Learning from many trajectories

Stephen Tu[1], Roy Frostig[1], and Mahdi Soltanolkotabi[2]

[1]Google Research, Brain team
[2]University of Southern California

April 1, 2022. Revised: February 1, 2023

Abstract

We initiate a study of supervised learning from many independent sequences ("trajectories") of non-independent covariates, reflecting tasks in sequence modeling, control, and reinforcement learning. Conceptually, our multi-trajectory setup sits between two traditional settings in statistical learning theory: learning from independent examples and learning from a single auto-correlated sequence. Our conditions for efficient learning generalize the former setting—trajectories must be non-degenerate in ways that extend standard requirements for independent examples. Notably, we do not require that trajectories be ergodic, long, nor strictly stable.

For linear least-squares regression, given n-dimensional examples produced by m trajectories, each of length T, we observe a notable change in statistical efficiency as the number of trajectories increases from a few (namely $m \lesssim n$) to many (namely $m \gtrsim n$). Specifically, we establish that the worst-case error rate of this problem is $\Theta(n/mT)$ whenever $m \gtrsim n$. Meanwhile, when $m \lesssim n$, we establish a (sharp) lower bound of $\Omega(n^2/m^2T)$ on the worst-case error rate, realized by a simple, marginally unstable linear dynamical system. A key upshot is that, in domains where trajectories regularly reset, the error rate eventually behaves as if *all* of the examples were independent, drawn from their marginals. As a corollary of our analysis, we also improve guarantees for the linear system identification problem.

Contents

1 Introduction

Statistical learning theory aims to characterize the worst-case efficiency of learning from example data. Its most common setup assumes that examples are independently and identically distributed (*iid*) draws from an underlying data distribution, but various branches of theory—not to mention deployed applications of machine learning—consume non-independent data as well. An especially fruitful setting, and the focus of this paper, is in learning from sequential data, where examples are generated by some ordered stochastic process that renders them possibly correlated. Naturally, sequential processes describe application domains spanning engineering and the sciences, such as robotics [NTP11], data center cooling (e.g. [LBL+18]), language (e.g. [SVL14, BK15]), neuroscience (e.g. [LJM+17, GWC+20]), and economic forecasting [MSS17]. Learning over sequential data can also capture some formulations of imitation learning [OPN+18] and reinforcement learning [CLR+21, JLL21].

In supervised learning, one learns to predict output *labels* from input *covariates*, given example pairings of the two. Formal treatments of learning from sequential data typically concern a *single* inter-dependent chain of covariates. Where these treatments vary is in their assumptions about the underlying process that generates the covariate chain. For instance, some assume that the process is auto-regressive (e.g. [LW83, GZ01, GR20]) or ergodic (e.g. [Yu94, DAJJ12]). Others assume that it is a linear dynamical system (e.g. [SMT+18, FTM18, SR19]).

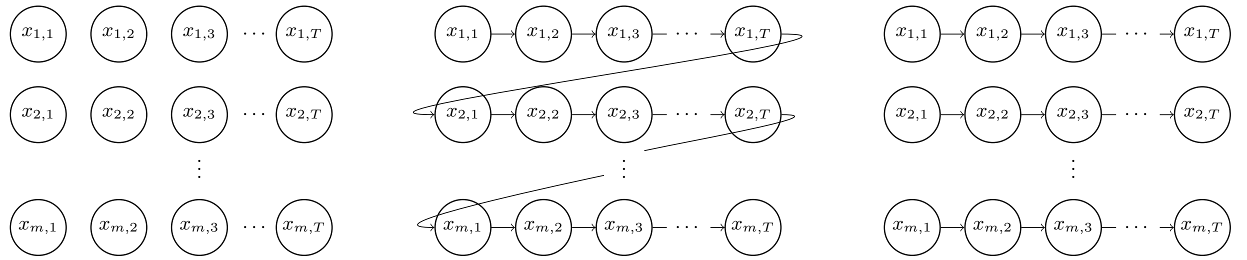
In this paper, we examine what happens when we learn from *many* independent chains rather than from one, as one does anyway in many applications (e.g. [Pom89, KZB11, BPX+07, JVS+16]). Figure 1 depicts the data dependence structure of our setup in comparison with its two natural counterparts. Learning from a dataset of many short (constant length) chains ought to be similar to independent learning, even if each chain is highly intra-dependent. On the other hand, for any non-trivial chain length, intuition suggests that the error can degrade relative to the total sample size in the worst case, since a greater proportion of the data may contain correlations. Lower bounds even show that, when one sees only a single chain, this degradation is outright necessary in the worst case [BJN+20]. Do we see any such effect with many chains?

We study this question by sharply characterizing worst-case error rates of a fundamental task—linear regression—imposed over a general sequential data model. Our findings reveal a remarkable phenomenon: after seeing sufficiently many chains (m) relative to the example dimension n, no matter the chain length T, *the error rate matches that of learning from the same total number mT of independent examples*, drawn from their respective marginal distributions.

In our data model, each chain, called a *trajectory*, comprises a sequence of covariates $\{x_t\}$ generated from a stochastic process. Each covariate is accompanied by a noisy linear response y_t as its label. A training set $\{(x_t^{(i)}, y_t^{(i)})\}_{i=1,t=1}^{m,T}$ comprises m independent chains, each of length T. From such a training set, an estimator produces a hypothesis that predicts the label of any covariate. The resulting hypothesis is evaluated according to its mean-squared prediction error over a fresh chain of length T', possibly unequal to T—a notion of risk defined naturally over a trajectory. All of our risk upper bounds are guarantees for the ordinary least-squares estimator in particular.

A concrete, recurring example in this paper takes the covariate-generating process to be a linear dynamical system (LDS). Specifically, fixing matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times d}$, and $W_\star \in \mathbb{R}^{p\times n}$, a single trajectory $\{(x_t, y_t)\}_{t\geqslant 1}$ is generated as follows. Let $x_0 = 0$, and for $t \geqslant 1$:

$$\begin{aligned} x_t &= Ax_{t-1} + Bw_t, && \text{(linear dynamics)} \\ y_t &= W_\star x_t + \xi_t, && \text{(linear regression)} \end{aligned}$$



(a) independent covariates **(b)** one long trajectory **(c)** independent trajectories

Figure 1: The covariate dependence structure induced by three data models on mT many training examples. In (a): independent examples, typical of basic statistical learning. In (b): the data models often considered in the sequential learning literature, comprising a long auto-correlated chain of examples. Learning in this setting can be infeasible in general, so oftentimes ergodicity is assumed in order to rule out strong long-range dependencies, essentially inducing an "independent resetting" effect across time. The effective reset frequency then factors uniformly into error bounds, in a way suggesting that one learns only one independent example's worth within each effective reset window (cf. Section 2). In (c): our multi-trajectory data model. Our accompanying assumptions allow for non-ergodic chains, and for arbitrary chain lengths T, while introducing *explicit* independent resets. Decoupling the m resets from the sequential data model lets us vary the training set dimensions (m, T) freely, without affecting other data assumptions, as we study their effect on error rates. We find that with enough trajectories m, the worst-case error rate behaves the same as in the independent setting depicted in (a); i.e., one learns as though every example were independently drawn from its marginal distribution. Some recent work in system identification assumes a data model related to ours (specifically linear dynamical data) and likewise avoids ergodicity; our bounds improve these guarantees T-fold where applicable, and upgrade the regimes in which they apply (cf. Section 2).

where the $\{w_t\}_{t\geqslant 1}$ are iid centered isotropic Gaussian draws and $\{\xi_t\}_{t\geqslant 1}$ is a sub-Gaussian martingale difference sequence (with respect to past covariates $\{x_k\}_{k=1}^{t}$ and noise variables $\{\xi_k\}_{k=1}^{t-1}$). Incidentally, combining linear dynamical systems with linear regression captures the basic problem of linear system identification (as in [SMT+18]) as a special case.

In other instantiations of learning from trajectories, the covariates $\{x_t\}$ may be generated by a different process; what remains common is the superimposed regression task set up by the ground truth $W_\star$ and the noise $\{\xi_t\}$. The key condition that we will introduce, which renders a covariate process amenable to regression, is that it satisfies a *trajectory small-ball* criterion (Definition 4.1). Section 4.1 shows that LDS-generated data conforms to the trajectory small-ball condition in particular, as do many other distributions.

Our main results (Sections 5 and 6) sharply characterize worst-case rates of learning from trajectory data as a function of the training trajectory count m, the training trajectory length T, the evaluation length T', the covariate and response dimensions n and p, and scale parameters of noise in the data model (such as the variance of the noise $\{\xi_t\}$). Restricting only to terms of covariate dimension n, training set size m and T, and evaluation length T', our bounds imply the following summary statement:

Theorem 1.1 (informal; error rate with many small-ball trajectories, $T' \leqslant T$)**.** *If $m \gtrsim n$, $T' \leqslant T$, and covariate trajectories are drawn from a* trajectory small-ball *distribution, then the worst-case excess prediction risk (over evaluation horizon T') for linear regression from m many trajectories of n-dimensional covariates, each of length T, is $\Theta(n/(mT))$.*

In drawing comparisons to learning from independent examples, it makes sense to consider training and evaluations lengths T and T' equal (cf. Section 3), rendering Theorem 1.1 applicable. The theorem thus echoes our main point above: the same rate of $\Theta(n/(mT))$ describes regression on mT independent examples (details on this point are expanded in Section 3.3).

Further structural assumptions are needed (cf. Section 3.3) in order to cover the remaining range of problem dimensions, namely few trajectories ($m \lesssim n$) or extended evaluations ($T' > T$), and to that end we return to linear dynamical systems as a focus. Our remaining risk upper bounds, targeting learning under linear dynamics, require that the dynamics matrix A be *marginally unstable* (meaning that its spectral radius $\rho(A)$ is at most one) and diagonalizable. When trajectories are longer at test time than during training (i.e., $T' > T$), marginal instability is practically necessary, otherwise the risk can scale exponentially in $T' - T$. The assumption otherwise still allows for unstable—and therefore non-ergodic—systems at $\rho(A) = 1$. For simplicity, we also require that the control matrix B have full row rank. Our bounds then imply the following summary statement about regression when the number of trajectories is limited:

Theorem 1.2 (informal; error rate with few LDS trajectories). *If $m \lesssim n$, $mT \gtrsim n$, and covariate trajectories are drawn from a linear dynamical system whose dynamics A are marginally unstable and diagonalizable, then the worst-case excess prediction risk (over evaluation horizon T') for linear regression from m many trajectories of n-dimensional covariates, each of length T, is $\tilde{\Theta}(n/(mT) \cdot \max\{nT'/(mT), 1\})$.*

If the evaluation horizon T' is a constant, the rate in Theorem 1.2 recovers that of Theorem 1.1, up to log factors and extra assumptions. To draw further comparison, suppose that the training and evaluation horizons are equal, i.e., that $T' = T$. On the face of it, the rate in Theorem 1.2 is evidently weaker than that of Theorem 1.1, by up to a factor of the covariate dimension n. But the varying premises—of many vs. few trajectories—necessarily constrain the risk definitions to differ. Under a fixed data budget $N := mT = mT'$, fewer trajectories m imply a longer horizon T' over which the risk is evaluated. Intuitively, a longer evaluation horizon makes for a different problem, and renders the rate comparison invalid.

A more sound comparison across regimes is possible by first normalizing the notion of performance within a problem instance. To this end, we can consider the worst-case risk of learning from trajectories *relative* to that of learning from independent examples *in the same regime*. Constructing the latter baseline is somewhat subtle (cf. Section 3.2). To decorrelate the problem of learning from trajectories while maintaining its temporal structure otherwise, we can imagine drawing from its marginal distributions independently at each time step. The resulting dataset is independent, but not identically distributed. Although the rates for the sequential and decorrelated regression problems are—as already highlighted—remarkably the same under many trajectories, the few-trajectory rate in Theorem 1.2 is indeed weaker than the $\Theta(n/(mT))$ rate that we prove for its decorrelated baseline (cf. Theorem 5.7).

Since the more general Theorem 1.1 already describes what happens under many trajectories ($m \gtrsim n$) and a strict evaluation horizon ($T' \leqslant T$), what remains is a somewhat niche regime: many trajectories and an extended evaluation horizon $T' > T$. For completeness, our bounds supply the following summary statement:

Theorem 1.3 (informal; error rate with many LDS trajectories). *If $m \gtrsim n$ and covariate trajectories are drawn from a linear dynamical system whose dynamics A are marginally unstable and diagonalizable, then the worst-case excess prediction risk (over evaluation horizon T')*

for linear regression from m many trajectories of n-dimensional covariates, each of length T, is $\Theta(n/(mT) \cdot \max\{T'/T, 1\})$.

Using the tools of our analysis, we also develop upper bounds for parameter error instead of prediction risk, which inform recovery of the ground truth $W_\star$ and (by reduction) of the dynamics matrix A in LDS. The latter captures the linear system identification problem. Our upper bounds improve on its worst-case guarantees by a factor of $1/T$ where applicable, and extend the parameter ranges in which guarantees hold at all.

2 Related work

Linear regression is a basic and well-studied problem. The two treatments most closely related to our work are [HKZ14] and [Mou22], who develop sharp finite-sample characterizations of the risk of random design linear regression (i.e., from iid examples). Discussion and references therein cover the broader problem over its long history.

A common approach to studying dependent covariates is to assume that the data-generating process is ergodic (see e.g. [Yu94, Mei00, MR08, SC09, MR10, DAJJ12, KM17, MSS17, Sha21] and references therein). The key phenomenon at play is that N correlated examples are statistically similar to N/τ_{mix} independent examples, where τ_{mix} is the process *mixing-time*. Relying on this idea, generalization bounds informing independent data can typically be ported to the ergodic setting, where the effective sample size is simply "deflated" by a factor of τ_{mix}. Since mixing-based bounds become vacuous as $\tau_{\mathsf{mix}} \to \infty$, they do not present an effective strategy for studying dynamics that do not mix. A critical instance of this arises in linear dynamical systems: in LDS, the ergodicity condition amounts to *stability* of the dynamics matrix A (i.e., $\rho(A) < 1$), where $\tau_{\mathsf{mix}} \to \infty$ as $\rho(A) \to 1$ [e.g. MT93, Thm. 17.6.2]. Marginally unstable systems, in which $\rho(A) = 1$, are thus not captured.

