Learning with little mixing

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study square loss in a realizable time-series framework with martingale difference noise. Our main result is a fast rate excess risk bound which shows that whenever a *trajectory hypercontractivity* condition holds, the risk of the least-squares estimator on dependent data matches the iid rate order-wise after a burn-in time. In comparison, many existing results in learning from dependent data have rates where the effective sample size is deflated by a factor of the mixing-time of the underlying process, even after the burn-in time. Furthermore, our results allow the covariate process to exhibit long range correlations which are substantially weaker than geometric ergodicity. We call this phenomenon *learning with little mixing*, and present several examples for which this phenomenon occurs: bounded function classes for which the L^2 and $L^{2+\varepsilon}$ norms are equivalent, finite state irreducible and aperiodic Markov chains, and a broad family of infinite dimensional function classes based on subsets of $\ell^2(\mathbb{N})$ ellipsoids.

1 Introduction

Consider regression in the context of the time-series model:

$$Y_t = f_\star(X_t) + W_t, \qquad t \in \mathbb{N}. \tag{1}$$

Such models are ubiquitous in applications of machine learning, signal processing, econometrics, and control theory. In our setup, the learner is given access to $T \in \mathbb{N}_+$ pairs $\{(X_t, Y_t)\}_{t=0}^{T-1}$ drawn from (1), and is asked to output a hypothesis $\widehat{f}$ from a hypothesis class $\mathscr{F}$ which best approximates the (realizable) regression function $f_\star \in \mathscr{F}$ in terms of square loss.

In this work, we study the least-squares estimator (LSE). This procedure minimizes the empirical risk associated to the square loss over the class $\mathscr{F}$. When each pair of observations (X_t, Y_t) is drawn iid from some fixed distribution, this procedure is minimax optimal over a broad set of hypothesis classes [1–4]. However, much less is known about the optimal rate of convergence for the general time-series model (1), as correlations across time in the covariates $\{X_t\}$ complicate the analysis.

With this in mind, we seek to extend our understanding of the minimax optimality of the LSE for the time-series model (1). We show that for a broad class of function spaces and covariate processes, the effects of data dependency across time enter the LSE excess risk only as a higher order term, whereas the leading term in the excess risk remains order-wise identical to that in the iid setting. Hence, after a sufficiently long, but finite *burn-in time*, the LSE's excess risk scales as if all T samples are independent. This behavior applies to covariate processes that exhibit correlations which decay slower than geometrically. We refer to this double phenomenon, where (a) mixing-times only enter as a burn-in time, and (b) where the mixing requirement can be mild, as *learning with little mixing*.

Our result stands in contrast to a long line of work on learning from dependent data (see e.g., [5–12] and the references within), where the blocking technique [5] is used to create independence amongst the dependent covariates, so that tools to analyze independent learning can be applied. While these

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

aforementioned works differ in their specific setups, the main commonality is that the resulting dependent data rates mimic the corresponding independent rates, but with the caveat that the sample size is replaced by an "effective" sample size that is *decreased* in some way by the mixing-time, even after any necessary burn-in time. Interestingly, work on the model (1) by Ziemann et al. [13], which does not rely on the blocking technique, also suffers from such sample degradation.

The model (1) captures learning dynamical systems by setting $Y_t = X_{t+1}$, so that the regression function $f_\star$ describes the dynamics of the state variable X_t. Recent progress in system identification shows that the lack of ergodicity does not necessarily degrade learning rates. Indeed, when the states evolve as a linear dynamical system (i.e., the function $f_\star$ is linear), learning rates are not deflated by any mixing times, and match existing rates for iid linear regression [14–17]. Kowshik et al. [18] extend results of this flavor to parameter recovery of dynamics driven by a generalized linear model. The extent to which this phenomenon—less ergodicity not impeding learning rates—generalizes beyond linear and generalized linear models is a key motivation for our work.

Contributions We consider the realizable setting, where $f_\star$ is assumed to be contained in a known function space $\mathscr{F}$. Our results rest on two assumptions regarding both the covariate process $\{X_t\}$ and the function space $\mathscr{F}$. The first assumption posits that the process $\{X_t\}$ exhibits some mild form of ergodicity (that is significantly weaker than the typical geometric ergodicity assumption). The second assumption is a hypercontractivity condition that holds uniformly in $\mathscr{F}$ along the trajectory $\{X_t\}$, extending contractivity assumptions for iid learning [3] to dependent processes.

Informally, our main result (Theorem 1, presented in Section 4), shows that under these two assumptions, letting $\mathrm{comp}(\mathscr{F})$ denote some (inverse) measure of complexity of $\mathscr{F}$, the LSE $\widehat{f}$ satisfies:

$$\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \lesssim \left(\frac{\text{dimensional factors} \times \sigma_W^2}{T}\right)^{\mathrm{comp}(\mathscr{F})} + \text{higher order } o(1/T^{\mathrm{comp}(\mathscr{F})}) \text{ terms.} \quad (2)$$

The first term in (2) matches existing LSE risk bounds for iid learning order-wise, and most importantly, does not include any dependence on the mixing-time of the process. Indeed, all mixing-time dependencies enter only in the higher order term. Since this term scales as $o(1/T^{\mathrm{comp}(\mathscr{F})})$, it becomes negligible after a finite burn-in time. This captures the crux of our results: on a broad class of problems, given enough data, the LSE applied to time-series model (1) behaves as if all samples are independent.

We conclude in Section 6 with several examples for which the trajectory hypercontractivity assumption holds. When the covariate process $\{X_t\}$ is generated by a finite-state irreducible and aperiodic Markov chain, then any function class $\mathscr{F}$ satisfies the requisite condition. More broadly, the condition is satisfied for any bounded function classes for which the L^2 and $L^{2+\varepsilon}$ norms (along trajectories) are equivalent. Our last example concerns infinite dimensional function spaces based on $\ell^2(\mathbb{N})$ ellipsoids, and demonstrates that our results are not inherently limited to finite-dimensional hypothesis classes.

