac{4}{\delta - 2(m-1)\beta_\mu(\tau)}}{m}}.$$

To bound the second and third terms, we can use Theorem 2 of this paper with the description in the covariance estimation section (for further details, see propositions in the next section). The result then follows by a union bound. □

Remark 1. *Theorem 9 can be used in order to derive an excess risk bound for the well-specified case, following the same reasoning as (Kostic et al., 2022; Luise et al., 2019).*

B.3 Adaptation of risk bound to Tikhonov regularization

It is known that there is an equivalence between Ivanov and Tikhonov regularization, in the sense that for each class of estimators satisfying the conditions of Ivanov regularization, there exists a corresponding Tikhonov regularization problem. However, to extend this result, we require additional techniques. The next step involves the following lemma (a restatement of Lemma 15.6 from Anthony and Bartlett (1999)):

Lemma 3. *Suppose* Pr *is a probability distribution and*

$$\{E(\alpha_1, \alpha_2, \delta) : 0 < \alpha_1, \alpha_2, \delta \leq 1\}$$

is a set of events, such that
(i) For all $0 < \alpha \leq 1$ *and* $0 < \delta \leq 1$,

$$\Pr\{E(\alpha, \alpha, \delta)\} \leq \delta$$

(ii) For all $0 < \alpha_1 \leq \alpha \leq \alpha_2 \leq 1$ *and* $0 < \delta_1 \leq \delta \leq 1$

$$E(\alpha_1, \alpha_2, \delta_1) \subseteq E(\alpha, \alpha, \delta)$$

Then for $0 < a, \delta < 1$,

$$\Pr \bigcup_{\alpha \in (0,1]} E(\alpha a, \alpha, \delta\alpha(1-a)) \leq \delta$$

The following proposition exemplifies how Lemma 3 can be used to convert Ivanov-type uniform bounds to Tykhonov-type bounds at the expense of a logarithmic penalty in the reguralizer.

Proposition 3. *Suppose* $\mathcal{H}$ *is a space of hypotheses,* X *is a random variable and* A *and* $R \geq 0$ *are functions such that for all* $h \in \mathcal{H}$ *with* $R(h) \leq K$ *we have with probability at least* $1 - \delta$ *in* X *for all* h *with* $R(h, X) \leq K$

$$A(h, X) \leq K + \ln(1/\delta) \tag{16}$$

Then, with probability at least $1 - \delta$ *in* X *we have for all* h *with* $R(h, X) \geq 1$ *that*

$$A(h, X) > 2R(h, X) + \ln(2R(h, X)/\delta) \tag{17}$$

Proof. Define the events

$$E(\alpha_1, \alpha_2, \delta) := \left\{\exists h, R(h, X) \leq \alpha_2^{-1}, A(h, X) > \alpha_1^{-1} + \ln(1/\delta)\right\}$$

Then by assumption, these events satisfy (i) of the lemma. It is also easy to check (ii). Using $a = 1/2$ we get, with $R(h, X)$ playing the role of α^{-1},

$$\begin{aligned}&\Pr\{\exists h, R(h,X) \geq 1, A(h,X) > 2R(h,X) + \ln(2R(h,X)/\delta)\}\\ \leq &\Pr \bigcup_{\alpha^{-1}\in[1,\infty)} \left\{\exists h, R(h,X) \leq \alpha^{-1}, A(h,X) > 2\alpha^{-1} + \ln\left(2\alpha^{-1}/\delta\right)\right\} \leq \delta\end{aligned}$$

□

In Proposition 3, Eq. 16 represents the "Ivanov"-type bound (a uniform bound with a hard constraint on hypotheses), while Eq. 17 represents the "Tikhonov"-type bound (valid for all hypotheses). Here, $R(h, X)$ is the (potentially data-dependent) regularizer—often something like $|h|$, and X represents the sample, while $A(h, X)$ denotes the difference between the true and empirical risk.

Theorem 4. *Let $\mathbf{X} = (X_t)_{t=1}^n$ be a stationary Markov chain with distribution μ, and the risk definitions as above noting that $\pi{=}\mu_1$. Denote $\widehat{G}_{r,\lambda}$ be a minimizer of reduced-rank Tikhonov regularized empirical risk and $\mathbf{Y}{=}(\phi(X_t)\otimes\phi(X_t))_{t=1}^n$, $\mathbf{Z}{=}(\phi(X_t)\otimes\phi(X_{t+1}))_{t=1}^n$, and $\mathbf{W}{=}(\|\phi(X_t)\|^2)_{t=1}^n$. Assume $n{=}2m\tau$, and exists $c_{\mathcal{H}}>0$ such that $\|\phi(X_t)\|^2{\leq}c_{\mathcal{H}}$ a.s. for all t. Let $\delta{\geq}0$ and assume $\widehat{\delta}_\mu(\tau,\lambda) := 0.5\,\delta/\|\widehat{G}_{r,\lambda}\|{-}2(\frac{n}{2\tau}{-}1)\beta_\mu(\tau){>}0$.*
Then, with probability at least $1{-}\delta$ we have for every $\widehat{G}_{r,\lambda}$ such that [2] $\|\widehat{G}_{r,\lambda}\|_{HS}{\geq}1$

