Learning with little mixing

Ingvar Ziemann
KTH Royal Institute of Technology
ziemann@kth.se

Stephen Tu
Robotics at Google
stephentu@google.com

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 when it occurs: bounded function classes for which the L^2 and $L^{2+\varepsilon}$ norms are equivalent, ergodic finite state Markov chains, various parametric models, and a broad family of infinite dimensional $\ell^2(\mathbb{N})$ ellipsoids. By instantiating our main result to system identification of nonlinear dynamics with generalized linear model transitions, we obtain a nearly minimax optimal excess risk bound after only a polynomial burn-in time.

1 Introduction

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

$$Y_t = f_\star(X_t) + W_t, \qquad t = 0, 1, 2, \ldots. \tag{1}$$

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 the model (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 processes that exhibit correlations which decay slower than geometrically. We refer to this double phenomenon, where the mixing-time only enters as a burn-in time, and where the mixing requirement is 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., [5–14] and the references within), where the blocking technique [5] is used to create independence amongst

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

the dependent covariates, so that tools to analyze independent learning can be applied. While these 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, the results of Ziemann et al. [15] studying the LSE on the model (1) also suffer from such sample degradation, but do not rely on the blocking technique.

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 [16–19]. Kowshik et al. [20], Gao and Raskutti [21] 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—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 4.1, 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.

Section 5 provides 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. Next, we show that many infinite dimensional function spaces based on $\ell^2(\mathbb{N})$ ellipsoids satisfy our hypercontractivity condition, demonstrating that our results are not inherently limited to finite-dimensional hypothesis classes.

To demonstrate the broad applicability of our framework, Section 6 instantiates our main result on two system identification problems that have received recent attention in the literature: linear dynamical systems (LDS), and systems with generalized linear model (GLM) transitions. For stable LDS, after a polynomial burn-in time, we recover an excess risk bound that matches the iid rate. A more general form of this result was recently established by Tu et al. [19]. For stable GLMs, also after a polynomial burn-in time, we obtain the first excess risk bound for this problem which matches the iid rate, up to logarithmic factors in various problem constants including the mixing-time. In both of these settings, our excess risk bounds also yield nearly optimal rates for parameter recovery, matching known results for LDS [16] and GLMs [20] in the stable case. In Appendix A, we show experimentally, using the stable GLM model, that the trends predicted by our theory are indeed realized in practice.

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. [16]. Mendelson [3] shows that for nonparametric iid regression, only minimal assumptions are required for one-sided isometry, and thus learning. Simchowitz et al. [16] 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, 22, 23] 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 [24] in the context of iid regression. Indeed, we show that a martingale analogue of the offset complexity (described in [15]) 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. [25] prove a lower bound for linear regression stating that in a worst case agnostic model, this deflation is unavoidable. Moreover, if the linear regression problem is realizable, Nagaraj et al. [25] provide upper and lower bounds showing that the mixing-time only affects the burn-in time, but not the final risk. We note that their upper bound is an algorithmic result that holds only for a specific modification of SGD. Our work can be interpreted as an upper bound in the more general nonparametric setting, where we put forth sufficient conditions to recover the iid rate after a burn-in time. Our result is algorithm agnostic and directly applies to the empirical risk minimizer. Ziemann et al. [15] 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. Roy et al. [13] extend the results from Mendelson [3] to the dependent data setting. While following Mendelson's argument allows their results to handle non-realizability and heavy-tailed noise, their proof ultimately still relies on two-sided concentration for both the "version space" and the "noise interaction". Hence, their rates end up degrading for slower mixing processes. We note that this is actually expected in the non-realizable setting in light of the lower bounds in Nagaraj et al. [25].

The measure of dependencies we use for the process $\{X_t\}$ is due to Samson [26]. Recently, Dagan et al. [27] 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 degrade as the measure of correlation increases.

Results in linear system identification show that lack of ergodicity does not degrade parameter recovery rates [16–19, 28, 29]. Beyond linear system identification, Kowshik et al. [20], Gao and Raskutti [21], Sattar and Oymak [30], Foster et al. [31] prove parameter recovery bounds for dynamical systems driven by a generalized linear model (GLM) transition. Most relevant are Kowshik et al. [20] and Gao and Raskutti [21], who again show that the lack of ergodicity does not hamper rates. Indeed, Gao and Raskutti [21] even manage to do so in a semiparametric setting with an unknown link function. As mentioned previously, our main result instantiated to these problems in the stable case matches existing excess risk and parameter recovery bounds for linear system identification, and actually provides the sharpest known excess risk bound for the GLM setting (when the link function is known). A more detailed comparison to existing LDS results is given in Appendix I.4, and to existing GLM results in Appendix J.1.

3 Problem formulation

The time-series (1) evolves on two subsets of Euclidean space, $\mathsf{X} \subset \mathbb{R}^{d_\mathsf{X}}$ and $\mathsf{Y} \subset \mathbb{R}^{d_\mathsf{Y}}$, 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, $\sigma_{\min}(M)$ the smallest non-zero singular value, and $\mathrm{cond}(M) = \|M\|_{\mathsf{op}}/\sigma_{\min}(M)$ the condition number. When the matrix M is symmetric, $\lambda_{\min}(M)$ will be used to denote its minimum eigenvalue.

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 $\mathbb{R}^{d_{\mathsf{X}}}$ to $\mathbb{R}^{d_{\mathsf{Y}}}$. 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] although we will see that this is not too restrictive. For any function $f : \mathsf{X} \to \mathbb{R}^{d_{\mathsf{Y}}}$, 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 $X_{0:T-1} \triangleq (X_0, \dots, X_{T-1})$ 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. For a radius $r > 0$, we let $B(r)$ denote the closed ball of $\mathscr{F}_\star$ with radius r in L^2, and we let $\partial B(r)$ denote its boundary: $B(r) \triangleq \left\{f \in \mathscr{F}_\star \,\middle|\, \|f\|_{L^2}^2 \le r^2\right\}$ and $\partial B(r) \triangleq \left\{f \in \mathscr{F}_\star \,\middle|\, \|f\|_{L^2}^2 = r^2\right\}$.

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 $\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2$.

4 Results

This section presents 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, 23].

Definition 4.1 (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*

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

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 5.2 (Section 5) shows that the condition holds for a nontrivial $\alpha = 1 + \varepsilon/2 \in (1, 2)$. More examples are given in Section 5.

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

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

the mixing-time. Our results instantiated for the $\alpha = 1$ case directly correspond to existing work by Ziemann et al. [15], 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 $\mathcal{A}$, the total-variation norm is defined as $\|\mu - \nu\|_{\mathsf{TV}} \triangleq \sup_{A \in \mathcal{A}} |\mu(A) - \nu(A)|$.

Definition 4.2 (Dependency matrix, Samson [26, Section 2])**.** *The* dependency matrix *of a process* $\{Z_t\}_{t=0}^{T-1}$ *with distribution* 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* $\mathcal{Z}_{0:i}$ *denote the* σ*-algebra generated by* $\{Z_t\}_{t=0}^{i}$*. For indices* $i < j$*, let*

$$\Gamma_{ij} = \sqrt{2 \sup_{A \in \mathcal{Z}_{0:i}} \|\mathsf{P}_{Z_{j:T-1}}(\cdot \mid A) - \mathsf{P}_{Z_{j:T-1}}\|_{\mathsf{TV}}}. \tag{4}$$

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

Given the dependency matrix from Definition 4.2, 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 $1 \leq \|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}} \lesssim T$. 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 $t \in \mathbb{N}$.

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 $\mathscr{F}$. If the process $\{X_t\}$ 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 [26, Section 2]. Other examples, such as processes satisfying Doeblin's condition [32], are given in Samson [26, Section 2]. When $\{X_t\}$ is a stationary time-homogenous Markov chain with invariant distribution π, the coefficients Γ_{ij} simplify to $\Gamma_{ij}^2 = 2 \sup_{A \in \mathcal{X}_\infty} \|\mathsf{P}_{X_{j-i}}(\cdot \mid A) - \pi\|_{\mathsf{TV}}$ for indices $j > i$, where $\mathcal{X}_\infty$ is the σ-algebra generated by $X_\infty \sim \pi$ (cf. Proposition F.1). Hence, the requirement $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2 \lesssim T^\beta$ for $\beta \in (0,1)$ then corresponds to $\sup_{A \in \mathcal{X}_\infty} \|\mathsf{P}_{X_t}(\cdot \mid A) - \pi\|_{\mathsf{TV}} \lesssim 1/t^{1-\beta}$ for $t \in \mathbb{N}_+$. Jarner and Roberts [33] give various examples and conditions to check polynomial convergence rates for Markov chains. We also provide further means to verify $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}} = O(1)$ in Appendix F and Appendix G.

4.2 Learning with little mixing

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

Definition 4.3 (Martingale offset complexity, cf. Liang et al. [24], Ziemann et al. [15])**.** *For the regression problem (1), the martingale offset complexity of a function space* $\mathscr{F}$ *is given by:*

$$\mathsf{M}_T(\mathscr{F}) \triangleq \sup_{f \in \mathscr{F}} \left\{ \frac{1}{T} \sum_{t=0}^{T-1} 4\langle W_t, f(X_t)\rangle - \|f(X_t)\|_2^2 \right\}. \tag{5}$$

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

Theorem 4.1. *Fix* $B > 0$*,* $C : (0, B] \to \mathbb{R}_+$*,* $\alpha \in [1, 2]$*, and* $r \in (0, B]$*. 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 (cf. Definition 4.1). Then:*

$$\mathbf{E}\|\widehat{f} - f_\star\|_{L_2}^2 \leq 8\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star) + r^2 + B^2 |\mathscr{F}_r| \exp\left(\frac{-T r^{4-2\alpha}}{8C(r)\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2} \right). \tag{6}$$

The assumption that $\mathscr{F}_\star$ is star-shaped in Theorem 4.1 is not particularly restrictive. Indeed, Theorem 4.1 still holds if we replace $\mathscr{F}_\star$ by its star-hull $\mathrm{star}(\mathscr{F}_\star) \triangleq \{\gamma f \mid \gamma \in [0,1],\ f \in \mathscr{F}_\star\}$, and $\partial B(r)$ with the boundary of the r-sphere of $\mathrm{star}(\mathscr{F}_\star)$. In this case, we note that (a) the metric entropy of $\mathrm{star}(\mathscr{F}_\star)$ is well controlled by the metric entropy of $\mathscr{F}_\star$,[2] and (b) the trajectory hypercontractivity conditions over a class $\mathscr{F}_\star$ and its star-hull $\mathrm{star}(\mathscr{F}_\star)$ are equivalent. Hence, at least whenever we are

[2] Specifically, $\log \mathcal{N}_\infty(\mathrm{star}(\mathscr{F}_\star), \varepsilon) \leq \log(2B/\varepsilon) + \log \mathcal{N}_\infty(\mathscr{F}_\star, \varepsilon/2)$ [34, Lemma 4.5].

able to verify hypercontractivity over the entire class $\mathscr{F}_\star$, little generality is lost. While most of our examples are star-shaped, we will need the observations above when we work with generalized linear model dynamics in Section 6.2.

To understand Theorem 4.1, 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,[3] a chaining argument detailed in Ziemann et al. [15, Lemma 4] shows that:

$$\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star) \lesssim \inf_{\gamma>0,\delta\in[0,\gamma]} \left\{ \frac{\sigma_W^2 \log \mathcal{N}_\infty(\mathscr{F}_\star,\gamma)}{T} + \sigma_W \sqrt{d_\mathsf{Y}}\delta + \frac{\sigma_W}{\sqrt{T}} \int_\delta^\gamma \sqrt{\log \mathcal{N}_\infty(\mathscr{F}_\star,s)}ds + \gamma^2 \right\}. \tag{7}$$

In particular, this bound only depends on $\mathscr{F}_\star$ and is *independent* of $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2$. Furthermore, (7) coincides with the corresponding risk bound for the LSE with iid covariates [24].

Given that $\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star)$ corresponds to the rate of learning from T iid covariates, the form of (6) suggests that we choose $r^2 \lesssim \mathbf{E}\mathsf{M}_T(\mathscr{F}_\star)$, so that the dominant term in (6) 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 (6) 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 in (6) that captures the interplay between the function class $\mathscr{F}_\star$ and the dependency measure $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}$. 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 that (b) $\mathscr{F}_\star$ is nonparametric, but not too large:

$$\exists\, p>0,\ q\in(0,2) \text{ s.t. } \log \mathcal{N}_\infty(\mathscr{F}_\star,\varepsilon) \le p\left(\frac{1}{\varepsilon}\right)^q \text{ for all } \varepsilon\in(0,1). \tag{8}$$

Covering numbers of the form (8) 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 $k\ge\lceil d_\mathsf{X}/2\rceil$ [35]. If condition (8) holds and the noise process $\{W_t\}$ is a sub-Gaussian MDS, inequality (7) 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 4.1. *Fix $B\ge 1$, $C>0$, $p>0$, $q\in(0,2)$, and $\gamma\in(0,\frac{q}{2+q})$. Suppose that $\mathscr{F}_\star$ is star-shaped, B-bounded, satisfies (8), and $(\mathscr{F}_\star,\mathsf{P}_X)$ satisfies the trajectory $(C,2)$-hypercontractivity condition. Suppose that T satisfies:*

$$T \ge \max\left\{ \left[8(32p+1)C\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2\right]^{\frac{1}{1-\frac{q}{2}\left(\frac{2}{2+q}+\gamma\right)}}, \left[4\log B \vee \frac{8}{q}\log\left(\frac{16}{q}\right)\right]^{\frac{1}{\frac{q}{2}\left(\frac{2}{2+q}+\gamma\right)}} \right\}. \tag{9}$$

Then, we have that:

$$\mathbf{E}\|\widehat{f}-f_\star\|_{L^2}^2 \le 8\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star) + 2T^{-\left(\frac{2}{2+q}+\gamma\right)}. \tag{10}$$

The rate (10) of Corollary 4.1 highlights the fact that the first order term of the excess risk is bounded by the martingale offset complexity $\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star)$. This behavior arises since the dependency matrix $\Gamma_{\mathsf{dep}}(\mathsf{P}_X)$ only appears as the burn-in requirement (9). Here, the value of q constrains how fast $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2$ is allowed to grow. In particular, condition (9) 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 $\gamma\in(0,\frac{q}{2+q})$.

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 hypercontractivity 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 $\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star) \le \tilde{O}(T^{-1})$, and hence we set $r^2 = o(T^{-1})$.

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

Corollary 4.2. *Fix $B \geq 1$, $C : (0, 1] \to \mathbb{R}_+$, $\alpha \in (1, 2]$, $b_1 \in [0, 1)$, $b_2 \in [0, 2)$, $\gamma \in (0, 1)$, 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 $\partial B(r)$ 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 $\psi := 1 - b_1 - \frac{(1+\gamma)(4-2\alpha+b_2)}{2} > 0$, 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 5.4 in Section 5 gives an example of an $\ell^2(\mathbb{N})$ ellipsoid which satisfies the assumptions in Corollary 4.2. Corollary 4.2 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, the $\psi > 0$ condition 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 $\psi > 0$. 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 $\psi > 0$ simplifies to $b_1 < \alpha - 1$. Compared to Corollary 4.1, 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 nonparametric 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. [15]– see Appendix H for a discussion.

5 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. Then 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 verified to be hypercontractive.

Proposition 5.1. *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 $\{\psi_1, \dots, \psi_K\} \subset \mathbb{R}^{d_{\mathsf{X}}}$, and that $\min_{0\leq t\leq T-1} \min_{1\leq k\leq K} \mu_t(\psi_k) \geq \underline{\mu}$. For any class of functions $\mathscr{F}$ mapping $\{\psi_1, \dots, \psi_K\} \to \mathbb{R}^{d_{\mathsf{Y}}}$, we have that $\mathscr{F}$ satisfies the trajectory $(1/\underline{\mu}, 2)$-hypercontractivity condition.*

We remark that when P_X is an aperiodic and irreducible Markov chain over a finite state space, the condition $\underline{\mu} > 0$ is always valid even as $T \to \infty$ [36]. In this case, our findings are related to Wolfer and Kontorovich [37, 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 then satisfies a nontrivial hypercontractivity constant, $\alpha > 1$ (cf. Mendelson [23]).

Proposition 5.2. *Fix $\varepsilon \in (0, 2]$ and $c > 0$. Suppose that $\mathscr{F}$ is B-bounded and that $\|f\|_{L^{2+\varepsilon}} \leq c\|f\|_{L^2}$ for all $f \in \mathscr{F}$. Then $\mathscr{F}$ is trajectory $(B^{2-\varepsilon}c^{2+\varepsilon}, 1 + \varepsilon/2)$-hypercontractive.*

Next, we show that for processes $\{X_t\}$ which converge fast enough to a stationary distribution, it suffices to verify the hypercontractivity condition only over the stationary distribution. This mimics existing results in iid learning, where hypercontractivity is assumed over the covariate distribution [3, 24]. 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 $\chi^2(\mu,\nu) \triangleq \mathbf{E}_\nu\left[\left(\frac{d\mu}{d\nu}-1\right)^2\right]$.

Proposition 5.3. *Fix positive r, C_{χ^2}, C_{TV}, and $C_{8\to 2}$. Suppose that the process $\{X_t\}$ has a stationary distribution π. Let $\{\mu_t\}$ denote the marginal distributions of $\{X_t\}$, and suppose that the marginals $\{\mu_t\}$ are absolutely continuous w.r.t. π. Assume the process is ergodic in the sense that:*

$$\sup_{t\in\mathbb{N}} \chi^2(\mu_t, \pi) \le C_{\chi^2}, \quad \frac{1}{T}\sum_{t=0}^{T-1} \|\mu_t - \pi\|_{\mathsf{TV}} \le C_{\mathsf{TV}} r^2. \tag{11}$$

Suppose also that for all $f \in \mathscr{F}_\star$: $\mathbf{E}_\pi\|f(X)\|_2^8 \le C_{8\to 2}(\mathbf{E}_\pi\|f(X)\|_2^2)^4$. Then the set $\partial B(r)$ satisfies $(C,2)$-trajectory hypercontractivity with $C = (1+\sqrt{C_{\chi^2}})\sqrt{C_{8\to 2}}(1+C_{\mathsf{TV}}B^2)^2$.

We further discuss the ergodicity condition (11) in Appendix E.3.1.

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 constant 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 reproducing kernel Hilbert spaces [4, Corollary 12.26].

Proposition 5.4. *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 $\|\phi_n\|_\infty \le Bn^q$, $\forall n \in \mathbb{N}$. Suppose $\mu_j \le e^{-2\beta j}$ 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 $\varepsilon > 0$, and let m_ε denote the smallest positive integer solution to $m \ge \frac{2}{\beta}\left|\log\left(\frac{8B}{\beta\varepsilon}\right)\right|$ subject to $\frac{m}{\log m} \ge \frac{q}{\beta}$. Let $P \subset \mathscr{P}$ be an arbitrary subset. There exists an ε-cover P_ε of P in the $\|\cdot\|_\infty$-norm satisfying $\log|P_\varepsilon| \le m_\varepsilon \log\left(1+\frac{8Bm_\varepsilon^q}{\varepsilon}\right)$. Further, let $\{\mu_t\}_{t=0}^{T-1}$ be the marginal distributions of P_X and suppose 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, as long as $\varepsilon \le \inf_{f\in P}\|f\|_{L^2(\mathsf{P}_X)}$, $(P_\varepsilon, \mathsf{P}_X)$ is trajectory $(C_\varepsilon, 2)$-hypercontractive with $C_\varepsilon = (1+7K^3B^4m_\varepsilon^{4q+2})$.*

Proposition 5.4 states that when $\mathscr{F}_\star \subseteq \mathscr{P}$, then $(\partial B(r), \mathsf{P}_X)$ is $(C(r), 2)$-hypercontractive where $C(r) = C_r$ only grows *poly-logarithmically* in $1/r$ and thus verifies the assumptions of Corollary 4.2.

6 System identification in parametric classes

To demonstrate the sharpness of our main result, we instantiate Theorem 4.1 on two parametric system identification problems which have received recent attention in the literature: linear dynamical systems (LDS) and generalized linear model (GLM) dynamics.

6.1 Linear dynamical systems

Consider the setting where the process $\{X_t\}_{t\ge 0}$ is described by a linear dynamical system:

$$X_{t+1} = A_\star X_t + HV_t, \;\; X_0 = HV_0, \;\; V_t \in \mathbb{R}^{d_{\mathsf{V}}}, \;\; V_t \sim N(0, I), \;\; V_t \perp V_{t'}\ \forall\, t \ne t'. \tag{12}$$

In this setting, the system identification problem is to recover the dynamics matrix $A_\star$ from $\{X_t\}_{t=0}^{T-1}$ evolving according to (12). We derive rates for recovering $A_\star$ by first deriving an excess risk bound on the least-squares estimator via Theorem 4.1, and then converting the risk bound to a parameter error bound. Since Theorem 4.1 relies on the process being ergodic, we consider the case when $A_\star$ is stable. We start by stating a few standard definitions.

Definition 6.1. *Fix a $k \in \{1,\dots,d_{\mathsf{X}}\}$. The pair (A, H) is k-step* controllable *if* $\operatorname{rank}\left(\begin{bmatrix} H & AH & A^2H & \dots & A^{k-1}H\end{bmatrix}\right) = d_{\mathsf{X}}$.

For $t \in \mathbb{N}$, let the t-step *controllability gramian* be defined as $\Gamma_t \triangleq \sum_{k=0}^{t} A^k H H^\mathsf{T} (A^k)^\mathsf{T}$. Since the noise in (12) serves as the "control" in this setting, the controllability gramian also coincides with the covariance at time t, i.e., $\mathbf{E}[X_t X_t^\mathsf{T}] = \Gamma_t$.

Definition 6.2. *Fix a $\tau \geq 1$ and $\rho \in (0,1)$. A matrix A is called (τ,ρ)-*stable *if for all $k \in \mathbb{N}$ we have $\|A^k\|_{\mathsf{op}} \leq \tau \rho^k$.*

With these definitions in place, we now state our result for linear dynamical system.