2 Related work

While regression is a fundamental problem studied across many disciplines, our work draws its main inspiration from Mendelson [3] and Simchowitz et al. [14]. Mendelson [3] shows that for nonparametric iid regression, only minimal assumptions are required for one-sided isometry, and thus learning. Simchowitz et al. [14] build on this intuition and provide mixing-free rates for linear regression over trajectories generated by a linear dynamical system. We continue this trend, by leveraging one-sided isometry to show that mixing only enters as a higher order term in the rates of the nonparametric LSE. More broadly, the technical developments we follow synthesize techniques from two lines of work: nonparametric regression with iid data, and learning from dependent data.

Nonparametric regression with iid data Beyond the seminal work of Mendelson [3], the works [2, 19, 20] all study iid regression with square loss under various moment equivalence conditions. In addition to moment equivalence, we build on the notion of offset Rademacher complexity defined by Liang et al. [21] in the context of iid regression. Indeed, we show that a martingale analogue of the offset complexity (described in Ziemann et al. [13]) characterizes the LSE rate in (1).

Learning from dependent data As discussed previously, many existing results for learning from dependent data reduce the problem to independent learning via the blocking technique [5], at the

expense of sample complexity deflation by the mixing-time. Nagaraj et al. [22] prove a lower bound stating that in a worst case agnostic model, this deflation is unavoidable. However, in a realizable setting, Nagaraj et al. [22] show that the mixing-time only needs to affect the burn-in time, but not necessarily the risk, leaving open the question of whether the risk after burn-in should take on the iid rate or not. Our work can be interpreted as the "upper bound" for this result, where we put forth sufficient conditions to recover the iid rate after a burn-in time. Ziemann et al. [13] also study the model (1), and provide an information-theoretic analysis of the nonparametric LSE. However, their approach fundamentally reduces to showing two-sided concentration—something our work evades—and therefore their bounds incur worst case dependency on the mixing-time.

The measure of dependencies we use for the process $\{X_t\}$ is due to Samson [23]. Recently, Dagan et al. [24] use a similar measure to study learning when the covariates have no obvious sequential ordering (e.g., a graph structure or Ising model). However, our results are not directly comparable, other than noting that their risk bounds increase with the measure of correlation.

Results in linear system identification show that lack of ergodicity does not degrade parameter recovery rates [14–17, 25, 26]. Beyond linear system identification, the works [18, 27, 28] prove parameter recovery bounds for a dynamical system driven by a generalized linear model. Most relevant is Kowshik et al. [18], who again show that the lack of ergodicity does not hamper rates. We defer a detailed comparison of our results with these aforementioned works to Section 4.3.

3 Problem formulation

The time-series (1) evolves on two subsets of Euclidean space, $\mathsf{X} \subset \mathbb{R}^{\mathsf{d_X}}$ and $\mathsf{Y} \subset \mathbb{R}^{\mathsf{d_Y}}$, with $X_t \in \mathsf{X}$ and $Y_t, W_t \in \mathsf{Y}$. Expectation (resp. probability) with respect to all the randomness of the underlying probability space is denoted by $\mathbf{E}$ (resp. $\mathbf{P}$). The Euclidean norm on $\mathbb{R}^d$ is denoted $\|\cdot\|_2$, and the unit sphere in $\mathbb{R}^d$ is denoted $\mathbb{S}^{d-1}$. For a matrix $M \in \mathbb{R}^{d_1 \times d_2}$, $\|M\|_{\mathsf{op}}$ denotes the largest singular value.

We assume there exists a filtration $\{\mathcal{F}_t\}$ such that a) $\{W_t\}$ is a square integrable martingale difference sequence (MDS) with respect to this filtration, and b) $\{X_t\}$ is adapted to $\{\mathcal{F}_{t-1}\}$. Further, tail conditions on this MDS will be imposed as necessary later on.

Let $\mathscr{F}$ be a hypothesis space of functions mapping $\mathbb{R}^{\mathsf{d_X}}$ to $\mathbb{R}^{\mathsf{d_Y}}$. We assume that the true regression function is an element of $\mathscr{F}$ (i.e., $f_\star \in \mathscr{F}$), and that $\mathscr{F}$ is known to the learner. Given two compatible function spaces $\mathscr{F}_1, \mathscr{F}_2$, let $\mathscr{F}_1 - \mathscr{F}_2 \triangleq \{f_1 - f_2 \mid f_1 \in \mathscr{F}_1, f_2 \in \mathscr{F}_2\}$. A key quantity in our analysis is the shifted function class $\mathscr{F}_\star \triangleq \mathscr{F} - \{f_\star\}$. Our results will be stated under the assumption that $\mathscr{F}_\star$ is *star-shaped*.[1] For any function $f : \mathsf{X} \to \mathbb{R}^{\mathsf{d_Y}}$, we define $\|f\|_\infty \triangleq \sup_{x \in \mathsf{X}} \|f\|_2$. A function f is B-bounded if $\|f\|_\infty \le B$. Similarly, a hypothesis class is B-bounded if each of its elements is B-bounded. For a bounded class $\mathscr{F}$ and resolution $\varepsilon > 0$, the quantity $\mathcal{N}_\infty(\mathscr{F}, \varepsilon)$ denotes the size of the minimal ε-cover of $\mathscr{F}$ (contained in $\mathscr{F}$) in the $\|\cdot\|_\infty$-norm.

We fix a $T \in \mathbb{N}_+$, indicating the number of labeled observations $\{(X_t, Y_t)\}_{t=0}^{T-1}$ from the time-series (1) that are available to the learner. The joint distribution of $X_{0:T-1} \triangleq (X_0, \ldots, X_{T-1})$ is denoted P_X. For $p \ge 1$, we endow $\mathscr{F} - \mathscr{F}$ with $L^p(\mathsf{P}_X)$ norms: $\|f - g\|_{L^p}^p \triangleq \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{E}\|f(X_t) - g(X_t)\|_2^p$, where expectation is taken with respect to P_X. We will mostly be interested in $L^2(\mathsf{P}_X)$, hereafter often just referred to as L^2. This is the L^2 space associated to the law of the uniform mixture over $X_{0:T-1}$ and thus, for iid data, coincides with the standard L^2 space often considered in iid regression.