A recent line of work uncovers ways to sharpen generalization bounds based on the specific structure of *realizable* least-squares regression problems over an ergodic trajectory. For realizable linear regression with stationary covariates, results from [BJN+20] imply that, after the trajectory length exceeds an initial burn-in time scaling as $\tau_{\mathsf{mix}} n$, the minimax (excess) risk coincides with the classic iid rates. Additionally, [ZT22] show that the empirical risk minimizer exhibits similar behavior in realizable nonparametric regression problems, provided certain small-ball assumptions of the underlying process hold. While these results sharpen our understanding of how the mixing time τ_{mix} affects regression risk bounds, they ultimately rely on ergodicity. Since learning from a single trajectory is generally impossible without ergodicity, we are led to study other sequential learning configurations. The two, however, are not mutually exclusive: our results actually apply when mixing, and in fact show that the empirical risk minimizer is minimax optimal (after a burn-in time scaling with the mixing time). This eschews the need for algorithmic modifications to learning from mixing trajectory data [BJN+20]. We give details on this in Appendix B.7.

Non-temporal dependency structures. Covariates and responses can be inter-dependent in many ways, not only via temporal structure. A recent resurgence of work investigates learning under an Ising model structure over covariates [Bre15, DDDJ19, GM20, DDDK21], as well as over responses [DDP19, DDD+21] (conditioned on the covariates). At a conceptual level, the extension from a single temporally dependent trajectory to multiple trajectories is analogous to the extension from single observations to Ising models with multiple independent observations. Incidentally, in this area, investigations *began* by studying learning under *multiple* independent observations, and progressed towards guarantees on learning from a single one. Relating these two data models—trajectories and Ising grids—under intercompatible assumptions may reveal interesting connections between these results.

System identification. A special case of our LDS-specific data model captures *linear system identification* with full state observation: the task of recovering the dynamical system parameters A from observations of trajectories. While classic results are asymptotic in nature (see e.g. [LW82, LW83, Lju98]), recent work gives finite-sample guarantees for recovery of linear systems with fully observed states [SMT+18, DMM+20, JP20, FTM18, SR19, JP19, TP21], and also partially observed states [OO19, SBR19, TP19, SRD21, ZL21]. The proof of our upper bounds builds on the "small-ball" arguments from [SMT+18] (that, in turn, extend [Men15, KM15]), which do not require ergodicity.

To the best of our knowledge, our results are the first to quantify the trade-offs between few long trajectories and many short trajectories. Nearly all finite-sample guarantees for linear system identification consider a *single* trajectory, with a few notable exceptions. First, [DMM+20] allow for $m \geqslant 1$ trajectories with fully observed states and make no assumptions on the dynamics matrix A. However, their analysis discards all but the last state transition within a trajectory, reducing to iid learning over only m examples. Second, [ZL21, XCS22] study the recovery of Markov parameters from partially observed states over many trajectories. However, their error bounds do not decrease with longer training horizons T, since the number of Markov parameters one must recover scales with the trajectory length. Third, [XGH+21] consider multiple trajectories where the noise enters *multiplicatively* instead of additively. Their main finite-sample parameter recovery result (Theorem 2) states that the operator norm of the parameter error scales as $\sqrt{T/m}$, with the additional restriction that $T \gtrsim n^2$. To achieve consistency, this result fixes the trajectory length T and takes the trajectory count $m \to \infty$. By contrast, our analysis varies the two quantities T and m independently. Finally, a line of work concurrent to ours investigates learning from multiple sources of linear dynamical systems [CP22, MFTM22]. This is a latent variable model, where the underlying index of the LDS must be disambiguated from data. This model is more general than the one studied in this paper, and specializing the corresponding results to our setup yields sub-optimal bounds and unnecessary requirements. We discuss this in Section 5.2, after presenting upper bounds in detail.

Furthermore, our LDS setup (Section 3.4) decouples the covariate dynamics model A from the observation model $W_\star$, and our risk definition additionally allows for an arbitrary evaluation horizon T'. The risk over an arbitrary evaluation horizon is harder to control than parameter error, which corresponds to an evaluation length of one. This is because the larger signal-to-noise ratio accrued by a less stable system magnifies the prediction error over the entire evaluation horizon. Although the observation model that we consider is mentioned in [SMT+18], the general setup with matching upper and lower bounds are all, to the best of our knowledge, new contributions.

A complementary line of work studies the problem of online sequence prediction in a no-regret framework, where the baseline expert class comprises of trajectories generated by a linear dynamical system [HSZ17, HLS+18, GLS+20]. These results also allow for marginally unstable dynamics but are otherwise not directly comparable. Other efforts look beyond linear systems to identifying various non-linear classes, such as exponentially stable non-linear systems [SO20, FRS20] and marginally unstable non-linear systems [JKNN21]. These results again learn from a single trajectory. We believe that elements of our analysis can be ported over to offer many-trajectory bounds for these particular classes of non-linear systems.

3 Problem formulation

Notation. The real eigenvalues of a Hermitian matrix $M \in \mathbb{C}^{k\times k}$ are $\lambda_{\max}(M) = \lambda_1(M) \geqslant \ldots \geqslant \lambda_k(M) = \lambda_{\min}(M)$. For a square matrix $M \in \mathbb{C}^{k\times k}$, M^* denotes its conjugate transpose, and $\rho(M)$ denotes its spectral radius: $\rho(M) = \max\{|\lambda| \mid \lambda \text{ is an eigenvalue of } M\}$. The space of $n \times n$ real-valued symmetric positive semidefinite (resp. positive definite) matrices is denoted $\mathsf{Sym}^n_{\geqslant 0}$ (resp. $\mathsf{Sym}^n_{>0}$). The non-negative (resp. positive) orthant in $\mathbb{R}^n$ is denoted as $\mathbb{R}^n_{\geqslant 0}$ (resp. $\mathbb{R}^n_{>0}$), and $\mathbb{S}^{n-1}$ denotes the unit sphere in $\mathbb{R}^n$. Finally, the set of positive integers is denoted by $\mathbb{N}_+$.

3.1 Linear regression from sequences

Regression model. A *covariate sequence* is an indexed set $\{x_t\}_{t\geqslant 1} \subset \mathbb{R}^n$. Any distribution P_x over covariate sequences is assumed to have bounded second moments, i.e., that $\mathbb{E}[x_t x_t^\mathsf{T}]$ exists and is finite for all $t \geqslant 1$. Also for such a distribution P_x, let $\mathsf{P}_\xi[\mathsf{P}_x]$ be a distribution over *observation noise* sequences $\{\xi_t\}_{t\geqslant 1} \subset \mathbb{R}^p$. Denoting by $\{\mathcal{F}_t\}_{t\geqslant 0}$ the filtration with $\mathcal{F}_t = \sigma(\{x_k\}_{k=1}^{t+1}, \{\xi_k\}_{k=1}^t)$, we assume that $\{\xi_t\}_{t\geqslant 1}$ is a σ_ξ-sub-Gaussian martingale difference sequence (MDS), i.e., for $t \geqslant 1$:

$$\mathbb{E}[\langle v, \xi_t\rangle \mid \mathcal{F}_{t-1}] = 0, \quad \mathbb{E}[\exp(\lambda\langle v, \xi_t\rangle) \mid \mathcal{F}_{t-1}] \leqslant \exp(\lambda^2\|v\|_2^2\sigma_\xi^2/2) \text{ a.s. } \forall\lambda \in \mathbb{R}, v \in \mathbb{R}^p.$$

Given a *ground truth model* $W_\star \in \mathbb{R}^{p\times n}$, define the *observations* (a.k.a. "responses" or "labels"):

$$y_t = W_\star x_t + \xi_t, \quad t \geqslant 1. \tag{3.1}$$

Denote by $\mathsf{P}^{W_\star}_{x,y}[\mathsf{P}_x, \mathsf{P}_\xi]$ the joint distribution over covariates and their observations $\{(x_t, y_t)\}_{t\geqslant 1}$.

Regression task. Fix a ground truth model $W_\star \in \mathbb{R}^{p\times n}$, a covariate distribution P_x, an observation noise model P_ξ, a training horizon T, and a test horizon T'. Draw m independent sequences $\{(x_t^{(i)}, y_t^{(i)})\}_{i\in[m],t\geqslant 1}$ from $\mathsf{P}^{W_\star}_{x,y}[\mathsf{P}_x, \mathsf{P}_\xi]$, and call their length-$T$ prefixes $\{(x_t^{(i)}, y_t^{(i)})\}_{i=1,t=1}^{m,T}$ the training *examples*. From these examples, the regression task is to find a hypothesis $\hat{f}_{m,T} : \mathbb{R}^n \to \mathbb{R}^p$ that matches ground truth predictions $f_{W_\star}(x) := W_\star x$ in expectation over unseen trajectories of length T'. Specifically, the excess *risk* of a hypothesis $\hat{f}$ is:

$$L(\hat{f}; T', \mathsf{P}_x) := \mathbb{E}_{\mathsf{P}_x}\left[\frac{1}{T'}\sum_{t=1}^{T'}\|\hat{f}(x_t) - f_{W_\star}(x_t)\|_2^2\right]. \tag{3.2}$$

We say that the evaluation horizon T' is *strict* if $T' \leqslant T$ and *extended* if $T' > T$. When the hypothesis class is linear, meaning the hypotheses $\hat{f}$ are of the form $\hat{f}(x) = \hat{W}x$ with $\hat{W} \in \mathbb{R}^{p\times n}$, the risk expression (3.2) simplifies as follows. For a positive definite matrix $\Sigma \in \mathbb{R}^{n\times n}$, define the weighted square norm $\|M\|_\Sigma^2 := \operatorname{tr}(M\Sigma M^\mathsf{T})$ for $M \in \mathbb{R}^{p\times n}$. Denoting, for $t \geqslant 1$:

$$\Sigma_t(\mathsf{P}_x) := \mathbb{E}_{\mathsf{P}_x}[x_t x_t^\mathsf{T}], \qquad \Gamma_t(\mathsf{P}_x) := \frac{1}{t}\sum_{k=1}^t \Sigma_k(\mathsf{P}_x), \tag{3.3}$$

we overload notation and write:

$$L(\hat{W}; T', \mathsf{P}_x) = \|\hat{W} - W_\star\|^2_{\Gamma_{T'}(\mathsf{P}_x)}. \tag{3.4}$$

The risk (3.2), being a notion of error averaged over time steps, relates to that of [ZSM22] in the study of learning dynamics (the difference lies in whether the error norm is squared).

By allowing unequal training and test horizons $T \neq T'$, we cover two related scenarios at once: system identification in linear dynamical systems (when $T' = 1$) and predicting past the end of a sequence (when $T' > T$). For the latter, the risk definition (3.2) is closely related to a commonly studied notion of "final step" generalization (see e.g. [KM17, Eq. 5], [MSS17, Def. 10]) that measures the performance of a hypothesis at $T' - T$ time steps beyond the training horizon: $L_{\mathrm{end}}(\hat{f}; T', \mathsf{P}_x) := \mathbb{E}_{\mathsf{P}_x}[\|\hat{f}(x_{T'}) - f_{W_\star}(x_{T'})\|_2^2]$. Linear hypotheses enjoy the identity $L_{\mathrm{end}}(\hat{W}; T', \mathsf{P}_x) = \|\hat{W} - W_\star\|^2_{\Sigma_{T'}(\mathsf{P}_x)}$. In turn:

$$L_{\mathrm{end}}(\hat{W}; T', \mathsf{P}_x) \geqslant L(\hat{W}; T', \mathsf{P}_x) \gtrsim L_{\mathrm{end}}(\hat{W}; \lfloor T'/2 \rfloor, \mathsf{P}_x).$$

In other words, provided the scale of the covariances $\Sigma_t(\mathsf{P}_x)$ does not grow substantially over time t, our risk definition L is comparable to the final-step risk L_{end}.

Minimax risk. To compare the hardness of learning across problem classes (i.e., families of covariate distributions P_x), we measure the *minimax rate* of the risk L—i.e., the behavior of the best estimator's worst-case risk over valid problem instances—as a function of the amount of training data m, T and other problem parameters such as n, p, σ_ξ, and T'. Recall that $\mathsf{P}_{x,y}^{W_\star}$ denotes the distribution over labeled trajectories $\{(x_t, y_t)\}_{t \geqslant 1}$. For a collection of covariate sequence distributions $\mathcal{P}_x$, the minimax risk over problem instances consistent with $\mathcal{P}_x$ is:

$$\mathsf{R}(m, T, T'; \mathcal{P}_x) := \inf_{\mathsf{Alg}} \sup_{\mathsf{P}_x \in \mathcal{P}_x} \sup_{W_\star, \mathsf{P}_\xi} \mathbb{E}_{\otimes_{i=1}^m \mathsf{P}_{x,y}^{W_\star}[\mathsf{P}_x, \mathsf{P}_\xi]} \left[L\left(\mathsf{Alg}(\{(x_t^{(i)}, y_t^{(i)})\}_{i=1,t=1}^{m,T}); T', \mathsf{P}_x \right) \right], \tag{3.5}$$

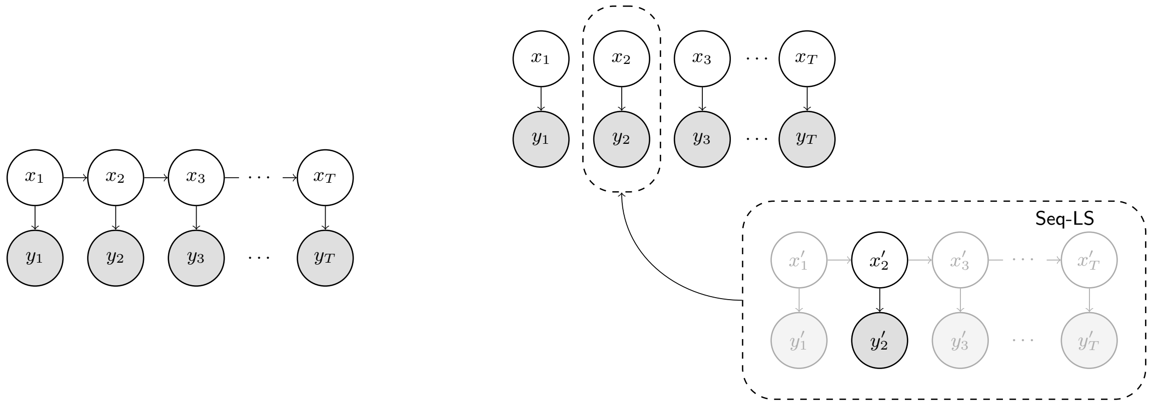
where the infimum ranges over estimators $\mathsf{Alg} : (\mathbb{R}^n \times \mathbb{R}^p)^{mT} \to (\mathbb{R}^n \to \mathbb{R}^p)$ that map training samples to hypotheses, the supremum over $W_\star$ is over all $p \times n$ ground truth models, and the supremum over P_ξ is over all σ_ξ-sub-Gaussian MDS processes determining the observation noise.