Theorem 6.1. *Suppose that the matrix $A_\star$ in (12) is (τ,ρ)-stable (cf. Definition 6.2), and that the pair $(A_\star, H)$ is κ-step controllable (cf. Definition 6.1). Suppose also that $\|A_\star\|_F \leq B$ for some $B \geq 1$. Consider the linear hypothesis class and true regression function:*

$$\mathscr{F} \triangleq \{f(x) = Ax \mid A \in \mathbb{R}^{d_\mathsf{X} \times d_\mathsf{X}},\ \ \|A\|_F \leq B\},\ \ f_\star(x) = A_\star x. \tag{13}$$

Suppose that model (1) follows the process described in (12) with $Y_t = X_{t+1}$. There exists T_0 such that the LSE with hypothesis class $\mathscr{F}$ achieves for all $T \geq T_0$:

$$\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \leq 8\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star) + \frac{4\|H\|_{\mathsf{op}}^2 d_\mathsf{X}^2}{T}. \tag{14}$$

Furthermore, T_0 satisfies for a universal positive constant c_0:

$$T_0 = c_0 \frac{\tau^4 \|H\|_{\mathsf{op}}^4 d_\mathsf{X}^2}{(1-\rho)^2 \lambda_{\min}(\Gamma_{\kappa-1})^2} \left[\kappa^2 \vee \frac{1}{(1-\rho)^2}\right] \operatorname{polylog}\left(B, d_\mathsf{X}, \tau, \|H\|_{\mathsf{op}}, \frac{1}{\lambda_{\min}(\Gamma_{\kappa-1})}, \frac{1}{1-\rho},\right). \tag{15}$$

Appendix I.4 contains a more detailed discussion about the results in Theorem 6.1. There, we argue that the term $\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star)$ in (14) is proportional to $\|H\|_{\mathsf{op}}^2 d_\mathsf{X}^2/T$ implying that the final rate is proportional to the minimax rate, i.e., $\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \lesssim \|H\|_{\mathsf{op}}^2 d_\mathsf{X}^2/T$.

6.2 Generalized linear models

We next consider the following non-linear dynamical system:

$$X_{t+1} = \sigma(A_\star X_t) + HV_t,\ \ X_0 = HV_0,\ \ V_t \in \mathbb{R}^{d_\mathsf{X}},\ \ V_t \sim N(0,I),\ \ V_t \perp V_{t'}\ \forall\, t \neq t'. \tag{16}$$

Here, $A_\star \in \mathbb{R}^{d_\mathsf{X} \times d_\mathsf{X}}$ is the dynamics matrix and $\sigma : \mathbb{R}^{d_\mathsf{X}} \to \mathbb{R}^{d_\mathsf{X}}$ is a coordinate wise link function. The notation σ will also be overloaded to refer to the individual coordinate function mapping $\mathbb{R} \to \mathbb{R}$. We study the system identification problem where the link function σ is assumed to be known, but the dynamics matrix $A_\star$ is unknown and to be recovered from $\{X_t\}_{t=0}^{T-1}$. We will apply Theorem 4.1 to derive a nearly optimal excess risk bound for the LSE on this problem in the stable case.

We start by stating a few assumptions that are again standard in the literature [20, 31].

Assumption 6.1. *Suppose that $A_\star$, H, and σ from the GLM process (16) satisfy:*

1. *(One-step controllability). The matrix $H \in \mathbb{R}^{d_\mathsf{X} \times d_\mathsf{X}}$ is full rank.*
2. *(Link function regularity). The link function $\sigma : \mathbb{R} \to \mathbb{R}$ is 1-Lipschitz, satisfies $\phi(0) = 0$, and there exists a $\zeta \in (0,1]$ such that $|\sigma(x) - \sigma(y)| \geq \zeta |x - y|$ for all $x, y \in \mathbb{R}$.*
3. *(Lyapunov stability). There exists a positive definite diagonal matrix $P_\star \in \mathbb{R}^{d_\mathsf{X} \times d_\mathsf{X}}$ satisfying $P_\star \succcurlyeq I$ and a $\rho \in (0,1)$ such that $A_\star^\mathsf{T} P_\star A_\star \preccurlyeq \rho P_\star$.*

With our assumptions in place, we are ready to instantiate our main result on the process (16).

Theorem 6.2. *Suppose the model (1) follows the process described in (16) with $Y_t = X_{t+1}$. Assume that the process (16) satisfies Assumption 6.1. Fix a $B \geq 1$, and suppose that $\|A_\star\|_F \leq B$. Consider the hypothesis class and true regression function:*

$$\mathscr{F} \triangleq \{f(x) = \sigma(Ax) \mid A \in \mathbb{R}^{d_\mathsf{X} \times d_\mathsf{X}},\ \ \|A\|_F \leq B\},\ \ f_\star(x) = \sigma(A_\star x). \tag{17}$$

There exists a T_0 and a universal positive constant c_0 such that the LSE with hypothesis class $\mathscr{F}$ achieves for all $T \geq T_0$:

$$\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \leq c_0 \frac{\|H\|_{\mathsf{op}}^2 d_\mathsf{X}^2}{T} \log\left(\max\left\{T, B, d_\mathsf{X}, \|P_\star\|_{\mathsf{op}}, \|H\|_{\mathsf{op}}, \frac{1}{1-\rho}\right\}\right). \tag{18}$$

Furthermore, for a universal constant $c_1 > 0$, we may choose T_0 that satisfies:

$$T_0 = c_1 \max\left(\frac{\|P_\star\|_{\mathsf{op}}^2 \mathrm{cond}(H)^4 d_{\mathsf{X}}^4}{\zeta^4(1-\rho)^6}, \frac{1}{\|H\|_{\mathsf{op}}^{1/3}}\right) \mathrm{polylog}\left(B, d_{\mathsf{X}}, \|P_\star\|_{\mathsf{op}}, \mathrm{cond}(H), \frac{1}{\zeta}, \frac{1}{1-\rho}\right). \tag{19}$$

Further discussion regarding Assumption 6.1 and Theorem 6.2, including a more detailed comparison with existing results, can be found in Appendix J.3.

7 Conclusion

We developed a framework for showing when 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 4.1). As a byproduct of our framework, by instantiating our results to system identification for dynamics with generalized linear model transitions (Section 6.2), we derived the sharpest known excess risk rate for this problem; our rates are nearly minimax optimal after only a polynomial burn-in time.

To arrive at Theorem 4.1, we leveraged insights from Mendelson [3] via a one-sided concentration inequality (Theorem B.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 had already been reported—this suggests an interesting direction for future work: are such conditions actually necessary for learning with little mixing?

Acknowledgements

We thank Dheeraj Nagaraj for helpful discussions regarding the results in Nagaraj et al. [25], and Abhishek Roy for clarifying the results in Roy et al. [13]. We also thank Mahdi Soltanolkotabi for an informative discussion about the computational aspects of empirical risk minimization for generalized linear models under the square loss.

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] Abhishek Roy, Krishnakumar Balasubramanian, and Murat A. Erdogdu. On empirical risk minimization with dependent and heavy-tailed data. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[14] Alessio Sancetta. Estimation in reproducing kernel hilbert spaces with dependent data. *IEEE Transactions on Information Theory*, 67(3):1782–1795, 2021.

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

[16] 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.

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

[18] 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.

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

[20] 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.

[21] Yue Gao and Garvesh Raskutti. Improved prediction and network estimation using the monotone single index multi-variate autoregressive model. *arXiv preprint arXiv:2106.14630*, 2021.

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

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

[24] 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.

[25] 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.

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

[27] 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.

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

[29] 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.

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

[31] 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.

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

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

[34] Shahar Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48(7):1977–1991, 2002.

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

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

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

[38] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

[39] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[40] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014.

[41] Jan R. Magnus. The moments of products of quadratic forms in normal variables. *Statistica Neerlandica*, 32(4):201–210, 1978.

[42] Minwoo Chae and Stephen G. Walker. Wasserstein upper bounds of the total variation for smooth densities. *Statistics & Probability Letters*, 163:108771, 2020.

[43] Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

[44] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011.

[45] Anastasios Tsiamis and George J. Pappas. Linear systems can be hard to learn. *arXiv preprint arXiv:2104.01120*, 2021.

[46] Randal Douc, Eric Moulines, and Jeffrey S. Rosenthal. Quantitative bounds on convergence of time-inhomogeneous markov chains. *The Annals of Applied Probability*, 14(4):1643–1665, 2004.

[47] Martin Hairer and Jonathan C. Mattingly. Yet another look at harris' ergodic theorem for markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, 2011.

[48] Duc N. Tran, Björn S. Rüffer, and Christopher M. Kellett. Incremental stability properties for discrete-time systems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 477–482, 2016.

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 6.
   (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)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (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)? [Yes]
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]

Contents

A Numerical experiments

We conduct a simple numerical simulation to illustrate the phenomenon of learning with little mixing empirically. We consider system identification of the GLM dynamics described in Section 6.2.

We first describe how the covariate process $\{X_t\}$ is generated. We set $d_{\mathsf{X}} = 25$. The true dynamic matrix $A_\star$ is randomly sampled from the distribution described in Section 7 of Kowshik et al. [20]. Specifically, $A_\star = U \Sigma U^\mathsf{T}$, where U is uniform from the Haar measure on the space of orthonormal $d_{\mathsf{X}} \times d_{\mathsf{X}}$ matrices, and $\Sigma = \mathrm{diag}(\underbrace{\rho, \dots, \rho}_{\lfloor d_{\mathsf{X}}/2 \rfloor \text{ times}}, \rho/3, \dots, \rho/3)$. We vary $\rho \in \{0.9, 0.99\}$ for this experiment. Next, we set the activation function σ to be the LeakyReLU with slope 0.5, i.e., $\sigma(x) = 0.5x\mathbf{1}\{x < 0\} + x\mathbf{1}\{x \geq 0\}$. Observe that these dynamics satisfy Assumption 6.1 with $\zeta = 0.5$, and where the Lyapunov matrix P can be taken to be identity, since $\|A_\star\|_{\mathsf{op}} = \rho < 1$. Next, we generate $X_0 \sim N(0, I_{d_{\mathsf{X}}})$, and $X_{t+1} = \sigma(A_\star X_t) + W_t$ with $W_t \sim N(0, 0.01 I_{d_{\mathsf{X}}})$ and $W_t \perp W_{t'}$ for $t \neq t'$. From this trajectory, the labelled dataset is $\{(X_t, Y_t)\}_{t=0}^{T-1}$ with $Y_t = X_{t+1}$.

To study the effects of the correlation from a single trajectory $\{X_t\}$ for learning, we consider the following *independent baseline* motivated by the Ind-Seq-LS baseline described in Tu et al. [19]. Let μ_t denote the marginal distribution of X_t. We sample $\bar{X}_t \sim \mu_t$ independently across time, and sample $\bar{Y}_t \mid \bar{X}_t$ from the conditional distribution $N(\sigma(A_\star \bar{X}_t), 0.01 I_{d_{\mathsf{X}}})$; the labelled dataset is $\{(\bar{X}_t, \bar{Y}_t)\}_{t=0}^{T-1}$. This ensures that the L^2 risk of a fixed hypothesis $f(x) = \sigma(Ax)$ is the same under both the independent baseline and the single trajectory distribution, so that our experiment singles out the effect of learning from correlated data. In practice, each $\bar{X}_t$ is sampled from a new independent rollout up to time t.

Given a dataset $\{(X_t, Y_t)\}_{t=0}^{T-1}$, we search for the empirical risk minimizer (ERM) of the loss

$$\hat{A} = \mathrm{argmin}_{A \in \mathbb{R}^{d_{\mathsf{X}} \times d_{\mathsf{X}}}} \left\{ \frac{1}{T} \sum_{t=0}^{T-1} \|\sigma(AX_t) - Y_t\|_2^2 \right\} \tag{20}$$

by running `scipy.optimize.minimize` with the `L-BFGS-B` method, using the default linesearch and termination criteria options. To calculate the L^2 excess risk $\frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|\sigma(\hat{A}X_t) - \sigma(A_\star X_t)\|_2^2$ of a hypothesis $\hat{A}$, we draw 1000 new trajectories and average the excess risk over these trajectories. The experimental code is implemented with `jax` [38], and run using the CPU backend with `float64` precision on a single machine.[4]

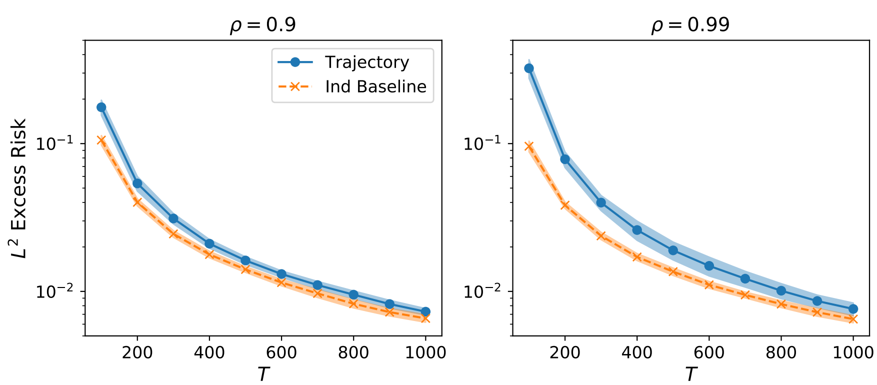

Figure 1: L^2 excess risk as a function of dataset length T of the empirical risk minimizer on the single trajectory (Trajectory) dataset versus the independent baseline (Ind Baseline) dataset.

The results of this experiment are shown in Figure 1 and Figure 2. In Figure 1, we plot the L^2 excess risk of the ERM $\hat{A}$ from (20) on both the trajectory dataset $\{(X_t, Y_t)\}$ and the independent baseline dataset $\{(\bar{X}_t, \bar{Y}_t)\}$, varying $\rho \in \{0.9, 0.99\}$. The shaded region indicates $\pm$ one standard deviation from the mean over 20 training datasets. In Figure 2, we plot the L^2 excess risk *ratio* of the estimator

[4] Code available at: https://github.com/google-research/google-research/tree/master/learning_with_little_mixing

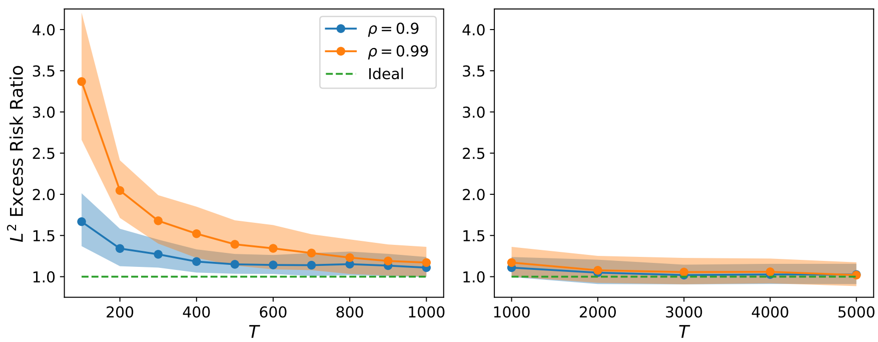

Figure 2: Ratio of the L^2 excess risk as a function of dataset length T of the empirical risk minimizer (ERM) on the single trajectory dataset over the ERM on the independent baseline dataset. The dashed green curve (Ideal) marks a ratio of exactly one.

$\hat{A}$ from the single trajectory dataset over the estimator $\hat{A}$ from the independent baseline trajectory, again varying $\rho \in \{0.9, 0.99\}$. Here, the shaded region is constructed using $\pm$ one standard deviation of the numerator and denominator taken over 20 training datasets.

Figure 1 and Figure 2 illustrate two different trends, which are both predicted by our theory. First, for a fixed ρ, as T increases, the L^2 excess risk of the ERM on the trajectory dataset approaches that of the ERM on the independent dataset. This illustrates the learning with little mixing phenomenon, where despite correlations in the covariates $\{X_t\}$ of the trajectory dataset across time, the statistical behavior of the ERM approaches that of the ERM on the independent dataset where the covariates $\{\bar{X}_t\}$ are independent across time. Next, for a fixed T, the burn-in time increases as ρ approaches one. That is, systems that mix slower have longer burn-in times.

B Proof techniques

In this section, we highlight the main ideas behind the proof of Theorem 4.1. 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. Recall that the closed ball of radius r is $B(r) = \left\{ f \in \mathscr{F}_\star \;\middle|\; \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2 \le r^2 \right\}$. We identify conditions which depend mildly on $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}$, so that with high probability we have:

$$\forall f \in \mathscr{F}_\star \setminus B(r), \;\; \mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right] \lesssim \frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2. \tag{21}$$

Once this *lower isometry* condition (21) holds, we bound the empirical excess risk, the RHS of inequality (21), by a version of the basic inequality of least squares [15, 24]. This leads to an upper bound of the empirical excess risk by the martingale complexity term (Definition 4.3). As our innovation mainly lies in establishing the lower isometry condition (21), 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 $\Gamma_{\mathsf{dep}}(\mathsf{P}_X)$.

Theorem B.1 (Samson [26, Theorem 2])**.** *Let $g : \mathsf{X} \to \mathbb{R}$ be non-negative. Then for any $\lambda > 0$:*

$$\mathbf{E}\exp\left(-\lambda\sum_{t=0}^{T-1} g(X_t)\right) \le \exp\left(-\lambda\sum_{t=0}^{T-1}\mathbf{E}g(X_t) + \frac{\lambda^2\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2 \sum_{t=0}^{T-1}\mathbf{E}g^2(X_t)}{2}\right). \tag{22}$$

We note that Samson [26, Theorem 2] actually proves a much stronger statement than Theorem B.1 (a Talagrand-style uniform concentration inequality), from which Theorem B.1 is a byproduct. With Theorem B.1 in hand, the following proposition allows us to relate hypercontractivity to lower isometry.

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

$$\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} g^2(X_t)\right] \leq C\left(\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} g(X_t)\right]\right)^{\alpha}. \tag{23}$$

Then we have:

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

Now fix an $f \in \mathscr{F}_\star \setminus B(r)$, and put $g(x) = \|f(x)\|_2^2$. Substituting g into (23) yields the trajectory hypercontractivity condition (3) in Definition 4.1. Thus, Proposition B.1 establishes the lower isometry condition (21) 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 the boundary $\partial B(r)$ since $\mathscr{F}_\star$ is star-shaped. Carrying out these details leads to the main lower isometry result.

Theorem B.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:*

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

B.1 Handling unbounded trajectories

Our main result Theorem 4.1 requires boundedness of both the hypothesis class $\mathscr{F}$ and the covariate process $\{X_t\}$ to hold. However, when this does not hold, Theorem 4.1 can often still be applied via a careful truncation argument. In this section, we outline the key ideas of this argument, with the full details given in Appendix G.

For concreteness, let us consider a Markovian process driven by Gaussian noise. Let $\{W_t\}_{t\geq 0}$ and $\{W_t'\}_{t\geq 0}$ be sequences of iid $N(0, I)$ vectors in $\mathbb{R}^{d_{\mathsf{X}}}$. Fix a dynamics function $f : \mathbb{R}^{d_{\mathsf{X}}} \to \mathbb{R}^{d_{\mathsf{X}}}$ and truncation radius $R > 0$. Define the truncated noise process $\{\bar{W}_t\}_{t\geq 0}$ as $\bar{W}_t \triangleq W_t' \mathbf{1}\{\|W_t'\|_2 \leq R\}$, and denote the original process and its truncated process by:

$$X_{t+1} = f(X_t) + HW_t, \;\; X_0 = HW_0, \qquad \text{(original process)} \tag{24a}$$

$$\bar{X}_{t+1} = f(\bar{X}_t) + H\bar{W}_t, \;\; \bar{X}_0 = H\bar{W}_0. \qquad \text{(truncated process)} \tag{24b}$$

Setting R appropriately, it is clear that the original process (24a) coincides with the truncated process (24b) with high probability by standard Gaussian concentration inequalities. Furthermore, the truncated noise process $\{H\bar{W}_t\}$ remains a martingale difference sequence due to the symmetry of the truncation. Additionally, since $\{H\bar{W}_t\}$ is bounded, if f is appropriately Lypaunov stable then the process $\{\bar{X}_t\}$ becomes bounded. In turn any class $\mathscr{F}$ containing continuous functions is bounded as well on (24b). Hence, the LSE $\hat{f}$ on (24a) can be controlled by the LSE $\bar{f}$ on (24b).

So far, this is a straightforward reduction. However, a subtle point arises in applying Theorem 4.1 to the LSE $\bar{f}$ on (24b): the dependency matrix Γ_{dep} now involves the truncated process (24b) instead of the original process (24a). This is actually *necessary* for this strategy to work, as the supremum in the dependency matrix coefficients (4) is now over the truncated process $\{\bar{X}_t\}$, instead of the original process $\{X_t\}$ which is unbounded. However, there is a trade-off, as bounding the coefficients for $\{\bar{X}_t\}$ is generally more complex than for $\{X_t\}$.[5] Nevertheless, a coupling argument allows us to switch back to bounding the dependency matrix coefficients for $\{X_t\}$, but crucially keep the supremum over the truncated $\{\bar{X}_t\}$. This reduction substantially broadens the scope of Theorem 4.1 without any modification to the proof.

[5]The clearest example of this is when the dynamics function f is linear: in this case, $\{X_t\}$ is jointly Gaussian (and hence (4) can be bounded by closed-form expressions), whereas $\{\bar{X}_t\}$ is not due to the truncation operator.