$$\begin{aligned}|\mathcal{R}(\widehat{G}){-}\widehat{\mathcal{R}}(\widehat{G})| \leq &\frac{128c_{\mathcal{H}}\tau\|\widehat{G}_{r,\lambda}\|(\sqrt{r}{+}\|\widehat{G}_{r,\lambda}\|)}{3n}\ln\frac{12}{\widehat{\delta}_\mu(\tau,\lambda)}\\ &+\frac{14c_{\mathcal{H}}\tau^2}{3n-2\tau}\ln\frac{12}{\widehat{\delta}_\mu(\tau,\lambda)}\\ &+\sqrt{\frac{32\|\widehat{G}_{r,\lambda}\|^4\widetilde{V}_\tau(\mathbf{Y})\tau}{n}\left(1{+}2\ln\frac{12}{\widehat{\delta}_\mu(\tau,\lambda)}\right)}\\ &+\sqrt{\frac{8r\|\widehat{G}_{r,\lambda}\|^2\widetilde{V}_\tau(\mathbf{Z})\tau}{n}\left(1{+}2\ln\frac{12}{\widehat{\delta}_\mu(\tau,\lambda)}\right)}\\ &+\sqrt{\frac{2\overline{V}_\tau(\mathbf{W})\tau}{n}\ln\frac{12}{\widehat{\delta}_\mu(\tau,\lambda)}},\end{aligned}$$

where $\overline{V}_\tau(\mathbf{W}) = \frac{1}{m(m-1)\tau^2}\sum_{1\leq i<j\leq m}(\overline{W}_i-\overline{W}_j)^2+(\overline{W'}_i-\overline{W'}_j)^2$, $\overline{W}_j = \sum_{i\in I_j}W_i$ and $\overline{W'}_j = \sum_{i\in I'_j}W_i$. $\widetilde{V}_\tau(\mathbf{Y})$ and $\widetilde{V}_\tau(\mathbf{Z})$ were defined before.

Proof. Based on Theorem 19 in (Luise et al., 2019), we know that $\widehat{G}_{r,\lambda}$, the minimizer of the rank-reduced Tikhonov regularized problem,

$$\min_{G\in\mathrm{HS}r(\mathcal{H})}\widehat{\mathcal{R}}^\lambda(G),$$

with λ as a regularization parameter, is also the minimizer of the corresponding Ivanov problem, where $\gamma(\lambda) = \|\widehat{G}r,\lambda\|_{\mathrm{HS}} = \|\widehat{C}_\lambda^{-1/2}[\![\widehat{C}_\lambda^{-1/2}\widehat{T}]\!]_r\|_{\mathrm{HS}}$. However, we cannot simply substitute $\|\widehat{G}r,\lambda\|_{\mathrm{HS}}$ in the formula, as it is a random variable and depends on the data, but using the method outlined in Lemma 3 we can obtain the proof with the choices $\gamma(\lambda) = 2\|\widehat{G}_{r,\lambda}\|_{\mathrm{HS}}$ and the substitution of $\frac{\delta}{2\|\widehat{G}r,\lambda\|_{\mathrm{HS}}}$ in place of δ. Indeed, let $Q(\gamma,\delta)$ be the right hand side of the inequality in Theorem 9, and define the events

$$E(\alpha_1,\alpha_2,\delta) := \left\{\exists G : \|G\|_{HS}\leq\alpha_2^{-1}, \left|\mathcal{R}\left(\hat{G}\right)-\hat{\mathcal{R}}\left(\hat{G}\right)\right| > Q\left(\alpha_1^{-1},\delta\right)\right\}.$$

Then by Theorem 9 these events satisfy (i) of Lemma 3 and they also have the monotonicity property (ii). Using the Lemma with $a = 1/2$ then gives the inequality for every G satisfying $\|G\|_{HS}\geq 1$. □

[2] This is an artifact of the conversion from Ivanov to Tikhonov-type bounds. In practice, we expect the Hilbert-Schmidt norm to be much larger than 1.

Remark 2. *If the process is known to be stationary, we can use the unbiased estimators for both the covariance, $\widehat{V}_\tau(\mathbf{Y})$, and the cross-covariance, $\widehat{V}_\tau(\mathbf{Z})$, respectively*

Remark 3. *We can easily adapt the above results for non-stationary processes by simply modifying the definition of risk.*

C Experiments

To facilitate the reproducibility of our main experimental results, we have made the code publicly available in a GitHub repository, which can be accessed via the link `https://github.com/erfunmirzaei/EBI4LDS`.

C.1 Covariance estimation using samples from Ornstein-Uhlenbeck process

C.1.1 Theoretical results

As mentioned in the section on covariance estimation for estimating the concentration of covariance operators in the Hilbert-Schmidt norm, the observed vectors are operators and X_t should be replaced by the operator $\phi(X_t) \otimes \phi(X_t)$.

Recent studies have relied on Pinelis and Sakhanenko's inequality, see (Caponnetto and De Vito, 2007, Proposition 2). As mentioned earlier, the method of blocks and β-mixing extends this i.i.d. analysis of transfer operator regression to more realistic settings, such as learning from data trajectories of a stationary process. We begin by restating Pinelis and Sakhanenko's inequality for convenience.

Proposition 4. *Let A_i, $i \in [n]$ be independent random variables with values in a separable Hilbert space with norm $\|\cdot\|$. If there exist constants $L > 0$ and $\sigma > 0$ such that $\forall i \in [n]$, $\|A_i\| \leq \frac{L}{2}$ and $\mathbb{E}[\|A_i\|^2] \leq \sigma^2$, then with probability at least $1 - \delta$*

$$\left\| \frac{1}{n} \sum_{i \in [n]} A_i - \mathbb{E}[A_i] \right\| \leq 2\left(\frac{L}{n} + \frac{\sigma}{\sqrt{n}}\right) \log\left(\frac{2}{\delta}\right) \tag{18}$$

We now aim to adapt Pinelis and Sakhanenko's inequality for covariance estimation in data-dependent scenarios.