The ordinary least-squares estimator. Much like its classical role in iid learning, the *ordinary least-squares* (OLS) estimator will be key to bounding the minimax risk (3.5) from above. We define the OLS estimator to be the linear hypothesis $\hat{W}_{m,T} \in \mathbb{R}^{p \times n}$ that satisfies:

$$\hat{W}_{m,T} \in \operatorname*{argmin}_{W \in \mathbb{R}^{p \times n}} \sum_{i=1}^m \sum_{t=1}^T \|W x_t^{(i)} - y_t^{(i)}\|_2^2. \tag{3.6}$$

For $i = 1, \ldots, m$, let $X_{m,T}^{(i)} \in \mathbb{R}^{T \times n}$ be the data matrix for the i-th trajectory (i.e., the t-th row of $X_{m,T}^{(i)}$ is $x_t^{(i)}$ for $t = 1, \ldots, T$). Define $Y_{m,T}^{(i)} \in \mathbb{R}^{T \times p}$ and $\Xi_{m,T}^{(i)} \in \mathbb{R}^{T \times p}$ analogously. Put $X_{m,T} \in \mathbb{R}^{mT \times n}$ as the vertical concatenation of $X_{m,T}^{(1)}, \ldots, X_{m,T}^{(m)}$, and similarly for $Y_{m,T} \in \mathbb{R}^{mT \times p}$ and $\Xi_{m,T} \in \mathbb{R}^{mT \times p}$. Whenever $X_{m,T}$ has full column rank, then we can write $\hat{W}_{m,T}$ as:

$$\hat{W}_{m,T} = Y_{m,T}^\mathsf{T} X_{m,T} (X_{m,T}^\mathsf{T} X_{m,T})^{-1}. \tag{3.7}$$



(a) The Seq-LS problem (Problem 3.1): covariates $\{x_t\}$ are drawn from a sequence distribution, and noisy observations $\{y_t\}$ are drawn conditioned on these covariates.

(b) The corresponding baseline Ind-Seq-LS problem (Problem 3.2): independent covariate-observation pairs $\{(x_t, y_t)\}$ are drawn, each from the marginal distribution of the corresponding t'th step in the Seq-LS problem.

Figure 2: Formulations of regression from sequential data, illustrated as graphical models. Specifically, these graphs depict a simplified special case of our data model, in which the observations $\{y_t\}$ across time are independent conditioned on the covariates $\{x_t\}$. In our general definitions (Problem 3.1 and Problem 3.2), the observations $\{y_t\}$ can be conditionally interdependent, via a martingale difference sequence on the observation noise (Section 3.1).

3.2 Problem classes

We formalize linear regression from sequential data generally as follows:

Problem 3.1 (Seq-LS)**.** *Assume a covariate sequence distribution P_x in the linear regression model* (3.1)*. Fix an evaluation horizon T'. On input m labeled trajectories of length T drawn from this model, in the form of examples $\{(x_t^{(i)}, y_t^{(i)})\}_{i=1,t=1}^{m,T}$, output a hypothesis $\hat{f}_{m,T}$ that minimizes excess risk $L(\hat{f}_{m,T}; T', \mathsf{P}_x)$.*

Our topmost goal is to study the effect of learning from sequentially dependent covariates *in comparison with* learning in the classical iid setup. Linear regression is well understood in the latter setting. Focusing on *well-specified* linear regression further simplifies our presentation, allowing us to isolate the effects of what interests us most—dependent covariates. Generalizing the supervision aspect of Seq-LS (say, to unrealizable and non-parametric regression, or to classification) is left to future work. We return to discuss this in Section 9.

To study how dependent data affects learning, we need to establish an "independent data" baseline. The natural comparison point for Seq-LS is to remove all correlations across time. Namely, instead of drawing covariates sequentially from the distribution P_x, consider learning separately from the marginals of P_x at each time step. The resulting decorrelated distribution generates *independent* examples, but typically not iid ones. We formalize linear regression from independent data generally as follows:

Problem 3.2 (Ind-Seq-LS)**.** *Fix a sequence of distributions $\{\mathsf{P}_{x,t}\}_{t\geqslant 1}$. Consider their product over time $\otimes_{t\geqslant 1}\mathsf{P}_{x,t}$ as the covariate sequence distribution in the linear regression model* (3.1)*. Fix an evaluation horizon T'. On input m labeled trajectories of length T drawn from this model, in the form of examples $\{(x_t^{(i)}, y_t^{(i)})\}_{i=1,t=1}^{m,T}$, output a hypothesis $\hat{f}_{m,T}$ that minimizes $L(\hat{f}_{m,T}; T', \otimes_{t\geqslant 1}\mathsf{P}_{x,t})$.*

This $\mathsf{Ind\text{-}Seq\text{-}LS}$ problem generalizes the canonical iid learning setup slightly. Existing theory can still characterize its minimax risk, provided the covariances of the distributions $\{\mathsf{P}_{x,t}\}$ are roughly equal in scale across time t. However, this equal-scale requirement rules out the marginals of interesting applications, such as dynamical systems that are not stable or ergodic. We therefore extend, in later sections, characterizations of the regression risk to handle covariances that can scale *polynomially* across time instead.

3.3 Problem separations

To set up a baseline for a $\mathsf{Seq\text{-}LS}$ problem, we will specifically instantiate $\mathsf{Ind\text{-}Seq\text{-}LS}$ over its marginals. Namely, for a sequence distribution P_x over $\{x_t\}_{t\geqslant 1}$, let $\mu_t[\mathsf{P}_x]$ be the marginal distribution of x_t at time $t \geqslant 1$, and consider $\mathsf{Ind\text{-}Seq\text{-}LS}$ with covariates drawn from the sequence $\{\mu_t[\mathsf{P}_x]\}_{t\geqslant 1}$. Figure 2 illustrates such a $\mathsf{Seq\text{-}LS}$ problem and the corresponding $\mathsf{Ind\text{-}Seq\text{-}LS}$ instance over its marginals.

This decorrelated baseline is a hypothetical benchmark: in a practical context, collecting independent marginal data, when nature only supplies its dependent form, can be expensive or infeasible. However, we can expect that having such data on hand would make learning easier, with risk rates that resemble iid learning. In what follows, we outline scenarios where a sequential learning problem and its decorrelated baseline coincide in difficulty, and others in which they diverge. We then outline the possible assumptions that would allow us to always relate the two.

The iid special case. When $T = T' = 1$, the example trajectories $\{x_1^{(i)}\}_{i=1}^m$ are trivially a set of iid covariates. The problems $\mathsf{Seq\text{-}LS}$ and $\mathsf{Ind\text{-}Seq\text{-}LS}$ thus coincide, and reduce to the well-specified random design linear regression problem over m iid covariates. It is well-known that under iid data, and mild regularity conditions, the minimax risk scales as $\sigma_\xi^2 pn/m$, and is achieved by the OLS estimator [HKZ14, Mou22, Wai19].

Extending the horizon. Considering nontrivial horizons $T = T' > 1$, both $\mathsf{Seq\text{-}LS}$ and its corresponding $\mathsf{Ind\text{-}Seq\text{-}LS}$ baseline become more involved, but for different reasons.

The $\mathsf{Ind\text{-}Seq\text{-}LS}$ problem, as we show in Section 6, is not generally learnable with polynomially many examples. Specifically, the minimax rate scales exponentially in the dimension n provided the trajectory count m is constant. To address this, we will require that the covariances of its constituent distributions $\{\mathsf{P}_{x,t}\}$ grow at most polynomially with time t. Under this constraint, the problem's minimax risk again scales as the iid-like rate $\sigma_\xi^2 pn/(mT)$ times, at most, a factor determined exponentially by the covariance growth.

The $\mathsf{Seq\text{-}LS}$ problem inherits the same growth limitation. Even then, it is still not generally learnable without further assumptions on the dependence structure of covariates: the minimax risk is otherwise bounded away from zero as the horizon T tends to infinity, provided the trajectory count m is constant. To realize this, consider $x_1 \sim N(0, I_n)$ and $x_t = x_{t-1}$ for $t \geqslant 2$, a sequence of identical covariates whose marginals are all independent Gaussians. The resulting dataset presents an underdetermined regression problem if $m < n$. In essence, its covariates lack sufficient "excitation" across time. To rein $\mathsf{Seq\text{-}LS}$ back in to the realm of learnability, one must:

(a) make further modeling assumptions about covariates, or

(b) introduce excitation via independent resets.

For (a), as detailed in Section 2, the most common modeling assumption considers sequences that mix rapidly to a stationary distribution. Another avenue—recently active in the literature, and sometimes overlapping with the mixing approach—considers sequences generated by linear dynamical systems. Among these two, mixing implies risk bounds that tend to zero with T, but only hold in the worst case after a burn-in time that scales proportionally to the mixing time [BJN+20]. This prevents a characterization of minimax risk uniformly across the full range of problem instances P_x that mix, unless one caps the mixing time to a fixed constant. Narrowing instead to LDS models in the sequel, we manage to succinctly carve out a basic problem family, with *unbounded* mixing time, and to characterize its minimax risk uniformly. One still pays a price for sequential dependency, as this minimax risk turns out to be larger than its Ind-Seq-LS counterpart by a factor of the dimension n.

Turning in addition to (b), by introducing (sufficiently many) resets, we can expand our data model substantially: we manage to lift most of our LDS assumptions and extend to other dynamical systems. Remarkably, we even show that for any controllable LDS—including ones that are unstable and hence grow exponentially in time—having sufficiently many resets guarantees that the risk exhibits, once again, the iid-like behavior of $\sigma_\xi^2 pn/(mT)$, up to mere constants.

3.4 Linear dynamical trajectories

Fix a *dynamics matrix* $A \in \mathbb{R}^{n\times n}$ and a *control matrix* $B \in \mathbb{R}^{n\times d}$. Consider the n-dimensional trajectory $\{x_t\}_{t\geqslant 1}$ defined by the linear dynamical system:

$$x_t = Ax_{t-1} + Bw_t, \text{ where } w_t \sim N(0, I_d), \text{ for } t \geqslant 1, \tag{3.8}$$

taking $x_0 = 0$ by convention. We assume that the noise process $\{w_t\}_{t\geqslant 1}$ is independent across time, i.e., that $w_t \perp w_{t'}$ whenever $t \neq t'$. Overloading notation, let the matrix $\Sigma_t(A,B) := \sum_{k=0}^{t-1} A^k BB^\mathsf{T}(A^k)^\mathsf{T}$ denote the covariance of x_t, and let the matrix $\Gamma_t(A,B) := \frac{1}{t}\sum_{k=1}^{t}\Sigma_k(A,B)$ denote the average covariance. Denote by $\mathsf{P}_x^{A,B}$ the distribution over the trajectory $\{x_t\}_{t\geqslant 1}$, and let $\{x_t^{(i)}\}_{t\geqslant 1}$ for $i \geqslant 1$ denote independent draws from $\mathsf{P}_x^{A,B}$. When $B = I_n$, we use the respective shorthand notation $\Sigma_t(A)$, $\Gamma_t(A)$, and P_x^A.

Modeling regression covariates as linear dynamical trajectories gives us the LDS-LS problem, a specialization of Seq-LS (Problem 3.1):

Problem 3.3 (LDS-LS)**.** *Assume a dynamics matrix $A \in \mathbb{R}^{n\times n}$, a control matrix $B \in \mathbb{R}^{n\times d}$, and a corresponding linear dynamical covariate distribution $\mathsf{P}_x^{A,B}$ in the linear regression model (3.1). Fix an evaluation horizon T'. On input m labeled trajectories of length T, drawn from this model, in the form of examples $\{(x_t^{(i)}, y_t^{(i)})\}_{i=1,t=1}^{m,T}$, output a hypothesis $\hat{f}_{m,T}$ that minimizes $L(\hat{f}_{m,T}; T', \mathsf{P}_x^{A,B})$.*

Let $\mathsf{P}_{x,t}^{A,B}$ be the marginal distribution of x_t under $\mathsf{P}_x^{A,B}$ at each $t \geqslant 1$. The natural decorrelated baseline for LDS-LS is a corresponding specialization of Ind-Seq-LS (Problem 3.2) to LDS trajectories:

Problem 3.4 (Ind-LDS-LS)**.** *Assume a dynamics matrix $A \in \mathbb{R}^{n\times n}$, a control matrix $B \in \mathbb{R}^{n\times d}$, and a corresponding trajectory distribution $\mathsf{P}_x^{A,B}$. Consider covariates drawn independently from its marginals, i.e., assume the linear regression model (3.1) under the covariate sequence distribution $\otimes_{t\geqslant 1}\mathsf{P}_{x,t}^{A,B}$. Fix an evaluation horizon T'. On input m labeled trajectories of length T, drawn from this model, in the form of examples $\{(x_t^{(i)}, y_t^{(i)})\}_{i=1,t=1}^{m,T}$, output a hypothesis $\hat{f}_{m,T}$ that minimizes $L(\hat{f}_{m,T}; T', \otimes_{t\geqslant 1}\mathsf{P}_{x,t}^{A,B})$.*

Learning dynamical systems. $\mathsf{LDS\text{-}LS}$ generalizes *linear system identification*, the problem of recovering the dynamics A from data. The reduction follows by setting $W_\star = A$ and $\xi_t^{(i)} = Bw_{t+1}^{(i)}$, so that $y_t^{(i)} = x_{t+1}^{(i)}$. Note that when B has full row rank, the squared parameter error in the weighted BB^T norm $\|\cdot\|_{BB^\mathsf{T}}$ is simply the risk $L(\hat{A}; T', \mathsf{P}_x^{A,B})$ when $T' = 1$. Recent related work typically assumes that B indeed has full row rank, but in later sections we touch on the more general case where this is not required, so long as the pair (A, B) is controllable. Bounds in operator norm are also easily obtainable from our proof techniques. However, our lower bounds will not inform the system identification problem specifically; our hardness results rely on decoupling $W_\star$ from A and $\xi_t^{(i)}$ from $w_{t+1}^{(i)}$, whereas this reduction naturally ties them.