C Proof of Theorem 4.1

C.1 Proof of Lemma B.1

Let us abbreviate $\Gamma = \Gamma_{\mathsf{dep}}(\mathsf{P}_X)$. A Chernoff argument yields

$$\begin{aligned}
&\mathbf{P}\left(\sum_{t=0}^{T-1} g(X_t) \leq \frac{1}{2}\sum_{t=0}^{T-1} \mathbf{E} g(X_t)\right) \\
&\leq \inf_{\lambda\geq 0} \mathbf{E} \exp\left(\frac{\lambda}{2}\sum_{t=0}^{T-1} \mathbf{E} g(X_t) - \lambda \sum_{t=0}^{T-1} g(X_t)\right) && \text{(Chernoff)} \\
&\leq \inf_{\lambda\geq 0} \exp\left(-\frac{\lambda}{2}\sum_{t=0}^{T-1} \mathbf{E} g(X_t) + \frac{\lambda^2 \|\Gamma\|_{\mathsf{op}}^2 \sum_{t=0}^{T-1} \mathbf{E} g^2(X_t)}{2}\right) && \text{(Proposition B.1)} \\
&= \exp\left(-\frac{\left(\sum_{t=0}^{T-1} \mathbf{E} g(X_t)\right)^2}{8\|\Gamma\|_{\mathsf{op}}^2 \sum_{t=0}^{T-1} \mathbf{E} g^2(X_t)}\right) && \left(\lambda = \frac{\sum_{t=0}^{T-1} \mathbf{E} g(X_t)}{2\|\Gamma\|_{\mathsf{op}}^2 \sum_{t=0}^{T-1} \mathbf{E} g^2(X_t)}\right) \\
&\leq \exp\left(-\frac{T}{8C\|\Gamma\|_{\mathsf{op}}^2} \times \left(\frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E} g(X_t)\right)^{2-\alpha}\right), && \text{(Using (23))}
\end{aligned}$$

as per requirement. ■

C.2 Proof of Theorem B.2

The hypothesis that $\mathscr{F}_\star$ is star-shaped allows us to rescale, so it suffices to prove the result for $f \in \partial B(r)$. Namely, if $f \in \mathscr{F}_\star \setminus B(r)$ then $\frac{r}{\|f\|_{L^2}} < 1$ and so $rf/\|f\|_{L^2} \in \partial B(r)$ by the star-shaped hypothesis. Recall that $\mathscr{F}_r \subset \partial B(r)$ is a $r/\sqrt{8}$-net of $\partial B(r)$ in the supremum norm. Hence, by construction and parallellogram, for every $f \in \partial B(r)$, there exists $f_i \in \mathscr{F}_r$ such that:

$$\frac{1}{T}\sum_{t=0}^{T-1} \|f(X_t)\|_2^2 \geq \frac{1}{2T}\sum_{t=0}^{T-1} \|f_i(X_t)\|_2^2 - \frac{r^2}{8}. \tag{25}$$

Define the event:

$$\mathcal{E} \triangleq \bigcup_{f\in\mathscr{F}_r} \left\{\frac{1}{T}\sum_{t=0}^{T-1} \|f(X_t)\|_2^2 \leq \mathbf{E}\frac{1}{2T}\sum_{t=0}^{T-1} \|f(X_t)\|_2^2\right\}.$$

Invoking Lemma B.1 with $g(x) = \|f(x)\|_2^2$ for $f \in \mathscr{F}_r$, by a union bound it clear that

$$\mathbf{P}(\mathcal{E}) \leq |\mathscr{F}_r| \exp\left(\frac{-Tr^{4-2\alpha}}{8C\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2}\right). \tag{26}$$

Fix now arbitrary $f \in \partial B(r)$. On the complement $\mathcal{E}^c$ it is true that

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} \|f(X_t)\|_2^2 &\geq \frac{1}{2T}\sum_{t=0}^{T-1} \|f_i(x_t)\|_2^2 - \frac{r^2}{8} && \text{(we may find such an } f_i \text{ by observation (25))} \\
&\geq \mathbf{E}\frac{1}{4T}\sum_{t=0}^{T-1} \|f_i(X_t)\|_2^2 - \frac{r^2}{8} && \text{(by definition of } \mathcal{E}) \\
&= \frac{r^2}{4} - \frac{r^2}{8} && (f_i \in \partial B(r)) \\
&\geq \frac{r^2}{8}.
\end{aligned}$$

Since $f \in \partial B(r)$ was arbitrary, by virtue of the estimate (26) we have that:

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

The result follows by rescaling. ■

C.3 Proof of Theorem 4.1

Define the event:

$$\mathcal{B}_r \triangleq \left(\sup_{f\in\mathscr{F}_\star\setminus B(r)}\left\{\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2 - \mathbf{E}\frac{1}{8T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right\} \le 0\right).$$

By definition, on the complement of $\mathcal{B}_r$ we have that:

$$\|\widehat{f} - f_\star\|_{L^2}^2 \le r^2 \vee \frac{8}{T}\sum_{t=0}^{T-1}\|\widehat{f}(X_t) - f_\star(X_t)\|_2^2 \le r^2 + \frac{8}{T}\sum_{t=0}^{T-1}\|\widehat{f}(X_t) - f_\star(X_t)\|_2^2. \tag{27}$$

Therefore, we can decompose $\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2$ as follows:

$$\begin{aligned}\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 &= \mathbf{E}\mathbf{1}_{\mathcal{B}_r}\|\widehat{f} - f_\star\|_{L^2}^2 + \mathbf{E}\mathbf{1}_{\mathcal{B}_r^c}\|\widehat{f} - f_\star\|_{L^2}^2 \\ &\le B^2\mathbf{P}(\mathcal{B}_r) + r^2 + 8\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\widehat{f}(X_t) - f_\star(X_t)\|_2^2\right]. \quad (B\text{-bdd \& ineq. (27)})\end{aligned} \tag{28}$$

Theorem B.2 informs us that:

$$\mathbf{P}(\mathcal{B}_r) \le |\mathscr{F}_r| \exp\left(\frac{-Tr^{4-2\alpha}}{8C\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2}\right). \tag{29}$$

On the other hand, we have by the basic inequality (as in [24]):

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\widehat{f}(X_t) - f_\star(X_t)\|_2^2 \le \frac{1}{T}\sup_{f\in\mathscr{F}_\star}\sum_{t=0}^{T-1} 4\langle W_t, f(X_t)\rangle - \|f(X_t)\|_2^2 \tag{30}$$

Combining inequalities (28), (29) and (30) we conclude:

$$\begin{aligned}\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \le 8\mathbf{E}&\left[\sup_{f\in\mathscr{F}_\star}\frac{1}{T}\sum_{t=0}^{T-1} 4\langle W_t, f(X_t)\rangle - \|f(X_t)\|_2^2\right] \\ &+ r^2 + B^2|\mathscr{F}_r| \exp\left(\frac{-Tr^{4-2\alpha}}{8C\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2}\right),\end{aligned}$$

as per requirement. ■

D Proofs for corollaries in Section 4

D.1 Proof of Corollary 4.1

We set $r^2 = \frac{1}{T^{\frac{2}{2+q}+\gamma}}$. We first use Vershynin [39, Exercise 4.2.10] followed by (8) to bound:

$$\log\mathcal{N}_\infty(\partial B(r), r/\sqrt{8}) \le \log\mathcal{N}_\infty(\mathscr{F}_\star, r/(2\sqrt{8})) \le p\left(\frac{2\sqrt{8}}{r}\right)^q.$$

Therefore:

$$B^2\mathcal{N}_\infty(\partial B(r), r/\sqrt{8})\exp\left(\frac{-T}{8C\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2}\right) \le B^2\exp\left(32pT^{\frac{q}{2+q}+\frac{q\gamma}{2}} - \frac{T}{8C\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2}\right).$$

We want to solve for T such that:

$$B^2 \exp\left(32pT^{\frac{q}{2+q}+\frac{q\gamma}{2}} - \frac{T}{8C\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2}\right) \leq \frac{1}{T^{\frac{2}{2+q}+\gamma}}.$$

To do this, we first require that:

$$\begin{aligned}
32pT^{\frac{q}{2+q}+\frac{q\gamma}{2}} - \frac{T}{8C\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2} \leq -T^{\frac{q}{2+q}+\frac{q\gamma}{2}} &\iff T^{1-\left(\frac{q}{2+q}+\frac{q\gamma}{2}\right)} \geq 8(32p+1)C\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2 \\
&\iff T \geq \left[8(32p+1)C\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2\right]^{\frac{1}{1-\frac{q}{2}\left(\frac{2}{2+q}+\gamma\right)}}.
\end{aligned}$$

Now, with this requirement, we are left with the sufficient condition:

$$B^2 \exp\left(-T^{\frac{q}{2+q}+\frac{q\gamma}{2}}\right) \leq \frac{1}{T^{\frac{2}{2+q}+\gamma}} \iff T^{\frac{q}{2+q}+\frac{q\gamma}{2}} \geq \log(B^2 T^{\frac{2}{2+q}+\gamma}).$$

It suffices to require that:

$$\begin{aligned}
T^{\frac{q}{2+q}+\frac{q\gamma}{2}} &\geq 4\log B \iff T \geq (4\log B)^{\frac{1}{\frac{q}{2}\left(\frac{2}{2+q}+\gamma\right)}}, \\
T^{\frac{q}{2+q}+\frac{q\gamma}{2}} &\geq 2\log(T^{\frac{2}{2+q}+\gamma}) = \frac{4}{q}\log(T^{\frac{q}{2+q}+\frac{q\gamma}{2}}).
\end{aligned}$$

By Simchowitz et al. [16, Lemma A.4], the bottom inequality holds when:

$$T^{\frac{q}{2+q}+\frac{q\gamma}{2}} \geq \frac{8}{q}\log\left(\frac{16}{q}\right).$$

The claim now follows from Theorem 4.1. ■

D.2 Proof of Corollary 4.2

We set $r^2 = 1/T^{1+\gamma}$. By the given assumptions, we can construct a $r/\sqrt{8}$-net $\mathscr{F}_r$ of $\partial B(r)$ in the $\|\cdot\|_\infty$-norm that (a) satisfies

$$\log|\mathscr{F}_r| \leq p\log^q\left(\frac{\sqrt{8}}{r}\right),$$

and (b) satisfies the trajectory $(C(r/\sqrt{8}), \alpha)$-hypercontractivity condition. Recalling the bounds $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2 \leq T^{b_1}$ and $C(r) \leq (1/r)^{b_2}$, we have:

$$\begin{aligned}
B^2|\mathscr{F}_r|\exp\left(\frac{-Tr^{4-2\alpha}}{8C(r/\sqrt{8})\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2}\right) &\leq B^2\exp\left\{p\log^q\left(\frac{\sqrt{8}}{r}\right) - \frac{T^{1-b_1}r^{4-2\alpha+b_2}}{8^{1+b_2/2}}\right\} \\
&= B^2\exp\left\{p\log^q\left(\sqrt{8}T^{\frac{1+\gamma}{2}}\right) - \frac{T^{1-b_1-\frac{(1+\gamma)(4-2\alpha+b_2)}{2}}}{8^{1+b_2/2}}\right\} \\
&= B^2\exp\left\{p\log^q\left(\sqrt{8}T^{\frac{1+\gamma}{2}}\right) - \frac{T^{\psi}}{8^{1+b_2/2}}\right\} \\
&\leq B^2\exp\left\{p\log^q\left(\sqrt{8}T^{\frac{1+\gamma}{2}}\right) - \frac{T^{\psi}}{64}\right\}.
\end{aligned}$$

Above, the last inequality holds since $b_2 < 2$. Now, we choose T large enough so that:

$$\begin{aligned}
p\log^q\left(\sqrt{8}T^{\frac{1+\gamma}{2}}\right) - \frac{T^{\psi}}{64} \leq -\frac{T^{\psi}}{128} &\iff T^{\psi} \geq 128p\log^q\left(\sqrt{8}T^{\frac{1+\gamma}{2}}\right) \\
&\iff T^{\psi/q} \geq (128p)^{1/q}\left(\log(\sqrt{8}) + \frac{1+\gamma}{2\psi/q}\log(T^{\psi/q})\right).
\end{aligned}$$

Thus, it suffices to require that:

$$T^{\psi/q} \geq (128p)^{1/q}\log 8, \;\; T^{\psi/q} \geq (128p)^{1/q}\frac{1+\gamma}{\psi/q}\log(T^{\psi/q}).$$

By Simchowitz et al. [16, Lemma A.4], the right hand side inequality holds when:

$$T^{\psi/q} \geq 2(128p)^{1/q} \frac{1+\gamma}{\psi/q} \log\left(4(128p)^{1/q} \frac{1+\gamma}{\psi/q}\right).$$

We finish the proof by finding T such that:

$$\begin{aligned} B^2 \exp\left(\frac{-T^\psi}{128}\right) \leq \frac{1}{T^{1+\gamma}} &\Longleftrightarrow T^\psi \geq 128 \log(B^2 T^{1+\gamma}) \\ &\Longleftrightarrow T^\psi \geq 256 \log B + 128 \frac{(1+\gamma)}{\psi} \log(T^\psi). \end{aligned}$$

Thus, it suffices to require that:

$$T^\psi \geq 512 \log B, \;\; T^\psi \geq 256 \frac{(1+\gamma)}{\psi} \log(T^\psi).$$

Another application of Simchowitz et al. [16, Lemma A.4] yields that the latter inequality holds if:

$$T^\psi \geq 512 \frac{1+\gamma}{\psi} \log\left(1024 \frac{1+\gamma}{\psi}\right).$$

Combining all our requirements on T, we require that $T \geq \max\{T_1, T_2\}$, with:

$$\begin{aligned} T_1 &\triangleq \max\left\{(128p)^{1/\psi}(\log 8)^{q/\psi}, (128p)^{1/\psi}\left[\frac{4q}{\psi}\log\left((128p)^{1/q}\frac{8q}{\psi}\right)\right]^{q/\psi}\right\}, \\ T_2 &\triangleq \max\left\{(512\log B)^{1/\psi}, \left[\frac{1024}{\psi}\log\left(\frac{2056}{\psi}\right)\right]^{1/\psi}\right\}. \end{aligned}$$

■

E Proofs for Section 5

E.1 Proof of Proposition 5.1

For notational brevity, we make the identification of the atoms $\{\psi_1, \ldots, \psi_K\}$ with the integers $\{1, \ldots, K\}$. Fix a function $f : \{1, \ldots, K\} \to \mathbb{R}^{d_\mathsf{Y}}$. For any time indices $t_1, t_2 \in \{0, \ldots, T-1\}$:

$$\begin{aligned} \mathbf{E}\|f(X_{t_1})\|_2^2 \mathbf{E}\|f(X_{t_2})\|_2^2 &= \left(\sum_{k=1}^K \|f(k)\|_2^2 \mu_{t_1}(k)\right)\left(\sum_{k=1}^K \|f(k)\|_2^2 \mu_{t_2}(k)\right) \\ &= \sum_{k_1=1}^K \sum_{k_2=1}^K \|f(k_1)\|_2^2 \|f(k_2)\|_2^2 \mu_{t_1}(k_1)\mu_{t_2}(k_2) \geq \sum_{k_1=1}^K \|f(k_1)\|_2^4 \mu_{t_1}(k_1)\mu_{t_2}(k_1) \\ &\geq \underline{\mu} \sum_{k_1=1}^K \|f(k_1)\|_2^4 \mu_{t_1}(k_1) = \underline{\mu}\mathbf{E}\|f(X_{t_1})\|_2^4. \end{aligned}$$

Therefore:

$$\begin{aligned} \left(\frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2\right)^2 &= \frac{1}{T^2}\sum_{t_1=0}^{T-1}\sum_{t_2=0}^{T-1} \mathbf{E}\|f(X_{t_1})\|^2 \mathbf{E}\|f(X_{t_2})\|_2^2 \\ &\geq \frac{\underline{\mu}}{T^2}\sum_{t_1=0}^{T-1}\sum_{t_2=0}^{T-1} \mathbf{E}\|f(X_{t_1})\|_2^4 = \frac{\underline{\mu}}{T}\sum_{t_1=0}^{T-1} \mathbf{E}\|f(X_{t_1})\|_2^4. \end{aligned}$$

The claim now follows since we assume $\underline{\mu} > 0$. ■

E.2 Proof of Proposition 5.2

Recall that $\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^p\right] = \|f\|_{L^p}^p$ for $p \geq 1$. We estimate the left hand side of inequality (3) as follows:

$$\begin{aligned}
\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^4\right] &= \mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^{2-\varepsilon}\|f(X_t)\|_2^{2+\varepsilon}\right] \\
&\leq B^{2-\varepsilon}\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^{2+\varepsilon}\right] && (B\text{-bounded}) \\
&= B^{2-\varepsilon}\|f\|_{L^{2+\varepsilon}}^{2+\varepsilon} \\
&\leq B^{2-\varepsilon}(c\|f\|_{L^2})^{2+\varepsilon} && (L^2 - L^{2+\varepsilon}\text{-equivalence}) \\
&= B^{2-\varepsilon}c^{2+\varepsilon}\|f\|_{L^2}^{2+\varepsilon} \\
&= B^{2-\varepsilon}c^{2+\varepsilon}\left(\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right]\right)^{1+\varepsilon/2}.
\end{aligned}$$

The result now follows. ■

E.3 Proof of Proposition 5.3

We first state an auxiliary proposition.

Proposition E.1. *Let μ and ν be distributions satisfying $\mu \ll \nu$. Let g be any measurable function such that $\mathbf{E}_\nu g^2 < \infty$. We have:*

$$\mathbf{E}_\mu g - \mathbf{E}_\nu g \leq \sqrt{\mathbf{E}_\nu g^2}\sqrt{\chi^2(\mu,\nu)}.$$

Proof. By Cauchy-Schwarz:

$$\mathbf{E}_\mu g - \mathbf{E}_\nu g = \int g\left(\frac{d\mu}{d\nu} - 1\right)d\nu \leq \sqrt{\int g^2 d\nu}\sqrt{\int\left(\frac{d\mu}{d\nu} - 1\right)^2 d\nu} = \sqrt{\mathbf{E}_\nu g^2}\sqrt{\chi^2(\mu,\nu)}.$$

■

We can now complete the proof of Proposition 5.3. Fix any $f \in \mathscr{F}_\star$. First, we note that the the condition (5.3) implies:

$$\mathbf{E}_\pi\|f\|_2^4 \leq (\mathbf{E}_\pi\|f\|_2^8)^{1/2} \leq (C_{8\to 2}(\mathbf{E}_\pi\|f\|_2^2)^4)^{1/2} = \sqrt{C_{8\to 2}}(\mathbf{E}_\pi\|f\|_2^2)^2. \quad (31)$$

Therefore, for any $f \in \mathscr{F}_\star$ and any $t \in \mathbb{N}$:

$$\begin{aligned}
\mathbf{E}\|f(X_t)\|_2^4 &\leq \mathbf{E}_\pi\|f\|_2^4 + \sqrt{\mathbf{E}_\pi\|f\|_2^8}\sqrt{\chi^2(\mu_t,\pi)} && \text{using Proposition E.1} \\
&\leq (1 + \sqrt{C_{\chi^2}})\sqrt{C_{8\to 2}}(\mathbf{E}_\pi\|f\|_2^2)^2 && \text{using (11), (5.3), and (31).} \quad (32)
\end{aligned}$$

Now let $f \in \partial B(r)$. By Kuznetsov and Mohri [12, Lemma 1], since $\|f(x)\|_2^2 \in [0, B^2]$, we have:

$$\mathbf{E}_\pi\|f\|_2^2 - \mathbf{E}\|f(X_t)\|_2^2 \leq B^2\|\mu_t - \pi\|_{\mathsf{TV}}. \quad (33)$$

Therefore:

$$\begin{aligned}
\mathbf{E}_\pi\|f\|_2^2 &= \frac{1}{T}\sum_{t=0}^{T-1}(\mathbf{E}_\pi\|f\|_2^2 - \mathbf{E}\|f(X_t)\|_2^2) + r^2 && \text{since } f \in \partial B(r) \\
&\leq \frac{B^2}{T}\sum_{t=0}^{T-1}\|\mu_t - \pi\|_{\mathsf{TV}} + r^2 && \text{using (33)} \\
&\leq (1 + C_{\mathsf{TV}}B^2)r^2 && \text{using (11).} \quad (34)
\end{aligned}$$

Combining these inequalities:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|f(X_t)\|_2^4 &\leq (1+\sqrt{C_{\chi^2}})\sqrt{C_{8\to 2}}(\mathbf{E}_\pi\|f\|_2^2)^2 && \text{using (32)} \\
&\leq (1+\sqrt{C_{\chi^2}})\sqrt{C_{8\to 2}}(1+C_{\mathsf{TV}}B^2)^2 r^4 && \text{using (34)} \\
&= (1+\sqrt{C_{\chi^2}})\sqrt{C_{8\to 2}}(1+C_{\mathsf{TV}}B^2)^2\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|f(X_t)\|_2^2\right)^2 && \text{since } f\in\partial B(r).
\end{aligned}$$

The claim now follows. ■

E.3.1 Further discussion related to Proposition 5.3

Let us discuss the ergodicity conditions in Proposition 5.3. The condition $\sup_{t\in\mathbb{N}}\chi^2(\mu_t,\pi)<\infty$ from (11) is quite mild. To illustrate this point, suppose that $\{X_t\}$ are regularly spaced samples in time from the Itô stochastic differential equation:

$$dZ_t = f(Z_t)\,dt + \sqrt{2}\,dB_t,$$

where (B_t) is standard Brownian motion in $\mathbb{R}^{d_{\mathsf{X}}}$. Assume the process (Z_t) admits a stationary distribution π, and let ρ_t denote the measure of Z_t at time t. A standard calculation [40, Theorem 4.2.5] shows that $\frac{d}{dt}\chi^2(\rho_t,\pi) = -2\mathbf{E}_\pi\left\|\nabla\left(\frac{\rho_t}{\pi}\right)\right\|_2^2 \leq 0$, and hence $\sup_{t\geq 0}\chi^2(\rho_t,\pi)\leq\chi^2(\rho_0,\pi)$. Thus, as long as the initial measure ρ_0 has finite divergence with π, then this condition holds. One caveat is that $\chi^2(\rho_0,\pi)$ can scale as $e^{d_{\mathsf{X}}}$, resulting in a hypercontractivity constant that scales exponentially in dimension. This however only affects the burn-in time and not the final rate.

The second condition in (11) is $\frac{1}{T}\sum_{t=0}^{T-1}\|\mu_t-\pi\|_{\mathsf{TV}}\lesssim r^2$. A typical setting is $r^2\asymp 1/T^\beta$ for some $\beta\in(0,1]$, where β is dictated by the function class $\mathscr{F}$. Hence, this requirement reads:

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\mu_t-\pi\|_{\mathsf{TV}}\lesssim\frac{1}{T^\beta}\iff\sum_{t=0}^{T-1}\|\mu_t-\pi\|_{\mathsf{TV}}\lesssim T^{1-\beta}.$$

Therefore, the setting of β determines the level of ergodicty required. For example, if $\beta=1$ (which corresponds to the parametric function case), then this condition necessitates geometric ergodicity, since it requires that $\sum_{t=0}^{T-1}\|\mu_t-\pi\|_{\mathsf{TV}}=O(1)$. On the other hand, suppose that $\beta\in(0,1)$. Then this condition is satisfied if $\|\mu_t-\pi\|_{\mathsf{TV}}\lesssim 1/t^\beta$, allowing for slower mixing rates.

E.4 Proof of Proposition 5.4

Covering: We first approximate $\mathscr{P}$ by a finite-dimensional ellipsoid at resolution $\varepsilon/4$. To this end, fix an integer $m\in\mathbb{N}_+$ and define:

$$\mathscr{P}_m = \left\{ f=\sum_{j=1}^m\theta_j\phi_j \;\middle|\; \sum_{j=1}^\infty\frac{\theta_j^2}{\mu_j}\leq 1\right\}.$$

Fix now an element $f \in \mathscr{P}$ with coordinates θ. Let f' be the orthogonal projection onto the subspace of the first m-many coordinates ($f' \in \mathscr{P}_m$). Then:

$$\begin{aligned}
\|f - f'\|_\infty = \left\| \sum_{j=m+1}^{\infty} \theta_j \phi_j \right\|_\infty &\leq \left\| \underbrace{\sqrt{\sum_{j=m+1}^{\infty} \frac{\theta_j^2}{\mu_j}}}_{\leq 1} \sqrt{\sum_{j=m+1}^{\infty} \mu_j \|\phi_j\|_2^2} \right\|_\infty && \text{(Cauchy-Schwarz)} \\
&\leq B \sqrt{\sum_{j=m+1}^{\infty} j^{2q} e^{-2\beta j}} && (\|\phi_j\|_\infty \leq B j^q, \mu_j \leq e^{-2\beta j}) \\
&\leq B \sqrt{\sum_{j=m+1}^{\infty} e^{-\beta j}} && \left(\text{if } \frac{m}{\log m} \geq \frac{q}{\beta}\right) \\
&= B \frac{e^{-\beta m/2}}{\sqrt{e^\beta - 1}} \leq 2B \frac{e^{-\beta m/2}}{\beta}. && (\sqrt{e^{2x} - 1} \geq e^x - 1 \geq x, x \geq 0)
\end{aligned} \tag{35}$$

Hence, we can take m_ε to be the smallest integer solution to $m \geq \frac{2}{\beta} \left| \log \left(\frac{8B}{\beta \varepsilon} \right) \right|$ to guarantee that for every $f \in \mathscr{P}$ there exists $f' \in \mathscr{P}_m$ at most $\varepsilon/4$ removed from f, i.e., $\|f - f'\|_\infty \leq \varepsilon/4$.