Proposition 5. *Let $\mathbf{X} = (X_t)_{t=1}^n$ be a Markov chain with distribution μ and $\mathbf{Y}=(\phi(X_t) \otimes \phi(X_t))_{t=1}^n$. Assume $n=2m\tau$, and exists $c_{\mathcal{H}} > 0$ such that $\|\phi(X_t)\|^2 \leq c_{\mathcal{H}}$ a.s. for all t. Let $\delta \geq 0$ and assume $\delta(\tau) = \delta - 2(\frac{n}{2\tau} - 1)\beta_\mu(\tau) > 0$, with probability at least $1-\delta$ we have*

$$\|\widehat{C} - C\| \leq \frac{4c_{\mathcal{H}}}{m} \ln \frac{4}{\delta(\tau)} + \frac{2c_{\mathcal{H}}}{\sqrt{m}} \ln \frac{4}{\delta(\tau)},$$

where $\widehat{C} = \frac{1}{n} \sum_{t \in [n]} \phi(x_t) \otimes \phi(x_t)$ and in general $C = \frac{1}{n} \sum_{t \in [n]} \mathbb{E}[\phi(X_t) \otimes \phi(X_t)]$.

Proof. As we mentioned before, for covariance estimation, our random variables are the rank-one operators $Y_t = \phi(X_t) \otimes \phi(X_t)$ instead of X_t. To prove we need to use the same procedure as section A.4 with these new random variables and use Pinelis and Sakhanenko's inequality as stated above.

Now, write $\overline{Y}_j = \sum_{i \in I_j} Y_i$ and $\overline{Y'}_j = \sum_{i \in I'_j} Y_i$. Note the bounds, $\left\|\overline{Y}_k\right\| \leq \tau c_{\mathcal{H}}$ and $\left\|\overline{Y'}_k\right\| \leq \tau c_{\mathcal{H}}$ since there exists $c_{\mathcal{H}} > 0$ such that $\|\phi(X_t)\|^2 \leq c_{\mathcal{H}}$ a.s. for all t . In addition, we know that the $\overline{Y}_k$ are independent under

$\Pr_I$, as are the $\overline{Y'}_k$ under $\Pr_{I'}$. Then also we have the following:

$$\begin{aligned}\|\overline{Y}_k\|_{HS}^2 &= \langle \sum_{i=1}^{\tau} \phi(X_i)\otimes\phi(X_i), \sum_{j=1}^{\tau} \phi(X_j)\otimes\phi(X_j)\rangle_{HS} \\ &= \sum_{i=1}^{\tau}\sum_{j=1}^{\tau} \langle \phi(X_i)\otimes\phi(X_i), \phi(X_j)\otimes\phi(X_j)\rangle_{HS} \\ &= \sum_{i=1}^{\tau}\sum_{j=1}^{\tau} \operatorname{tr}((\phi(X_i)\otimes\phi(X_i))(\phi(X_j)\otimes\phi(X_j))) \\ &= \sum_{i=1}^{\tau}\sum_{j=1}^{\tau} \langle \phi(X_i),\phi(X_j)\rangle^2 \leq \tau^2 c_{\mathcal{H}}^2\end{aligned}$$

Therefore, we can set $L = L' = 2\tau c_{\mathcal{H}}$ and $\sigma = \sigma' = \tau c_{\mathcal{H}}$ for using for $\overline{Y}_k$ and $\overline{Y'}_k$.

Define

$$F(\mathbf{Y}) = F'(\mathbf{Y}) = (4\tau c_{\mathcal{H}} + 2\tau c_{\mathcal{H}}\sqrt{m})\log(\frac{2}{\delta'})$$

Lemma 1 gives

$$\begin{aligned}&\Pr\left\{\left\|\sum_{i=1}^{n}(Y_i - \mathbb{E}[Y_i])\right\| > F(\mathbf{Y}) + F'(\mathbf{Y})\right\} \\ &\leq \Pr_I\left\{\left\|\sum_{k=1}^{m}(\overline{Y}_k - \mathbb{E}\left[\overline{Y}_k\right])\right\| > F(\mathbf{Y})\right\} + \Pr_{I'}\left\{\left\|\sum_{k=1}^{m}(\overline{Y'}_k - \mathbb{E}\left[\overline{Y'}_k\right])\right\| > F'(\mathbf{Y})\right\} + 2(m-1)\beta_\mu(\tau), \\ &\leq \delta/2 + \delta/2 + 2(m-1)\beta_\mu(\tau),\end{aligned}$$

where the second inequality follows from Prop. 4. Substitution of the expressions for $F(\mathbf{Y})$ and $F'(\mathbf{Y})$ and using $\delta = 2\delta'$, we obtain

$$\Pr\left\{\left\|\sum_{i=1}^{n}(Y_i - \mathbb{E}[Y_i]\right\| > (8\tau c_{\mathcal{H}} + 4\tau c_{\mathcal{H}}\sqrt{m})\log(\frac{4}{\delta})\right\} < \delta + 2(m-1)\beta_\mu(\tau)$$

Now, if we divide both sides in the probability by n and substitute $\frac{2\tau}{n} = m$ the proof is finished. □

For comparison, we propose the following proposition, where Theorem 5 is used in place of Pinelis and Sakhanenko's inequality.

Proposition 6. *Let $\mathbf{X} = (X_t)_{t=1}^n$ be a Markov chain with distribution μ and $\mathbf{Y} = (\phi(X_t)\otimes\phi(X_t))_{t=1}^n$. Assume $n = 2m\tau$, and exists $c_{\mathcal{H}} > 0$ such that $\|\phi(X_t)\|^2 \leq c_{\mathcal{H}}$ a.s. for all t. Let $\delta \geq 0$ and assume $\delta(\tau) = \delta - 2(\frac{n}{2\tau} - 1)\beta_\mu(\tau) > 0$, with probability at least $1-\delta$ we have*

$$\|\widehat{C} - C\| \leq \frac{4c_{\mathcal{H}}}{3m}\ln\frac{2}{\delta(\tau)} + \sqrt{\frac{2c_{\mathcal{H}}^2}{m}\left(1 + 2\ln\frac{2}{\delta(\tau)}\right)},$$

where $\widehat{C} = \frac{1}{n}\sum_{t\in[n]}\phi(x_t)\otimes\phi(x_t)$ and in general $C = \frac{1}{n}\sum_{t\in[n]}\mathbb{E}[\phi(X_t)\otimes\phi(X_t)]$.