4 Trajectory small-ball definition and examples

We establish risk upper bounds by studying the behavior of the ordinary least-squares estimator. The key technical definition that drives the analysis is a "small-ball" condition on covariate sequences:

Definition 4.1 (Trajectory small-ball (TrajSB))**.** *Fix a trajectory length $T \in \mathbb{N}_+$, a parameter $k \in \{1, \ldots, T\}$, positive definite matrices $\{\Psi_j\}_{j=1}^{\lfloor T/k \rfloor} \subset \mathsf{Sym}_{>0}^n$, and constants $c_{\mathsf{sb}} \geqslant 1$, $\alpha \in (0, 1]$. The distribution P_x satisfies the* $(T, k, \{\Psi_j\}_{j=1}^{\lfloor T/k \rfloor}, c_{\mathsf{sb}}, \alpha)$-trajectory-small-ball (TrajSB) *condition if:*

1. $\frac{1}{\lfloor T/k \rfloor} \sum_{j=1}^{\lfloor T/k \rfloor} \Psi_j \preccurlyeq \Gamma_T(\mathsf{P}_x)$,
2. *$\{x_t\}_{t \geqslant 1}$ is adapted to a filtration $\{\mathcal{F}_t\}_{t \geqslant 1}$, and*
3. *for all $v \in \mathbb{R}^n \setminus \{0\}$, $j \in \{1, \ldots, \lfloor T/k \rfloor\}$ and $\varepsilon > 0$:*

$$\mathbb{P}_{\{x_t\} \sim \mathsf{P}_x} \left\{ \frac{1}{k} \sum_{t=(j-1)k+1}^{jk} \langle v, x_t \rangle^2 \leqslant \varepsilon \cdot v^\mathsf{T} \Psi_j v \;\middle|\; \mathcal{F}_{(j-1)k} \right\} \leqslant (c_{\mathsf{sb}} \varepsilon)^\alpha \;\; a.s. \tag{4.1}$$

Above, $\mathcal{F}_0$ is understood to be the minimal σ-algebra. Additionally, the distribution P_x satisfies the $(T, k, \Psi, c_{\mathsf{sb}}, \alpha)$-TrajSB condition if it satisfies $(T, k, \{\Psi_j\}_{j=1}^{\lfloor T/k \rfloor}, c_{\mathsf{sb}}, \alpha)$-TrajSB with $\Psi_j = \Psi$. Finally, we call the parameter k the excitation window*.*

In Definition 4.1, we typically consider the matrices Ψ_j to be the sharpest almost-sure lower bound that we can specify (in the Loewner order) on the quantity $\mathbb{E}[\frac{1}{k} \sum_{t=(j-1)k+1}^{jk} x_t x_t^\mathsf{T} \mid \mathcal{F}_{(j-1)k}]$. Section 4.1 lists examples of covariate sequence distributions P_x that satisfy the TrajSB condition.

Definition 4.1 draws inspiration from the block martingale small-ball condition from [SMT+18]. There are two main differences: (a) we consider the small-ball probability of the *entire* block $\frac{1}{k} \sum_{t=(j-1)k+1}^{jk} \langle v, x_t \rangle^2$ at once, instead of the *average* of small-ball probabilities:

$$\frac{1}{k} \sum_{t=(j-1)k+1}^{jk} \mathbb{P} \left\{ \langle v, x_t \rangle^2 \leqslant \varepsilon \cdot v^\mathsf{T} \Psi_j v \mid \mathcal{F}_{(j-1)k} \right\},$$

and (b) equation (4.1) is required to hold at all scales $\varepsilon > 0$, instead of at a single resolution. We need the first modification (a) to prove optimal rates under many trajectories without assuming stability or ergodicity. Furthermore, condition (4.1) is implied by a bound on the average of small-ball probabilities (Proposition 4.2). We need the second modification (b) in order to bound the expected value of the OLS risk. Under weaker conditions, we would only have risk bounds that hold with high probability, as detailed in the following remark:

Remark 4.1. In Appendix B.7, we consider the following modification to Definition 4.1, where we instead suppose that (4.1) holds for *some* fixed ε (such that the inequality's right-hand side is strictly less than one), rather than for all ε; we refer to this modification as *weak trajectory small-ball* (Definition B.1). As described above, a consequence of the weak trajectory small-ball condition is that the main OLS risk bounds now hold with high probability (i.e., polylogarithmic in $1/\delta$) rather than in expectation (Lemma B.23).[1] A key upshot (Proposition B.24), however, is that this change allows for an ergodic covariate sequence with ϕ-mixing time bounded by τ_{mix} to be considered (weak) trajectory small-ball (with excitation window $k \asymp \tau_{\mathsf{mix}}$), provided the stationary distribution μ satisfies a standard (weak) small-ball condition [Men15, KM15, Oli16]:

$$\sup_{v \in \mathbb{S}^{n-1}} \mathbb{P}_\mu\{\langle v, x\rangle^2 \leqslant \varepsilon \cdot \mathbb{E}_\mu[\langle v, x\rangle^2]\} < 1 \text{ for some } \varepsilon > 0.$$

This in turn yields upper bounds for $\mathsf{Seq\text{-}LS}$ in the few trajectories ($m \lesssim n$) regime of the following form: if $mT \geqslant \tilde{\Omega}(\tau_{\mathsf{mix}} n)$, then

$$L(\hat{W}_{m,T}; T, \mathsf{P}_x) \leqslant \tilde{O}\left(\sigma_\xi^2 \frac{pn}{mT}\right),$$

with high probability. This statement generalizes the risk bound for a single ergodic trajectory from [BJN+20, Theorem 1] to the ordinary least-squares estimator (3.6). We can interpret the condition on mT as a "burn-in time" requirement. Meanwhile, at least in the single-trajectory ($m = 1$) setting, [BJN+20, Theorem 9] tells us that that such a burn-in assumption ($T \gtrsim \tau_{\mathsf{mix}} n$) is necessary for a non-trivial risk guarantee.

4.1 Examples of trajectory small-ball distributions

We now turn to specific examples of distributions P_x which satisfy the trajectory small-ball condition. First is the example introduced in Section 3.3, where x_1 is drawn from a multivariate Gaussian and subsequently copied as $x_t = x_{t-1}$ for all $t \geqslant 2$:

Example 4.1 (Copies of a Gaussian draw)**.** *Let $\Sigma \in \mathsf{Sym}^n_{>0}$, and let P_x denote the process $x_1 \sim N(0, \Sigma)$ and $x_t = x_{t-1}$ for $t \geqslant 2$. Fix any $T \in \mathbb{N}_+$. Then P_x satisfies the $(T, T, \Sigma, e, \frac{1}{2})$-TrajSB condition.*

Note that this process only satisfies the trajectory small-ball condition with excitation window $k = T$. In other words, the conditional distribution $x_{t+k} \mid x_t$ for $k \geqslant 1$ (a Dirac distribution on x_t) contains no excitation as needed for learning. This example can actually be generalized to arbitrary Gaussian processes indexed by time:

[1] Such a high probability bound does not, in turn, imply a bound on the expected risk via integration over the tail. The reason is that the high probability bound (Lemma B.23) requires that the number of data points mT grow as $\log(1/\delta)$, where δ is the failure probability.

Example 4.2 (Gaussian processes). *Let P_x be a Gaussian process indexed by time, i.e., for every finite index set $I \subset \mathbb{N}_+$, the collection of random variables $(x_t)_{t \in I}$ is jointly Gaussian. Let $T_{\mathsf{nd}} := \inf\{t \in \mathbb{N}_+ \mid \det(\mathbb{E}[x_t x_t^\mathsf{T}]) \neq 0\}$, and suppose T_{nd} is finite. Fix a $T \in \mathbb{N}_+$ satisfying $T \geqslant T_{\mathsf{nd}}$. Then P_x satisfies the $(T, T, \Gamma_T(\mathsf{P}_x), 2e, \frac{1}{2})$-TrajSB condition.*

Our next example involves independent, but not identically distributed, covariates:

Example 4.3 (Independent Gaussians). *Let $\{\Sigma_t\}_{t \geqslant 1} \subset \mathsf{Sym}^n_{>0}$, and let $\mathsf{P}_x = \otimes_{t \geqslant 1} N(0, \Sigma_t)$. Fix a $T \in \mathbb{N}_+$. Then P_x satisfies the $(T, 1, \{\Sigma_t\}_{t=1}^T, e, \frac{1}{2})$-TrajSB condition.*

Example 4.3 allows us to select $k = 1$, reflecting the independence of the covariates across time.

We can also craft an example around a process that does not mix, but that still exhibits an excitation window of $k = 2$:

Example 4.4 (Alternating halfspaces). *Suppose that $n \geqslant 4$ is even, and let $u_1, \ldots, u_n$ be a fixed orthonormal basis of $\mathbb{R}^n$. Put $U_0 = \operatorname{span}(u_1, \ldots, u_{n/2})$ and $U_1 = \operatorname{span}(u_{n/2+1}, \ldots, u_n)$. Let $i_1 \sim \operatorname{Bern}(\frac{1}{2})$, $i_{t+1} = i_t \mod 2$ for $t \in \mathbb{N}_+$, and let P_x denote the process with conditional distribution $x_t \mid i_t$ uniform over the spherical measure on $U_{i_t} \cap \mathbb{S}^{n-1}$. For any $T \geqslant 2$, the process P_x satisfies the $(T, 2, I_n/(2n), e, \frac{1}{2})$-TrajSB condition.*

To see that the covariate distribution $\{x_t\}$ does not mix, observe that the marginal distribution for all t is uniform on $\mathbb{S}^{n-1}$, whereas the conditional distribution $x_{t+k} \mid x_t$ for any $k \in \mathbb{N}_+$ is either uniform on $U_0 \cap \mathbb{S}^{n-1}$ or uniform on $U_1 \cap \mathbb{S}^{n-1}$. Although it does not mix at all, the trajectory supplies ample excitation for learning in any mere two steps.

Even for a process that does mix, it may exhibit an excitation window far smaller than its mixing time. The following sets up such an example, where again where sufficient excitation is provided with $k = 2$ steps:

Example 4.5 (Normal subspaces). *Suppose that $n \geqslant 3$. Let $u_1, \ldots, u_n$ be a fixed orthonormal basis in $\mathbb{R}^n$, and let $U_{\neg i} := \operatorname{span}(\{u_j\}_{j \neq i})$ for $i \in \{1, \ldots, n\}$. Consider the Markov chain $\{i_t\}_{t \geqslant 1}$ defined by $i_1 \sim \operatorname{Unif}(\{1, \ldots, n\})$, and $i_{t+1} \mid i_t \sim \operatorname{Unif}(\{1, \ldots, n\} \setminus \{i_t\})$. Let P_x denote the process with conditional distribution $x_t \mid i_t$ uniform over the spherical measure on $U_{\neg i_t} \cap \mathbb{S}^{n-1}$. For any $T \geqslant 2$, the process P_x satisfies the $(T, 2, I_n/(4n-4), e, \frac{1}{2})$-TrajSB condition.*

In this example, a straightforward computation (detailed in Proposition B.11) shows that the mixing time $\tau_{\mathsf{mix}}(\varepsilon)$ of the Markov chain $\{i_t\}_{t \geqslant 1}$ scales as $\log_n(1/\varepsilon)$.[2] In most analyses which rely on mixing time arguments, one requires that the mixing time resolution ε tends to zero as either the amount of data and/or probability of success increases; as a concrete example, [DAJJ12, Eq. 3.2] suggests to set $\varepsilon = 1/\sqrt{T}$, where T is the number of samples drawn from the underlying distribution. On the other hand, the trajectory small-ball condition in Example 4.5 holds with a short excitation window of length $k = 2$, independently of T.

Next we consider linear dynamical systems. As setup, we first define the notion of controllability for a pair of dynamics matrices (A, B):

[2]For concreteness, given a discrete-time Markov chain over a finite state-space S with transition matrix P and stationary distribution π, we define the mixing time as: $\tau_{\mathsf{mix}}(\varepsilon) := \inf\{k \in \mathbb{N} \mid \sup_{\mu \in \mathcal{P}(S)} \|\mu P^k - \pi\|_{\mathrm{tv}} \leqslant \varepsilon\}$. Here, $\mathcal{P}(S)$ denotes the set of all probability distributions over S, and $\|\cdot\|_{\mathrm{tv}}$ denotes the total variation norm over distributions.

Definition 4.2 (Controllability)**.** *Let (A, B) be a pair of matrices with $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times d}$. For $k \in \{1, \dots, n\}$, we say that (A, B) is* k-step controllable *if the matrix:*

$$\begin{bmatrix} B & AB & A^2B & \cdots & A^{k-1}B \end{bmatrix} \in \mathbb{R}^{n\times kd}$$

has full row rank.

The classical definition of controllability in linear systems [cf. Rug96, Chapter 25] is equivalent to n-step controllability. Definition 4.2 allows the system to be controllable in fewer than n steps. Also note that k is restricted to $\{1, \dots, n\}$, since if a system is not n-step controllable, it will not be n'-step controllable for any $n' > n$ (by the Cayley-Hamilton theorem). A few special cases of interest to note are as follows. If B has rank n, then (A, B) is trivially one-step controllable for any A. On the other hand, if (A, B) are in canonical controllable form (i.e., A is the companion matrix associated with the polynomial $p(z) = a_0 + a_1z + \cdots + a_{n-1}z^{n-1} + z^n$ and B is the n-th standard basis vector), then (A, B) is n-step controllable. The latter corresponds directly to the state-space representation of autoregressive processes of order n, e.g. $\mathsf{AR}(n)$.

Example 4.6 (Linear dynamical systems)**.** *Let (A, B) with $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times d}$ be k_{c}-step-controllable (Definition 4.2). Let $\mathsf{P}_x^{A,B}$ be the linear dynamical system defined in (3.8). Fix any $T, k \in \mathbb{N}_+$ satisfying $T \geqslant k \geqslant k_{\mathsf{c}}$. Then, $\mathsf{P}_x^{A,B}$ satisfies the $(T, k, \Gamma_k(A, B), e, \frac{1}{2})$-TrajSB condition.*

In all of the examples so far, the time-t marginal distribution of covariates x_t has either been a multivariate Gaussian or a spherical measure. To underscore the generality of the small-ball method, we can create additional examples where this is not the case. In what follows, we consider Volterra series [MS00], which generalize the classical Taylor series to causal sequences. Analogous to how polynomials can approximate continuous functions arbitrarily well on a compact set, Volterra series can approximate signals that depend continuously (and solely) on their history over a bounded set of inputs [cf. Rug81, Section 1.5].