Next, we construct an $\varepsilon/4$-covering of the set $\mathscr{P}_m$. Observe now that the set of parameters of Θ_m defining $\mathscr{P}_m$ satisfies:

$$\Theta_m \triangleq \left\{ \theta \in \mathbb{R}^m \;\middle|\; \sum_{j=1}^{m} \frac{\theta_j^2}{\mu_j} \leq 1 \right\}.$$

Using this, we obtain a covering of $(\mathscr{P}_m, \|\cdot\|_\infty)$ by regarding it as a subset of $\mathbb{R}^m$. More precisely, Θ_m is the unit ball in the norm $\|\theta\|_\mu \triangleq \sqrt{\sum_{i=1}^m \theta_i^2/\mu_i}, \theta \in \mathbb{R}^m$. Hence, by a standard volumetric argument, we need no more than $(1 + 2/\delta)^m$ points to cover Θ_m at resolution δ in $\|\cdot\|_\mu$. Let now $\delta > 0$ and choose $N \in \mathbb{N}_+$ so that $\{\theta^1, \ldots, \theta^N\}$ is an optimal δ-covering of Θ_m. We thus obtain the cover $\mathscr{P}_m^N \triangleq \{(\theta^1)^\top \phi(\cdot), \ldots, (\theta^N)^\top \phi(\cdot)\} \subset \mathscr{P}_m$ where $\phi(\cdot) = (\phi_1(\cdot), \ldots, \phi_m(\cdot))$. Let $f' = (\theta')^\top \phi \in \mathscr{P}_m$ be arbitrary. It remains to verify the resolution of $\mathscr{P}_m^N$:

$$\begin{aligned}
\min_{n \in [N]} \|f' - (\theta^n)^\top \phi\|_\infty &= \min_{n \in [N]} \left\| \sum_{j=1}^{m} (\theta'_j - \theta_j^n) \phi_j \right\|_\infty \\
&\leq \min_{n \in [N]} \left\| \sqrt{\sum_{j=1}^{m} \frac{(\theta'_j - \theta_j^n)^2}{\mu_j}} \sqrt{\sum_{j=1}^{m} \mu_j \|\phi_j\|_2^2} \right\|_\infty && \text{(Cauchy-Schwarz)} \\
&\leq \delta B m^q && (\|\phi_j\|_\infty \leq B m^q \text{ if } j \leq m).
\end{aligned}$$

Hence, if we take N large enough so that $\delta \leq \frac{\varepsilon}{4Bm^q}$, $\mathscr{P}_m^N$ is a cover of $\mathscr{P}_m$ at resolution $\varepsilon/4$. Hence, since we may take $m \leq m_\varepsilon$:

$$N \leq \left(1 + \frac{8Bm_\varepsilon^q}{\varepsilon}\right)^{m_\varepsilon}.$$

Now, we can immediately convert the covering $\mathscr{P}_m^N$ into an *exterior cover*[6] of the set P. For every $f \in P$, by the approximation property of $\mathscr{P}_m$, there exists an $f' \in \mathscr{P}_m$ such that $\|f - f'\|_\infty \leq \varepsilon/4$. But since $f' \in \mathscr{P}_m$, there exists an $f'' \in \mathscr{P}_m^N$ such that $\|f' - f''\|_\infty \leq \varepsilon/4$. By triangle inequality, $\|f - f''\|_\infty \leq \varepsilon/2$. Thus, $\mathscr{P}_m^N$ forms an exterior cover of P at resolution $\varepsilon/2$. By Vershynin [39, Exercise 4.2.9], this means that there exists a (proper) cover of P at resolution ε with cardinality bounded by N.

[6] An exterior cover of a set T is a cover where the elements are not restricted to T.

Hypercontractivity: We first show that every $f \in \mathscr{P}_m$ is hypercontractive. First, observe by orthogonality that the second moment takes the form:

$$\int \left\| \sum_{i=1}^{m_\varepsilon} \theta_i \phi_i \right\|_2^2 d\lambda = \sum_{i=1}^{m_\varepsilon} \theta_i^2 \|\phi_i\|_{L_2(\lambda)}^2 = \|\theta\|_2^2.$$

On the other hand by the eigenfunction growth condition:

$$\begin{aligned}
\int \left\| \sum_{i=1}^{m_\varepsilon} \theta_i \phi_i \right\|_2^4 d\lambda &\leq \int \left(\sum_{i=1}^{m_\varepsilon} |\theta_i| \|\phi_i\|_2 \right)^4 d\lambda \\
&\leq B^4 m_\varepsilon^{4q} \left(\sum_{i=1}^{m_e} |\theta_i| \right)^4 \\
&\leq B^4 m_\varepsilon^{4q+2} \|\theta\|_2^4 \\
&= B^4 m_\varepsilon^{4q+2} \left(\int \left\| \sum_{i=1}^{m_\varepsilon} \theta_i \phi_i \right\|_2^2 d\lambda \right)^2.
\end{aligned}$$

Now, for any $t \in \mathbb{N}$, by a change of measure, with $f = \sum_{i=1}^{m_\varepsilon} \theta_i \phi_i$,

$$\mathbf{E}_{\mu_t} \|f\|_2^4 = \int \|f\|_2^4 \frac{d\mu_t}{d\lambda} d\lambda \leq K \int \|f\|_2^4 \, d\lambda \leq K B^4 m_\varepsilon^{4q+2} \left(\int \|f\|_2^2 \, d\lambda \right)^2.$$

Hence, applying the previous inequality and another change of measure:

$$\begin{aligned}
\frac{1}{T} \sum_{t=0}^{T-1} \mathbf{E}_{\mu_t} \|f\|_2^4 &\leq K B^4 m_\varepsilon^{4q+2} \left(\int \|f\|_2^2 \, d\lambda \right)^2 \\
&= K B^4 m_\varepsilon^{4q+2} \left(\frac{1}{T} \int \sum_{t=0}^{T-1} \|f\|_2^2 \frac{d\lambda}{d\mu_t} d\mu_t \right)^2 \\
&\leq K^3 B^4 m_\varepsilon^{4q+2} \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathbf{E}_{\mu_t} \|f\|_2^2 \right)^2.
\end{aligned} \tag{36}$$

Next, fix a $f \in P$. We will show that f is hypercontractive. First, recall that f' is the element in $\mathscr{P}_m$ satisfying $\|f - f'\|_\infty \leq \varepsilon/4$. Hence, we have for every x:

$$\|f(x)\|_2^4 \leq 8(\|f(x) - f'(x)\|_2^4 + \|f'(x)\|_2^4) \leq \frac{\varepsilon^4}{32} + 8\|f'(x)\|_2^4, \tag{37}$$

$$\|f'(x)\|_2^2 \leq 2(\|f(x) - f'(x)\|_2^2 + \|f(x)\|_2^2) \leq \frac{\varepsilon^2}{2} + 2\|f(x)\|_2^2. \tag{38}$$

We now bound:

$$\begin{aligned}
\frac{1}{T} \sum_{t=0}^{T-1} \mathbf{E}_{\mu_t} \|f\|_2^4 &\overset{(a)}{\leq} \frac{\varepsilon^4}{32} + \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{E}_{\mu_t} \|f'\|_2^4 \overset{(b)}{\leq} \frac{\varepsilon^4}{32} + K^3 B^4 m_\varepsilon^{4q+2} \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathbf{E}_{\mu_t} \|f'\|_2^2 \right)^2 \\
&\overset{(c)}{\leq} \frac{\varepsilon^4}{32} + K^3 B^4 m_\varepsilon^{4q+2} \left(\frac{\varepsilon^2}{2} + \frac{2}{T} \sum_{t=0}^{T-1} \mathbf{E}_{\mu_t} \|f\|_2^2 \right)^2 \\
&\overset{(d)}{\leq} \left(\frac{1}{32} + \frac{25}{4} K^3 B^4 m_\varepsilon^{4q+2} \right) \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathbf{E}_{\mu_t} \|f\|_2^2 \right)^2.
\end{aligned}$$

Above, (a) uses the inequality (37), (b) uses the fact that $f' \in \mathscr{P}_m$ and (36), (c) uses (38), and (d) uses the assumption that $\varepsilon \leq \inf_{f \in P} \|f\|_{L^2(\mathsf{P}_X)}$, which implies that $\varepsilon^2 \leq \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{E}_{\mu_t} \|f\|_2^2$ and $\varepsilon^4 \leq \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathbf{E}_{\mu_t} \|f\|_2^2 \right)^2$. Since $f \in P$ is arbitrary, the claim follows. ■

F Basic tools for analyzing the dependency matrix

In this section, we outline some basic tools used to analyze the dependency matrix $\Gamma_{\mathsf{dep}}(\mathsf{P}_X)$. We will introduce the following shorthand. Given a process $\{Z_t\}_{t\geq 0}$ and indices $0 \leq i \leq j \leq k$, we will write $\mathsf{P}_{Z_{j:k}}(\cdot \mid Z_{0:i} = z_{0:i})$ as shorthand for $\mathsf{P}_{Z_{j:k}}(\cdot \mid A)$ for $A \in \mathcal{Z}_{0:i}$, where we recall that $\mathcal{Z}_{0:i}$ denotes the σ-algebra generated by $Z_{0:i}$. We will also write $\operatorname{ess\,sup}_{z_{0:i} \in \mathsf{Z}_{0:i}}$ as shorthand for $\sup_{A \in \mathcal{Z}_{0:i}}$.

Before we proceed, we recall the coupling representation of the total-variation norm:

$$\|\mu - \nu\|_{\mathsf{TV}} = \inf\{\mathbf{P}(X \neq Y) \mid (X, Y) \text{ is a coupling of } (\mu, \nu)\}. \tag{39}$$

Proposition F.1. *Suppose that $\{Z_t\}_{t\geq 0}$ is a Markov chain. For any integers $0 \leq i \leq j \leq k$:*

$$\|\mathsf{P}_{Z_{j:k}}(\cdot \mid Z_i = z) - \mathsf{P}_{Z_{j:k}}\|_{\mathsf{TV}} = \|\mathsf{P}_{Z_j}(\cdot \mid Z_i = z) - \mathsf{P}_{Z_j}\|_{\mathsf{TV}}.$$

Proof. Let us first prove the upper bound. Let (Z_j, Z_j') be a coupling of $(\mathsf{P}_{Z_j}(\cdot \mid Z_i = z), \mathsf{P}_{Z_j})$. We can construct a coupling $(\bar{Z}_{j:k}, \bar{Z}'_{j:k})$ of $(\mathsf{P}_{Z_{j:k}}(\cdot \mid Z_i = z), \mathsf{P}_{Z_{j:k}})$ by first setting $\bar{Z}_j = Z_j$, $\bar{Z}'_j = Z'_j$, and then evolving the chains onward via the following process. If $\bar{Z}_j = \bar{Z}'_j$, we evolve $\bar{Z}_{j+1:k}$ onwards according to the dynamics, and copy $\bar{Z}'_{j+1:k} = \bar{Z}_{j+1:k}$. Otherwise if $\bar{Z}_j \neq \bar{Z}'_j$, then we evolve both chains separately. Observe that $\bar{Z}_{j:k} \neq \bar{Z}'_{j:k}$ iff $Z_j \neq Z'_j$. Hence $\|\mathsf{P}_{Z_{j:k}}(\cdot \mid Z_i = x) - \mathsf{P}_{Z_{j:k}}\|_{\mathsf{TV}} \leq \mathbf{P}(Z_j \neq Z'_j)$. Since the coupling (Z_j, Z'_j) is arbitrary, taking the infimum over all couplings of $(\mathsf{P}_{Z_j}(\cdot \mid Z_i = z), \mathsf{P}_{Z_j})$ yields the upper bound via (39).

We now turn to the lower bound. Let $(Z_{j:k}, Z'_{j:k})$ be a coupling of $(\mathsf{P}_{Z_{j:k}}(\cdot \mid Z_i = z), \mathsf{P}_{Z_{j:k}})$. Since projection (Z_j, Z'_j) is a coupling for $(\mathsf{P}_{Z_j}(\cdot \mid Z_i = z), \mathsf{P}_{Z_j})$, and $Z_j \neq Z'_j$ implies $Z_{j:k} \neq Z'_{j:k}$, we have again by (39):

$$\|\mathsf{P}_{Z_j}(\cdot \mid Z_i = x) - \mathsf{P}_{Z_j}\|_{\mathsf{TV}} \leq \mathbf{P}(Z_j \neq Z'_j) \leq \mathbf{P}(Z_{j:k} \neq Z'_{j:k}).$$

Taking the infimum over all couplings of $(\mathsf{P}_{Z_{j:k}}(\cdot \mid Z_i = z), \mathsf{P}_{Z_{j:k}})$ yields the lower bound. ■

Proposition F.2. *Let M, N be two size conforming matrices with all non-negative entries. Suppose that $M \leq N$, where the inequality holds elementwise. Then, $\|M\|_{\mathsf{op}} \leq \|N\|_{\mathsf{op}}$.*

Proof. Let Q be a matrix with non-negative entries, and let q_i denote the rows of Q. The variational form of the operator norm states that $\|Q\|_{\mathsf{op}} = \sup_{\|v\|_2 \leq 1} \|Qv\|_2 = \sup_{\|v\|_2 \leq 1} \sqrt{\sum_i \langle q_i, v\rangle^2}$. Since each q_i only has non-negative entries. The supremum must be attained by a vector v with non-negative entries, otherwise flipping the sign of the negative entries in v would only possibly increase the value of $\|Qv\|_2$, and never decrease the value.

Now let m_i, n_i denote the rows of M, N, and let v be a vector with non-negative entries. Since $0 \leq m_i \leq n_i$ (elementwise), it is clear that $\langle m_i, v\rangle^2 \leq \langle n_i, v\rangle^2$. Hence the claim follows. ■

Proposition F.3. *Let $a_1, \ldots, a_n \in \mathbb{R}$, and let $M \in \mathbb{R}^{n\times n}$ be the upper triangular Toeplitz matrix:*

$$M = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 & \cdots & a_n \\ 0 & a_1 & a_2 & a_3 & \cdots & a_{n-1} \\ 0 & 0 & a_1 & a_2 & \cdots & a_{n-2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \vdots & a_1 \end{bmatrix}.$$

We have that:

$$\|M\|_{\mathsf{op}} \leq \sum_{i=1}^{n} |a_i|.$$

Proof. Let E_i, for $i = 1, \ldots, n$, denote the shift matrix where E_i has ones along the $(i-1)$-th super diagonal and is zero everywhere else (the zero-th diagonal refers to the main diagonal). It is not hard to see that $\|E_i\|_{\mathsf{op}} \leq 1$ for all i, since it simply selects (and shifts) a subset of the coordinates of the input. With this notation, $M = \sum_{i=1}^n a_i E_i$. The claim now follows by the triangle inequality. ■

Proposition F.4. *Let $\{Z_t\}_{t\geq 0}$ be a Markov process, and let P_Z denote the joint distribution of $\{Z_t\}_{t=0}^{T-1}$. We have that:*

$$\|\Gamma_{\mathsf{dep}}(\mathsf{P}_Z)\|_{\mathsf{op}} \leq 1 + \sqrt{2} \sum_{k=1}^{T-1} \max_{t=0,\ldots,T-1-k} \operatorname*{ess\,sup}_{z \in \mathsf{Z}_t} \sqrt{\|\mathsf{P}_{Z_{t+k}}(\cdot \mid Z_t = z) - \mathsf{P}_{Z_{t+k}}\|_{\mathsf{TV}}}.$$

Proof. For any indices $0 \leq i < j$, by the Markov property and Proposition F.1:

$$\begin{aligned}
\operatorname*{ess\,sup}_{z_{0:i} \in \mathsf{Z}_{0:i}} \|\mathsf{P}_{Z_{j:T-1}}(\cdot \mid Z_{0:i} = z_{0:i}) - \mathsf{P}_{Z_{j:T-1}}\|_{\mathsf{TV}} &= \operatorname*{ess\,sup}_{z \in \mathsf{Z}_i} \|\mathsf{P}_{Z_{j:T-1}}(\cdot \mid Z_i = z) - \mathsf{P}_{Z_{j:T-1}}\|_{\mathsf{TV}} \\
&= \operatorname*{ess\,sup}_{z \in \mathsf{Z}_i} \|\mathsf{P}_{Z_j}(\cdot \mid Z_i = z) - \mathsf{P}_{Z_j}\|_{\mathsf{TV}}.
\end{aligned}$$

Therefore:

$$\begin{aligned}
\Gamma_{\mathsf{dep}}(\mathsf{P}_Z)_{ij} &= \sqrt{2} \operatorname*{ess\,sup}_{z \in \mathsf{Z}_i} \sqrt{\|\mathsf{P}_{Z_j}(\cdot \mid Z_i = z) - \mathsf{P}_{Z_j}\|_{\mathsf{TV}}} \\
&\leq \sqrt{2} \max_{t=0,\ldots,T-1-(j-i)} \operatorname*{ess\,sup}_{z \in \mathsf{Z}_t} \sqrt{\|\mathsf{P}_{Z_{t+j-i}}(\cdot \mid Z_t = z) - \mathsf{P}_{Z_{t+j-i}}\|_{\mathsf{TV}}} \triangleq a_{j-i}.
\end{aligned}$$

Thus, we can construct a matrix Γ' such that for all indices $0 \leq i < j$, we have $\Gamma'_{ij} = a_{j-i}$, and the other entries are identical to $\Gamma_{\mathsf{dep}}(\mathsf{P}_Z)$. This gives us the entry-wise bound $\Gamma_{\mathsf{dep}}(\mathsf{P}_Z) \leq \Gamma'$. Applying Proposition F.2 and Proposition F.3, we conclude $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_Z)\|_{\mathsf{op}} \leq \|\Gamma'\|_{\mathsf{op}} \leq 1 + \sum_{k=1}^{T-1} a_k$. ■

G Mixing properties of truncated Gaussian processes

We first recall the notation from Appendix B.1. Let $\{W_t\}_{t\geq 0}, \{W'_t\}_{t\geq 0}$ be sequences of iid $N(0, I)$ vectors in $\mathbb{R}^{d_{\mathsf{X}}}$. Fix a dynamics function $f : \mathbb{R}^{d_{\mathsf{X}}} \to \mathbb{R}^{d_{\mathsf{X}}}$ and radius $R > 0$. Define the truncated Gaussian noise process $\{\bar{W}_t\}_{t\geq 0}$ as $\bar{W}_t \triangleq W'_t \mathbf{1}\{\|W'_t\|_2 \leq R\}$. Now, consider the two processes:

$$X_{t+1} = f(X_t) + HW_t, \;\; X_0 = HW_0, \tag{40a}$$

$$\bar{X}_{t+1} = f(\bar{X}_t) + H\bar{W}_t, \;\; \bar{X}_0 = H\bar{W}_0. \tag{40b}$$

We develop the necessary arguments in this section to transfer mixing properties of the original process (40a) to the truncated process (40b). This will let us apply our results in Section 4 to unbounded processes of the form (40a), by studying their truncated counterparts (40b).

The main tool to do this is the following coupling argument.

Proposition G.1. *Fix a $\delta \in (0, 1)$. Let $k \in \{1, \ldots, T-1\}$ and $t \in \{0, \ldots, T-1-k\}$. Consider the processes $\{X_t\}_{t\geq 0}$ and $\{\bar{X}_t\}_{t\geq 0}$ described in (40a) and (40b) with R satisfying the inequality $R \geq \sqrt{d_{\mathsf{X}}} + \sqrt{2\log(T/\delta)}$. The following bound hold for any $x \in \mathbb{R}^{d_{\mathsf{X}}}$:*

$$\|\mathsf{P}_{X_{t+k}}(\cdot \mid X_t = x) - \mathsf{P}_{\bar{X}_{t+k}}(\cdot \mid \bar{X}_t = x)\|_{\mathsf{TV}} \leq \delta.$$

The following bound also holds for any $t \in \{0, \ldots, T-1\}$:

$$\|\mathsf{P}_{X_t} - \mathsf{P}_{\bar{X}_t}\|_{\mathsf{TV}} \leq \delta.$$

Proof. Let (Z_{t+k}, Z'_{t+k}) be a coupling of $(\mathsf{P}_{X_{t+k}}(\cdot \mid X_t = x), \mathsf{P}_{\bar{X}_{t+k}}(\cdot \mid \bar{X}_t = x))$ defined as follows. We initialize both $X_t = \bar{X}_t = x$. We let $\{W_s\}_{s=t}^{t+k-1}$ be iid draws from $N(0, I)$, we set $\bar{W}_s = W_s \mathbf{1}\{\|W_s\|_2 \leq R\}$, and we evolve $X_t, \bar{X}'_t$ forward to $X_{t+k}, \bar{X}'_{t+k}$ according to their laws (40a) and (40b), respectively. Let $\mathcal{E}$ denote the event $\mathcal{E} = \{\max_{s=t,\ldots,t+k-1} \|W_s\|_2 \leq R\}$. A standard Gaussian concentration plus union bound yields $\mathbf{P}(\mathcal{E}^c) \leq \delta$, since $t + k - 1 \leq T - 2$. By the coupling representation (39) of the total-variation norm:

$$\begin{aligned}
\|\mathsf{P}_{X_{t+k}}(\cdot \mid X_t = x) - \mathsf{P}_{\bar{X}_{t+k}}(\cdot \mid \bar{X}_t = x)\|_{\mathsf{TV}} &\leq \mathbf{P}\{Z_{t+k} \neq Z'_{t+k}\} \\
&= \mathbf{P}(\{Z_{t+k} \neq Z'_{t+k}\} \cap \mathcal{E}) + \mathbf{P}(\{Z_{t+k} \neq Z'_{t+k}\} \cap \mathcal{E}^c) \\
&\leq \mathbf{P}(\mathcal{E}^c) \leq \delta.
\end{aligned}$$

The second inequality holds since on $\mathcal{E}$, $Z_{t+k} = Z'_{t+k}$ because the truncation is inactive the entire duration of the process. This establishes the first inequality.