The learning task is to produce an estimate $\widehat{f}$ of $f_\star$, which renders the excess risk $\|\widehat{f} - f_\star\|_{L^2}^2$ as small as possible. We emphasize that $\|\widehat{f} - f_\star\|_{L^2}^2 = \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{E}_{\tilde{X}_{0:T-1}} \|\widehat{f}(\tilde{X}_t) - f_\star(\tilde{X}_t)\|_2^2$ where $\tilde{X}_{0:T-1}$ is a fresh, statistically independent, sample with the same law P_X as $X_{0:T-1}$. Namely, $\|\widehat{f} - f_\star\|_{L^2}^2$ is a random quantity, still depending on the internal randomness of the learner and that of the sample $X_{0:T-1}$ used to generate $\widehat{f}$. We study the performance of the least-squares estimator (LSE) defined as $\widehat{f} \in \operatorname{argmin}_{f \in \mathscr{F}} \left\{ \frac{1}{T} \sum_{t=0}^{T-1} \|Y_t - f(X_t)\|_2^2 \right\}$, and measure the excess risk $\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2$.

[1] A function class $\mathscr{F}$ is star-shaped if for any $\alpha \in [0, 1]$, $f \in \mathscr{F}$ implies $\alpha f \in \mathscr{F}$.

4 Results

This section present our main result. We first detail the definitions behind our main assumptions in Section 4.1. The main result and two corollaries are then presented in Section 4.2.

4.1 Hypercontractivity and the dependency matrix

Hypercontractivity We first state our main trajectory hypercontractivity condition, which we will use to establish lower isometry. The following definition is heavily inspired by recent work on learning without concentration [3, 20].

Definition 1 (Trajectory (C, α)-hypercontractivity)**.** *Fix constants $C > 0$ and $\alpha \in [1, 2]$. We say that the tuple $(\mathscr{F}, \mathsf{P}_X)$ satisfies the trajectory (C, α)-hypercontractivity condition if*

$$\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^4\right] \le C\left(\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right]\right)^{\alpha} \text{ for all } f \in \mathscr{F}. \tag{3}$$

Here, the expectation is with respect to P_X, the joint law of $X_{0:T-1}$.

Condition (3) interpolates between boundedness and small-ball behavior. Indeed, if the class $\mathscr{F}$ is B-bounded, then it satisfies trajectory $(B^2, 1)$-hypercontractivity trivially. On the other hand, for $\alpha = 2$, (3) asks that $\|f\|_{L^4} \le C^{1/4}\|f\|_{L^2}$ for trajectory-wise L^p-norms; by the Paley-Zygmund inequality, this implies that a small-ball condition holds. Moreover, if for some $\varepsilon \in (0, 2)$, the trajectory-wise L^2 and $L^{2+\varepsilon}$ norms are equivalent on $\mathscr{F}$, then Proposition 3 (Section 6) shows that the condition holds for a nontrivial $\alpha = 1 + \varepsilon/2 \in (1, 2)$. More examples are given in Section 6.

Our main results assume that $(\mathscr{F}_\star, \mathsf{P}_X)$ (or a particular subset of $\mathscr{F}_\star$) satisfies the trajectory (C, α)-hypercontractivity condition with $\alpha > 1$, which we refer to as the *hypercontractive* regime. The condition $\alpha > 1$ is required in our analysis for the lower order excess risk term to not depend on the mixing-time. Our results instantiated for the $\alpha = 1$ case directly correspond to existing work by Ziemann et al. [13], and exhibit a lower order term that depends on the mixing-time.

Ergodicity via the dependency matrix We now state the main definition we use to measure the stochastic dependency of a process. Recall that for two measures μ, ν on the same measurable space with σ-algebra $\mathscr{A}$, the total-variation norm is defined as $\|\mu - \nu\|_{\mathsf{TV}} \triangleq \sup_{A \in \mathscr{A}} |\mu(A) - \nu(A)|$.

Definition 2 (Dependency matrix, Samson [23, Section 2])**.** *The dependency matrix of a process $(Z_t)_{t=0}^{T-1}$ with measure P_Z is the (upper-triangular) matrix $\Gamma_{\mathsf{dep}}(\mathsf{P}_Z) = (\Gamma_{ij})_{i,j=0}^{T-1} \in \mathbb{R}^{T \times T}$ defined as follows. Let $\mathscr{Z}_{0:i}$ denote the σ-algebra generated by $(Z_t)_{t=0}^{i}$. For indices $i < j$, let*

$$\Gamma_{ij} = \sqrt{2 \sup_{A \in \mathscr{Z}_{0:i}} \|\mathsf{P}_{Z_{j:T-1}}(\cdot \mid A) - \mathsf{P}_{Z_{j:T-1}}\|_{\mathsf{TV}}}.$$

For the remaining indices $i \ge j$, let $\Gamma_{ii} = 1$ and $\Gamma_{ij} = 0$ when $i > j$ (below the diagonal).

Given the dependency matrix from Definition 2, we measure the dependency of the process P_X by the quantity $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}$. Notice that this quantity always satisfies $1 \le \|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}} \lesssim T$. The lower bound indicates that the process P_X is independent across time. The upper bound indicates that the process is fully dependent, e.g., $X_{t+1} = X_t$ for all t.

Our results apply to cases where $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2$ grows sub-linearly in T– the exact requirement depends on the specific function class $\mathscr{F}_\star$. If the process is geometrically ϕ-mixing, then $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2$ is upper bounded by a constant that depends on the mixing-time of the process, and is independent of T [23, Section 2]. Other examples, such as processes satisfying Doeblin's condition [29], are given in Samson [23, Section 2]. When $\{Z_t\}$ is a time-homogenous Markov chain, $\Gamma_{ij}^2 = 2\,\mathrm{ess\,sup}_{z \in \mathsf{Z}} \|\mathsf{P}_{Z_{j-i}}(\cdot \mid Z_0 = z) - \mathsf{P}_{Z_{j-i}}\|_{\mathsf{TV}}$ for indices $j > i$. Hence, the requirement $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2 \lesssim T^\beta$ for $\beta \in (0, 1)$ then corresponds to $\mathrm{ess\,sup}_{z \in \mathsf{Z}} \|\mathsf{P}_{Z_t}(\cdot \mid Z_0 = z) - \mathsf{P}_{Z_t}\|_{\mathsf{TV}} \lesssim 1/t^{1-\beta}$ for $t \in \mathbb{N}_+$. Jarner and Roberts [30] give various examples and conditions to check polynomial convergence rates for Markov chains.