Example 4.7 (Degree-D Volterra series)**.** *Fix a $D \in \mathbb{N}_+$. Let $\{c^{(d,\ell)}_{i_1,\dots,i_d}\}_{i_1,\dots,i_d\in\mathbb{N}}$ for $d \in \{1, \dots, D\}$ and $\ell \in \{1, \dots, n\}$ denote arbitrary rank-d arrays. Let $\{w^{(\ell)}_t\}_{t\geqslant 0}$ be iid $N(0, 1)$ random variables for $\ell \in \{1, \dots, n\}$. Consider the process P_x where for $t \geqslant 1$, the ℓ-th coordinate of x_t, denoted $(x_t)_\ell$, is:*

$$(x_t)_\ell = \sum_{d=1}^{D} \sum_{i_1,\dots,i_d=0}^{t-1} c^{(d,\ell)}_{i_1,\dots,i_d} \prod_{d'=1}^{d} w^{(\ell)}_{t-i_{d'}-1}. \tag{4.2}$$

Let $T_{\mathsf{nd}} := \inf\{t \in \mathbb{N}_+ \mid \det(\Gamma_t(\mathsf{P}_x)) \neq 0\}$, and suppose T_{nd} is finite. There is a constant $c_D > 0$, depending only on D, such that for any $T \geqslant T_{\mathsf{nd}}$, P_x satisfies the $(T, T, \Gamma_T(\mathsf{P}_x), c_D, 1/(2D))$-TrajSB condition.

The main idea behind Example 4.7 is that, while x_t is certainly not Gaussian, the quadratic form $\sum_{t=1}^{T} \langle v, x_t\rangle^2$ is a degree at most $2D$ polynomial in $\{w^{(\ell)}_t\}_{t=0}^{T-1}$. It will hence exhibit anti-concentration, according to a landmark result from [CW01]. The same result actually provides an immediate extension of this example—as well as the previous examples—to noise distributions with log-concave densities, such as Laplace or uniform noise.

We next present a special case of the Volterra series, where we can choose the excitation window k in the small-ball definition strictly between the endpoints 1 and T. To set up, a few more definitions are needed:

Definition 4.3. *Fix an integer $d \in \mathbb{N}_+$. A rank-d array of coefficients $\{c_{i_1,\dots,i_d}\}_{i_1,\dots,i_d \in \mathbb{N}}$ is called:*

(a) symmetric *if $c_{i_1,\dots,i_d} = c_{\pi(i_1,\dots,i_d)}$ for any permutation π of indices $i_1,\dots,i_d \in \mathbb{N}$,*

(b) traceless *if $c_{i,\dots,i} = 0$ for all $i \in \mathbb{N}$, and*

(c) non-degenerate *if there exists an $k_{\mathsf{nd}} \in \mathbb{N}_+$ such that the following set is non-empty:*

$$\{(i_1,\dots,i_d) \mid c_{i_1,\dots,i_d} \neq 0, i_1,\dots,i_d \in \{0,\dots,k_{\mathsf{nd}}-1\}\}.$$

The smallest k_{nd} such that $\{c_{i_1,\dots,i_d}\}$ is the non-degeneracy index.

Example 4.8 (Degree-2 Volterra series)**.** *Consider the following process P_x. Let $\{c_{i,j}^{(\ell)}\}_{i,j\geqslant 0}$ for $\ell \in \{1,\dots,n\}$ be symmetric, traceless, non-degenerate arrays (Definition 4.3). Let $\{w_t^{(\ell)}\}_{t\geqslant 0}$ be iid $N(0,1)$ random variables for $\ell \in \{1,\dots,n\}$. For $t \geqslant 1$, the ℓ-th coordinate of x_t, denoted $(x_t)_\ell$, is:*

$$(x_t)_\ell = \sum_{i=0}^{t-1}\sum_{j=i}^{t-1} c_{i,j}^{(\ell)} w_{t-i-1}^{(\ell)} w_{t-j-1}^{(\ell)}. \tag{4.3}$$

Let $k_{\mathsf{nd}} \in \mathbb{N}_+$ denote the smallest non-degeneracy index for all n arrays. There is a universal positive constant c such that for any T and k satisfying $T \geqslant k \geqslant k_{\mathsf{nd}}$, P_x satisfies the $(T, k, \Gamma_k(\mathsf{P}_x), c, \frac{1}{4})$-TrajSB condition.

The assumptions pulled in from Definition 4.3 help simplify the construction of an almost sure lower bound for conditional covariances $\mathbb{E}[\frac{1}{k}\sum_{t=(j-1)k+1}^{jk} x_t x_t^{\mathsf{T}} \mid \mathcal{F}_{(j-1)k}]$, to establish that Example 4.8 satisfies the trajectory small-ball condition. We believe that generalizations to higher degree Volterra series with k strictly between 1 and T are possible by more involved calculations.

Of course, many other examples are possible. To help in recognizing them, the following statement shows that condition (4.1) in the trajectory small-ball definition can be verified by separately establishing small-ball probabilities for the conditional distributions:

Proposition 4.2 (Average small-ball implies trajectory small-ball)**.** *Fix $T \in \mathbb{N}_+$, $k \in \{1,\dots,T\}$, $\{\Psi_j\}_{j=1}^{\lfloor T/k \rfloor} \subset \mathsf{Sym}^n_{>0}$, and $\alpha, \beta \in (0,1)$. Let P_x be a covariate distribution, with $\{x_t\}_{t\geqslant 1}$ adapted to a filtration $\{\mathcal{F}_t\}_{t\geqslant 1}$. Suppose for all $v \in \mathbb{R}^n \setminus \{0\}$ and $j \in \{1,\dots,\lfloor T/k \rfloor\}$:*

$$\frac{1}{k}\sum_{t=(j-1)k+1}^{jk} \mathbb{P}_{x_t\sim\mathsf{P}_x}\left\{\langle v, x_t\rangle^2 \leqslant \alpha \cdot v^{\mathsf{T}}\Psi_j v \;\middle|\; \mathcal{F}_{(j-1)k}\right\} \leqslant \beta \text{ a.s.}, \tag{4.4}$$

where $\mathcal{F}_0$ is the minimal σ-algebra. Then, for all $v \in \mathbb{R}^n \setminus \{0\}$, $j \in \{1,\dots,\lfloor T/k \rfloor\}$, and $\varepsilon \in (0,\alpha)$

$$\mathbb{P}_{\{x_t\}\sim\mathsf{P}_x}\left\{\frac{1}{k}\sum_{t=(j-1)k+1}^{jk}\langle v, x_t\rangle^2 \leqslant \varepsilon \cdot v^{\mathsf{T}}\Psi_j v \;\middle|\; \mathcal{F}_{(j-1)k}\right\} \leqslant \frac{\beta}{1-\varepsilon/\alpha} \text{ a.s.} \tag{4.5}$$

An immediate corollary of Proposition 4.2 is the following: suppose that for all $v \in \mathbb{R}^n \setminus \{0\}$, $j \in \{1,\dots,\lfloor T/k \rfloor\}$, and $\varepsilon > 0$,

$$\frac{1}{k}\sum_{t=(j-1)k+1}^{jk} \mathbb{P}_{x_t\sim\mathsf{P}_x}\left\{\langle v, x_t\rangle^2 \leqslant \varepsilon \cdot v^{\mathsf{T}}\Psi_j v \;\middle|\; \mathcal{F}_{(j-1)k}\right\} \leqslant (c_{\mathsf{sb}}\varepsilon)^\alpha \text{ a.s.} \tag{4.6}$$

Then, the $(T, k, \{\Psi_j\}_{j=1}^{\lfloor T/k \rfloor}, 2^{1+1/\alpha} c_{\mathsf{sb}}, \alpha)$-TrajSB condition holds. Equation (4.6) can be easier to verify than (4.1), since the former allows one to reason about each conditional distribution individually, whereas the latter requires reasoning about the entire excitation window altogether.

The following two sections present upper and lower bounds for learning from trajectories, involving various instances of the trajectory small-ball assumption where applicable. All main results are summarized in Table 1.

5 Risk upper bounds

The trajectory small-ball definition allows us to carve out conditions for learnability. A key quantity for what follows is the minimum eigenvalue of the ratio of two positive definite matrices:

$$\underline{\lambda}(A, B) := \lambda_{\min}(B^{-1/2} A B^{-1/2}), \quad A, B \in \mathsf{Sym}^n_{>0}. \tag{5.1}$$

Our various upper bounds statements build on the following general lemma:

Lemma 5.1 (General OLS upper bound)**.** *There are universal positive constants c_0 and c_1 such that the following holds. Suppose that P_x satisfies the $(T, k, \{\Psi_j\}_{j=1}^{\lfloor T/k \rfloor}, c_{\mathsf{sb}}, \alpha)$-TrajSB condition (Definition 4.1). Put $S := \lfloor T/k \rfloor$ and $\Gamma_T := \Gamma_T(\mathsf{P}_x)$. Fix any $\underline{\Gamma} \in \mathsf{Sym}^n_{>0}$ satisfying $\frac{1}{S}\sum_{j=1}^S \Psi_j \preccurlyeq \underline{\Gamma} \preccurlyeq \Gamma_T$, and let $\underline{\mu}(\{\Psi_j\}_{j=1}^S, \underline{\Gamma})$ denote the geometric mean of the minimum eigenvalues $\{\underline{\lambda}(\Psi_j, \underline{\Gamma})\}_{j=1}^S$, i.e.,*

$$\underline{\mu}(\{\Psi_j\}_{j=1}^S, \underline{\Gamma}) := \left[\prod_{j=1}^S \underline{\lambda}(\Psi_j, \underline{\Gamma})\right]^{1/S}. \tag{5.2}$$

Suppose that:

$$n \geqslant 2, \quad \frac{mT}{kn} \geqslant \frac{c_0}{\alpha} \log\left(\frac{\max\{e, c_{\mathsf{sb}}\}}{\alpha \underline{\lambda}(\underline{\Gamma}, \Gamma_T)\underline{\mu}(\{\Psi_j\}_{j=1}^S, \underline{\Gamma})}\right). \tag{5.3}$$

Then, for any $\Gamma' \in \mathsf{Sym}^n_{>0}$:

$$\mathbb{E}[\|\hat{W}_{m,T} - W_\star\|^2_{\Gamma'}] \leqslant c_1 c_{\mathsf{sb}} \sigma_\xi^2 \cdot \frac{pn}{mT\alpha\underline{\lambda}(\underline{\Gamma}, \Gamma')\underline{\mu}(\{\Psi_j\}_{j=1}^S, \underline{\Gamma})} \cdot \log\left(\frac{\max\{e, c_{\mathsf{sb}}\}}{\alpha \underline{\lambda}(\underline{\Gamma}, \Gamma_T)\underline{\mu}(\{\Psi_j\}_{j=1}^S, \underline{\Gamma})}\right). \tag{5.4}$$

The proof of Lemma 5.1 blends ideas from the analysis of random design linear regression [HKZ14, Oli16, Mou22] with techniques from linear system identification with full state observation [SMT⁺18, SR19, FTM18, DMM⁺20]. Note that Lemma 5.1 makes no explicit assumptions on the ergodicity of the process P_x. The role of P_x is instead succinctly captured by the trajectory small-ball condition, together with the minimum eigenvalue quantities that appear in the bound. The proof of Lemma 5.1 also yields, with some straightforward modifications, bounds on the risk that hold with high probability; we only present bounds in expectation for simplicity. Finally, if the square norm $\|X\|_M^2$ is defined to be $\lambda_{\max}(XMX^\mathsf{T})$ instead of $\mathrm{tr}(XMX^\mathsf{T})$, then (5.4) holds with the expression $p + n$ replacing pn in the numerator.

As long as the process P_x satisfies the trajectory small-ball condition with excitation window $k = T$, Lemma 5.1 (with $\Psi_1 = \underline{\Gamma} = \Gamma_T(\mathsf{P}_x)$) immediately yields the following result for learning from many trajectories in the $\mathsf{Seq\text{-}LS}$ problem:

Problem	Many?	Upper	Lower	Assumptions
Seq-LS	Y	$\frac{pn}{mT}$	✓	TrajSB[(a)]
Seq-LS	N	$\frac{pn}{mT}$	$e^{-T/(\tau_{\mathsf{mix}} n)} + \frac{pn}{T}$	Ergodicity of covariates[(b)]
LDS-LS	Y	$\frac{pn}{mT}$	✓	k_{c}-step controllability[(c)]
LDS-LS	N	$\gamma \frac{pn^2}{m^2 T}$	$\frac{pn^2}{m^2 T}$	Marginal stability, etc.[(d)]
Ind-Seq-LS	N	$\frac{pn}{mT}$	✓	Small-ball, poly variance growth[(e)]
Ind-LDS-LS	N	$\gamma \frac{pn}{mT}$	$\frac{pn}{mT}$	Diagonalizable[(f)]
LDS-SysID	Y	$\frac{n^2}{mT\lambda_{\min}(\Gamma_T)}$	-	Same as LDS-LS (Y)[(g)]
LDS-SysID	N	$\gamma \frac{n^2}{mT\lambda_{\min}(\Gamma_{mT/n})}$	$\frac{n^2}{T^2}$	Same as LDS-LS (N)[(h)]

Table 1: Summary of main results presented in Section 5 and Section 6. All upper/lower bounds shown suppress constant and polylogarithmic factors. The **Many?** column indicates whether the bounds apply in the many trajectories regime (Y) where $m \gtrsim n$, or the few trajectories regime (N) where $m \lesssim n$. A checkmark (✓) in the **Lower** column indicates that the lower bound matches the upper bound, up to polylogarithmic factors. The LDS-SysID problem is classic linear system identification: recover the unknown dynamics matrix A from linear dynamical trajectories, with error measured in squared Frobenius norm (see Equation (3.8) and the discussion at the end of Section 3.4). Elaborations on assumptions:

(a) Upper bound follows from Theorem 5.2, treating as $O(1)$ all constants related to trajectory small-ball (Definition 4.1). Lower bound follows from iid linear regression being a special case.

(b) Upper bound follows from combining (i) Lemma B.23, a general OLS high probability upper bound that utilizes a simple modification (Definition B.1) to our trajectory small-ball definition, with (ii) Proposition B.24, which shows that a ϕ-mixing (Definition B.2) covariate process (where the marginal distributions also fulfill a weak small-ball condition (B.29)) satisfies our modified trajectory small-ball condition. The upper bound holds in high probability instead of in expectation, and requires a burn-in time that satisfies $T \gtrsim \tau_{\mathsf{mix}} n \log(1/\delta)$, where δ denotes the failure probability. The lower bound is from [BJN+20, Theorems 1 and 3], and holds for the single trajectory ($m = 1$).