Proof. As we mentioned before, for covariance estimation, our random variables are the rank-one operators $Y_t = \phi(X_t)\otimes\phi(X_t)$ instead of X_t. To prove we need to use the same procedure as section A.4 with these new random variables and use Theorem 5.

Now, write $\overline{Y}_j = \sum_{i\in I_j} Y_i$ and $\overline{Y'}_j = \sum_{i\in I'_j} Y_i$. Note the bounds $\left\|\overline{Y}_k\right\| \leq \tau c_{\mathcal{H}}$ and $\left\|\overline{Y'}_k\right\| \leq \tau c_{\mathcal{H}}$ since there exists $c_{\mathcal{H}} > 0$ such that $\|\phi(X_t)\|^2 \leq c_{\mathcal{H}}$ a.s. for all t. In addition, we know that the $\overline{Y}_k$ are independent under $\Pr_I$, as are the $\overline{Y'}_k$ under $\Pr_{I'}$.

Define

$$F(\mathbf{Y}) = \sqrt{\sum_{k=1}^{m} \mathbb{E}\left\|\overline{Y}_k - \mathbb{E}\left[\overline{Y}_k\right]\right\|^2}\left(1+\sqrt{2\ln(1/\delta)}\right)+\frac{4c_{\mathcal{H}}\tau}{3}\ln(1/\delta)$$

$$F'(\mathbf{Y}) = \sqrt{\sum_{k=1}^{m} \mathbb{E}\left\|\overline{Y'}_k - \mathbb{E}\left[\overline{Y'}_k\right]\right\|^2}\left(1+\sqrt{2\ln(1/\delta)}\right)+\frac{4c_{\mathcal{H}}\tau}{3}\ln(1/\delta).$$

Lemma 1 gives

$$\begin{aligned}
&\Pr\left\{\left\|\sum_{i=1}^{n}(Y_i - \mathbb{E}[Y_i])\right\| > F(\mathbf{Y}) + F'(\mathbf{Y})\right\} \\
&\leq \Pr_{I}\left\{\left\|\sum_{k=1}^{m}(\overline{Y}_k - \mathbb{E}\left[\overline{Y}_k\right])\right\| > F(\mathbf{Y})\right\} + \Pr_{I'}\left\{\left\|\sum_{k=1}^{m}(\overline{Y}'_k - \mathbb{E}\left[\overline{Y}'_k\right])\right\| > F'(\mathbf{Y})\right\} + 2(m-1)\beta_\mu(\tau), \\
&\leq \delta/2+\delta/2+2(m-1)\beta_\mu(\tau),
\end{aligned}$$

where the second inequality follows after dropping the mean-zero assumption of Thm. 5. Substitution of the expressions for $F(\mathbf{Y})$ and $F'(\mathbf{Y})$ and using $\sqrt{a}+\sqrt{b} \leq \sqrt{2}\sqrt{a+b}$, and $\delta = 2\delta'$, we obtain

$$\begin{aligned}
&\Pr\left\{\left\|\sum_{i=1}^{n}(Y_i - \mathbb{E}[Y_i]\right\| > \sqrt{2\sum_{k=1}^{m}(\mathbb{E}\left\|\overline{Y}_k - \mathbb{E}\left[\overline{Y}_k\right]\right\|^2 + \mathbb{E}\left\|\overline{Y'}_k - \mathbb{E}\left[\overline{Y'}_k\right]\right\|^2)}\left(1+\sqrt{2\ln(2/\delta)}\right)+\frac{8c_{\mathcal{H}}\tau}{3}\ln(2/\delta)\right\} \\
&< \delta + 2(m-1)\beta_\mu(\tau)
\end{aligned}$$

We know $\sum_{k=1}^{m}(\mathbb{E}\left\|\overline{Y}_k - \mathbb{E}\left[\overline{Y}_k\right]\right\|^2 + \mathbb{E}\left\|\overline{Y'}_k - \mathbb{E}\left[\overline{Y'}_k\right]\right\|^2) = 2m\tau^2 c_{\mathcal{H}}^2$. Now, if we divide both sides in the probability by n and substitute $\frac{2\tau}{n} = m$ and using $\sqrt{a}+\sqrt{b} \leq \sqrt{2}\sqrt{a+b}$ the proof is finished. □

C.1.2 Experimental results

As stated in the paper, the following experimental results demonstrate the consistent monotonic increase of the covariance upper bound with respect to τ.