The second inequality holds by a nearly identical coupling argument, where we set $\{W_s\}_{s=0}^{t-1}$ to be iid draws from $N(0, I)$, we set $\bar{W}_s = W_s \mathbf{1}\{\|W_s\|_2 \leq R\}$, and we initialize the processes at $X_0 = HW_0$ and $\bar{X}_0 = H\bar{W}_0$. ■

The next result states that as long as we set the failure probability δ in R as $1/T^2$, then we can bound the dependency matrix appropriately.

Proposition G.2. *Let P_X denote the joint distribution of $\{X_t\}_{t=0}^{T-1}$ from (40a), and let $\mathsf{P}_{\bar{X}}$ denote the joint distribution of $\{\bar{X}_t\}_{t=0}^{T-1}$ from (40b), with $R \geq \sqrt{d_{\mathsf{X}}} + \sqrt{6 \log T}$. We have that:*

$$\|\Gamma_{\mathsf{dep}}(\mathsf{P}_{\bar{X}})\|_{\mathsf{op}} \leq 3 + \sqrt{2} \sum_{k=1}^{T-1} \max_{t=0,\dots,T-1-k} \operatorname*{ess\,sup}_{x \in \bar{\mathsf{X}}_t} \sqrt{\|\mathsf{P}_{X_{t+k}}(\cdot \mid X_t = x) - \mathsf{P}_{X_{t+k}}\|_{\mathsf{TV}}}. \tag{41}$$

Proof. First, we invoke Proposition F.4 to obtain:

$$\|\Gamma_{\mathsf{dep}}(\mathsf{P}_{\bar{X}})\|_{\mathsf{op}} \leq 1 + \sqrt{2} \sum_{k=1}^{T-1} \max_{t=0,\dots,T-1-k} \operatorname*{ess\,sup}_{x \in \bar{\mathsf{X}}_t} \sqrt{\|\mathsf{P}_{\bar{X}_{t+k}}(\cdot \mid \bar{X}_t = x) - \mathsf{P}_{\bar{X}_{t+k}}\|_{\mathsf{TV}}}.$$

Now fix $k \in \{1, \dots, T-1\}$, $t \in \{0, \dots, T-1-k\}$, and $x \in \bar{\mathsf{X}}_t$. By triangle inequality:

$$\begin{aligned}
\|\mathsf{P}_{\bar{X}_{t+k}}(\cdot \mid \bar{X}_t = x) - \mathsf{P}_{\bar{X}_{t+k}}\|_{\mathsf{TV}} &\leq \|\mathsf{P}_{X_{t+k}}(\cdot \mid X_t = x) - \mathsf{P}_{X_{t+k}}\|_{\mathsf{TV}} \\
&\quad + \|\mathsf{P}_{\bar{X}_{t+k}}(\cdot \mid \bar{X}_t = x) - \mathsf{P}_{X_{t+k}}(\cdot \mid X_t = x)\|_{\mathsf{TV}} \\
&\quad + \|\mathsf{P}_{\bar{X}_{t+k}} - \mathsf{P}_{X_{t+k}}\|_{\mathsf{TV}}.
\end{aligned}$$

By setting $\delta = 1/T^2$ in Proposition G.1, the last two terms are bounded by $1/T^2$. Hence:

$$\|\mathsf{P}_{\bar{X}_{t+k}}(\cdot \mid \bar{X}_t = x) - \mathsf{P}_{\bar{X}_{t+k}}\|_{\mathsf{TV}} \leq \|\mathsf{P}_{X_{t+k}}(\cdot \mid X_t = x) - \mathsf{P}_{X_{t+k}}\|_{\mathsf{TV}} + \frac{2}{T^2}.$$

The claim now follows. ■

Crucially, the essential supremum in (41) is over $\bar{\mathsf{X}}_t$ and *not* X_t, of which the latter is unbounded.

The next condition that we need to check for the truncated process (40b) is that the noise process $\{H\bar{W}_t\}_{t\geq 0}$ is still a zero-mean sub-Gaussian martingale difference sequence. By symmetry of the truncation, it is clear that the noise process remains zero-mean. To check sub-Gaussianity, we use the following result.

Proposition G.3. *Let $A \subseteq \mathbb{R}^{d_{\mathsf{X}}}$ be any set that is symmetric about the origin. Let $W \sim N(0, I)$, and let $\bar{W} := W\mathbf{1}\{W \in A\}$. We have that $\bar{W}$ is 4-sub-Gaussian. Hence for any H, $H\bar{W}$ is $4\|H\|_{\mathsf{op}}^2$-sub-Gaussian.*

Proof. Since A is symmetric about the origin, $\bar{W}$ inherits the symmetry of W, i.e., $\mathbf{E}[\bar{W}] = 0$. Now fix a unit vector $u \in \mathbb{R}^{d_{\mathsf{X}}}$, and $\lambda \in \mathbb{R}$. First, let us assume that $\lambda^2 \leq 1/2$. Let ε denote a Rademacher random variable[7] that is independent of $\bar{W}$. Since $\bar{W}$ is a symmetric zero-mean distribution, we have that $\langle u, \bar{W}\rangle$ has the same distribution as $\varepsilon\langle u, \bar{W}\rangle$. Therefore:

$$\begin{aligned}
\mathbf{E}\exp(\lambda\langle u, \bar{W}\rangle) &= \mathbf{E}_{\bar{W}}\mathbf{E}_{\varepsilon}\exp(\lambda\varepsilon\langle u, \bar{W}\rangle) \\
&\leq \mathbf{E}_{\bar{W}}\exp(\lambda^2\langle u, \bar{W}\rangle^2/2) && \cosh(x) \leq \exp(x^2/2)\ \forall x \in \mathbb{R} \\
&\leq \mathbf{E}_{\bar{W}}\exp(\lambda^2\langle u, W\rangle^2/2) \\
&= \frac{1}{(1-\lambda^2)^{1/2}} && \text{since } \langle u, W\rangle \sim N(0,1) \text{ and } \lambda^2 < 1 \\
&\leq \exp(\lambda^2) && \frac{1}{1-x} \leq \exp(2x)\ \forall x \in [0, 1/2].
\end{aligned}$$

[7] That is, $\mathbf{P}(\varepsilon = 1) = \mathbf{P}(\varepsilon = -1) = 1/2$.

Now, let us assume $\lambda^2 > 1/2$. We have:

$$\begin{aligned}
\mathbf{E}\exp(\lambda\langle u, \bar{W}\rangle) &= \mathbf{E}\exp(\lambda\langle u, W\rangle)\mathbf{1}\{W \in A\} + \mathbf{P}(W \notin A) \\
&\leq \mathbf{E}\exp(\lambda\langle u, W\rangle) + 1 \\
&= \exp(\lambda^2/2) + 1 && \text{since } \langle u, W\rangle \sim N(0,1) \\
&\leq \exp(\log 2 + \lambda^2/2) && \text{since } 1 \leq \exp(\lambda^2/2) \\
&\leq \exp((2\log 2 + 1/2)\lambda^2) && \text{since } \lambda^2 > 1/2 \\
&\leq \exp(2\lambda^2).
\end{aligned}$$

The claim now follows. ∎

The following result will be useful later on. It states that the truncation does not affect the isotropic nature of the noise, as long as the truncation probability is a sufficiently small constant.

Proposition G.4. *Let $A \subseteq \mathbb{R}^{d_{\mathsf{X}}}$ be any set. Let $W \sim N(0, I)$ and $\bar{W} = W\mathbf{1}\{W \in A\}$, and suppose that $\mathbf{P}(W \notin A) \leq 1/12$. We have that:*

$$\frac{1}{2}I \preccurlyeq \mathbf{E}[\bar{W}\bar{W}^{\mathsf{T}}] \preccurlyeq I.$$

Proof. The upper bound is immediate. For the lower bound, fix a $v \in \mathbb{S}^{d_{\mathsf{X}}-1}$. We have:

$$\begin{aligned}
\mathbf{E}[\langle v, \bar{W}\rangle^2] &= \mathbf{E}[\langle v, \bar{W}\rangle^2\mathbf{1}\{W \in A\}] + \mathbf{E}[\langle v, \bar{W}\rangle^2\mathbf{1}\{W \notin A\}] \\
&= \mathbf{E}[\langle v, \bar{W}\rangle^2\mathbf{1}\{W \in A\}] && \text{since } \bar{W} = W\mathbf{1}\{W \in A\} \\
&= \mathbf{E}[\langle v, W\rangle^2] - \mathbf{E}[\langle v, W\rangle^2\mathbf{1}\{W \notin A\}] \\
&\geq 1 - \sqrt{\mathbf{E}[\langle v, W\rangle^4]\mathbf{P}(W \notin A)} && \text{since } \langle v, W\rangle \sim N(0,1) \text{ and Cauchy-Schwarz} \\
&\geq 1 - \sqrt{3\delta} \\
&\geq 1/2 && \text{since } \mathbf{P}(W \notin A) \leq 1/12.
\end{aligned}$$

Since $v \in \mathbb{S}^{d_{\mathsf{X}}-1}$ is arbitrary, the claim follows. ∎

Proposition G.5. *Let $w \sim N(0, I)$ and let M be positive semidefinite. We have:*

$$\mathbf{E}[(w^{\mathsf{T}}Mw)^2] \leq 3(\mathbf{E}[w^{\mathsf{T}}Mw])^2.$$

Proof. This is a standard calculation [see e.g. 41, Lemma 6.2]. ∎

We will also need the following result which states that the square of quadratic forms under $\bar{W}$ can be upper bounded by the square of the same quadratic form under the original noise W.

Proposition G.6. *Let $A \subseteq \mathbb{R}^{d_{\mathsf{X}}}$ be any set. Let $W \sim N(0, I)$ and $\bar{W} = W\mathbf{1}\{W \in A\}$. Fix a $k \geq 1$. Let $M \in \mathbb{R}^{d_{\mathsf{X}}k \times d_{\mathsf{X}}k}$ be a positive semidefinite matrix, and let $\{W_i\}_{i=1}^k$ and $\{\bar{W}_i\}_{i=1}^k$ be iid copies of W and $\bar{W}$, respectively. Let $W_{1:k} \in \mathbb{R}^{d_{\mathsf{X}}k}$ denote the stacked column vector of $\{W_i\}_{i=1}^k$ and similarly for $\bar{W}_{1:k} \in \mathbb{R}^{d_{\mathsf{X}}k}$. We have that:*

$$\mathbf{E}[(\bar{W}_{1:k}^{\mathsf{T}}M\bar{W}_{1:k})^2] \leq \mathbf{E}[(W_{1:k}^{\mathsf{T}}MW_{1:k})^2].$$

Proof. Let $\{M_{ij}\}_{i,j=1}^k \subset \mathbb{R}^{d_{\mathsf{X}} \times d_{\mathsf{X}}}$ denote the blocks of M. We have:

$$\mathbf{E}[(\bar{W}_{1:k}^{\mathsf{T}}M\bar{W}_{1:k})^2] = \sum_{a,b,c,d}\mathbf{E}[(\bar{W}_a^{\mathsf{T}}M_{ab}\bar{W}_b)(\bar{W}_c^{\mathsf{T}}M_{cd}\bar{W}_d)].$$

Since $\bar{W}$ is zero-mean, the only terms that are non-zero in the summation have the following form $\mathbf{E}[(\bar{W}_a^{\mathsf{T}}M_{aa}\bar{W}_a)(\bar{W}_b^{\mathsf{T}}M_{bb}\bar{W}_b)]$. Hence:

$$\begin{aligned}
\mathbf{E}[(\bar{W}_{1:k}^{\mathsf{T}}M\bar{W}_{1:k})^2] &= \sum_a \mathbf{E}[(\bar{W}_a^{\mathsf{T}}M_{aa}\bar{W}_a)^2] + \sum_{a \neq b}\mathbf{E}[(\bar{W}_a^{\mathsf{T}}M_{aa}\bar{W}_a)(\bar{W}_b^{\mathsf{T}}M_{bb}\bar{W}_b)] \\
&\overset{(a)}{\leq} \sum_a \mathbf{E}[(W_a^{\mathsf{T}}M_{aa}W_a)^2] + \sum_{a \neq b}\mathbf{E}[(W_a^{\mathsf{T}}M_{aa}W_a)(W_b^{\mathsf{T}}M_{bb}W_b)] \\
&= \mathbf{E}[(W_{1:k}^{\mathsf{T}}MW_{1:k})^2].
\end{aligned}$$

Above, (a) holds since the matrix M is positive semidefinite and therefore so are its diagonal sub-blocks M_{aa}. This ensures that each of the quadratic forms are non-negative, and hence we can upper bound the first expression by removing the indicators. ■

We conclude this section with a result that will be useful for analyzing the mixing properties of the Gaussian process (40a), when the dynamics function f is nonlinear. First, recall the definition of the 1-Wasserstein distance:

$$W_1(\mu, \nu) \triangleq \inf\{\mathbf{E}\|X - Y\|_2 \mid (X, Y) \text{ is a coupling of } (\mu, \nu)\}. \tag{42}$$

The following result uses the smoothness of the Gaussian transition kernel to upper bound the TV norm via the 1-Wasserstein distance. This result is inspired by the work of Chae and Walker [42].

Lemma G.1. *Let X_0, Y_0 be random vectors in $\mathbb{R}^p$, and let $f : \mathbb{R}^p \to \mathbb{R}^n$ be an L-Lipschitz function. Suppose that X_0, Y_0 are both absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^p$. Let $\Sigma \in \mathbb{R}^{n \times n}$ be positive definite, and let X_1, Y_1 be random vectors in $\mathbb{R}^n$ defined conditionally: $X_1 \mid X_0 = N(f(X_0), \Sigma)$ and $Y_1 \mid Y_0 = N(f(Y_0), \Sigma)$. Then:*

$$\|\mathsf{P}_{X_1} - \mathsf{P}_{Y_1}\|_{\mathsf{TV}} \le \frac{L\sqrt{\operatorname{tr}(\Sigma^{-1})}}{2} W_1(\mathsf{P}_{X_0}, \mathsf{P}_{Y_0}).$$

Proof. Since X_0, Y_0 are absolutely continuous, the Radon-Nikodym theorem ensures that there exists densities p_0, q_0 for X_0, Y_0, respectively. Let ϕ denote the density of the $N(0, \Sigma)$ distribution. Let p_1, q_1 denote the densities of X_1, Y_1, respectively. We have the following convolution expressions:

$$p_1(x) = \int \phi(x - f(x_0)) p_0(x_0) dx_0 = \int \phi(x - f(X_0)) dX_0,$$
$$q_1(x) = \int \phi(x - f(x_0)) q_0(x_0) dx_0 = \int \phi(x - f(Y_0)) dY_0.$$

Now, let π be a coupling of (X_0, Y_0). We can equivalently write p_1, q_1 as a convolution over π:

$$p_1(x) = \int \phi(x - f(X_0)) d\pi(X_0, Y_0),$$
$$q_1(x) = \int \phi(x - f(Y_0)) d\pi(X_0, Y_0).$$

Hence:

$$\begin{aligned}(p_1 - q_1)(x) &= \int [\phi(x - f(X_0)) - \phi(x - f(Y_0))] d\pi(X_0, Y_0) \\ &= \mathbf{E}_{\pi(X_0,Y_0)}[\phi(x - f(X_0)) - \phi(x - f(Y_0))].\end{aligned}$$

Now by the L^1 representation of total-variation norm [see e.g. 1, Lemma 2.1]:

$$\begin{aligned}\|\mathsf{P}_{X_1} - \mathsf{P}_{Y_1}\|_{\mathsf{TV}} &= \frac{1}{2}\int |p_1(x) - q_1(x)| dx \\ &= \frac{1}{2}\int |\mathbf{E}_{\pi(X_0,Y_0)}[\phi(x - f(X_0)) - \phi(x - f(Y_0))]| dx \\ &\overset{(a)}{\le} \frac{1}{2}\int \mathbf{E}_{\pi(X_0,Y_0)}[|\phi(x - f(X_0)) - \phi(x - f(Y_0))|] dx \\ &\overset{(b)}{=} \frac{1}{2}\mathbf{E}_{\pi(X_0,Y_0)}\left[\int |\phi(x - f(X_0)) - \phi(x - f(Y_0))| dx\right].\end{aligned}$$

The inequality (a) is Jensen's inequality, and the equality (b) is Tonelli's theorem since the integrand is non-negative. We now focus on the inner integral inside the expectation over π. By the mean-value theorem, since ϕ is continuously differentiable, fixing x, X_0, Y_0:

$$\begin{aligned}|\phi(x - f(X_0)) - \phi(x - f(Y_0))| &= \left|\int_0^1 \nabla\phi((1-s)(x - f(Y_0)) + s(x - f(X_0)))^{\mathsf{T}}(f(Y_0) - f(X_0)) ds\right| \\ &\le \|f(Y_0) - f(X_0)\|_2 \int_0^1 \|\nabla\phi(x - (sf(X_0) - (1-s)f(Y_0))\|_2 ds.\end{aligned}$$

Hence, by another application of Tonelli's theorem:

$$\begin{aligned}\int |\phi(x - f(X_0)) - \phi(x - f(Y_0))|dx &\leq \|f(X_0) - f(Y_0)\|_2 \int \int_0^1 \|\nabla\phi(x - (sf(X_0) - (1-s)f(Y_0))\|_2 dsdx \\ &= \|f(X_0) - f(Y_0)\|_2 \int_0^1 \int \|\nabla\phi(x - (sf(X_0) - (1-s)f(Y_0))\|_2 dxds \\ &\leq \|f(X_0) - f(Y_0)\|_2 \sqrt{\mathrm{tr}(\Sigma^{-1})}.\end{aligned}$$

The last inequality follows from the following computation. Observe that $\nabla\phi(x) = -\Sigma^{-1}x\phi(x)$ and define $\mu = sf(X_0) - (1-s)f(Y_0)$. Since μ does not depend on x, by the translation invariance of the Lebesgue integral:

$$\begin{aligned}\int \|\nabla\phi(x - (sf(X_0) - (1-s)f(Y_0))\|_2 dx &= \int \|\Sigma^{-1}(x-\mu)\|_2 \phi(x-\mu)dx \\ &= \int \|\Sigma^{-1}x\|_2 \phi(x)dx \\ &= \mathbf{E}_{x\sim N(0,\Sigma^{-1})}[\|x\|_2] \\ &\leq \sqrt{\mathrm{tr}(\Sigma^{-1})}.\end{aligned}$$

The last inequality above is another application of Jensen's inequality. Therefore, combining the inequalities thus far, and using the L-Lipschitz property of f:

$$\begin{aligned}\|\mathsf{P}_{X_1} - \mathsf{P}_{Y_1}\|_{\mathsf{TV}} &\leq \frac{\sqrt{\mathrm{tr}(\Sigma^{-1})}}{2}\mathbf{E}_{\pi(X_0,Y_0)}\left[\|f(X_0) - f(Y_0)\|_2\right] \\ &\leq \frac{L\sqrt{\mathrm{tr}(\Sigma^{-1})}}{2}\mathbf{E}_{\pi(X_0,Y_0)}[\|X_0 - Y_0\|_2].\end{aligned}$$

Since the coupling π of (X_0, Y_0) was arbitrary, the result now follows by taking the infimum of the right hand side over all valid couplings. ■

H Recovering Ziemann et al. [15] via boundedness

Here we show how to recover the results for mixing systems from Ziemann et al. [15, Theorem 1 combined with Proposition 2], corresponding to $\alpha = 1$. This rests on the observation that $(B^2, 1)$-hypercontractivity is automatic by B-boundedness.

Corollary H.1. *Suppose that $\mathscr{F}_\star$ is star-shaped and B-bounded. Fix also $p \in \mathbb{R}_+$ and $q \in (0, 2)$ and suppose further that $\mathscr{F}_\star$ satisfies condition (8). Then we have that:*

$$\begin{aligned}\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \leq 8\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star) + \frac{1}{\sqrt{8}}\left(16B^2 p\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2\right)^{2/q} T^{-2/(2+q)} \\ + \exp\left(\frac{-T^{q/(2+q)}}{16B^2\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2}\right). \qquad (43)\end{aligned}$$

The first two terms in inequality (43) are both of order $T^{-2/(2+q)}$ if $\|\Gamma_{\mathsf{dep}}\|_{\mathsf{op}}^2 = O(1)$. Note that, without further control of the moments of $f \in \mathscr{F}_\star$, the bound in Theorem 4.1 thus degrades by a factor of the dependency matrix through the second term.