4.2 Learning with little mixing

A key quantity appearing in our bounds is a martingale variant of the notion of Gaussian complexity.

Definition 3 (Martingale offset complexity, cf. Liang et al. [21], Ziemann et al. [13])**.** *For the regression problem (1), the (P_X) martingale offset complexity of a function space $\mathscr{F}$ is given by:*

$$\mathsf{M}_T(\mathscr{F}) \triangleq \sup_{f \in \mathscr{F}} \left\{ \frac{1}{T} \sum_{t=0}^{T-1} 4\langle W_t, f(X_t)\rangle - \|f(X_t)\|_2^2 \right\}. \tag{4}$$

Recall that $\mathscr{F}_\star = \mathscr{F} - \{f_\star\}$ is the centered function class, and define $\partial B(r) \triangleq \{f \in \mathscr{F}_\star \mid \|f\|_{L^2} = r\}$. The following theorem is the main result of this paper.

Theorem 1. *Fix $B > 0$, $r \in (0, B]$, $\alpha \in [1, 2]$, and $C : (0, B] \to \mathbb{R}_+$. Suppose that $\mathscr{F}_\star$ is star-shaped and B-bounded. Let $\mathscr{F}_r \subset \mathscr{F}_\star$ be a $r/\sqrt{8}$-net of $\partial B(r)$ in the supremum norm $\|\cdot\|_\infty$, and suppose that $(\mathscr{F}_r, \mathsf{P}_X)$ satisfies the trajectory $(C(r), \alpha)$-hypercontractivity condition. Then:*

$$\mathbf{E}\|\widehat{f} - f_\star\|_{L_2}^2 \le 8\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star) + r^2 + B^2|\mathscr{F}_r| \exp\left(\frac{-Tr^{4-2\alpha}}{8C(r)\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2} \right). \tag{5}$$

To understand Theorem 1, we will proceed in a series of steps. We first need to understand the martingale complexity term $\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star)$. Since $\mathscr{F}_\star$ is B-bounded, if one further imposes the tail conditions that the noise process $\{W_t\}$ is a σ_W^2-sub-Gaussian MDS,[2] a chaining argument detailed in Ziemann et al. [13, Lemma 4] shows that:

$$\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star) \lesssim \inf_{\gamma > 0} \left\{ \frac{\sigma_W^2 \log \mathcal{N}_\infty(\mathscr{F}_\star, \gamma)}{T} + \frac{\sigma_w}{\sqrt{T}} \int_0^\gamma \sqrt{\log \mathcal{N}_\infty(\mathscr{F}_\star, s)} ds \right\}. \tag{6}$$

In particular, this bound only depends on $\mathscr{F}$ and is *independent* of $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2$. Furthermore, (6) coincides with the corresponding risk bound for the LSE with iid covariates [21].

Given that $\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star)$ corresponds to the rate of learning from T iid covariates, the form of (5) suggests that we choose $r^2 \lesssim \mathbf{E}\mathsf{M}_T(\mathscr{F}_\star)$, so that the dominant term in (5) is equal to $\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star)$ in scale. Given that r has been set, the only remaining degree of freedom in (5) is to set T large enough (the burn-in time) so that the third term is dominated by r^2. Thus, it is this third term captures the interplay between the function class $\mathscr{F}_\star$ and the dependency structure $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2$. We will now consider specific examples to illustrate how the burn-in time can be set.

Our first example supposes that (a) $\mathscr{F}_\star$ satisfies the trajectory $(C, 2)$-hypercontractivity condition, and (b) that $\mathscr{F}_\star$ is non-parametric, but not too large:

$$\exists\, p \in \mathbb{R}_+,\ q \in (0, 2) \text{ s.t. } \log \mathcal{N}_\infty(\mathscr{F}_\star, \varepsilon) \le p \left(\frac{1}{\varepsilon}\right)^q \text{ for all } \varepsilon \in (0, 1). \tag{7}$$

Covering numbers of the form (7) are typical for sufficiently smooth function classes, e.g. the space of k-times continuously differentiable functions mapping $\mathsf{X} \to \mathsf{Y}$ for any $k \ge \lceil \mathsf{d}_\mathsf{X}/2 \rceil$ [31]. If condition (7) holds and the noise process $\{W_t\}$ is a sub-Gaussian MDS, inequality (6) yields $\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star) \lesssim T^{-\frac{2}{2+q}}$, and hence we want to set $r^2 = o(T^{-\frac{2}{2+q}})$. Carrying out this program yields the following corollary.

Corollary 1. *Fix $C > 0$, $B \ge 1$, $q \in (0, 2)$, $p > 0$, and $\gamma \in (0, \frac{q}{2+q})$. Suppose that $\mathscr{F}_\star$ is star-shaped, B-bounded, satisfies (7), and $(\mathscr{F}_\star, \mathsf{P}_X)$ satisfies the trajectory $(C, 2)$-hypercontractivity condition. Suppose that T satisfies:*

$$T \ge \max\left\{ \left[8(8p+1)C\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2\right]^{\frac{1}{1-\frac{q}{2}\left(\frac{2}{2+q}+\gamma\right)}}, \left[\log B + \frac{4}{q}\log\left(\frac{8}{q}\right)\right]^{\frac{1}{\frac{q}{2}\left(\frac{2}{2+q}+\gamma\right)}} \right\}. \tag{8}$$

Then, we have that:

$$\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \le 8\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star) + 2T^{-\left(\frac{2}{2+q}+\gamma\right)}. \tag{9}$$

[2] That is, for any $u \in \mathbb{S}^{\mathsf{d}_\mathsf{Y}-1}$, $\lambda \in \mathbb{R}$, and $t \in \mathbb{N}$, we have $\mathbf{E}[\exp(\lambda\langle W_t, u\rangle) \mid \mathcal{F}_{t-1}] \le \exp(\lambda^2\sigma_W^2/2)$.

Inequality (9) of Corollary 1 emphasizes the fact that the first order term of the excess risk is bounded by the martingale offset complexity $\mathsf{M}_T(\mathscr{F}_\star)$. This behavior arises since the dependency matrix $\Gamma_{\mathsf{dep}}(\mathsf{P}_X)$ only appears as the burn-in requirement (8). Here, the value of q constrains how fast $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2$ is allowed to grow. In particular, condition (8) requires that $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2 = o(T^{1-\frac{q}{2+q}})$, otherwise the burn-in condition cannot be satisfied for any valid $\gamma \in (0, \frac{q}{2+q})$.