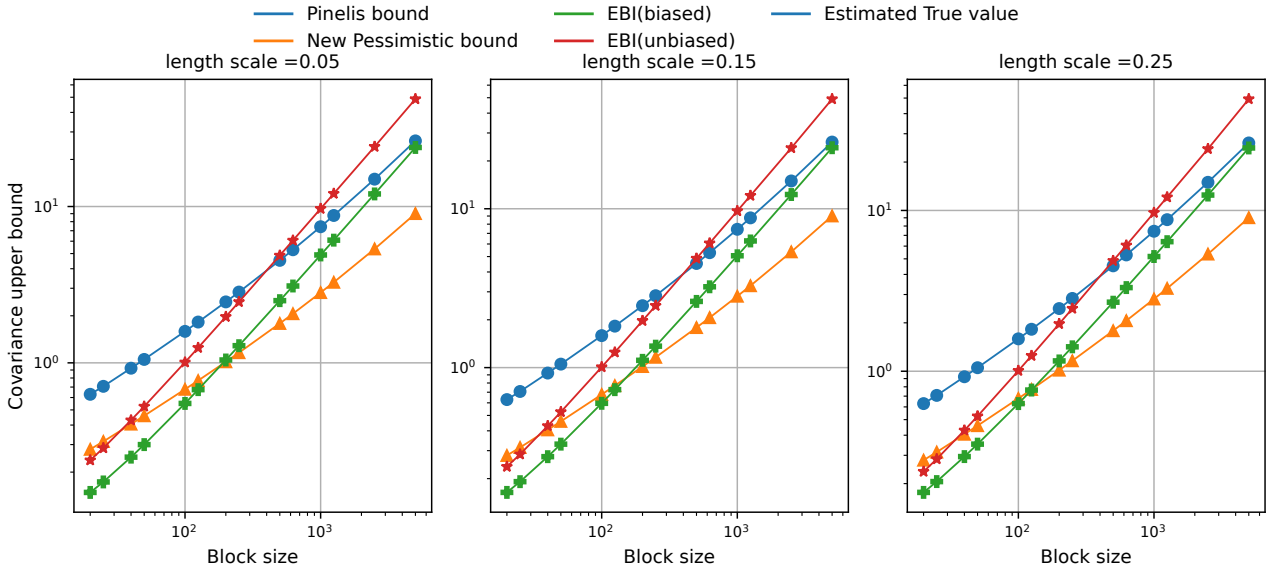

Figure 5: Covariance upper bound as a function of the block size for three different length scales of Gaussian kernel in logarithmic scale. The failure probability is set to 0.05, and we used 10k training points. The plots have averaged over 30 independent simulations.

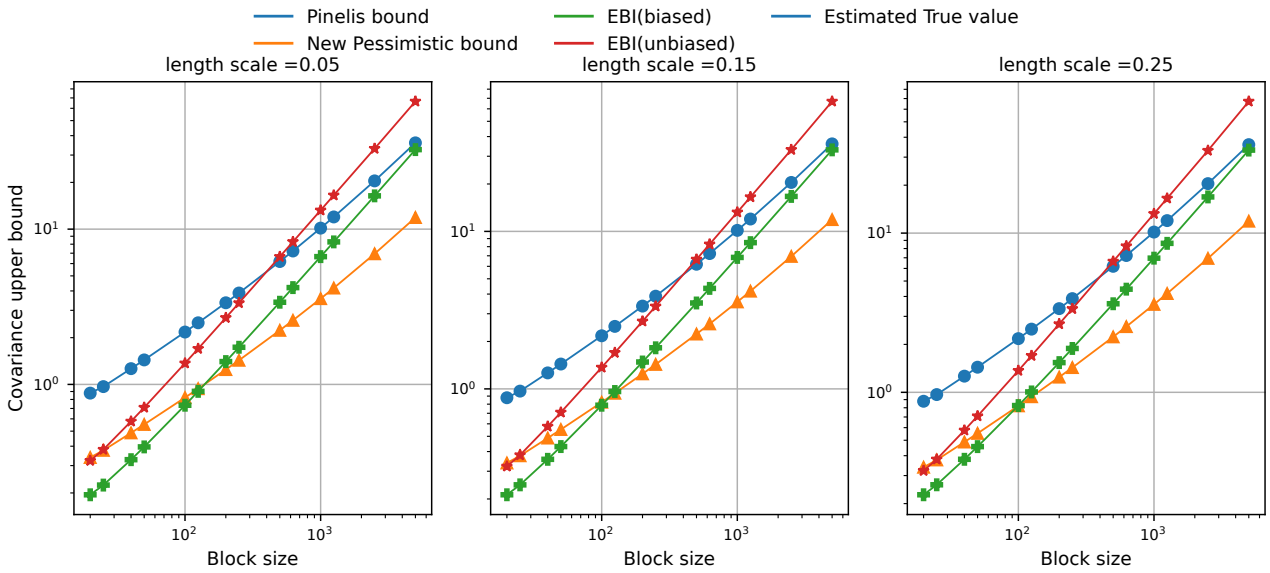

Figure 6: Covariance upper bound as a function of the block size for three different length scales of Gaussian kernel in logarithmic scale. The failure probability is set to 0.01, and we used 10k training points. The plots have averaged over 30 independent simulations.

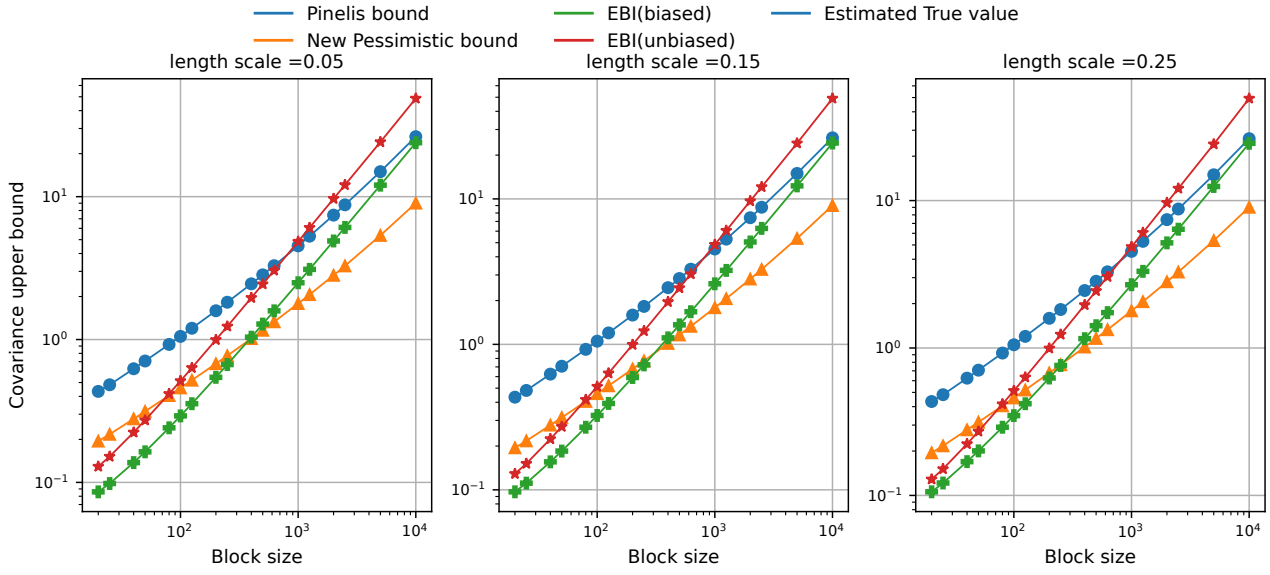

Figure 7: Covariance upper bound as a function of the block size for three different length scales of Gaussian kernel in logarithmic scale. The failure probability is set to 0.05, and we used 20k training points. The plots have averaged over 30 independent simulations.