(c) Upper bound follows from Theorem 5.5; see Definition 4.2 for definition of k_{c}-step controllability. Lower bound follows again from iid linear regression being a special case.

(d) Upper bound follows from Theorem 5.6, under Assumption 5.1 (marginal stability), Assumption 5.2 (diagonalizability), and Assumption 5.3 (one-step controllability). The condition number of the diagonalizing factor is denoted by γ (Definition 5.1). Lower bound follows from Theorem 6.3, and is realized by a decoupled noise sequence (Definition 7.1).

(e) Upper bound follows from Theorem 5.3, treating as $O(1)$ constants relating to small-ball (Equation (5.7)) and variance growth (Equation (5.8)). The necessity of the variance growth condition is shown in Theorem 6.2. Note that in the many trajectories regime, the Seq-LS upper bound applies. Lower bound again follows from iid linear regression.

(f) Upper bound follows from Theorem 5.7; γ is the condition number of the diagonalizing factor (Definition 5.1). Lower bound again follows from iid linear regression.

(g) Upper bound follows from Theorem 5.4; Γ_T is the T-step average covariance matrix (Equation (3.3)).

(h) Upper bound follows from Theorem 5.8; γ is the condition number of the diagonalizing factor (Definition 5.1), and $\Gamma_{mT/n}$ is the mT/n-step average covariance matrix (Equation (3.3)). Lower bound applies to the single trajectory ($m = 1$) setting and follows from [SMT+18, Theorem 2.3]. (Technically, their bound applies to the operator, instead of Frobenius norm, but the proof can be adjusted to apply.) Specializing the upper bound to one trajectory ($m = 1$), drawn from the lower bound's hard instance, implies a gap of n^3/T^2 (upper) versus n^2/T^2 (lower) as noted in [SMT+18, Section 2.2].

Theorem 5.2 (Upper bound for $\mathsf{Seq\text{-}LS}$, many trajectories)**.** *There are univeral positive constants c_0 and c_1 such that the following holds. Suppose that P_x satisfies the trajectory small-ball condition (Definition 4.1) with parameters $(T, T, \Gamma_T(\mathsf{P}_x), c_{\mathsf{sb}}, \alpha)$. If:*

$$n \geqslant 2, \quad m \geqslant \frac{c_0 n}{\alpha} \log\left(\frac{\max\{e, c_{\mathsf{sb}}\}}{\alpha}\right),$$

then, for any $\Gamma' \in \mathsf{Sym}^n_{>0}$:

$$\mathbb{E}[\|\hat{W}_{m,T} - W_\star\|^2_{\Gamma'}] \leqslant c_1 c_{\mathsf{sb}} \sigma_\xi^2 \cdot \frac{pn}{mT\alpha\underline{\lambda}(\Gamma_T(\mathsf{P}_x), \Gamma')} \cdot \log\left(\frac{\max\{e, c_{\mathsf{sb}}\}}{\alpha}\right). \tag{5.5}$$

This result provides the upper bound for the summary statement Theorem 1.1. To interpret the bound (5.5), suppose that c_{sb} and α are universal constants. Then, the requirement on m simplifies to $m \gtrsim n$. Under any strict evaluation horizon $T' \leqslant T$, taking $\Gamma' = \Gamma_{T'}(\mathsf{P}_x)$, the risk $\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x)]$ scales as $\sigma_\xi^2 pn/(mT)$. The lower bound for Theorem 1.1 follows from the fact that iid linear regression is a special case of $\mathsf{Seq\text{-}LS}$.

Meanwhile, to obtain guarantees for parameter recovery, consider taking $\Gamma' = I_n$. Then Theorem 5.2 implies that the parameter error $\mathbb{E}[\|\hat{W}_{m,T} - W_\star\|_F^2]$ scales as $\sigma_\xi^2 pn/[mT \cdot \lambda_{\min}(\Gamma_T(\mathsf{P}_x))]$. Note that operator norm bounds on parameters also hold, with the expression $p + n$ replacing pn in the bound.

Lemma 5.1 also yields a bound for $\mathsf{Ind\text{-}Seq\text{-}LS}$, assuming polynomial growth of the time-t covariances Σ_t (3.3). To state the result, let $\phi : [1, \infty) \times [0, \infty) \to [1, \infty)$ be defined as:

$$\phi(a, x) := \begin{cases} 1 & \text{if } x \leqslant 1, \\ ax & \text{otherwise.} \end{cases} \tag{5.6}$$

Note that $1 \leqslant \phi(a, x) \leqslant \max\{ax, 1\}$.

Theorem 5.3 (Upper bound for $\mathsf{Ind\text{-}Seq\text{-}LS}$)**.** *There are universal positive constants c_0 and c_1 such that the following holds. Fix any sequence of distributions $\{\mathsf{P}_{x,t}\}_{t\geqslant 1}$, and let $\Sigma_t := \mathbb{E}_{x_t \sim \mathsf{P}_{x,t}}[x_t x_t^\mathsf{T}]$ for $t \in \mathbb{N}_+$. Suppose there exists $c_{\mathsf{sb}} > 0$ and $\alpha \in (0, 1]$ such that for all $v \in \mathbb{R}^n \setminus \{0\}$, $\varepsilon > 0$ and $t \in \mathbb{N}_+$:*

$$\mathbb{P}_{x_t \sim \mathsf{P}_{x,t}}\left\{\langle v, x_t\rangle^2 \leqslant \varepsilon \cdot v^\mathsf{T} \Sigma_t v\right\} \leqslant (c_{\mathsf{sb}}\varepsilon)^\alpha. \tag{5.7}$$

Furthermore, suppose there exists a $c_\beta \geqslant 1$ and $\beta \geqslant 0$ such that for all $s, t \in \mathbb{N}_+$ satisfying $s \leqslant t$:

$$\frac{1}{\underline{\lambda}(\Sigma_s, \Sigma_t)} \leqslant c_\beta (t/s)^\beta. \tag{5.8}$$

If:

$$n \geqslant 2, \quad mT \geqslant \frac{c_0 n}{\alpha}\left(\beta + \log\left(\frac{\max\{e, c_{\mathsf{sb}}\} c_\beta}{\alpha}\right)\right),$$

then, for $\mathsf{P}_x = \otimes_{t\geqslant 1} \mathsf{P}_{x,t}$:

$$\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x)] \leqslant c_1 c_{\mathsf{sb}} \sigma_\xi^2 c_\beta e^\beta \cdot \frac{pn}{mT\alpha} \cdot \phi\left(c_\beta(\beta+1), (T'/T)^\beta\right)\left[\beta + \log\left(\frac{\max\{e, c_{\mathsf{sb}}\} c_\beta}{\alpha}\right)\right]. \tag{5.9}$$

Consider specializing Theorem 5.3 to the case when $\Sigma_t = \Sigma$ for all $t \in \mathbb{N}_+$. Doing so yields random design linear regression from mT covariates drawn iid from $\mathsf{P}_{x,1}$. The growth condition (5.8) is trivially satisfied with $c_\beta = 1$ and $\beta = 0$. The small-ball assumption (5.7) simplifies to $\mathbb{P}_{x_1 \sim \mathsf{P}_{x,1}}\{|\langle v, x_1\rangle| \leqslant \varepsilon\|v\|_\Sigma\} \leqslant (\sqrt{c_{\mathsf{sb}}}\varepsilon)^{2\alpha}$ for all $v \neq 0$ and $\varepsilon > 0$, which matches [Mou22, Assumption 1] up to a minor redefinition of the constants c_{sb}, α. Treating c_{sb} and α as constants, the conclusion of Theorem 5.3 in this setting is that $\mathbb{E}[\|\hat{W}_{m,T} - W_\star\|_\Sigma^2] \lesssim \sigma_\xi^2 pn/(mT)$ as long as $n \geqslant 2$ and $mT \gtrsim n$, which recovers [Mou22, Proposition 2].

On the other hand, Theorem 5.3 does not require that the covariates are drawn iid from the same distribution, allowing the time-t covariances Σ_t to grow polynomially. As an example, suppose that $\Sigma_t = t^\beta \cdot I_n$ for some $\beta > 0$. In this case, $1/\underline{\lambda}(\Sigma_s, \Sigma_t) = (t/s)^\beta$, so we can take $c_\beta = 1$ in (5.8). Again treating c_{sb} and α as constants and taking $T' \leqslant T$, we have $\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x)] \lesssim \sigma_\xi^2 \beta e^\beta \cdot pn/(mT)$ as long as $mT \gtrsim \beta n$. If β is also considered a constant, then we further have the risk bound $\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x)] \lesssim \sigma_\xi^2 pn/(mT)$. This matches the minimax rate for iid linear regression.

It is natural to ask if the covariance growth condition (5.8) is needed under strict evaluation horizons $T' \leqslant T$.[3] In Section 6, we show that if the covariances are set to $\Sigma_t = 2^t \cdot I_n$ and $\mathsf{P}_{x,t} = N(0, \Sigma_t)$ (satisfying (5.7)), then the minimax risk $\mathsf{R}(m, T, T; \{\otimes_{t \geqslant 1} \mathsf{P}_{x,t}\})$ must scale at least $2^{cn/m}/T$ whenever $m \lesssim n$, for some positive constant c. Sub-exponential growth rates are therefore necessary for polynomial sample complexity. Determining the optimal dependence of β in (5.9) is left to future work.

Note that Theorem 5.3 is most interesting either when trajectories are few ($m \lesssim n$) or evaluations are extended ($T' > T$). When $m \gtrsim n$ and $T' \leqslant T$, one can usually apply Theorem 5.2 with $\Gamma' = \Gamma_{T'}(\mathsf{P}_x)$ instead, and avoid placing any requirements on the growth of covariances.

Considering any of the small-ball examples in Section 4.1, recall that when the excitation window k and the horizon T are equal, Theorem 5.2 provides an upper bound on the risk of OLS estimation for the corresponding Seq-LS problem. Specifically, for Example 4.1 and Example 4.6 with $k = T$, if $T' = T$ and trajectories are abundant ($m \gtrsim n$), then the OLS estimator's rate $\sigma_\xi^2 pn/(mT)$ matches its behavior in iid linear regression. Meanwhile, for the degree-D Volterra series (Example 4.7), we require that $m \gtrsim c_D \cdot n$, and the OLS risk bound scales as $\sigma_\xi^2 c'_D \cdot pn/(mT)$, for constants c_D and c'_D that only depend on D.

In order to cover scenarios in which trajectories may be relatively scarce, namely $m \lesssim n$, we need additional structure. More technically, when the small-ball condition is satisfied with $k < T$, one needs to further control the various eigenvalues that appear in Lemma 5.1 in order to bound the risk of OLS. Specifically for Ind-Seq-LS, a covariate growth assumption suffices: Example 4.3 combined with Theorem 5.3 yields an OLS risk bound. Furthermore, both Example 4.4 and Example 4.5 can be immediately combined with Lemma 5.1, since the matrices Ψ_j in these examples are bounded above and below by $\Gamma_T(\mathsf{P}_x)$ up to universal constant factors. But arbitrarily large risk can still be realized in the general Seq-LS problem, even when the trajectory small-ball condition is satisfied. To study the behavior of OLS across all regimes of trajectory count m, example dimensions p and n, and trajectory lengths T and T', we focus specifically on linear dynamical systems and the LDS-LS problem for our remaining upper bounds.

[3] Some regularity is needed when under extended evaluations $T' > T$, otherwise the risk could be arbitrarily large.

5.1 Upper bounds for linear dynamical system

In this section, we focus exclusively on dynamics $\mathsf{P}_x^{A,B}$ described by a linear dynamical system (3.8). As discussed previously, in order to apply Lemma 5.1 in the few trajectories regime when $m \lesssim n$ (or when $m \gtrsim n$ and $T' > T$), we must (a) show that the process $\mathsf{P}_x^{A,B}$ satisfies the trajectory small-ball condition, and (b) bound the various eigenvalues which appear in Lemma 5.1. Example 4.6 establishes that $\mathsf{P}_x^{A,B}$ satisfies the $(T, k, \Gamma_k(A,B), e, \frac{1}{2})$-TrajSB condition, as long as (A,B) is k_{c}-step controllable and $k \geqslant k_{\mathsf{c}}$, thus taking care of (a). To handle (b), we introduce additional assumptions on the dynamics matrices (A,B):

Assumption 5.1 (Marginal instability)**.** *The dynamics matrix A in* LDS-LS *is* marginally unstable*. That is, $\rho(A) \leqslant 1$, where $\rho(A)$ denotes the spectral radius of A.*

Assumption 5.2 (Diagonalizability)**.** *The dynamics matrix A in* LDS-LS *is* complex diagonalizable *as $A = SDS^{-1}$, where $S \in \mathbb{C}^{n\times n}$ is invertible and $D \in \mathbb{C}^{n\times n}$ is a diagonal matrix comprising the eigenvalues of A.*

Assumption 5.3 (One-step controllability)**.** *The control matrix B in* LDS-LS *has full row rank, i.e.,* $\operatorname{rank}(B) = n$*. Equivalently, the pair (A,B) is one-step controllable (Definition 4.2).*

Assumption 5.1 is fairly standard in the literature. Going beyond the regime $\rho(A) = 1 + \varepsilon$, where $\varepsilon \lesssim 1/T$, requires additional technical assumptions on the dynamics matrix A that we choose to avoid in the interest of simplicity; the OLS estimator is in general not a consistent estimator when $\rho(A) > 1$ and $m = 1$ (cf. [PM13, SR19]). The condition $\rho(A) \leqslant 1$ is often referred to as *marginal stability* in other work. We choose to call it marginally *unstable* instead, to emphasize the fact that such systems, namely at $\rho(A) = 1$, may not be ergodic and that the state can grow unbounded (e.g. have magnitude roughly t^n at time t).

Diagonalizability (Assumption 5.2) is less standard in the literature. We use it together with Assumption 5.1 and Assumption 5.3 to establish that $\underline{\lambda}(k,t;A,B) := \underline{\lambda}(\Gamma_k(A,B), \Gamma_t(A,B)) \gtrsim c \cdot k/t$ whenever $k \leqslant t$, where c is a constant that depends only on A and B (and not k and t). In previous work on linear system identification, the term $\underline{\lambda}(k,t;A,B)$ only appears under a logarithm, and so coarser analyses in the general case can still establish polynomial rates (cf. [SMT+18, Proposition A.1] and [SR19, Proposition 7.6]).[4] However, by allowing for evaluation lengths $T' > 1$, the dependence on $\underline{\lambda}(k,t;A,B)$ is no longer entirely confined under a logarithm (cf. Lemma 5.1). A sharp characterization is hence critical for deriving optimal rates. In Appendix A, we conjecture the correct scaling of $\underline{\lambda}(k,t;A,B)$ as a function of the ratio k/t and the largest Jordan block size of A, based on numerical simulation.