In our next example, we consider both a variable hypercontractivity parameter $C(r)$ that varies with the covering radius r, and also allow $\alpha \in (1, 2]$ to vary. Since our focus is on the interaction of the parameters in the hypercontracitity definition, we will consider smaller function classes with logarithmic metric entropy. This includes parametric classes but also bounded subsets of certain reproducing kernel Hilbert spaces. For such function spaces, one expects $\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star) \asymp T^{-1}$, and hence we set $r^2 = o(T^{-1})$.

Corollary 2. *Fix a $B \geq 1$, $C : (0, 1] \to \mathbb{R}_+$, $\alpha \in (1, 2]$, $b_1 \in [0, 1)$, $b_2 \in [0, 2)$, $\gamma \in (0, 1)$, and $p, q \geq 1$. Suppose that $\mathscr{F}_\star$ is star-shaped and B-bounded, and that for every $r \in (0, 1)$, there exists a r-net $\mathscr{F}_r$ of $\mathscr{F}_\star$ in the $\|\cdot\|_\infty$-norm such that (a) $\log |\mathscr{F}_r| \leq p \log^q \left(\frac{1}{r}\right)$ and (b) $(\mathscr{F}_r, \mathsf{P}_X)$ satisfies the trajectory $(C(r), \alpha)$-hypercontractivity condition. Next, suppose the growth conditions hold:*

$$\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2 \leq T^{b_1}, \;\; C(r) \leq (1/r)^{b_2} \, \forall r \in (0, 1).$$

As long as the constants α, b_1, b_2, and γ satisfy:

$$\psi := 1 - b_1 - \frac{(1+\gamma)(4 - 2\alpha + b_2)}{2} > 0, \tag{10}$$

then for any $T \geq \mathsf{poly}_{\frac{q}{\psi}}\left(p, \log B, \frac{q}{\psi}\right)$, we have:

$$\mathbf{E}\|\widehat{f} - f_\star\|_{L_2}^2 \leq 8\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star) + 2\left(\frac{1}{T}\right)^{1+\gamma}.$$

Here $\mathsf{poly}_{q/\psi}$ denotes a polynomial of degree $O(q/\psi)$ in its arguments– the exact expression is given in the proof. Proposition 5 in Section 6 gives an example of an $\ell^2(\mathbb{N})$ ellipsoid which satisfies the assumptions in Corollary 2. Corollary 2 illustrates the interplay between the function class $\mathscr{F}_\star$, the data dependence of the covariate process $\{X_t\}$, and the hypercontractivity constant α. Let us consider a few cases. First, let us suppose that the process $\{X_t\}$ is geometrically ergodic and that $C(r)$ is a constant, so that we can set b_1 and b_2 arbitrarily close to zero (at the expense of a longer burn-in time). Then, inequality (10) simplifies to $\alpha > 2 - \frac{1}{1+\gamma}$. This illustrates that in the hypercontractivity regime ($\alpha > 1$), there exists a valid setting of (b_1, b_2, γ) that satisfies inequality (10). Next, let us consider the case where $C(r)$ is again a constant, but $\{X_t\}$ is not geometrically ergodic. Setting b_2 and γ arbitrarily close to zero, we have that (10) simplifies to $b_1 < \alpha - 1$. Compared to Corollary 1, we see that in the case when $\alpha = 2$, the parametric nature of $\mathscr{F}_\star$ allows the dependency requirement to be less strict: $o(T)$ in the parametric case versus $o(T^{1-\frac{q}{2+q}})$ in the non-parametric case.

We conclude with noting that when $\alpha = 1$, it is not possible to remove the dependence on $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2$ in the lowest order term. In this situation, our results recover existing risk bounds from Ziemann et al. [13]– see the appendix for a discussion.

4.3 Comparison to existing system identification literature

Linear dynamical systems As mentioned in Section 2, a special case of the model (1) which has received much recent interest are LDS, where $f_\star(x) = A_\star x$, $A_\star \in \mathbb{R}^{\mathsf{d_x} \times \mathsf{d_x}}$ and $Y_t = X_{t+1}$. While our results concern L^2 excess risk, in this particular case, Frobenius norm parameter recovery bounds and excess risk bounds are almost equivalent, and so a comparison with Theorem 1 is appropriate. Assume that $A_\star$ is stable, i.e., the spectral radius of $A_\star$, denoted ρ, satisfies $\rho < 1$. Let the hypothesis class be $\mathscr{F} = \{f(x) = Ax \mid A \in \mathbb{R}^{\mathsf{d_x} \times \mathsf{d_x}}, \operatorname{tr} A^\top A \leq c\}$ for some $c > 0$. Further suppose that the noise W_t distribution is a log-concave density on $\mathbb{R}^{\mathsf{d_x}}$. It is straightforward to verify that $(\mathscr{F}_\star, \mathsf{P}_X)$ is trajectory $(C, 2)$-hypercontractive for some $C \asymp 1/(1-\rho^2)^2$ via Carbery and Wright [32, Theorem 7]. Additionally, one can bound $\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star)$ by either chaining [13, Lemma 4], or by a more direct argument [21, Lemma 10]. We then recover, up to logarithmic factors, existing results [14, 16, 33] for stable systems. We do however note that this assertion holds with two caveats: (1) Theorem 1 in this

particular case requires bounded noise to ensure that the hypothesis class $\mathscr{F}$ is B-bounded for some $B > 0$. (2) Similar to Jedra and Proutiere [33], we suffer a burn-in depending on the mixing-time and thus cannot necessarily handle the marginally stable case (where $A_\star$ is allowed to have spectral radius 1) as in [14, 16]. Regarding (1), this limitation can be avoided if one exploits the fact that the function $x \mapsto Ax$ is uniformly Lipschitz over $\mathscr{F}$ (with high probability) by invoking a moment assumption on the covariates themselves. As for alleviating (2), one would likely require a different approach to lower isometry (cf. Section 5) that does not rely on the exponential inequality of Samson [23], which implicitly assumes some level of mixing. We leave this to future work.