Proof of Corollary H.1 Fix $c > 0$ to be determined later and choose $r = cT^{-1/(2+q)}$. We find:

$$\begin{aligned}&B^2\mathcal{N}_\infty(\mathscr{F}_\star, r/\sqrt{8})\exp\left(\frac{-Tr^2}{8B^2\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2}\right) \\ &\leq \exp\left(p\left(\frac{\sqrt{8}}{r}\right)^q - \frac{Tr^2}{8B^2\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2}\right) && \text{(Condition (8))} \\ &= \exp\left(\left[p\left(\frac{\sqrt{8}}{c}\right)^q - \frac{1}{8B^2\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2}\right]T^{q/(2+q)}\right). && (r = cT^{-1/(2+q)})\end{aligned}$$

Hence we may solve for $c = \frac{1}{\sqrt{8}} \left(16B^2 p \|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2\right)^{1/q}$ in

$$p \left(\frac{\sqrt{8}}{c}\right)^q = \frac{1}{16B^2 \|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2}$$

to arrive at the desired conclusion. ∎

I Linear dynamical systems

We define the truncated linear dynamics:

$$\bar{X}_{t+1} = A_\star \bar{X}_t + H\bar{V}_t, \;\; \bar{X}_0 = H\bar{V}_0, \;\; \bar{V}_t = V_t \mathbf{1}\{\|V_t\|_2 \le R\}. \tag{44}$$

We set $R = \sqrt{d_{\mathsf{X}}} + \sqrt{2(1+\beta)\log T}$ where $\beta > 4$ is a free parameter. Define the event $\mathcal{E}$ as:

$$\mathcal{E} := \left\{ \max_{0 \le t \le T-1} \|V_t\|_2 \le R \right\}. \tag{45}$$

Note that by the setting of R, we have $\mathbf{P}(\mathcal{E}^c) \le 1/T^\beta$ using standard Gaussian concentration results plus a union bound. Furthermore on $\mathcal{E}$, the original GLM process driven by Gaussian noise (12) coincides with the truncated process (44). Let $\widehat{f}$ denote the LSE on the original process (44), and let $\bar{f}$ denote the LSE on the truncated process (44). Hence:

$$\begin{aligned} \mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 &= \mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \mathbf{1}\{\mathcal{E}\} + \mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \mathbf{1}\{\mathcal{E}^c\} \\ &\le \mathbf{E}\|\bar{f} - f_\star\|_{L^2}^2 + \mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \mathbf{1}\{\mathcal{E}^c\}. \end{aligned}$$

Let us now control the error term $\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \mathbf{1}\{\mathcal{E}^c\}$. Since X_t is a linear function of the Gaussian noise $\{W_t\}$ process, by Proposition G.5 we have $\mathbf{E}\|X_t\|_2^4 \le 3(\mathbf{E}\|X_t\|_2^2)^2$. Write $\widehat{f}(x) = \widehat{A}x$, and put $\widehat{\Delta} = \widehat{A} - A_\star$. We have:

$$\begin{aligned} \mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \mathbf{1}\{\mathcal{E}^c\} &= \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|\widehat{\Delta} X_t\|_2^2 \mathbf{1}\{\mathcal{E}^c\} \overset{(a)}{\le} \frac{4B^2}{T}\sum_{t=0}^{T-1} \mathbf{E}\|X_t\|_2^2 \mathbf{1}\{\mathcal{E}^c\} \\ &\overset{(b)}{\le} \frac{4B^2}{T^{1+\beta/2}} \sum_{t=0}^{T-1} \sqrt{\mathbf{E}\|X_t\|_2^4} \overset{(c)}{\le} \frac{4\sqrt{3}B^2}{T^{1+\beta/2}} \sum_{t=0}^{T-1} \mathbf{E}\|X_t\|_2^2 \\ &= \frac{4\sqrt{3}B^2}{T^{1+\beta/2}} \sum_{t=0}^{T-1} \operatorname{tr}(\Gamma_t) \overset{(d)}{\le} \frac{4\sqrt{3}B^2 \operatorname{tr}(\Gamma_{T-1})}{T^{\beta/2}} \overset{(e)}{\le} \frac{4\sqrt{3}B^2 \|H\|_{\mathsf{op}}^2 \tau^2 d_{\mathsf{X}}}{(1-\rho)T^{\beta/2}}. \end{aligned} \tag{46}$$

Above, (a) follows from the definition of $\mathscr{F}$, (b) follows from Cauchy-Schwarz, (c) uses the hypercontractivity bound $\mathbf{E}\|X_t\|_2^4 \le 3(\mathbf{E}\|X_t\|_2^2)^2$, (d) uses the fact that Γ_t is monotonically increasing in the Loewner order, and (e) uses the following bound on $\operatorname{tr}(\Gamma_{T-1})$ using the (τ, ρ)-stability of $A_\star$:

$$\operatorname{tr}(\Gamma_{T-1}) \le \frac{\|H\|_{\mathsf{op}}^2 \tau^2 d_{\mathsf{X}}}{1-\rho^2} \le \frac{\|H\|_{\mathsf{op}}^2 \tau^2 d_{\mathsf{X}}}{1-\rho}.$$

The remainder of the proof is to bound the error of the LSE $\bar{f}$ using Theorem 4.1. This involves two main steps: showing the trajectory hypercontractivity condition Definition 4.1 holds, and bounding the dependency matrix $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_{\bar{X}})\|_{\mathsf{op}}$ (cf. Definition 4.2), where $\mathsf{P}_{\bar{X}}$ denotes the joint distribution of the process $\{\bar{X}_t\}_{t=0}^{T-1}$. Before we proceed, we define some reoccurring constants:

$$\mu \triangleq \lambda_{\min}(\Gamma_{\kappa-1}), \;\; B_{\bar{X}} \triangleq \frac{\|H\|_{\mathsf{op}} \tau (\sqrt{d_{\mathsf{X}}} + \sqrt{2(1+\beta)\log T})}{1-\rho}. \tag{47}$$

I.1 Trajectory hypercontractivity for truncated LDS

Proposition I.1. *Suppose that $T \ge \max\{6, 2\kappa\}$. The pair $(\mathscr{F}_\star, \mathsf{P}_{\bar{X}})$ with $\mathscr{F}$ given in (13) and $\mathsf{P}_{\bar{X}}$ as the joint distribution of $\{\bar{X}_t\}_{t=0}^{T-1}$ from (44) satisfies the $(C_{\mathsf{LDS}}, 2)$-trajectory hypercontractivity condition with $C_{\mathsf{LDS}} = \frac{108\tau^4\|H\|_{\mathsf{op}}^4}{(1-\rho)^2\mu^2}$.*

Proof. Fix any size-conforming matrix M. Let the noise process $\{\bar{V}_t\}_{t=0}^{T-1}$ be stacked into a noise vector $\bar{V}_{0:T-1} \in \mathbb{R}^{d_X T}$. Observe that we can write $MX_t = MT_t\bar{V}_{0:T-1}$ for some matrix T_t. We invoke the comparison inequality in Proposition G.6 followed by the Gaussian fourth moment identity in Proposition G.5 to conclude that:

$$\mathbf{E}\|M\bar{X}_t\|_2^4 = \mathbf{E}\|MT_t\bar{V}_{0:T-1}\|_2^4 \le \mathbf{E}\|MX_t\|_2^4 \le 3(\mathbf{E}\|MX_t\|_2^2)^2 = 3\operatorname{tr}(M^\mathsf{T} M\Gamma_t)^2.$$

By monotonicity of Γ_t and the assumption $T \ge 6$:

$$\frac{1}{T}\sum_{t=0}^{T-1}\Gamma_t \succcurlyeq \frac{1}{T}\sum_{t=\lfloor T/2\rfloor}^{T-1}\Gamma_t \succcurlyeq \frac{T-\lfloor T/2\rfloor}{T}\Gamma_{\lfloor T/2\rfloor} \succcurlyeq \frac{1}{3}\Gamma_{\lfloor T/2\rfloor}. \tag{48}$$

Since $T \ge 2\kappa$, the inequality $\Gamma_{\lfloor T/2\rfloor} \succcurlyeq \Gamma_{\kappa-1}$ holds, and therefore $\Gamma_{\lfloor T/2\rfloor}$ is invertible since $(A_\star, H)$ is κ-step controllable. Therefore:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|M\bar{X}_t\|_2^4 &\le 3\operatorname{tr}(M^\mathsf{T} M\Gamma_{T-1})^2 \\
&= 3\operatorname{tr}(M\Gamma_{\lfloor T/2\rfloor}^{1/2}\Gamma_{\lfloor T/2\rfloor}^{-1/2}\Gamma_{T-1}\Gamma_{\lfloor T/2\rfloor}^{-1/2}\Gamma_{\lfloor T/2\rfloor}^{1/2}M^\mathsf{T})^2 \\
&\le 3\|\Gamma_{\lfloor T/2\rfloor}^{-1}\Gamma_{T-1}\|_{\mathsf{op}}^2\operatorname{tr}(M^\mathsf{T} M\Gamma_{\lfloor T/2\rfloor})^2 \\
&\le 27\|\Gamma_{\lfloor T/2\rfloor}^{-1}\Gamma_{T-1}\|_{\mathsf{op}}^2\operatorname{tr}\left(M^\mathsf{T} M\cdot\frac{1}{T}\sum_{t=0}^{T-1}\Gamma_t\right)^2 \qquad \text{using (48)} \\
&= 27\|\Gamma_{\lfloor T/2\rfloor}^{-1}\Gamma_{T-1}\|_{\mathsf{op}}^2\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|MX_t\|_2^2\right)^2 \\
&\le 108\|\Gamma_{\lfloor T/2\rfloor}^{-1}\Gamma_{T-1}\|_{\mathsf{op}}^2\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|M\bar{X}_t\|_2^2\right)^2 \qquad \text{using Proposition G.4.}
\end{aligned}$$

Since the matrix M is arbitrary, the claim follows using the following bound for $\|\Gamma_{\lfloor T/2\rfloor}^{-1}\Gamma_{T-1}\|_{\mathsf{op}}^2$:

$$\|\Gamma_{\lfloor T/2\rfloor}^{-1}\Gamma_{T-1}\|_{\mathsf{op}}^2 \le \frac{\tau^4\|H\|_{\mathsf{op}}^4}{(1-\rho^2)^2\mu^2} \le \frac{\tau^4\|H\|_{\mathsf{op}}^4}{(1-\rho)^2\mu^2}.$$

■

I.2 Bounding the dependency matrix for truncated LDS

We control $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_{\bar{X}})\|_{\mathsf{op}}$ by a direct computation of the mixing properties of the original Gaussian process (12).

Proposition I.2. *Consider the process $\{\bar{X}_t\}_{t\ge0}$ from (44), and let $\mathsf{P}_{\bar{X}}$ denote the joint distribution of $\{\bar{X}_t\}_{t=0}^{T-1}$. We have that:*

$$\|\Gamma_{\mathsf{dep}}(\mathsf{P}_{\bar{X}})\|_{\mathsf{op}} \le 5\kappa + \frac{22}{1-\rho}\log\left(\frac{\tau^2}{4\mu}\left[B_{\bar{X}}^2 + \frac{d_X\|H\|_{\mathsf{op}}^2}{1-\rho}\right]\right).$$

Proof. We first construct an almost sure bound on the process $\{\bar{X}_t\}_{t\ge0}$. Indeed, for any $t \ge 0$, using the (τ,ρ)-stability of $A_\star$:

$$\|\bar{X}_t\|_2 \le \frac{\|H\|_{\mathsf{op}}\tau R}{1-\rho} = \frac{\|H\|_{\mathsf{op}}\tau(\sqrt{d_X} + \sqrt{2(1+\beta)\log T})}{1-\rho} = B_{\bar{X}}.$$

Also, by (τ,ρ)-stability, we have for any indices $s \le t$:

$$\|\Gamma_s - \Gamma_t\|_{\mathsf{op}} = \left\|\sum_{k=s+1}^{t} A^k HH^\mathsf{T}(A^k)^\mathsf{T}\right\|_{\mathsf{op}} \le \|H\|_{\mathsf{op}}^2\tau^2\sum_{k=s+1}^{t}\rho^{2k} \le \frac{\|H\|_{\mathsf{op}}^2\tau^2}{1-\rho^2}\rho^{2(s+1)}. \tag{49}$$

The marginal and conditional distributions of $\{X_t\}_{t\ge 0}$ are easily characterized. We have that $X_t \sim N(0, \Gamma_t)$. Furthermore, $X_t \mid X_0 = x$ for $t \ge 1$ is distributed as $N(A^t x, \Gamma_{t-1})$. So now for any $t \ge 0$ and $k \ge 1$:

$$\mathsf{P}_{X_{t+k}}(\cdot \mid X_t = x) = N(A^k x, \Gamma_{k-1}), \;\; \mathsf{P}_{X_{t+k}} = N(0, \Gamma_{t+k}).$$

Now suppose $k \ge \kappa$. The matrices Γ_{k-1} and Γ_{t+k} will both be invertible, so the two distributions are mutually absolutely continuous. We can then use the closed-form expression for the KL-divergence between two multivariate Gaussians:

$$\begin{aligned}
&\mathsf{KL}(N(A^k x, \Gamma_{k-1}), N(0, \Gamma_{t+k})) \\
&= \frac{1}{2}\left[\mathrm{tr}(\Gamma_{t+k}^{-1}\Gamma_{k-1}) + x^\mathsf{T}(A^k)^\mathsf{T}\Gamma_{t+k}^{-1}A^k x - d_\mathsf{X} + \log\det(\Gamma_{t+k}\Gamma_{k-1}^{-1})\right] \\
&\le \frac{1}{2}x^\mathsf{T}(A^k)^\mathsf{T}\Gamma_{t+k}^{-1}A^k x + \frac{d_\mathsf{X}}{2}\log\|\Gamma_{t+k}\Gamma_{k-1}^{-1}\|_{\mathsf{op}} && \text{since } \Gamma_{k-1} \preccurlyeq \Gamma_{t+k} \\
&\le \frac{\tau^2\rho^{2k}\|x\|_2^2}{2\mu} + \frac{d_\mathsf{X}}{2}\log\left(1 + \frac{\|\Gamma_{t+k} - \Gamma_{k-1}\|_{\mathsf{op}}}{\mu}\right) && \text{using } (\tau,\rho)\text{-stability} \\
&\le \frac{\tau^2\rho^{2k}\|x\|_2^2}{2\mu} + \frac{d_\mathsf{X}}{2}\log\left(1 + \frac{\|H\|_{\mathsf{op}}^2\tau^2}{(1-\rho^2)\mu}\rho^{2k}\right) && \text{using (49)} \\
&\le \left[\frac{\tau^2\|x\|_2^2}{2\mu} + \frac{d_\mathsf{X}\|H\|_{\mathsf{op}}^2\tau^2}{2(1-\rho^2)\mu}\right]\rho^{2k} && \log(1+x) \le x \,\forall x \ge 0.
\end{aligned}$$

Hence by Pinsker's inequality [see e.g. 1, Lemma 2.5], whenever $k \ge \kappa$:

$$\begin{aligned}
\|\mathsf{P}_{X_{t+k}}(\cdot \mid X_t = x) - \mathsf{P}_{X_{t+k}}\|_{\mathsf{TV}} &\le \sqrt{\mathsf{KL}(N(A^k x, \Gamma_{k-1}), N(0, \Gamma_{t+k}))/2} \\
&\le \sqrt{\frac{\tau^2\|x\|_2^2}{4\mu} + \frac{d_\mathsf{X}\|H\|_{\mathsf{op}}^2\tau^2}{4(1-\rho^2)\mu}}\rho^k.
\end{aligned}$$

By Proposition G.2 (which we can invoke since we constrained $\beta \ge 2$), for any $\ell \in \mathbb{N}$:

$$\begin{aligned}
\|\Gamma_{\mathsf{dep}}(\mathsf{P}_{\bar{X}})\|_{\mathsf{op}} &\le 3 + \sqrt{2}\sum_{k=1}^{T-1}\max_{t=0,\dots,T-1-k}\operatorname*{ess\,sup}_{x\in\bar{\mathsf{X}}_t}\sqrt{\|\mathsf{P}_{X_{t+k}}(\cdot \mid X_t = x) - \mathsf{P}_{X_{t+k}}\|_{\mathsf{TV}}} \\
&\le 3 + \sqrt{2}(\kappa - 1 + \ell) + \sum_{k=\kappa+\ell}^{T-1}\left[\frac{\tau^2 B_{\bar{X}}^2}{4\mu} + \frac{d_\mathsf{X}\|H\|_{\mathsf{op}}^2\tau^2}{4(1-\rho^2)\mu}\right]^{1/4}\rho^{k/2} \\
&\le 5(\kappa+\ell) + \left[\frac{\tau^2 B_{\bar{X}}^2}{4\mu} + \frac{d_\mathsf{X}\|H\|_{\mathsf{op}}^2\tau^2}{4(1-\rho^2)\mu}\right]^{1/4}\frac{\rho^{(\kappa+\ell)/2}}{1-\rho^{1/2}}.
\end{aligned}$$

Now, define $\psi \triangleq \frac{\tau^2 B_{\bar{X}}^2}{4\mu} + \frac{d_\mathsf{X}\|H\|_{\mathsf{op}}^2\tau^2}{4(1-\rho^2)\mu}$. We choose $\ell = \max\left\{\left\lceil\frac{\log(\psi^{1/4})}{1-\rho^{1/2}}\right\rceil - \kappa, 0\right\}$, so $\rho^{(\kappa+\ell)/2} \le 1/\psi^{1/4}$. With this choice of ℓ and the observation that $\inf_{x\in[0,1]}\frac{1-\sqrt{x}}{1-x} = \frac{1}{2}$,

$$\|\Gamma_{\mathsf{dep}}(\mathsf{P}_{\bar{X}})\|_{\mathsf{op}} \le 5\kappa + \frac{11\log\psi}{4(1-\rho^{1/2})} \le 5\kappa + \frac{22\log\psi}{1-\rho}.$$

The claim now follows. ■

I.3 Finishing the proof of Theorem 6.1

For what follows, c_i will denote universal positive constants whose values remain unspecified.

For any $\varepsilon > 0$ and $r > 0$, we now construct an ε-covering of $\partial B(r)$ with $\mathscr{F}_\star$ the offset class of $\mathscr{F}$ from (13). To this end, we let $\{A_1, \dots, A_N\}$ be a δ-cover of $\mathscr{A} \triangleq \{A \in \mathbb{R}^{d_\mathsf{X}\times d_\mathsf{X}} \mid \|A\|_F \le B\}$ for δ to be specified. By a volumetric argument we may choose $\{A_1, \dots, A_N\}$ such that $N \le \left(1 + \frac{2B}{\delta}\right)^{d_\mathsf{X}^2}$. Now, any realization of $\{\bar{X}_t\}$ will have norm less than $B_{\bar{X}}$, where $B_{\bar{X}}$ is given by (47) and satisfies

$$B_{\bar{X}} \le c_0\frac{\|H\|_{\mathsf{op}}\tau(\sqrt{d_\mathsf{X}} + \sqrt{(1+\beta)\log T})}{1-\rho}.$$

Let $A \in \mathscr{A}$, and let A_i denote an element in the covering satisfying $\|A - A_i\|_F \leq \delta$. For any x satisfying $\|x\|_2 \leq B_{\bar{X}}$:

$$\|(A_i x - A_\star x) - (Ax - A_\star x)\|_F = \|(A_i - A)x\|_2 \leq \|A_i - A\|_F \|x\|_2 \leq \delta B_{\bar{X}}.$$

Thus, it suffices to take $\delta = \varepsilon / B_{\bar{X}}$ to construct an ε-covering of $\mathscr{F}_\star$ over $\{\bar{X}_t\}$, which shows that $\mathcal{N}_\infty(\mathscr{F}_\star, \varepsilon) \leq \left(1 + \frac{2BB_{\bar{X}}}{\varepsilon}\right)^{d_{\mathsf{X}}^2}$. Since $\partial B(r) \subset \mathscr{F}_\star$, we have the following inequality [see e.g. 39, Exercise 4.2.10]:

$$\mathcal{N}_\infty(\partial B(r), \varepsilon) \leq \mathcal{N}_\infty(\mathscr{F}_\star, \varepsilon/2) \leq \left(1 + \frac{4BB_{\bar{X}}}{\varepsilon}\right)^{d_{\mathsf{X}}^2}.$$

By Proposition I.1, $(\mathscr{F}_\star, \mathsf{P}_{\bar{X}})$ is $(C_{\mathsf{LDS}}, 2)$-hypercontractive for all $T \geq \max\{6, 2\kappa\}$, with

$$C_{\mathsf{LDS}} = \frac{108\tau^4 \|H\|_{\mathsf{op}}^4}{(1-\rho)^2 \mu^2}.$$

Also by Proposition I.2,

$$\|\Gamma_{\mathsf{dep}}(\mathsf{P}_{\bar{X}})\|_{\mathsf{op}}^2 \leq c_1 \kappa^2 + \frac{c_2}{(1-\rho)^2} \log^2\left(\frac{\tau^2}{4\mu}\left[B_{\bar{X}}^2 + \frac{d_{\mathsf{X}} \|H\|_{\mathsf{op}}^2}{1-\rho}\right]\right) \triangleq \gamma^2.$$

Since $\mathscr{F}_\star$ is convex and contains the zero function, it is also star-shaped. Furthermore, on the truncated process (44), the class $\mathscr{F}_\star$ is $2BB_{\bar{X}}$-bounded. Invoking Theorem 4.1, we thus have for every $r > 0$ that

$$\mathbf{E}\|\bar{f} - f_\star\|_{L^2}^2 \leq 8\mathbf{E}\bar{\mathsf{M}}_T(\mathscr{F}_\star) + r^2 + 4B^2 B_{\bar{X}}^2 \left(1 + \frac{4\sqrt{8}BB_{\bar{X}}}{r}\right)^{d_{\mathsf{X}}^2} \exp\left(\frac{-T}{8C_{\mathsf{LDS}}\gamma^2}\right). \qquad (50)$$

Here, the notation $\mathbf{E}\bar{\mathsf{M}}_T(\mathscr{F}_\star)$ is meant to emphasize that the offset complexity is with respect to the truncated process $\mathsf{P}_{\bar{X}}$ and *not* the original process P_X. We now set $r^2 = \|H\|_{\mathsf{op}}^2 d_{\mathsf{X}}^2 / T$, and compute a T_0 such that the third term in (50) is also bounded by $\|H\|_{\mathsf{op}}^2 d_{\mathsf{X}}^2/T$. To do this, it suffices to compute T_0 such that for all $T \geq T_0$:

$$T \geq c_3 C_{\mathsf{LDS}} \gamma^2 d_{\mathsf{X}}^2 \log\left(\frac{TBB_{\bar{X}}}{\|H\|_{\mathsf{op}}\sqrt{d_{\mathsf{X}}}}\right).$$

Thus it suffices to set T_0 as (provided that β is at most polylogarithmic in the problem constants—we later make such a choice):

$$T_0 = c_4 \frac{\tau^4 \|H\|_{\mathsf{op}}^4 d_{\mathsf{X}}^2}{(1-\rho)^2 \mu^2}\left[\kappa^2 + \frac{1}{(1-\rho)^2}\right] \operatorname{polylog}\left(B, d_{\mathsf{X}}, \tau, \|H\|_{\mathsf{op}}, \frac{1}{\mu}, \frac{1}{1-\rho},\right). \qquad (51)$$

We do not attempt to compute the exact power of the polylog term; it can in principle be done via Du et al. [43, Lemma F.2].

Next, by (46), $\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \mathbf{1}\{\mathcal{E}^c\} \leq \frac{4\sqrt{3}B^2\|H\|_{\mathsf{op}}^2 \tau^2 d_{\mathsf{X}}}{(1-\rho)T^{\beta/2}}$. Thus we also need to set T_0 large enough so that this term is bounded by $\|H\|_{\mathsf{op}}^2 d_{\mathsf{X}}^2/T$. To do this, it suffices to constrain $\beta > 2$ and set $T_0 \geq c_5 \left[\frac{B^2\tau^2}{1-\rho}\right]^{\frac{1}{\beta/2-1}}$. Hence, setting $\beta = \max\{4, c_6 \log B\}$ implies that (51) suffices.