One-step controllability (Assumption 5.3) is also an assumption commonly made in linear system identification. It is clear that some form of controllability is needed, otherwise learning may be impossible (e.g. consider the extreme case of $B = 0$). General multi-step controllability does not suffice either: [TP21, Theorem 2] show that under a single trajectory ($m = 1$), n-step controllability (where n remains the state dimension) does not ensure finite risk, and even a more robust controllability definition [TP21, Definition 3] cannot ensure risk bounds better than exponential in the dimension n. Considering these barriers, we simply choose to rely on one-step controllability in the few-trajectory setting ($m \lesssim n$).

Finally, we introduce a condition number quantity that will feature commonly in our bounds:

[4] Note, however, that without diagonalizability, [SMT+18, Corollary A.2] can only guarantee a $\sqrt{n^2/T}$ rate for the operator norm of the parameter error in general, and this is likely not optimal.

Definition 5.1. *For dynamics matrices* (A, B) *in* LDS-LS *satisfying Assumption 5.2 and Assumption 5.3, the* condition number $\gamma(A, B)$ *is defined as:* $\gamma(A, B) := \frac{\lambda_{\max}(S^{-1}BB^{\mathsf{T}}S^{-*})}{\lambda_{\min}(S^{-1}BB^{\mathsf{T}}S^{-*})}$.

5.1.1 Many trajectories

Our first result instantiates Theorem 5.2 in the special case of $\Gamma' = I_n$, which yields a sharp bound for parameter recovery without requiring stability of the dynamics matrix A:

Theorem 5.4 (Parameter recovery upper bound for LDS-LS, many trajectories)**.** *There are universal positive constants* c_0 *and* c_1 *such that the following holds for any instance of* LDS-LS*. Suppose that* (A, B) *is* k_{c}*-step controllable, If* $n \geqslant 2$*,* $m \geqslant c_0 n$*, and* $T \geqslant k_{\mathsf{c}}$*, then:*

$$\mathbb{E}[\|\hat{W}_{m,T} - W_\star\|_F^2] \leqslant c_1\sigma_\xi^2 \cdot \frac{pn}{mT \cdot \lambda_{\min}(\Gamma_T(A, B))}.$$

Theorem 5.4 improves on existing linear system identification results in the following way: it replaces stability assumptions on the dynamics matrix A with a simpler assumption of relatively many trajectories ($m \gtrsim n$), and it guarantees a rate that is inversely proportional to the *total* number of examples mT instead of only one example per trajectory. In other words, our analysis does not need to "discard" the data within a trajectory, which is the case in [DMM+20, Proposition 1.1]. Additionally, although OLS is generally not a consistent estimator from one trajectory ($m = 1$) if the dynamics A are unstable, the results of [DMM+20] imply consistency as $m \to \infty$, i.e., that $\hat{W}_{m,T}$ converges in probability to $W_\star$ as $m \to \infty$. Theorem 5.4 adds that, provided $m \gtrsim n$, OLS is consistent under unstable systems as $T \to \infty$ as well, even if the trajectory count m remains finite. We will return to parameter recovery from relatively few trajectories ($m \lesssim n$) by this section's end.

We now look beyond an evaluation horizon of length one, and consider the setting with many trajectories ($m \gtrsim n$). As noted previously, in order to handle an arbitrary evaluation horizon T' (in particular those that extend past the training horizon T), some constraint on the admissible dynamics matrices is needed to ensure that the minimax risk remains finite. Without assumptions, the quantity $\underline{\lambda}(\Gamma_T(A, B), \Gamma_{T'}(A, B))$, whose inverse inevitably bounds the risk (3.2) from below, can be arbitrarily small whenever $T' > T$, resulting in arbitrarily large risk. We will use our stated assumptions from the beginning of this section. The following specializes Theorem 5.2 to LDS-LS:

Theorem 5.5 (Risk upper bound for LDS-LS, many trajectories)**.** *There are universal positive constants* c_0 *and* c_1 *such that the following holds for any instance of* LDS-LS*. Suppose that* (A, B) *is* k_{c}*-step controllable. If* $n \geqslant 2$*,* $m \geqslant c_0 n$*,* $T \geqslant k_{\mathsf{c}}$*, and the evaluation horizon is strict* ($T' \leqslant T$)*, then:*

$$\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x^{A,B})] \leqslant c_1\sigma_\xi^2 \cdot \frac{pn}{mT}.$$

On the other hand, suppose that (A, B) *satisfies Assumption 5.1, Assumption 5.2, and Assumption 5.3, with* $\gamma := \gamma(A, B)$ *(Definition 5.1). If* $n \geqslant 2$*,* $m \geqslant c_0 n$*, and the evaluation horizon is extended* ($T' > T$)*, then:*

$$\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x^{A,B})] \leqslant c_1\sigma_\xi^2 \cdot \frac{pn}{mT} \cdot \gamma\frac{T'}{T}.$$

Setting $T' = T$, Theorem 5.5 states that the risk of LDS-LS in the many trajectories regime satisfies $\mathbb{E}[L(\hat{W}_{m,T}; T, \mathsf{P}_x^{A,B})] \lesssim \sigma_\xi^2 pn/(mT)$. This rate matches the corresponding independent baseline Ind-LDS-LS in the many trajectories regime. To see this, first observe that the marginal distribution $\mathsf{P}_{x,t}^{A,B}$ at time $t \in \mathbb{N}_+$ is $N(0, \Sigma_t(A, B))$. Hence, the covariate distribution for Ind-LDS-LS corresponds to the product distribution $\otimes_{t \geqslant 1} N(0, \Sigma_t(A, B))$, which is an instance of a Gaussian process. Therefore, Example 4.2 combined with Theorem 5.2 yields that the Ind-LDS-LS problem also has a risk bound that scales as $\sigma_\xi^2 pn/(mT)$ whenever $m \gtrsim n$. Put differently, the dependent structure of the covariate distribution $\mathsf{P}_x^{A,B}$ in LDS-LS does not add any statistical overhead to the learning problem (compared to the independent learning problem Ind-LDS-LS), as long as $m \gtrsim n$.

5.1.2 Few trajectories

We now cover the regime in which relatively few training trajectories are available ($m \lesssim n$). Our first result bounds the OLS risk for the LDS-LS problem:

Theorem 5.6 (Risk upper bound for LDS-LS, few trajectories)**.** *There are universal positive constants c_0, c_1, and c_2 such that the following holds for any instance of* LDS-LS*. Suppose that (A, B) satisfies Assumption 5.1, Assumption 5.2, and Assumption 5.3, with $\gamma := \gamma(A, B)$ (Definition 5.1). If $n \geqslant 2$, $m \leqslant c_0 n$, and $mT \geqslant c_1 n \log(\max\{\gamma n/m, e\})$, then:*

$$\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x^{A,B})] \leqslant c_2 \sigma_\xi^2 \cdot \frac{pn \log(\max\{\gamma n/m, e\})}{mT} \cdot \phi\left(\gamma, \frac{c_1 n \log(\max\{\gamma n/m, e\})}{m} \cdot \frac{T'}{T}\right).$$

To interpret Theorem 5.6, consider γ a constant and suppose that $T' = T$. Then Theorem 5.6 states that $\mathbb{E}[L(\hat{W}_{m,T}; T, \mathsf{P}_x^{A,B})] \lesssim \sigma_\xi^2 \cdot pn/(mT) \cdot n \log^2(n/m)/m$. We now see that this LDS-LS risk is an extra $n \log^2(n/m)/m$ factor larger than the risk of the baseline problem Ind-LDS-LS:

Theorem 5.7 (Risk upper bound for Ind-LDS-LS)**.** *There are universal positive constants c_0 and c_1 such that the following holds for any instance of* Ind-LDS-LS*. Suppose that (A, B) satisfies Assumption 5.1, Assumption 5.2, and Assumption 5.3, with $\gamma := \gamma(A, B)$ (Definition 5.1). If $n \geqslant 2$ and $mT \geqslant c_0 n \log(\max\{\gamma, e\})$, then:*

$$\mathbb{E}[L(\hat{W}_{m,T}; T', \otimes_{t \geqslant 1} \mathsf{P}_{x,t}^{A,B})] \leqslant c_1 \sigma_\xi^2 \cdot \frac{pn\gamma \log(\max\{\gamma, e\})}{mT} \cdot \phi\left(\gamma, \frac{T'}{T}\right).$$

Treating γ as a constant and setting $T' = T$, Theorem 5.7 states that $\mathbb{E}[L(\hat{W}_{m,T}; T, \otimes_{t \geqslant 1} \mathsf{P}_{x,t}^{A,B})]$ scales as $\sigma_\xi^2 pn/(mT)$, matching the risk of iid linear regression up to constant factors. In Section 6, we will see that the result of Theorem 5.6 is sharp up to constants, and therefore the LDS-LS problem is fundamentally more difficult than its corresponding baseline problem Ind-LDS-LS when trajectories are relatively scarce.

We conclude with our final upper bound, using our assumptions to generalize [SMT+18, Theorem 2.1] to the few-trajectory setting:

Theorem 5.8 (Parameter recovery upper bound for LDS-LS, few trajectories)**.** *There are universal positive constants c_0, c_1, and c_2 such that the following holds for any instance of* LDS-LS*. Suppose that (A, B) satisfies Assumption 5.1, Assumption 5.2, and Assumption 5.3, with $\gamma := \gamma(A, B)$ (Definition 5.1). If $n \geqslant 2$, and $mT \geqslant c_0 n \log(\max\{\gamma n/m, e\})$, then:*

$$\mathbb{E}[\|\hat{W}_{m,T} - W_\star\|_F^2] \leqslant c_1 \sigma_\xi^2 \cdot \frac{pn \log(\max\{\gamma n/m, e\})}{mT \cdot \lambda_{\min}(\Gamma_{k_\star}(A, B))}, \quad k_\star := \left\lfloor \frac{c_2 T}{n/m \cdot \log(\max\{\gamma n/m, e\})} \right\rfloor.$$

Theorem 5.8 complements Theorem 5.4; together they cover parameter recovery across all problem regimes. Again, operator norm bounds also hold with $p + n$ in place of pn.

5.2 Comparison to learning from trajectories of multiple unknown systems

As mentioned in Section 2, [CP22, MFTM22] both study the setup where a learner observes multiple independent trajectories from K different unknown linear dynamical systems. The task is to identify the parameters of the K underlying systems. This is more general than the setting we consider, which is recovered by fixing $K = 1$. However, specializing these rates to our setting yield either unnecessary requirements, suboptimal bounds, or both.

To see this, first, if we specialize [CP22, Theorem 1] to our setup, we generate unnecessary assumptions. Specifically, Theorem 1 requires strict stability, one-step controllability, and $mT \gtrsim \max\{n^3, 1/(1-\rho)\}$, where ρ is the spectral radius of A. In comparison, Theorem 5.4 only requires k_{c}-step controllability, $T \geqslant k_{\mathsf{c}}$, and $m \gtrsim n$. However, note that Theorem 1, like Theorem 5.4, does have the property that the parameter error (in operator norm) scales as $\sqrt{n/(mT)}$, reflecting that all collected datapoints contribute to reducing error.

Next, we specialize [MFTM22, Theorem 2]. Theorem 2 bounds the error of an estimation procedure which outputs m different estimates $\{\hat{A}_i\}_{i=1}^m$, one for each observed trajectory (cf. Eq. (3)). Specifically, it gives an upper bound on the quantity $\frac{1}{m}\sum_{i=1}^m \|\hat{A}_i - A_i\|_F^2$, where A_i is the dynamics matrix associated with the i-th trajectory. To specialize this to our setting, we average the estimates and apply Jensen's inequality followed by Theorem 2. This yields the bound $\|\hat{A} - A\|_F^2 \lesssim 1/T + n^2/(mT)$, where $\hat{A} := \frac{1}{m}\sum_{i=1}^m \hat{A}_i$ is the averaged estimate. We see that, for a fixed T, as $m \to \infty$, the rate tends to $1/T$ instead of zero (compared with the $n^2/(mT)$ bound from Theorem 5.4). Additionally, Theorem 2 requires both that the dynamics are one-step controllable and that the spectral radius of A is bounded by $1 + O(1/T)$.

6 Risk lower bounds

Our lower bounds rely on the following statement, that the expected trace inverse covariance—a classic quantity in asymptotic statistics—bounds the minimax risk from below:

Lemma 6.1 (Expected trace of inverse covariance bounds risk from below)**.** *Fix $m, T \in \mathbb{N}_+$ and a set of covariate distributions $\mathcal{P}_x$. Suppose that for every $\mathsf{P}_x \in \mathcal{P}_x$, the data matrix $X_{m,T} \in \mathbb{R}^{mT \times n}$ drawn from $\otimes_{i=1}^m \mathsf{P}_x$ has full column rank almost surely. The minimax risk $\mathsf{R}(m, T, T'; \mathcal{P}_x)$ satisfies:*

$$\mathsf{R}(m, T, T'; \mathcal{P}_x) \geqslant \sigma_\xi^2 p \cdot \sup_{\mathsf{P}_x \in \mathcal{P}_x} \mathbb{E}_{\otimes_{i=1}^m \mathsf{P}_x}\left[\operatorname{tr}\left(\Gamma_{T'}^{1/2}(\mathsf{P}_x)(X_{m,T}^\mathsf{T} X_{m,T})^{-1}\Gamma_{T'}^{1/2}(\mathsf{P}_x)\right)\right].$$

Lemma 6.1 is well known, possibly considered folklore; we state and prove it for completeness. Our proof is inspired by a recent argument from [Mou22]. It smooths over problem instances according to a Gaussian prior, and analytically characterizes the posterior distribution of the parameter $W_\star$ under a simple Gaussian observation model detailed in Section 7.2.