Generalized linear model dynamical systems Moving beyond linear dynamics, Kowshik et al. [18], study parameter recovery of the dynamics $A_\star$ from state observations $\{X_t\}$ of the dynamical system $X_{t+1} = f_\star(X_t) + W_t$, $f_\star(x) = \sigma(A_\star x)$, where σ is a 1-Lipschitz link function. Their main result [18, Theorem 1] shows that when $\rho \triangleq \|A_\star\|_{\mathsf{op}}$ satisfies $\rho < 1$,[3] the parameter error $\|\hat{A} - A_\star\|_F^2$ of their gradient based estimator $\hat{A}$ only depends poly-logarithmically on $1/(1-\rho)$.

Our results are not directly comparable to this work for two main reasons. First, we are unable to verify our requisite hypercontractivity conditions on this GLM model. Second, for the GLM model, the relationship between the excess risk and the parameter error is not as straightforward as the linear case. A naïve upper bound using the 1-Lipschitz property of the link function is: $\|\hat{f} - f_\star\|_{L_2}^2 \le \|\hat{A} - A_\star\|_{\mathsf{op}}^2 \frac{1}{T}\sum_{t=0}^{T-1}\|X_t\|_2^2 \lesssim \|\hat{A} - A_\star\|_{\mathsf{op}}^2 \frac{1}{(1-\rho)^2}$. Hence, even if the parameter error grows poly-logarithmically in $1/(1-\rho)$, the excess risk bound still grows polynomially in $1/(1-\rho)$. Thus, for this GLM model, a more sophisticated analysis is necessary to determine if the excess risk (for any estimator) is independent (or poly-logarithmically dependent) on the mixing-time $1/(1-\rho)$.

5 Proof techniques

In this section, we highlight the main ideas behind the proof of Theorem 1. The full details can be found in the appendix. We start with a key insight from Mendelson [3]: establishing a one-sided inequality between the empirical versus true risk is substantially easier than the corresponding two-sided inequality. Define the closed ball of radius r as $B(r) \triangleq \left\{ f \in \mathscr{F}_\star \,\middle|\, \mathbf{E}\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2 \le r^2 \right\}$. We identify conditions which depend mildly on $\Gamma_{\mathsf{dep}}(\mathsf{P}_X)$, so that with high probability we have:

$$\forall f \in \mathscr{F}_\star \setminus B(r), \;\; \mathbf{E}\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2 \lesssim \frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2. \tag{11}$$

Once this *lower isometry* condition (11) holds, we bound the empirical excess risk, the RHS of inequality (11), by a version of the basic inequality of least squares [13, 21]. This leads to an upper bound of the empirical excess risk by the martingale complexity term (Definition 3). As our innovation mainly lies in establishing the lower isometry condition (11), we focus on this component for the remainder of the proof outline.

Lower isometry The key tool we use is the following exponential inequality, which controls the lower tail of sums of non-negative dependent random variables via the dependency matrix $\Gamma_{\mathsf{dep}}(\mathsf{P}_X)$.

Proposition 1 (Samson [23, Theorem 2])**.** *Let $g : \mathsf{X} \to \mathbb{R}$ be non-negative. Then for any $\lambda > 0$:*

$$\mathbf{E}\exp\left(-\lambda\sum_{t=0}^{T-1} g(X_t)\right) \le \exp\left(-\lambda\sum_{t=0}^{T-1}\mathbf{E}g(X_t) + \frac{\lambda^2\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2\sum_{t=0}^{T-1}\mathbf{E}g^2(X_t)}{2}\right). \tag{12}$$

We note that Samson [23, Theorem 2] actually proves a much stronger statement than Proposition 1 (a Talagrand-style uniform concentration inequality), from which Proposition 1 is a byproduct. With Proposition 1 in hand, the following lemma allows us to relate hypercontractivity to lower isometry.

Lemma 1. *Fix $C > 0$ and $\alpha \in [1, 2]$. Let $g : \mathsf{X} \to \mathbb{R}$ be a non-negative function satisfying*

$$\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} g^2(X_t)\right] \le C\left(\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} g(X_t)\right]\right)^\alpha. \tag{13}$$

[3] They allow for $\rho \le 1 + O(1/T)$, but the discussion is simplified by further assuming $\rho < 1$.

Then we have:

$$\mathbf{P}\left(\sum_{t=0}^{T-1} g(X_t) \le \frac{1}{2}\sum_{t=0}^{T-1} \mathbf{E}g(X_t)\right) \le \exp\left(-\frac{T}{8C\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}g(X_t)\right)^{2-\alpha}\right).$$

Now fix an $f \in \mathscr{F}_\star \setminus B(r)$, and put $g(x) = \|f(x)\|_2^2$. Substituting g into (13) yields the trajectory hypercontractivity condition (3) in Definition 1. Thus, Lemma 1 establishes the lower isometry condition (11) for any fixed function. Hence, it remains to take a union bound over a supremum norm cover of $\mathscr{F}_\star \setminus B(r)$ at resolution r. It turns out that it suffices to instead cover $\partial B(r)$ since $\mathscr{F}_\star$ is star-shaped. Carrying out these details leads to the main lower isometry result.

Theorem 2. *Fix constants $\alpha \in [1,2]$ and $C, r > 0$. Let $\mathscr{F}_\star$ be star-shaped, and suppose that there exists a $r/\sqrt{8}$-net $\mathscr{F}_r$ of $\partial B(r)$ in the $\|\cdot\|_\infty$-norm such that $(\mathscr{F}_r, \mathsf{P}_X)$ satisfies the trajectory (C,α)-hypercontractivity condition. Then the following lower isometry holds:*

$$\mathbf{P}\left(\sup_{f\in\mathscr{F}_\star\setminus B(r)}\left\{\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2 - \mathbf{E}\frac{1}{8T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right\} \le 0\right) \le |\mathscr{F}_r|\exp\left(\frac{-Tr^{4-2\alpha}}{8C\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2}\right).$$

6 Examples of trajectory hypercontractivity

In this section, we detail a few examples of trajectory hypercontractivity. Let us begin by considering the simplest possible example: a finite hypothesis class. Let $|\mathscr{F}| < \infty$. Define for any fixed $f \in \mathscr{F}_\star$ the constant $c_f \triangleq \mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^4\right] / \left(\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right]\right)^2$, where the ratio $0/0$ is taken to be 1. The class $\mathscr{F}_\star$ is trajectory $(\max_{f\in\mathscr{F}_\star} c_f, 2)$-hypercontractive.