C.1.3 Estimating the true error of covariance operator

For this experiment, we tried to estimate the norm of the difference between the sample covariance operator and the true one as follows:

$$\|\widehat{C} - C\|^2_{\mathrm{HS}(\mathcal{H})} = \mathrm{tr}(\widehat{C}^2) + \mathrm{tr}(C^2) - 2\,\mathrm{tr}(\widehat{C}C).$$

Since we have the $\widehat{C}$ operator based on the data finding its trace is easy. Furthermore, in this example, we know that the invariant distribution is the standard normal Gaussian. Thus, we can find the analytic form for the $\mathrm{tr}(C)$. First, recall $C = \mathbb{E}_{x\sim\pi}[\phi(x)\otimes\phi(x)]$. Then, we have

$$\begin{aligned}\mathrm{tr}(C^2) &= tr(\mathbb{E}_{x\sim\pi}[\phi(x)\otimes\phi(x)]\mathbb{E}_{y\sim\pi}[\phi(y)\otimes\phi(y)])\\ &= \mathbb{E}_{x\sim\mathcal{N}(0,1)}\mathbb{E}_{y\sim\mathcal{N}(0,1)}[\langle\phi(x),\phi(y)\rangle^2] = \mathbb{E}_{x\sim\mathcal{N}(0,1)}\mathbb{E}_{y\sim\mathcal{N}(0,1)}[k(x,y)^2]\\ &= \frac{1}{2\pi}\int_x\int_y (\exp(\frac{-(x-y)^2}{2l^2}))^2\exp(\frac{-y^2}{2})\exp(\frac{-x^2}{2})dydx = \sqrt{\frac{1}{1+\frac{4}{l^2}}},\end{aligned}$$

where l is the length scale of the Gaussian kernel. Finally, for the last term, we tried to estimate it by $\mathrm{tr}(\widehat{C},\tilde{C})$, where $\tilde{C} = \frac{1}{n}\sum_{i=1}^n \phi(\tilde{x}_i)\otimes\phi(\tilde{x}_i)$ and $\tilde{x}_i$s are new samples from OU process. Then we repeat the previous process many times, which means $\mathrm{tr}(\widehat{C}C) \approx \frac{1}{T}\sum_{t=1}^T k(\tilde{X}_t, X)^2$, where $\tilde{X}_t$ is a new batch of data with the size n at each time. We set $T = 100$ and $n = 10^4$ for this experiment.

C.1.4 Extreme examples

We introduced two empirical Bernstein's inequalities in which two different estimations of the variance proxy have been used. In this section, we will take a more precise look at these estimations for extreme examples in the case of covariance estimation.

1) $K^{\cdot 2} = c^2 I_{n\times n}$, This means that $K(x,y) = 0$ unless $x = y$. Thus, if we have a very small length scale for the Gaussian kernel (close to zero) this can happen. In this situation, $\widehat{V}_\tau(\mathbf{X}) = \widetilde{V}_\tau(\mathbf{X}) = \frac{c^2}{\tau}$.

2) $K^{\cdot 2} = \mathrm{diag}(c^2 1_{\tau\times\tau})$ this means that $K(x,y) = 0$ unless x and y are less than τ-time separated and the boundary is very sharp. In other words, if the distance of x and y is less than τ time steps, then $k(x,y) = c^2$ otherwise 0. In this situation, $\widehat{V}_\tau(\mathbf{X}) = \widetilde{V}_\tau(\mathbf{X}) = c^2$.

3) $K^{\cdot 2} = c^2 1_{n\times n}$, this means that $K(x,y) = c^2$ no matter what are x and y. Thus, if we have a very large length scale for the Gaussian kernel (close to infinity), this would be the case. In this situation, $\widehat{V}_\tau(\mathbf{X}) = c^2$ and $\widetilde{V}_\tau(\mathbf{X}) = 0$.

Thus we can conclude that $\frac{b^2}{\tau} \le \widehat{V}_\tau(\mathbf{X}) \le c^2$ and $0 \le \widetilde{V}_\tau(\mathbf{X}) \le c^2$ where b and c could be the infimum and supremum values of the kernel respectively.

C.2 Noisy ordered MNIST

The architecture of the Oracle network used for DPNet is given by $\phi_{\boldsymbol{\theta}}$: Conv2d $(1, 16; 5)$ $\to$ ReLU $\to$ MaxPool(2) $\to$ Conv 2 d$(16, 32; 5)$ $\to$ ReLU $\to$ MaxPool(2) $\to$ Dense$(1568, 5)$. We set the seed number 42 and used 2 as a padding hyperparameter. Here, the arguments of the convolutional layers are Conv2d(#in channel #out channel; kernel size). The Tikhonov regularization parameter for the CNN kernel $\gamma_{\mathrm{CNN}} = 10^{-4}$. The network $\phi_{\boldsymbol{\theta}}$ has been pre-trained as a digit classifier using the cross entropy loss function. The training was performed with the Adam optimizer (learning rate = 0.01) for 20 epochs (batch size = 64). The training dataset corresponds to the same 1000 images used to train the Koopman estimators.