Let us now upper bound $\mathbf{E}\bar{\mathsf{M}}_T(\mathscr{F}_\star)$ by $\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star)$ plus $\|H\|_{\mathsf{op}}^2 d_{\mathsf{X}}^2/T$. Recall the definition of $\mathcal{E}$ from (45). We first write:

$$\begin{aligned}
\mathbf{E}\bar{\mathsf{M}}_T(\mathscr{F}_\star) &= \mathbf{E}\bar{\mathsf{M}}_T(\mathscr{F}_\star)\mathbf{1}\{\mathcal{E}\} + \mathbf{E}\bar{\mathsf{M}}_T(\mathscr{F}_\star)\mathbf{1}\{\mathcal{E}^c\} \\
&= \mathbf{E}\mathsf{M}_T(\mathscr{F}_\star)\mathbf{1}\{\mathcal{E}\} + \mathbf{E}\bar{\mathsf{M}}_T(\mathscr{F}_\star)\mathbf{1}\{\mathcal{E}^c\} \\
&\leq \mathbf{E}\mathsf{M}_T(\mathscr{F}_\star) + \mathbf{E}\bar{\mathsf{M}}_T(\mathscr{F}_\star)\mathbf{1}\{\mathcal{E}^c\}.
\end{aligned}$$

The last inequality holds since it can be checked that $\mathsf{M}_T(\mathscr{F}_\star) \geq 0$ (i.e., by lower bounding the supremum with the zero function which is in $\mathscr{F}_\star$ since $\mathscr{F}$ contains $f_\star$). Furthermore, an elementary linear algebra calculation yields that we can upper bound $\bar{\mathsf{M}}_T(\mathscr{F}_\star)$ deterministically by:

$$\bar{\mathsf{M}}_T(\mathscr{F}_\star) \leq \frac{4}{T} \left\| \left(\left(\sum_{t=0}^{T-1} \bar{X}_t \bar{X}_t^\mathsf{T} \right)^\dagger \right)^{1/2} \sum_{t=0}^{T-1} \bar{X}_t \bar{V}_t^\mathsf{T} H^\mathsf{T} \right\|_F^2 \leq \frac{4}{T} \sum_{t=0}^{T-1} \|H\bar{V}_t\|_2^2$$

Here, the $\dagger$ notation refers to the Moore-Penrose pseudo-inverse. Therefore taking expectations:

$$\mathbf{E}\bar{M}_T(\mathscr{F}_\star)\mathbf{1}\{\mathcal{E}^c\} \leq \frac{4}{T} \sum_{t=0}^{T-1} \mathbf{E}\|H\bar{V}_t\|_2^2 \mathbf{1}\{\mathcal{E}^c\} \leq \frac{4}{T} \sum_{t=0}^{T-1} \mathbf{E}\|HV_t\|_2^2 \mathbf{1}\{\mathcal{E}^c\} \overset{(a)}{\leq} \frac{4}{T^{1+\beta/2}} \sum_{t=0}^{T-1} \sqrt{\mathbf{E}\|HV_t\|_2^4}$$

$$\overset{(b)}{\leq} \frac{4\sqrt{3}}{T^{1+\beta/2}} \sum_{t=0}^{T-1} \mathbf{E}\|HV_t\|_2^2 = \frac{4\sqrt{3}\|H\|_F^2}{T^{\beta/2}} \leq \frac{4\sqrt{3} d_\mathsf{X} \|H\|_{\mathsf{op}}^2}{T^{\beta/2}}.$$

Here, (a) is Cauchy-Schwarz, and (b) follows from Proposition G.5. This last term will be bounded by $\|H\|_{\mathsf{op}}^2 d_\mathsf{X}^2 / T$ as soon as $T \geq 4\sqrt{3}$, since we set $\beta \geq 4$. The claim now follows. ■

I.4 Further discussion related to Theorem 6.1

We first discuss the rate (14) prescribed by Theorem 6.1. A simple computation shows that the martingale complexity $\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star)$ can be upper bounded by $1/T$ times the self-normalized martingale term which typically appears in the analysis of least-squares [44]. Specifically, when the empirical covariance matrix $\sum_{t=0}^{T-1} X_t X_t^\mathsf{T}$ is invertible:

$$\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star) \leq \frac{4}{T}\mathbf{E} \left\| \left(\sum_{t=0}^{T-1} X_t X_t^\mathsf{T} \right)^{-1/2} \sum_{t=0}^{T-1} X_t V_t^\mathsf{T} H^\mathsf{T} \right\|_F^2.$$

A sharp analysis of this self-normalized martingale term [19, Lemma 4.1] shows that $\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star) \lesssim \frac{\|H\|_{\mathsf{op}}^2 d_\mathsf{X}^2}{T}$, and hence (14) yields the minimax optimal rate up to constant factors after a polynomial burn-in time.[8] This is unlike the chaining bound (7) which yields extra logarithmic factors [see e.g. 15, Lemma 4]. Note that the burn-in time of $\tilde{O}(d_\mathsf{X}^2)$ given by our result is sub-optimal by a factor of d_X. This extra factor comes from the union bound over a Frobenius norm ball of $d_\mathsf{X} \times d_\mathsf{X}$ matrices in Theorem 4.1.

To convert (14) into a parameter recovery bound, we simply lower bound the excess risk:

$$\mathbf{E}\|\hat{f} - f_\star\|_{L^2}^2 \geq \mathbf{E}\|\hat{A} - A_\star\|_F^2 \lambda_{\min}(\bar{\Gamma}_T) \Longrightarrow \mathbf{E}\|\hat{A} - A_\star\|_F^2 \lesssim \frac{\|H\|_{\mathsf{op}}^2 d_\mathsf{X}^2}{T\lambda_{\min}(\bar{\Gamma}_T)}, \tag{52}$$

where $\bar{\Gamma}_T \triangleq \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}[X_t X_t^\mathsf{T}]$ is the average covariance matrix. The rate (52) recovers, after the polynomial burn-in time, existing results [16, 18, 19, 29] for stable systems, with a few caveats. First, most of the existing results are given in operator instead of Frobenius norm. We ignore this issue, since the only difference is the extra unavoidable factor of d_X in the rate for the Frobenius norm compared to the operator norm rate. Second, since Theorem 4.1 ultimately relies on some degree of ergodicity for the covariate process $\{X_t\}_{t\geq 0}$, we cannot handle the marginally stable case (where $A_\star$ is allowed to have spectral radius equal to one) as in Simchowitz et al. [16], Sarkar and Rakhlin [18], Tu et al. [19], nor the unstable case as in Faradonbeh et al. [17], Sarkar and Rakhlin [18].

We conclude with a short discussion on the proof of Theorem 6.1. As the LDS process (12) is unbounded, we use the truncation argument outlined in Appendix B.1 so that Theorem 4.1 still applies. Furthermore, since the process (12) is jointly Gaussian, the dependency matrix coefficients are simple to bound, resulting in polynomial rates (15) for the burn-in time. A much wider variety of non-Gaussian noise distributions can be handled via ergodic theory for Markov chains [see e.g. 32, Chapter 15]. While these results typically do not offer explicit expressions for the mixing coefficients, both Douc et al. [46] and Hairer and Mattingly [47] provide a path forward for deriving explicit bounds. We however omit these calculations in the interest of simplicity.

[8] While the burn-in time is polynomial in the problem constants listed in (15), Tsiamis and Pappas [45] show that these constants (specifically $1/\lambda_{\min}(\Gamma_{\kappa-1})$) can scale exponentially in κ, the controllability index of the system.

J General linearized model dynamics

J.1 Comparison to existing results

We first compare the results of Theorem 6.2 to the existing bounds from Kowshik et al. [20], Sattar and Oymak [30], Foster et al. [31]. Before doing so, we note that these existing results bound the loss of specific gradient based algorithms. On the other hand, Theorem 6.2 directly applies to the empirical risk minimizer of the square loss. In general, the LSE optimization problem specialized to this setting is non-convex due to the composition of the square loss with the link function σ. However, we believe it should be possible to show that the quasi-Newton method described in Kowshik et al. [20, Algorithm 1] can be used to optimize the empirical risk to precision of order $\|H\|_{\mathsf{op}}^2 d_{\mathsf{X}}^2/T$, in which case a simple modification of Theorem 4.1 combined with the current analysis in Theorem 6.2 would apply to bound the excess risk of the final iterate of this algorithm. This is left to future work.

For our comparison, we will ignore all logarithmic factors, and assume any necessary burn-in times, remarking that the existing results all prescribe sharper burn-in times than Theorem 6.2. First, we compare with Sattar and Oymak [30, Corollary 6.2]. In doing so, we will assume that $H = (1+\sigma)I$ for some $\sigma > 0$, since this is the setting they study. When H is diagonal, (66) is actually invariant to the noise scale σ which is the correct behavior: $\mathbf{E}\|\widehat{A} - A_\star\|_F^2 \le \tilde{O}(1)\frac{d_{\mathsf{X}}^2}{\zeta^2 T}$. On the other hand, Sattar and Oymak [30, Corollary 6.2] gives a high probability bound of $\|\widehat{A} - A_\star\|_F^2 \le \tilde{O}(1)\frac{\sigma^2 d_{\mathsf{X}}^2}{\zeta^4(1-\rho)^3 T}$. Thus, (66) improves on this rate by not only a factor of $1/\zeta^2$, but also in moving the $1/(1-\rho)$ dependence into the log. We note that their result seems to improve as $\sigma \to 0$, but the probability of success also tends to 0 as $\sigma \to 0$.

Next, we turn our attention to Foster et al. [31, Theorem 2, fast rate]. This result actually gives both in-sample excess risk and parameter recovery bounds. For simplicity, we only compare to the parameter recovery bounds, as this is their sharper result. Their result yields a high probability bound that $\|\widehat{A} - A_\star\|_F^2 \le \tilde{O}(1)\frac{\|H\|_{\mathsf{op}}^2 \|P_\star\|_{\mathsf{op}} d_{\mathsf{X}}^2}{\zeta^4(1-\rho)T}$. Again, we see (66) improve this rate by a factor of $1/\zeta^2$, and moves the dependence on $\|P_\star\|_{\mathsf{op}}$ and $1/(1-\rho)$ into the logarithm. We note again, this rate seems to improve as $\|H\|_{\mathsf{op}} \to 0$, but the number of iterations m of GLMtron needed tends to ∞ as $\|H\|_{\mathsf{op}} \to 0$. We conclude by noting that the rate of Foster et al. [31] does not have any burn-in times.

Finally, we compare to Kowshik et al. [20, Theorem 1]. We will assume that $H = \sigma I$ again, as this is the setting of their work. Their parameter recovery bound states that with high probability, $\|\widehat{A} - A_\star\|_F^2 \le \tilde{O}(1)\frac{\sigma^2 d_{\mathsf{X}}^2}{\zeta^2 T}$, which matches (66) up to the log factors. As noted previously, their logarithmic dependencies are sharper than ours. Furthermore, their result can also handle the unstable regime when $\rho \le 1 + O(1/T)$, which ours cannot. However, we note that Theorem 6.2 also bounds L^2 excess risk with logarithmic dependence on $1/(1-\rho)$, which is not an immediate consequence of parameter error bounds. Indeed, a naïve upper bound using the 1-Lipschitz property of the link function yields: $\|\widehat{f} - f_\star\|_{L^2}^2 \le \|\widehat{A} - A_\star\|_{\mathsf{op}}^2 \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|X_t\|_2^2 \lesssim \|\widehat{A} - A_\star\|_{\mathsf{op}}^2 \frac{1}{(1-\rho)^2}$. Hence, even if the parameter error only depends logarithmically on $1/(1-\rho)$, it does not immediately translate over to excess risk.

J.2 Proof of Theorem 6.2

We first turn our attention to controlling the states $\{X_t\}_{t\ge 0}$ in expectation. Note that the Lyapunov assumption in Assumption 6.1 implies that for every $x \in \mathbb{R}^{d_{\mathsf{X}}}$:

$$\|\sigma(A_\star x)\|_{P_\star}^2 \le \rho \|x\|_{P_\star}^2, \tag{53}$$

and hence the function $x \mapsto \|x\|_{P_\star}^2$ is a Lyapunov function (recall that $P_\star \succcurlyeq I$) which certifies exponential stability to the origin for the deterministic system $x_+ = \sigma(A_\star x)$.

Proposition J.1. *Consider the GLM process $\{X_t\}_{t\ge 0}$ from (16). Under Assumption 6.1:*

$$\sup_{t\in\mathbb{N}} \mathbf{E}\|X_t\|_2^4 \le B_X^4, \;\; B_X \triangleq \frac{12\sqrt{2}\|H\|_{\mathsf{op}}\|P_\star\|_{\mathsf{op}}^{1/2}\sqrt{d_{\mathsf{X}}}}{1-\rho}. \tag{54}$$

Proof. For any $a, b \in \mathbb{R}$ and $\varepsilon > 0$, $(a+b)^4 \le (1+\varepsilon)^3 a^4 + (1+1/\varepsilon)^3 b^4$. Hence for any $\varepsilon > 0$:

$$\begin{aligned}
\mathbf{E}\|X_t\|_{P_\star}^4 &= \mathbf{E}\|\sigma(A_\star X_t) + HV_t\|_{P_\star}^4 \\
&\le (1+\varepsilon)^3 \mathbf{E}\|\sigma(A_\star X_t)\|_{P_\star}^4 + (1+1/\varepsilon)^3 \mathbf{E}\|HV_t\|_{P_\star}^4 \\
&\le (1+\varepsilon)^3 \rho^2 \mathbf{E}\|X_t\|_{P_\star}^4 + (1+1/\varepsilon)^3 \mathbf{E}\|HV_t\|_{P_\star}^4 \qquad \text{using (53)}. \qquad (55)
\end{aligned}$$

Now, we first assume that $\rho \in [1/2, 1)$. For any $\varepsilon \in [0,1]$, we have that $(1+\varepsilon)^3 \le 1 + 12\varepsilon$. Choosing $\varepsilon = \frac{1-\rho^2}{24\rho^2}$, we have that $\varepsilon \le 1$, and therefore continuing from (55):

$$\begin{aligned}
\mathbf{E}\|X_t\|_{P_\star}^4 &\le (1+12\varepsilon)\mathbf{E}\|X_t\|_{P_\star}^4 + (1+1/\varepsilon)^3 \mathbf{E}\|HV_t\|_{P_\star}^4 \\
&= \frac{1+\rho^2}{2}\mathbf{E}\|X_t\|_{P_\star}^4 + \frac{24^3}{(1-\rho^2)^3}\mathbf{E}\|HV_t\|_{P_\star}^4 \\
&\le \frac{1+\rho^2}{2}\mathbf{E}\|X_t\|_{P_\star}^4 + \frac{3\cdot 24^3}{(1-\rho^2)^3}(\mathbf{E}\|HV_t\|_{P_\star}^2)^2 \qquad \text{using Proposition G.5} \\
&\le \frac{1+\rho^2}{2}\mathbf{E}\|X_t\|_{P_\star}^4 + \frac{3\cdot 24^3}{(1-\rho^2)^3}\operatorname{tr}(H^\mathsf{T} P_\star H)^2.
\end{aligned}$$

Unrolling this recursion yields:

$$\mathbf{E}\|X_t\|_{P_\star}^4 \le \frac{6\cdot 24^3}{(1-\rho^2)^4}\operatorname{tr}(H^\mathsf{T} P_\star H)^2.$$

We now handle the case when $\rho \in [0, 1/2)$. Setting $\varepsilon = 2^{1/3} - 1$ and starting from (55):

$$\begin{aligned}
\mathbf{E}\|X_t\|_{P_\star}^4 &\le \frac{1}{2}\mathbf{E}\|X_t\|_{P_\star}^4 + 125\mathbf{E}\|HV_t\|_{P_\star}^4 \\
&\le \frac{1}{2}\mathbf{E}\|X_t\|_{P_\star}^4 + 375\operatorname{tr}(H^\mathsf{T} P_\star H)^2.
\end{aligned}$$

Unrolling this recursion yields:

$$\mathbf{E}\|X_t\|_{P_\star}^4 \le 750\operatorname{tr}(H^\mathsf{T} P_\star H)^2.$$

The claim now follows by taking the maximum of these two bounds and using the inequalities $\operatorname{tr}(H^\mathsf{T} P_\star H) \le \|H\|_{\mathsf{op}}^2 \|P_\star\|_{\mathsf{op}} d_{\mathsf{X}}$ and $1 - \rho^2 \ge 1 - \rho$. ∎

This proof proceeds quite similarly to the linear dynamical systems proof given in Appendix I. We start again by defining the truncated GLM dynamics:

$$\bar{X}_{t+1} = \sigma(A_\star \bar{X}_t) + H\bar{V}_t, \;\; \bar{X}_0 = H\bar{V}_0, \;\; \bar{V}_t = V_t \mathbf{1}\{\|V_t\|_2 \le R\}. \qquad (56)$$

We set $R = \sqrt{d_{\mathsf{X}}} + \sqrt{2(1+\beta)\log T}$ where $\beta \ge 2$ is a free parameter. Define the event $\mathcal{E}$ as:

$$\mathcal{E} := \left\{ \max_{0 \le t \le T-1} \|V_t\|_2 \le R \right\}. \qquad (57)$$

Note that by the setting of R, we have $\mathbf{P}(\mathcal{E}^c) \le 1/T^\beta$ using standard Gaussian concentration results plus a union bound. Furthermore on $\mathcal{E}$, the original GLM process driven by Gaussian noise (16) coincides with the truncated process (56). Let $\widehat{f}$ denote the LSE on the original process (56), and let $\bar{f}$ denote the LSE on the truncated process (56). Hence:

$$\begin{aligned}
\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 &= \mathbf{E}\|\bar{f} - f_\star\|_{L^2}^2 \mathbf{1}\{\mathcal{E}\} + \mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \mathbf{1}\{\mathcal{E}^c\} \\
&\le \mathbf{E}\|\bar{f} - f_\star\|_{L^2}^2 + \mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \mathbf{1}\{\mathcal{E}^c\}. \qquad (58)
\end{aligned}$$

Let us now control the error term $\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \mathbf{1}\{\mathcal{E}^c\}$. Write $\widehat{f}(x) = \sigma(\widehat{A}x)$, and put $\widehat{\Delta} = \widehat{A} - A_\star$. We have:

$$\begin{aligned}
\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \mathbf{1}\{\mathcal{E}^c\} &= \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|\sigma(\widehat{A}X_t) - \sigma(A_\star X_t)\|_2^2 \mathbf{1}\{\mathcal{E}^c\} \overset{(a)}{\le} \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|\widehat{\Delta}X_t\|_2^2 \mathbf{1}\{\mathcal{E}^c\} \\
&\overset{(b)}{\le} \frac{4B^2}{T}\sum_{t=0}^{T-1}\mathbf{E}\|X_t\|_2^2 \mathbf{1}\{\mathcal{E}^c\} \overset{(c)}{\le} \frac{4B^2}{T^{1+\beta/2}}\sum_{t=0}^{T-1}\sqrt{\mathbf{E}\|X_t\|_2^4} \overset{(d)}{\le} \frac{4B^2 B_X^2}{T^{\beta/2}}. \qquad (59)
\end{aligned}$$

Here, (a) follows since σ is 1-Lipschitz, (b) uses the definition of $\mathscr{F}$ in (17), (c) follows by Cauchy-Schwarz, and (d) uses Proposition J.1.

The remainder of the proof is to bound the LSE error $\mathbf{E}\|\bar{f} - f_\star\|_{L^2}^2$. First, we establish an almost sure bound on $\{\bar{X}_t\}_{t\geq 0}$.

Proposition J.2. *Consider the truncated GLM process* (56)*. Under Assumption 6.1, the process* $\{\bar{X}_t\}_{t\geq 0}$ *satisfies:*

$$\sup_{t\in\mathbb{N}} \|\bar{X}_t\|_{P_\star} \leq \frac{2\|P_\star\|_{\mathsf{op}}^{1/2}\|H\|_{\mathsf{op}}(\sqrt{d_{\mathsf{X}}} + \sqrt{2(1+\beta)\log T})}{1-\rho} \triangleq B_{\bar{X}}. \tag{60}$$

Proof. By triangle inequality and (53):

$$\begin{aligned}\|\bar{X}_{t+1}\|_{P_\star} &= \|\sigma(A_\star \bar{X}_t) + H\bar{V}_t\|_{P_\star} \leq \|\sigma(A_\star \bar{X}_t)\|_{P_\star} + \|H\bar{V}_t\|_{P_\star} \\ &\leq \rho^{1/2}\|\bar{X}_t\|_{P_\star} + \|H\bar{V}_t\|_{P_\star} \leq \rho^{1/2}\|\bar{X}_t\|_{P_\star} + \|P_\star^{1/2}H\|_{\mathsf{op}}R.\end{aligned}$$

Unrolling this recursion, and using the fact that $\inf_{x\in[0,1]} \frac{1-\sqrt{x}}{1-x} = 1/2$ yields the result. ■

We next establish uniform bounds for the covariance matrices of the truncated process.

Proposition J.3. *Suppose $T \geq 4$. Consider the truncated GLM process* (56)*, and let the covariance matrices for the process $\{\bar{X}_t\}_{t\geq 0}$ be denoted as $\bar{\Gamma}_t \triangleq \mathbf{E}[\bar{X}_t\bar{X}_t^\mathsf{T}]$. Under Assumption 6.1:*

$$\frac{1}{2}HH^\mathsf{T} \preccurlyeq \bar{\Gamma}_t \preccurlyeq B_{\bar{X}}^2 \cdot I.$$

Proof. The upper bound is immediate from Proposition J.2, since $\mathbf{E}[\bar{X}_t\bar{X}_t^\mathsf{T}] \preccurlyeq \mathbf{E}[\|\bar{X}_t\|_2^2]I \preccurlyeq B_{\bar{X}}^2 I$. For the lower bound, it is immediate when $t = 0$ using Proposition G.4. On the other hand, for $t \geq 1$, since $\bar{V}_t$ is zero-mean:

$$\begin{aligned}\mathbf{E}[\bar{X}_t\bar{X}_t^\mathsf{T}] &= \mathbf{E}[(\sigma(A_\star\bar{X}_{t-1}) + H\bar{V}_{t-1})(\sigma(A_\star\bar{X}_{t-1}) + H\bar{V}_{t-1})^\mathsf{T}] \\ &= \mathbf{E}[\sigma(A_\star\bar{X}_{t-1})\sigma(A_\star\bar{X}_{t-1})^\mathsf{T}] + \mathbf{E}[H\bar{V}_{t-1}\bar{V}_{t-1}^\mathsf{T}H^\mathsf{T}] \succcurlyeq \mathbf{E}[H\bar{V}_{t-1}\bar{V}_{t-1}^\mathsf{T}H^\mathsf{T}] \succcurlyeq \frac{1}{2}HH^\mathsf{T}.\end{aligned}$$

The last inequality again holds from Proposition G.4. ■

J.2.1 Trajectory hypercontractivity for truncated GLM

For our purposes, the link function assumption in Assumption 6.1 ensures the following approximate isometry inequality which holds for all $x \in \mathbb{R}^{d_{\mathsf{X}}}$ and all matrices $A, A' \in \mathbb{R}^{d_{\mathsf{X}}\times d_{\mathsf{X}}}$:

$$\zeta^2\|Ax - A'x\|_2^2 \leq \|\sigma(Ax) - \sigma(A'x)\|_2^2 \leq \|Ax - A'x\|_2^2. \tag{61}$$

This inequality is needed to establish trajectory hypercontractivity for $\mathscr{F}_\star$.