Similarly, processes evolving on a finite state space can also be readily verified to be hypercontractive.

Proposition 2. *Fix a $\underline{\mu} > 0$. Let $\{\mu_t\}_{t=0}^{T-1}$ denote the marginal distributions of P_X. Suppose that the μ_t's all share a common support of a finite set of atoms $\{\psi_1,\dots,\psi_K\} \subset \mathbb{R}^{\mathsf{d_X}}$, and that $\min_{0\le t\le T-1}\min_{1\le k\le K}\mu_t(\psi_k) \ge \underline{\mu}$. For any function class $\mathscr{F}$ of functions mapping $\{\psi_1,\dots,\psi_K\} \to \mathbb{R}^{\mathsf{d_Y}}$, we have that $\mathscr{F}$ satisfies the trajectory $(1/\underline{\mu}, 2)$-hypercontractivity condition.*

We remark that when P_X is an ergodic and irreducible Markov chain over a finite state space, the condition $\underline{\mu} > 0$ is always valid even as $T \to \infty$ [34]. In this case, our findings are related to Wolfer and Kontorovich [35, Theorem 3.1], who show that in the high accuracy regime (i.e., after a burn-in time), the minimax rate of estimating the transition probabilities of such a chain is not affected by the mixing time (in their case the pseudo-spectral gap).

The examples considered thus far rely on the fact that under a certain degree of finiteness, the fourth and second moment can be made uniformly equivalent. The next proposition relaxes this assumption. Namely, if for some $\varepsilon \in (0,2]$ the L^2 and $L^{2+\varepsilon}$ norms are equivalent on a bounded class $\mathscr{F}$, this class automatically enjoys a nontrivial hypercontractivity constant, $\alpha > 1$ (cf. Mendelson [20]).

Proposition 3. *Fix $\varepsilon \in (0,2]$ and $c > 0$. Suppose that $\mathscr{F}$ is B-bounded and that $\|f\|_{L^{2+\varepsilon}} \le c\|f\|_{L^2}$ for all $f \in \mathscr{F}$. Then $\mathscr{F}$ satisfies the trajectory $(B^{2-\varepsilon}c^{2+\varepsilon}, 1+\varepsilon/2)$-hypercontractivity condition.*

Next, we show that for processes $\{X_t\}$ which converge fast enough to a stationary distribution, it suffices to verify the hypercontracitivty condition only over the stationary distribution. This mimics existing results in iid learning, where hypercontractivity is assumed over the covariate distribution [3, 21]. We first recall the definition of the χ^2 divergence between two measures. Let μ and ν be two measures over the same probability space, and suppose that μ is absolutely continuous w.r.t. ν. The $\chi^2(\mu,\nu)$ divergence is defined as $\chi^2(\mu,\nu) := \mathbf{E}_\nu\left[\left(\frac{d\mu}{d\nu} - 1\right)^2\right]$.

Proposition 4. *Suppose that the process $\{X_t\}$ has a stationary distribution π. Let $\{\mu_t\}$ denote the marginal distributions of $\{X_t\}$, and suppose that $\{\mu_t\}$ and π are all mutually absolutely continuous. Assume the process is ergodic in the following sense: there exists a finite C_{χ^2} such that*

$$\sup_{t\in\mathbb{N}}\frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\chi^2(\mu_t,\pi)} \le C_{\chi^2}, \quad \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\chi^2(\pi,\mu_t) = 0. \tag{14}$$

Let $\mathscr{F}$ be a set of functions such that there exists $C_{8\to 2}$ satisfying for all $f \in \mathscr{F}$:

$$\mathbf{E}_\pi \|f(X)\|_2^8 \le C_{8\to 2}(\mathbf{E}_\pi \|f(X)\|_2^2)^4. \tag{15}$$

Let T_0 be such that for any $T \ge T_0$, $\frac{1}{T}\sum_{t=0}^{T-1} \chi^2(\pi, \mu_t) \le \frac{1}{4\sqrt{C_{8\to 2}}(1+C_{\chi^2})}$. For any $T \ge T_0$, $\mathscr{F}$ satisfies the trajectory $(4\sqrt{C_{8\to 2}}\left(1 + C_{\chi^2}\right), 2)$-hypercontractivity condition.

Several remarks on Proposition 4 are in order. First, ergodicity is usually specified in terms of total-variation norm convergence instead of χ^2-divergence convergence. Since the χ^2-divergence upper bounds the total-variation norm, (14) is a stronger ergodicity condition than is typically required. Nevertheless, (14) is exhibited for a broad class of processes, including Langevin diffusions where the stationary distribution π satisfies a Poincaré inequality of the form $\mathrm{Var}_\pi(f) \le C_{\mathsf{PI}}\mathbf{E}_\pi\|\mathrm{grad} f\|_2^2$ for all continuously differentiable functions f [36, 37].

6.1 Ellipsoids in $\ell^2(\mathbb{N})$

Given that equivalence of norms is typically a finite-dimensional phenomenon, one may wonder whether examples of hypercontractivity exist in an infinite-dimensional setting. Here we show that such examples are actually rather abundant. The key is that hypercontractivity need only be satisfied on an ε-cover of $\mathscr{F}_\star$. As discussed above, every finite hypothesis class (and thus every finite cover) is automatically $(C, 2)$-hypercontractive for some $C > 0$. The issue is to ensure that this C does not grow too fast as one refines the cover. The next result shows that the growth can be controlled for $\ell^2(\mathbb{N})$ ellipsoids of orthogonal expansions. By Mercer's theorem, these ellipsoids correspond to unit balls in various reproducing kernel Hilbert spaces (RKHS) [4, Corollary 12.26].