In Figure 8, t-SNE, a nonlinear dimension reduction, was applied to the concatenation of left and right eigenfunctions on the test dataset. The results indicate that models with superior representations tend to perform better and exhibit better generalization.

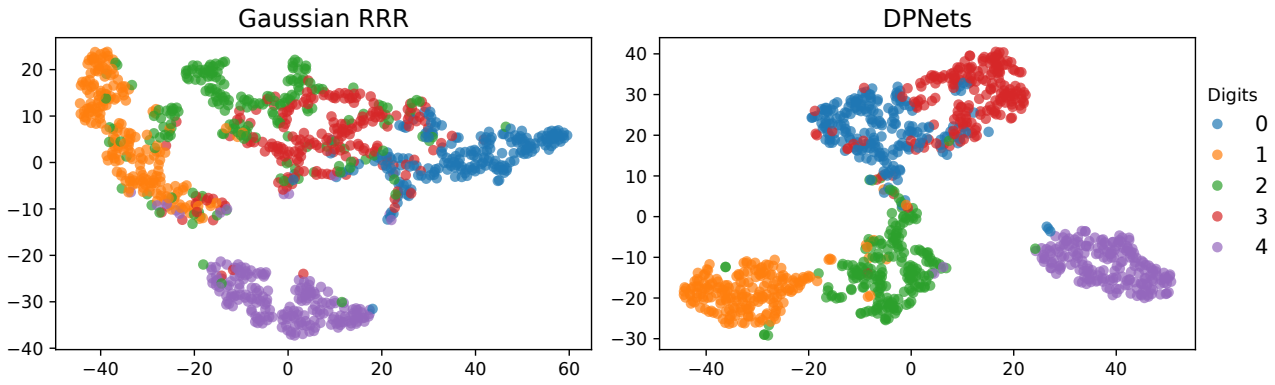

Figure 8: "t-SNE visualization of the concatenated left and right eigenfunctions on the MNIST test dataset.

Here are the results for two different choices of $\eta = 0.2$ and $\eta = 0.05$.

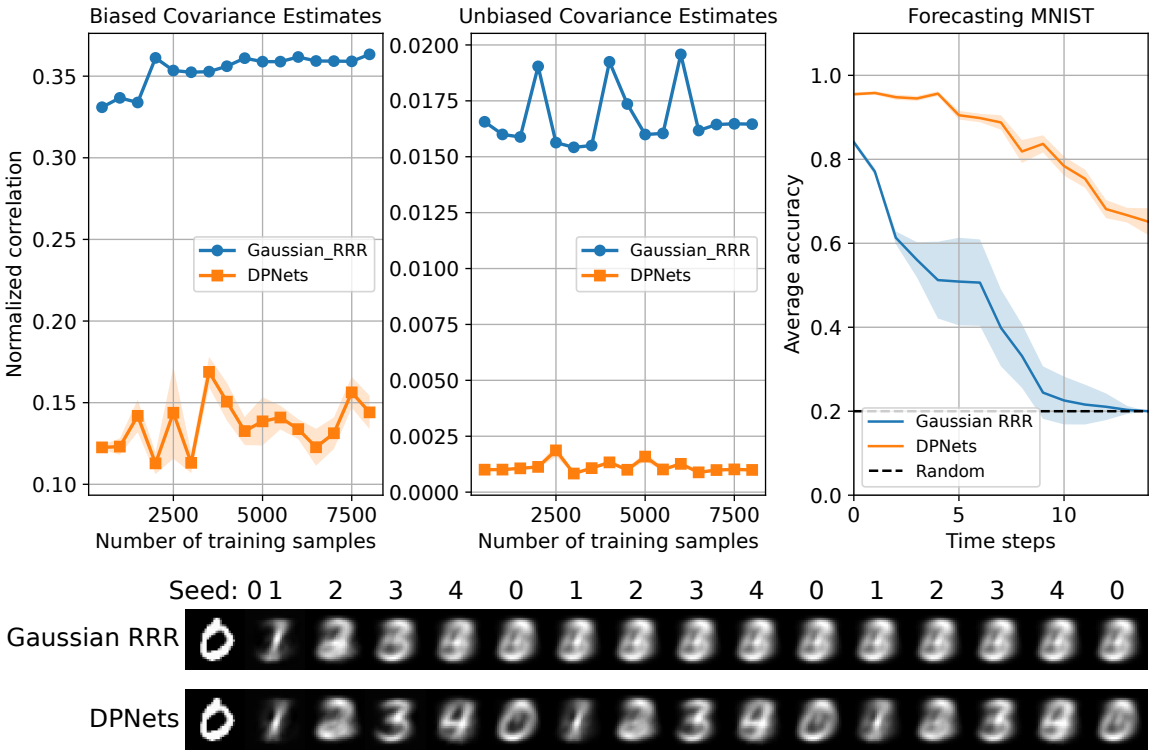

Figure 9: Performance evaluation of rank-5 RRR estimators using Gaussian and DPNet kernels on MNIST with $\eta = 0.2$: normalized correlations and forecast accuracy.