Proposition J.4. *Suppose that $T \geq 4$. Fix any matrix $A \in \mathbb{R}^{d_{\mathsf{X}}\times d_{\mathsf{X}}}$. Under Assumption 6.1, the truncated process* (56) *satisfies:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|\sigma(A\bar{X}_t) - \sigma(A_\star\bar{X}_t)\|_2^4 \leq \frac{4B_{\bar{X}}^4}{\sigma_{\min}(H)^4\zeta^4}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|\sigma(A\bar{X}_t) - \sigma(A_\star\bar{X}_t)\|_2^2\right)^2. \tag{62}$$

Hence, the function class $\mathscr{F}_\star$ with $\mathscr{F}$ defined in (17) *satisfies the $(C_{\mathsf{GLM}}, 2)$-trajectory hypercontractivity condition with $C_{\mathsf{GLM}} = \frac{4B_{\bar{X}}^4}{\sigma_{\min}(H)^4\zeta^4}$.*

Proof. Put $\Delta \triangleq A - A_\star$ and $M \triangleq \Delta^\mathsf{T}\Delta$. We have:

$$\begin{aligned}\mathbf{E}\|\Delta\bar{X}_t\|_2^4 &= \mathbf{E}[\bar{X}_t^\mathsf{T}M\bar{X}_t\bar{X}_t^\mathsf{T}M\bar{X}_t] \\ &\leq B_{\bar{X}}^2\operatorname{tr}(M^2\bar{\Gamma}_t) && \text{using Proposition J.2} \\ &\leq B_{\bar{X}}^2\|M\|_{\mathsf{op}}\operatorname{tr}(M\bar{\Gamma}_t) && \text{Hölder's inequality} \\ &\leq B_{\bar{X}}^2\operatorname{tr}(M)\operatorname{tr}(M\bar{\Gamma}_t) && \text{since } M \text{ is positive semidefinite} \\ &\leq B_{\bar{X}}^4\operatorname{tr}(M)^2 && \text{using Proposition J.3} \\ &\leq \frac{B_{\bar{X}}^4}{\lambda_{\min}(HH^\mathsf{T})^2}\operatorname{tr}(MHH^\mathsf{T})^2.\end{aligned}$$

On the other hand, by Proposition J.3:

$$\mathbf{E}\|\Delta \bar{X}_t\|_2^2 = \operatorname{tr}(M\bar{\Gamma}_t) \geq \frac{1}{2}\operatorname{tr}(MHH^\mathsf{T}).$$

Combining these bounds yields:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|\Delta \bar{X}_t\|_2^4 \leq \frac{B_{\bar{X}}^4}{\lambda_{\min}(HH^\mathsf{T})^2}\operatorname{tr}(MHH^\mathsf{T})^2 \leq \frac{4B_{\bar{X}}^4}{\lambda_{\min}(HH^\mathsf{T})^2}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|\Delta \bar{X}_t\|_2^2\right)^2.$$

The claim now follows via the approximate isometry inequality (61). ■

J.2.2 Bounding the dependency matrix for truncated GLM

We will use the result in Lemma G.1 to bound the total-variation distance by the 1-Wasserstein distance. This is where the non-degenerate noise assumption in Assumption 6.1 is necessary.

The starting point is the observation that the diagonal Lyapunov function in Assumption 6.1 actually yields *incremental stability* [48] in addition to Lyapunov stability. In particular, let $\{a_i\}$ denote the rows of $A_\star$. For any $x, x' \in \mathbb{R}^{d_\mathsf{X}}$:

$$\begin{aligned}
\|\sigma(A_\star x) - \sigma(A_\star x')\|_{P_\star}^2 &= \sum_{i=1}^{d_\mathsf{X}} (P_\star)_{ii}(\sigma(\langle a_i, x\rangle) - \sigma(\langle a_i, x'\rangle))^2 \\
&\leq \sum_{i=1}^{d_\mathsf{X}} (P_\star)_{ii}(\langle a_i, x\rangle - \langle a_i, x'\rangle)^2 \\
&= (x-x')^\mathsf{T} A_\star^\mathsf{T} P_\star A_\star (x-x') \\
&\leq \rho\|x-x'\|_{P_\star}^2. \qquad (63)
\end{aligned}$$

This incremental stability property allows us to control the dependency matrix as follows.

Proposition J.5. *Consider the truncated GLM process $\{\bar{X}_t\}_{t\geq 0}$ from (56). Let $\mathsf{P}_{\bar{X}}$ denote the joint distribution of $\{\bar{X}_t\}_{t=0}^{T-1}$. Under Assumption 6.1 and when $B \geq 1$, we have that:*

$$\|\Gamma_{\mathsf{dep}}(\mathsf{P}_{\bar{X}})\|_{\mathsf{op}} \leq \frac{22}{1-\rho}\log\left(\frac{B\sqrt{d_\mathsf{X}}(B_{\bar{X}} + B_X)}{2\sigma_{\min}(H)}\right).$$

Proof. Let $\{X_t\}_{t\geq 0}$ denote the original GLM dynamics from (16). Fix indices $t \geq 0$ and $k \geq 1$. We construct a coupling of $(\mathsf{P}_{X_{t+k}}(\cdot \mid X_t = x), \mathsf{P}_{X_{t+k}})$ as follows. Let $\{V_t\}_{t\geq 0}$ be iid $N(0, I)$. Let $\{Z_s\}_{s\geq t}$ be the process such that $Z_t = x$, and follows the GLM dynamics (16) using the noise $\{V_t\}_{t\geq 0}$ (we do not bother defining $Z_{t'}$ for $t' < t$ since we do not need it). Similarly, let $\{Z'_s\}_{s\geq 0}$ be the process following the GLM dynamics (16) using the same noise $\{V_t\}_{t\geq 0}$. Now we have:

$$\begin{aligned}
\mathbf{E}\|Z_{t+k} - Z'_{t+k}\|_{P_\star} &= \mathbf{E}\|\sigma(A_\star Z_{t+k-1}) - \sigma(A_\star Z'_{t+k-1})\|_{P_\star} \\
&\leq \rho^{1/2}\mathbf{E}\|Z_{t+k-1} - Z'_{t+k-1}\|_{P_\star} \qquad \text{using Equation (63).}
\end{aligned}$$

We now unroll this recursion down to t:

$$\mathbf{E}\|Z_{t+k} - Z'_{t+k}\|_{P_\star} \leq \rho^{k/2}\mathbf{E}\|Z_t - Z'_t\|_{P_\star} = \rho^{k/2}\mathbf{E}\|x - Z'_t\|_{P_\star}.$$

Since $P_\star \succcurlyeq I$, this shows that:

$$W_1(\mathsf{P}_{X_{t+k}}(\cdot \mid X_t = x), \mathsf{P}_{X_{t+k}}) \leq \rho^{k/2}(\|x\|_{P_\star} + \mathbf{E}\|X_t\|_{P_\star}) \leq \rho^{k/2}(\|x\|_{P_\star} + B_X),$$

where the last inequality follows from Proposition J.1 and Jensen's inequality.

Next, it is easy to see the map $x \mapsto \sigma(A_\star x)$ is $\|A\|_{\mathsf{op}}$-Lipschitz. Furthermore, since H is full rank by Assumption 6.1, then for any t and $k \geq 1$ both P_t and $\mathsf{P}_{X_{t+k}}(\cdot \mid X_t = x)$ are absolutely continuous w.r.t. the Lebesgue measure in $\mathbb{R}^{d_\mathsf{X}}$. Using Lemma G.1, we have for any $k \geq 2$:

$$\begin{aligned}
\|\mathsf{P}_{X_{t+k}}(\cdot \mid X_t = x) - \mathsf{P}_{X_{t+k}}\|_{\mathsf{TV}} &\leq \frac{\|A_\star\|_{\mathsf{op}}\sqrt{\operatorname{tr}((HH^\mathsf{T})^{-1})}}{2} W_1(\mathsf{P}_{X_{t+k-1}}(\cdot \mid X_t = x), \mathsf{P}_{X_{t+k-1}}) \\
&\leq \frac{\|A_\star\|_{\mathsf{op}}\sqrt{\operatorname{tr}((HH^\mathsf{T})^{-1})}}{2}\rho^{(k-1)/2}(\|x\|_{P_\star} + B_X).
\end{aligned}$$

Using Proposition G.2 to bound $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_{\bar{X}})\|_{\mathsf{op}}$ (which is valid because we constrained $\beta \geq 2$), and Proposition J.2 to bound $x \in \bar{\mathsf{X}}_t$, for any $\ell \geq 1$:

$$\begin{aligned}\|\Gamma_{\mathsf{dep}}(\mathsf{P}_{\bar{X}})\|_{\mathsf{op}} &\leq 3 + \sqrt{2}\sum_{k=1}^{T-1} \max_{t=0,\dots,T-1-k} \operatorname*{ess\,sup}_{x \in \bar{\mathsf{X}}_t} \sqrt{\|\mathsf{P}_{X_{t+k}}(\cdot \mid X_t = x) - \mathsf{P}_{X_{t+k}}\|_{\mathsf{TV}}} \\ &\leq 3 + \sqrt{2}\ell + \left[\frac{\|A_\star\|_{\mathsf{op}}\sqrt{\operatorname{tr}((HH^\mathsf{T})^{-1})}(B_{\bar{X}} + B_X)}{2}\right]^{1/2} \sum_{k=\ell+1}^{T-1} \rho^{(k-1)/4} \\ &\overset{(a)}{\leq} 5\ell + \left[\frac{B\sqrt{d_\mathsf{X}}(B_{\bar{X}} + B_X)}{2\sigma_{\min}(H)}\right]^{1/2} \frac{\rho^{\ell/4}}{1-\rho^{1/4}}.\end{aligned}$$

Above, (a) uses the bounds $\|A_\star\|_{\mathsf{op}} \leq B$ and $\operatorname{tr}(HH^{-1}) \leq d_\mathsf{X}/\sigma_{\min}(H)^2$. Now put $\psi \triangleq \frac{B\sqrt{d_\mathsf{X}}(B_{\bar{X}}+B_X)}{2\sigma_{\min}(H)}$. We choose $\ell = \left\lceil \frac{\log(\sqrt{\psi})}{1-\rho^{1/4}} \right\rceil$ so that $\rho^{\ell/4} \leq 1/\sqrt{\psi}$. This yields:

$$\|\Gamma_{\mathsf{dep}}(\mathsf{P}_{\bar{X}})\|_{\mathsf{op}} \leq \frac{11\log\psi}{2(1-\rho^{1/4})} \overset{(a)}{\leq} \frac{22\log\psi}{1-\rho} = \frac{22}{1-\rho}\log\left(\frac{B\sqrt{d_\mathsf{X}}(B_{\bar{X}}+B_X)}{2\sigma_{\min}(H)}\right).$$

Above, (a) follows from $\inf_{x\in[0,1]} \frac{1-x^{1/4}}{1-x} = 1/4$. ■

J.2.3 Finishing the proof of Theorem 6.2

Below, we let c_i be universal positive constants that we do not track precisely.

For any $\varepsilon > 0$ we now construct an ε-covering of $\mathscr{F}_\star \setminus B(r)$, with $\mathscr{F}_\star$ the offset class of $\mathscr{F}$ from (17). Note that we are not covering $\partial B(r)$ since the class $\mathscr{F}_\star$ is not star-shaped. However, an inspection of the proof of Theorem B.2 shows that one can remove the star-shaped assumption by instead covering the set $\mathscr{F}_\star \setminus B(r)$. To this end, we let $\{A_1,\dots,A_N\}$ be a δ-cover of $\mathscr{A} \triangleq \{A \in \mathbb{R}^{d_\mathsf{X}\times d_\mathsf{X}} \mid \|A\|_F \leq B\}$, for a δ to be specified. By a volumetric argument we may choose $\{A_1,\dots,A_N\}$ such that $N \leq \left(1+\frac{2B}{\delta}\right)^{d_\mathsf{X}^2}$. Now, any realization of $\{\bar{X}_t\}$ will have norm less than $B_{\bar{X}}$ from (60), where $B_{\bar{X}}$ is bounded by:

$$B_{\bar{X}} \leq \frac{c_0\|P_\star\|_{\mathsf{op}}^{1/2}\|H\|_{\mathsf{op}}(\sqrt{d_\mathsf{X}} + \sqrt{(1+\beta)\log T})}{1-\rho}.$$

Now fix any $A \in \mathscr{A}$, and let A_i be an element in the δ-cover satisfying $\|A - A_i\|_F \leq \delta$. We observe that for any x satisfying $\|x\|_2 \leq B_{\bar{X}}$:

$$\begin{aligned}\|(\sigma(A_i x) - \sigma(A_\star x)) - (\sigma(Ax) - \sigma(A_\star x))\|_2 &= \|\sigma(A_i x) - \sigma(Ax)\|_2 \leq \|(A_i - A)x\|_2 \\ &\leq \|A_i - A\|_F\|x\|_2 \leq \delta B_{\bar{X}}.\end{aligned}$$

Thus, it suffices to take $\delta = \varepsilon/B_{\bar{X}}$ to construct the ε cover of $\mathscr{F}_\star$, i.e., $\mathcal{N}_\infty(\mathscr{F}_\star,\varepsilon) \leq \left(1+\frac{2BB_{\bar{X}}}{\varepsilon}\right)^{d_\mathsf{X}^2}$. This then implies [see e.g. 39, Example 4.2.10]:

$$\mathcal{N}_\infty(\mathscr{F}_\star \setminus B(r),\varepsilon) \leq \mathcal{N}_\infty(\mathscr{F}_\star,\varepsilon/2) \leq \left(1+\frac{4BB_{\bar{X}}}{\varepsilon}\right)^{d_\mathsf{X}^2}.$$

Next, by Proposition J.4, $(\mathscr{F}_\star, \mathsf{P}_{\bar{X}})$ is $(C_{\mathsf{GLM}}, 2)$-hypercontractive for all $T \geq 4$, where

$$C_{\mathsf{GLM}} \leq \frac{4B_{\bar{X}}^4}{\sigma_{\min}(H)^4\zeta^4} \leq \frac{c_1\|P_\star\|_{\mathsf{op}}^2 \operatorname{cond}(H)^4(d_\mathsf{X}^2 + ((1+\beta)\log T)^2)}{\zeta^4(1-\rho)^4}.$$

Furthermore, by Proposition J.5:

$$\|\Gamma_{\mathsf{dep}}(\mathsf{P}_{\bar{X}})\|_{\mathsf{op}}^2 \leq \frac{c_2}{(1-\rho)^2}\log^2\left(\frac{B\sqrt{d_\mathsf{X}}(B_{\bar{X}}+B_X)}{2\sigma_{\min}(H)}\right) \triangleq \gamma^2.$$

The class $\mathscr{F}_\star$ is $2BB_{\bar{X}}$-bounded on (56). Invoking Theorem 4.1, for every $r > 0$:

$$\mathbf{E}\|\bar{f} - f_\star\|_{L^2}^2 \leq 8\mathbf{E}\bar{\mathsf{M}}_T(\mathscr{F}_\star) + r^2 + 4B^2B_{\bar{X}}^2\left(1+\frac{4\sqrt{8}BB_{\bar{X}}}{r}\right)^{d_\mathsf{X}^2}\exp\left(\frac{-T}{8C_{\mathsf{GLM}}\gamma^2}\right). \tag{64}$$

Here, the notation $\mathbf{E}\bar{\mathsf{M}}_T(\mathscr{F}_\star)$ is meant to emphasize that the offset complexity is with respect to the truncated process $\mathsf{P}_{\bar{X}}$ and *not* the original process P_X. We now set $r^2 = \|H\|_{\mathsf{op}}^2 d_{\mathsf{X}}^2/T$, and compute a T_0 such that the third term in (64) is also bounded by $\|H\|_{\mathsf{op}}^2 d_{\mathsf{X}}^2/T$. To do this, it suffices to compute T_0 such that for all $T \geq T_0$:

$$T \geq c_3 C_{\mathsf{GLM}} \gamma^2 d_{\mathsf{X}}^2 \log\left(\frac{TBB_{\bar{X}}}{\|H\|_{\mathsf{op}}\sqrt{d_{\mathsf{X}}}}\right).$$

It suffices to take (assuming β is at most polylogarithmic in any problem constants):

$$T_0 \geq c_4 \frac{\|P_\star\|_{\mathsf{op}}^2 \mathrm{cond}(H)^4 d_{\mathsf{X}}^4}{\zeta^4(1-\rho)^6} \mathrm{polylog}\left(B, d_{\mathsf{X}}, \|P_\star\|_{\mathsf{op}}, \mathrm{cond}(H), \frac{1}{\zeta}, \frac{1}{1-\rho}\right). \tag{65}$$

Again, we do not attempt to compute the exact power of the polylog term, but note it can in principle be done via Du et al. [43, Lemma F.2].

Next, from (59) we have that the error term $\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \mathbf{1}\{\mathcal{E}^c\} \leq \frac{4B^2B_X^2}{T^{\beta/2}}$. Thus if we further constrain $\beta > 2$ and require $T_0 \geq c_5 \left[\frac{B^2\|P_\star\|_{\mathsf{op}}}{(1-\rho)^2}\right]^{\frac{1}{\beta/2-1}}$, then $\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \mathbf{1}\{\mathcal{E}^c\} \leq \frac{\|H\|_{\mathsf{op}}^2 d_{\mathsf{X}}^2}{T}$. Note that setting $\beta = \max\{3, c_6 \log B\}$ suffices.

To finish the proof, it remains to bound $\mathbf{E}\bar{\mathsf{M}}_T(\mathscr{F}_\star)$. Now, unlike the linear dynamical systems case, there is no closed-form expression for $\mathbf{E}\bar{M}_T(\mathscr{F}_\star)$. Hence, we will bound it via the chaining bound (7). This computation is done in Ziemann et al. [15, Example 3]. Before we can use the result, however, we need to verify that the truncated noise process $\{H\bar{V}_t\}_{t\geq 0}$ is a sub-Gaussian MDS. The MDS part is clear since $\bar{V}_t \perp \bar{V}_{t'}$ for $t \neq t'$, and $\bar{V}_t$ is zero-mean. Furthermore, Proposition G.3 yields that $H\bar{V}_t$ is a $4\|H\|_{\mathsf{op}}^2$-sub-Gaussian random vector. Hence, we have:

$$\mathbf{E}\bar{\mathsf{M}}_T(\mathscr{F}_\star) \leq c_7 \frac{\|H\|_{\mathsf{op}}^2 d_{\mathsf{X}}^2}{T} \log(1 + \|H\|_{\mathsf{op}}\sqrt{d_{\mathsf{X}}} BB_{\bar{X}} T^2).$$

The claim now follows.

J.3 Further discussion related to Theorem 6.2

Several remarks on Assumption 6.1 are in order. First, the rank condition on H ensures that the noise process $\{HV_t\}_{t\geq 0}$ is non-degenerate. Viewing (16) as a control system mapping $\{V_t\}_{t\geq 0} \mapsto \{X_t\}_{t\geq 0}$, this condition ensures that this system is one-step controllable. Next, the link function assumption is standard in the literature (see e.g. Kowshik et al. [20], Sattar and Oymak [30], Foster et al. [31]). The expansiveness condition $|\sigma(x) - \sigma(y)| \geq \zeta|x-y|$ ensures that the link function is increasing at a uniform rate. For efficient parameter recovery, some extra assumption other than Lipschitzness and monotonicity is needed [20, Theorem 4], and expansiveness yields a sufficient condition. However, for excess risk, it is unclear if any extra requirements are necessary. We leave resolving this issue to future work. Finally, the Lyapunov stability condition is due to Foster et al. [31, Proposition 2], and yields a certificate for global exponential stability (GES) to the origin. It is weaker than requiring that $\|A_\star\|_{\mathsf{op}} < 1$, which amounts to taking $P_\star = I$. The assumption $P_\star \succcurlyeq I$ is without loss of generality by rescaling $P_\star$.

Theorem 6.2 states that after a polynomial burn-in time (which scales quite sub-optimally as $\tilde{O}(d_{\mathsf{X}}^4)$ in the dimension), the excess risk scales as the minimax rate $\|H\|_{\mathsf{op}}^2 d_{\mathsf{X}}^2/T$ times a logarithmic factor of various problem constants. To the best of our knowledge, this is the sharpest excess risk bound for this problem in the literature and is nearly minimax optimal. As noted previously, the logarithmic factor enters via the chaining inequality (7) when bounding the martingale offset complexity. We leave to future work a more refined analysis that removes this logarithmic dependence, and also improves the polynomial dependence of T_0 on d_{X}. The extra d_{X}^2 factor in (19) over the LDS burn-in time (15) comes from our analysis of the trajectory hypercontractivity constant for this problem, and should be removable.

Much like in (52), we can use the link function expansiveness in Assumption 6.1 to convert the excess risk bound (18) to a parameter recovery rate:

$$\mathbf{E}\|\hat{A} - A_\star\|_F^2 \leq \tilde{O}(1) \frac{\|H\|_{\mathsf{op}}^2 d_{\mathsf{X}}^2}{\zeta^2 T \lambda_{\min}(\bar{\Gamma}_T)}, \tag{66}$$

where again $\bar{\Gamma}_T \triangleq \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}[X_t X_t^\mathsf{T}]$ is the average covariance matrix of the GLM process (16). Note that the one-step controllability assumption in Assumption 6.1 ensures that the covariance matrix $\mathbf{E}[X_t X_t^\mathsf{T}] \succcurlyeq HH^\mathsf{T}$ is invertible for every $t \in \mathbb{N}$. A detailed comparison of the excess risk rate (18) and parameter recovery rate (66) with existing bounds in the literature is given in Appendix J.1.

Let us briefly discuss the proof of Theorem 6.2. As in the LDS case, we use the truncation argument described in Appendix B.1 that allows us apply Theorem 4.1 while still using bounds on the dependency matrix coefficients of the original unbounded process (16). However, an additional complication arises compared to the LDS case, as the covariates are not jointly Gaussian due to the presence of the link function. While at this point we could appeal to ergodic theory, we instead develop an alternative approach that still allows us to compute explicit constants. Building on the work of Chae and Walker [42], we use the smoothness of the Gaussian transition kernel to upper bound the TV distance by the 1-Wasserstein distance. This argument is where our analysis crucially relies on the non-degeneracy of H in Assumption 6.1, as the transition kernel corresponding to multiple steps of (16) is no longer Gaussian. The 1-Wasserstein distance is then controlled by using the *incremental stability* [48] properties of the deterministic dynamics $x_+ = \sigma(A_\star x)$. Since this technique only depends on the GLM dynamics through incremental stability, it is of independent interest as it applies much more broadly.