Proposition 5. *Fix positive constants β, B, K, and q. Fix a base measure λ on X and suppose that $\{\phi_n\}_{n\in\mathbb{N}_+}$ is an orthonormal system in $L^2(\lambda)$ satisfying $\|\phi_n\|_\infty \le Bm^q$, $\forall n \le m$. Set $\mu_j \le e^{-2\beta j}$ and define the ellipsoid:*

$$\mathscr{P} \triangleq \left\{ f = \sum_{j=1}^{\infty} \theta_j \phi_j \,\middle|\, \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} \le 1 \right\}.$$

Fix a $\varepsilon > 0$, and let m_e denote the smallest integer solution to $m - \frac{q}{\beta}\log m \ge \frac{1}{\beta}\left|\log\left(\frac{2B}{\beta\varepsilon}\right)\right|$. There exists an ε-cover $\mathscr{P}_\varepsilon \subset \mathscr{P}$ of $\mathscr{P}$ in the $\|\cdot\|_\infty$-norm satisfying:

$$\dim(\mathscr{P}_\varepsilon) \le m_\varepsilon, \;\; |\mathscr{P}_\varepsilon| \le \left(1 + \frac{4Bm_\varepsilon^q}{\varepsilon}\right)^{m_\varepsilon}.$$

Furthermore, let $\{\mu_t\}_{t=0}^{T-1}$ be the marginal distributions of P_X, and suppose that the μ_t's and λ are all mutually absolutely continuous. Assume that $\max_{0\le t\le T-1}\max\left\{\frac{d\mu_t}{d\lambda}, \frac{d\lambda}{d\mu_t}\right\} \le K$. Then, $(\mathscr{P}_\varepsilon, \mathsf{P}_X)$ is trajectory $(C_\varepsilon, 2)$-hypercontractive with $C_\varepsilon = K^3B^4m_\varepsilon^{4q+2}$.

Proposition 5 states that $(\mathscr{P}, \mathsf{P}_X)$ is $(C(\varepsilon), 2)$-hypercontractive where $C(\varepsilon) = C_\varepsilon$ only grows *poly-logarithmically* in $1/\varepsilon$. In particular, this example verifies the the assumptions of Corollary 2.

7 Conclusion

In a framework stretching far beyond contemporary results on this theme, we have established that the mixing-time of the covariates plays a relatively small role in the rate of convergence of the least-squares estimator. In many situations, after a finite burn-in time, this learning procedure exhibits an excess risk that scales as if all the samples were independent (Theorem 1). We have also provided numerous examples. Perhaps the most significant one can be found in Proposition 5: we show that learning with little mixing takes place with hypothesis classes as expressive as RKHS.

To arrive at Theorem 1, we leveraged insights from Mendelson [3] via a one-sided concentration inequality (Theorem 2). As mentioned in Section 4.1, hypercontractivity is closely related to the small-ball condition [3]. Such conditions can be understood as quantitative identifiability conditions by providing control of the "version space" (cf. Mendelson [3]). Given that identifiability conditions also play a key role in linear system identification—a setting in which a similar phenomenon as studied here has been reported—this suggests an interesting direction for future work: are such conditions actually necessary for learning with little mixing in the more general model (1)?

References

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

[2] Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. *arXiv preprint arXiv:1305.4825*, 2013.

[3] Shahar Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39. PMLR, 2014.

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

[5] Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994.

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

[7] Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. *Advances in Neural Information Processing Systems*, 21, 2008.

[8] Ingo Steinwart and Andreas Christmann. Fast learning from non-i.i.d. observations. *Advances in Neural Information Processing Systems*, 22, 2009.

[9] Bin Zou, Luoqing Li, and Zongben Xu. The generalization performance of erm algorithm with strongly mixing observations. *Machine Learning*, 75:275–295, 2009.

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

[11] John C. Duchi, Alekh Agarwal, Mikael Johansson, and Michael I. Jordan. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012.

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

[13] Ingvar Ziemann, Henrik Sandberg, and Nikolai Matni. Single trajectory nonparametric learning of nonlinear dynamics. *arXiv preprint arXiv:2202.08311*, 2022.

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

[15] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.

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

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

[18] Suhas Kowshik, Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. Near-optimal offline and streaming algorithms for learning non-linear dynamical systems. *Advances in Neural Information Processing Systems*, 34, 2021.

[19] Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

[20] Shahar Mendelson. On aggregation for heavy-tailed classes. *Probability Theory and Related Fields*, 168(3):641–674, 2017.

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

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

[23] Paul-Marie Samson. Concentration of measure inequalities for markov chains and ϕ-mixing processes. *The Annals of Probability*, 28(1):416–461, 2000.

[24] Yuval Dagan, Constantinos Daskalakis, Nishanth Dikkala, and Siddhartha Jayanti. Learning from weakly dependent data under dobrushin's condition. In *Conference on Learning Theory*, pages 914–928. PMLR, 2019.

[25] Anders Rantzer. Concentration bounds for single parameter adaptive control. In *2018 Annual American Control Conference (ACC)*, pages 1862–1866, 2018.

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

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

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

[29] Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.

[30] Søren F. Jarner and Gareth O. Roberts. Polynomial convergence rates of markov chains. *The Annals of Applied Probability*, 12(1):224–247, 2002.

[31] Vladimir M. Tikhomirov. *ϵ-Entropy and ϵ-Capacity of Sets In Functional Spaces*, pages 86–170. Springer Netherlands, 1993.

[32] Anthony Carbery and James Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical Research Letters*, 8:233–248, 2001.

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

[34] David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, 2008.

[35] Geoffrey Wolfer and Aryeh Kontorovich. Statistical estimation of ergodic markov chain kernel over discrete state space. *Bernoulli*, 27(1):532–553, 2021.

[36] Yu Cao, Jianfeng Lu, and Yulong Lu. Exponential decay of rényi divergence under fokker–planck equations. *Journal of Statistical Physics*, 176:1172–1184, 2019.

[37] Murat A. Erdogdu, Rasa Hosseinzadeh, and Matthew S. Zhang. Convergence of langevin monte carlo in chi-squared and rényi divergence. *arXiv preprint arXiv:2007.11612*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 4.1 and Section 4.3.
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our paper is a theoretical paper regarding the risk of the least-squares estimator on dependent data. We do not see any negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 4.1 for a discussion of the hypercontractivity and dependency matrix assumptions we use in our proofs. Furthermore, each result contains a full list of the necessary hypothesis so that the theorem holds.
   (b) Did you include complete proofs of all theoretical results? [Yes] See the appendices in the supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]