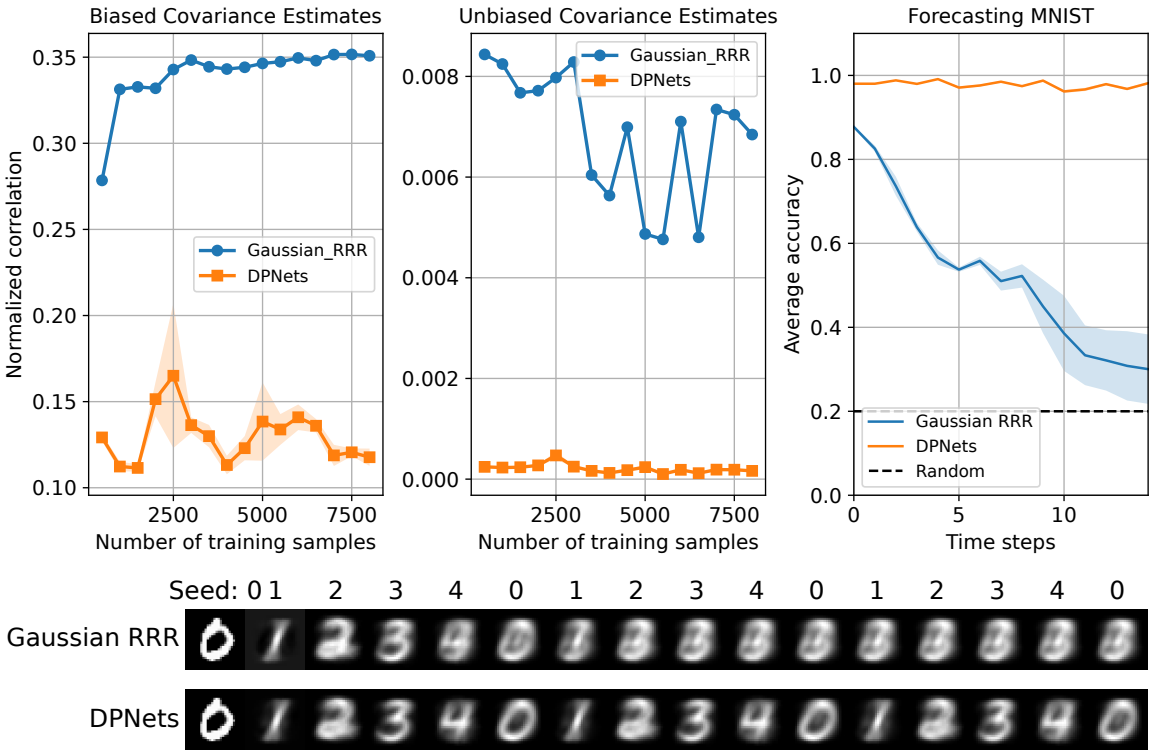

Figure 11: Performance evaluation of rank-5 RRR estimators using Gaussian and DPNet kernels on MNIST with $\eta = 0.05$: normalized correlations and forecast accuracy.

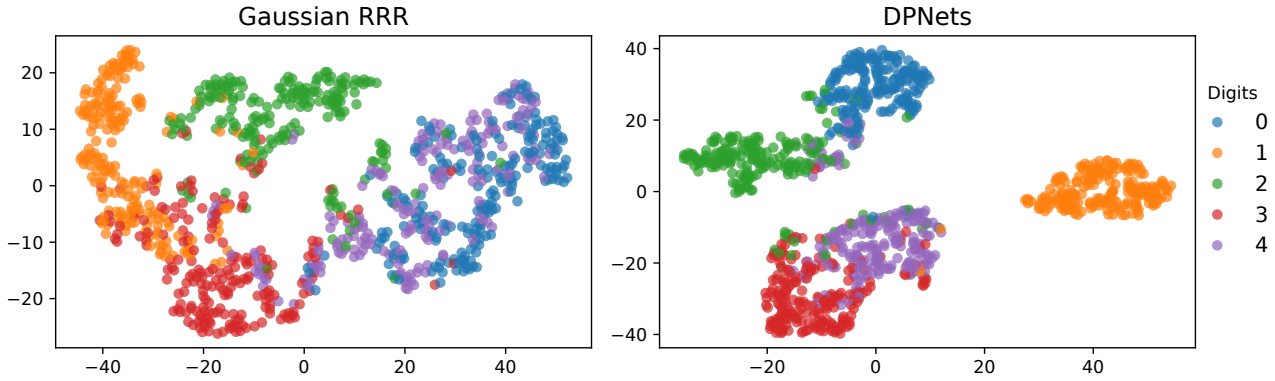

Figure 10: "t-SNE visualization of the concatenated left and right eigenfunctions on the MNIST test dataset with $\eta = 0.2$

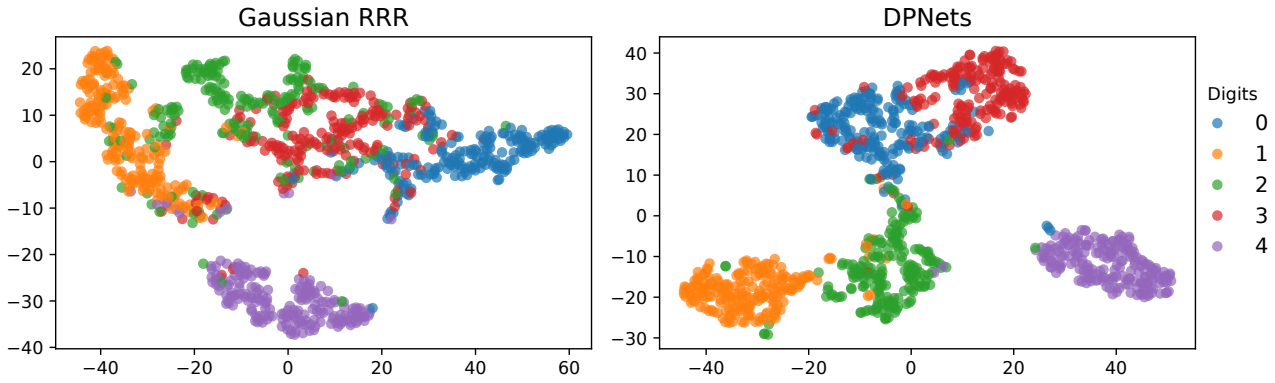

Figure 12: "t-SNE visualization of the concatenated left and right eigenfunctions on the MNIST test dataset with $\eta = 0.05$