Our first lower bound underscores the need to make variance growth assumptions (5.8), in Theorem 5.3, for $\mathsf{Ind\text{-}Seq\text{-}LS}$ in the few trajectories ($m \lesssim n$) regime:

Theorem 6.2 (Need for growth assumptions in $\mathsf{Ind\text{-}Seq\text{-}LS}$ when $m \lesssim n$)**.** *There exists universal constant c_0, c_1, and c_2 such that the following holds. Suppose that $\mathsf{P}_x = \otimes_{t \geqslant 1} N(0, 2^t \cdot I_n)$, $n \geqslant 6$, $mT \geqslant n$, and $m \leqslant c_0 n$. Then:*

$$\mathsf{R}(m, T, T; \{\mathsf{P}_x\}) \geqslant c_1 \sigma_\xi^2 \cdot \frac{p \cdot 2^{c_2 n/m}}{T}.$$

Theorem 6.2 states that if the variances Σ_t are allowed to grow exponentially in t, then the minimax risk of $\mathsf{Ind\text{-}Seq\text{-}LS}$ scales exponentially in n/m when $m \lesssim n$. Thus, some sub-exponential growth assumption is necessary in order to have the risk scale polynomially in n/m.

We now turn to a lower bound for $\mathsf{LDS\text{-}LS}$. We consider two particular hard instances for $\mathsf{LDS\text{-}LS}$ dynamics matrices (A, B), where we set $B = I_n$ and vary A. The first instance corresponds to iid covariates, i.e., $A = 0_{n\times n}$. The second instance corresponds to an isotropic Gaussian random walk, i.e., $A = I_n$. These two hard instances satisfy Assumption 5.1, Assumption 5.2, and Assumption 5.3. Together they show that our upper bounds are sharp up to logarithmic factors, treating the condition number $\gamma(A, B)$ from Definition 5.1 as a constant:

Theorem 6.3 (Risk lower bound)**.** *There are universal positive constants c_0, c_1, and c_2 such that the following holds. Recall that $\mathsf{P}_x^{I_n}$ (resp. $\mathsf{P}_x^{0_{n\times n}}$) denotes the covariate distribution for a linear dynamical system with $A = I_n$ and $B = I_n$ (resp. $A = 0_{n\times n}$ and $B = I_n$). If $T \geqslant c_0$, $n \geqslant c_1$, and $mT \geqslant n$, then:*

$$\mathsf{R}(m, T, T'; \{\mathsf{P}_x^{0_{n\times n}}, \mathsf{P}_x^{I_n}\}) \geqslant c_2\sigma_\xi^2 \cdot \frac{pn}{mT} \cdot \max\left\{\frac{nT'}{mT}, \frac{T'}{T}, 1\right\}.$$

We can interpret this lower bound by a breakdown of the quantity $\varphi := \max\{nT'/(mT), T'/T, 1\}$ across various regimes. When trajectories are limited ($m \lesssim n$), $\varphi \asymp \max\{nT'/(mT), 1\}$, and therefore the minimax risk is bounded below by $\sigma_\xi^2 \cdot pn/(mT) \cdot \max\{nT'/(mT), 1\}$. This is the same rate prescribed by the OLS upper bound of Theorem 5.6, up to the condition number $\gamma(A, B)$ and logarithmic factors in n/m. We have thus justified the summary statement Theorem 1.2. On the other hand, under many trajectories ($m \gtrsim n$), $\varphi \asymp \max\{T'/T, 1\}$ and the minimax risk is bounded below by $\sigma_\xi^2 \cdot pn/(mT) \cdot \max\{T'/T, 1\}$. By Theorem 5.5, the OLS risk is bounded above by the same quantity times $\gamma(A, B)$, justifying the summary statement Theorem 1.3.

7 Key proof ideas

In this section, we highlight some of the key ideas behind our results. Proofs of the upper bounds are in Appendix B, and proofs of the lower bounds are in Appendix C.

Additional notation. For $r \in \mathbb{N}_+$ and $M \in \mathbb{R}^{n\times n}$, let $J_r \in \mathbb{R}^{r\times r}$ denote the Jordan block of size r with ones along its diagonal, let $\mathsf{BDiag}(M, r) \in \mathbb{R}^{nr\times nr}$ denote the block diagonal matrix with diagonal blocks M, and let $\mathsf{BToep}(M, r) \in \mathbb{R}^{nr\times nr}$ denote the block Toeplitz matrix with first column $(I_n, M^\mathsf{T}, \ldots, (M^{r-1})^\mathsf{T})^\mathsf{T}$.

7.1 Upper bounds

The proof of Lemma 5.1 decomposes the risk using a standard basic inequality, which we now describe. While Lemma 5.1 is stated quite generally, for simplicity of exposition we restrict ourselves in this section to the case when the matrix parameters $\{\Psi_j\}_{j=1}^S$ in Definition 4.1 are all set to $\underline{\Gamma}$. Under this simplification, we have that $\underline{\mu}(\{\Psi_j\}_{j=1}^S, \underline{\Gamma}) = 1$.

Equation (3.1) yields the identity $Y_{m,T} = X_{m,T}W_\star^\mathsf{T} + \Xi_{m,T}$. Plugging this relationship into the formula (3.6) for $\hat{W}_{m,T}$ gives $\hat{W}_{m,T} - W_\star = \Xi_{m,T}^\mathsf{T} X_{m,T}(X_{m,T}^\mathsf{T} X_{m,T})^{-1}$. Define the whitened version of $X_{m,T}$ as $\tilde{X}_{m,T} := X_{m,T}\underline{\Gamma}^{-1/2}$. From these definitions and after some basic manipulations, for any $\Gamma' \in \mathsf{Sym}^n_{>0}$:

$$\|\hat{W}_{m,T} - W_\star\|_{\Gamma'}^2 \leqslant \min\{n,p\}\frac{\|(\tilde{X}_{m,T}^\mathsf{T}\tilde{X}_{m,T})^{-1/2}\tilde{X}_{m,T}^\mathsf{T}\Xi_{m,T}\|_{\mathrm{op}}^2}{\lambda_{\min}(\tilde{X}_{m,T}^\mathsf{T}\tilde{X}_{m,T})\cdot\underline{\lambda}(\underline{\Gamma},\Gamma')}. \tag{7.1}$$

This decomposes the analysis into two parts: (a) upper-bounding $\|(\tilde{X}_{m,T}^\mathsf{T}\tilde{X}_{m,T})^{-1/2}\tilde{X}_{m,T}^\mathsf{T}\Xi_{m,T}\|_{\mathrm{op}}^2$, which is a self-normalized martingale term, and (b) lower-bounding the term $\lambda_{\min}(\tilde{X}_{m,T}^\mathsf{T}\tilde{X}_{m,T})$. The analysis for the martingale term is fairly standard [cf. AYPS11, Corollary 1], so for the remainder of this section we focus on the minimum eigenvalue bound, which contains much of what is novel in our analysis.

We first demonstrate how the trajectory small-ball definition (Definition 4.1) can be used to establish *pointwise* convergence of the quadratic form $\chi(v) := \sum_{i=1}^m\sum_{t=1}^T\langle v,\tilde{x}_t^{(i)}\rangle^2$ for $v\in\mathbb{S}^{n-1}$, where $\tilde{x}_t^{(i)} := \underline{\Gamma}^{-1/2}x_t^{(i)}$ is a whitened state vector. Specifically, we show that for a fixed $v\in\mathbb{S}^{n-1}$, the probability of the event $\{\chi(v)\leqslant\psi\cdot\varepsilon\}$ is small, for ψ,ε to be specified.

The key idea is that for any non-negative random variable X satisfying $\mathbb{P}\{X\leqslant\varepsilon\}\leqslant(c\varepsilon)^\alpha$ for all $\varepsilon>0$, the moment generating function satisfies $\mathbb{E}[\exp(-\eta X)]\leqslant(c/\eta)^\alpha$ for all $\eta>0$ (cf. Proposition B.4). Hence, by condition (4.1) from Definition 4.1, for any $\eta>0$:

$$\mathbb{E}\left[\exp\left(-\frac{\eta}{k}\sum_{t=(j-1)k+1}^{jk}\langle v,\tilde{x}_t^{(i)}\rangle^2\right)\;\middle|\;\mathcal{F}_{(j-1)k}\right]\leqslant\left(\frac{c_{\mathsf{sb}}}{\eta}\right)^\alpha\ \text{a.s.},\ i=1,\dots m,\ j=1,\dots S.$$

By a Chernoff bound, the tower property of conditional expectation, and the independence of the trajectories $\{\tilde{x}_t^{(i)}\}_{t\geqslant1}$ and $\{\tilde{x}_t^{(i')}\}_{t\geqslant1}$ when $i\neq i'$, for any $\psi>0$:

$$\begin{aligned}\mathbb{P}\left(\frac{1}{k}\sum_{i=1}^m\sum_{t=1}^T\langle v,\tilde{x}_t^{(i)}\rangle^2\leqslant\zeta\right)&\leqslant\inf_{\eta\geqslant0}e^{\eta\zeta}\mathbb{E}\exp\left(-\frac{\eta}{k}\sum_{i=1}^m\sum_{t=1}^T\langle v,\tilde{x}_t^{(i)}\rangle^2\right)\\&\leqslant\inf_{\eta\geqslant0}e^{\eta\zeta}\left(\frac{c_{\mathsf{sb}}}{\eta}\right)^{mS\alpha}\\&=\exp\left(-mS\alpha\left(\log\left(\frac{mS\alpha}{c_{\mathsf{sb}}\zeta}\right)-1\right)\right).\end{aligned}$$

Now with a change of variables $t:=\log\left(\frac{mS\alpha}{c_{\mathsf{sb}}\zeta}\right)-1$, we obtain:

$$\mathbb{P}\left(\sum_{i=1}^m\sum_{t=1}^T\langle v,\tilde{x}_t^{(i)}\rangle^2\leqslant\frac{mT\alpha}{2c_{\mathsf{sb}}}e^{-(t+1)}\right)\leqslant\exp(-mS\alpha t)\quad\forall t>0. \tag{7.2}$$

The key upshot of (7.2) is that it controls tail probability at all scales. This control is needed in order to bound the expected value of (7.1) by integration. At this point, it remains to upgrade (7.2) from pointwise to uniform over $\mathbb{S}^{n-1}$. A natural approach is to use standard covering and union bound arguments, as is done in [SMT+18]. However, straightforward covering argument yields un-necessary logarithmic factors in the covariate dimension n. In order to circumvent this issue, we utilize the PAC-Bayes argument from [Mou22] (which itself is an extension of [Oli16]) to establish uniform concentration. The details are given in Appendix B.4.

7.2 Lower bounds

7.2.1 Observation noise behind Lemma 6.1

Our definition of minimax risk $\mathsf{R}(m, T, T'; \mathcal{P}_x)$ in (3.5) involves a supremum over the worst case σ_ξ-sub-Gaussian MDS distribution that models the observation noise. The proof of Lemma 6.1 bounds this supremum from below by considering a noise model that decouples the observation noise $\{\xi_t\}_{t\geqslant 1}$ from the randomness that drives the trajectory $\{x_t\}_{t\geqslant 1}$:

Definition 7.1 (Gaussian observation noise)**.** *The* Gaussian observation noise model *holds when* $\xi_t \sim N(0, \sigma_\xi^2 I_p)$*,* $\xi_t \perp \xi_{t'}$ *if* $t \neq t'$*, and the process* $\{\xi_t\}_{t\geqslant 1}$ *is independent from the process* $\{x_t\}_{t\geqslant 1}$*.*

Decoupling the noise processes orthogonalizes the two problems simultaneously present in $\mathsf{Seq\text{-}LS}$: learning the dynamics of covariates and learning the responses from covariates. Definition 7.1 draws attention to the latter. It will unfortunately exclude us from addressing linear system identification specifically with our lower bound, but it allows a sharp and simple characterization of the minimax risk in general. The proof of Lemma 6.1 is given in Appendix C.2.

7.2.2 An analysis of non-isotropic gramian matrices

A key technical challenge for our analysis lies in constructing a sharp lower bound on the expected trace inverse of a gramian matrix formed by random non-isotropic Gaussian random vectors. Specifically, for integers $q, n \in \mathbb{N}_+$ with $q \geqslant n$, and for a fixed positive definite matrix $\Sigma \in \mathsf{Sym}^q_{>0}$, we are interested in a lower bound on the quantity $\mathbb{E}\,\mathrm{tr}((W^\mathsf{T}\Sigma W)^{-1})$, where $W \in \mathbb{R}^{q\times n}$ has iid $N(0,1)$ entries. The matrix $W^\mathsf{T}\Sigma W$ is equal in distribution to the gramian matrix $Y \in \mathbb{R}^{n\times n}$ of the vectors $g_1, \ldots, g_n \in \mathbb{R}^q$, which are drawn iid from $N(0, \Sigma)$, i.e., $Y_{ij} = \langle g_i, g_j\rangle$.

The main tool we use to analyze $\mathbb{E}\,\mathrm{tr}((W^\mathsf{T}\Sigma W)^{-1})$ is the convex Gaussian min-max theorem (CGMT) from [TOH14], which allows us to bound from below the expected trace inverse by studying a two dimensional min-max game that is more amenable to analysis. The key idea is to cast the expected trace inverse as a least-norm optimization problem, and apply CGMT to the value of the optimization problem. We believe the following result to be of independent interest.

Lemma 7.1. *Let* q, n *be positive integers with* $q \geqslant n$ *and* $n \geqslant 2$*. Let* $W \in \mathbb{R}^{q\times n}$ *have iid* $N(0,1)$ *entries, and let* $\Sigma \in \mathbb{R}^{q\times q}$ *be positive definite. Let* $g \sim N(0, I_q)$ *and* $h \sim N(0, I_{n-1})$*, with* g *and* h *independent. Also, let* $\{e_i\}_{i=1}^q$ *be the standard basis vectors in* $\mathbb{R}^q$*. We have:*

$$\mathbb{E}\,\mathrm{tr}((W^\mathsf{T}\Sigma W)^{-1}) \geqslant \frac{n}{\sum_{i=1}^q \mathbb{E}\min_{\beta\geqslant 0}\max_{\tau\geqslant 0}\left[-\frac{\beta\|h\|_2}{\tau} + \|\beta g - e_i\|^2_{(\Sigma^{-1}+\beta\|h\|_2\tau I_q)^{-1}}\right]}. \tag{7.3}$$

The proof of Lemma 7.1 appears in Appendix C.4. We now discuss how to analyze the two-dimensional min-max game appearing in Lemma 7.1. We first start by heuristically replacing it with a *stylized problem*, where the random quantities which appear in (7.3) are replaced by their expected scaling:

$$\mathsf{SP}(\Sigma, n) := \sum_{i=1}^q \min_{\beta\geqslant 0}\max_{\tau\geqslant 0} \underbrace{\left[-\frac{\beta\sqrt{n}}{\tau} + \beta^2\,\mathrm{tr}((\Sigma^{-1}+\beta\sqrt{n}\tau I_q)^{-1}) + (\Sigma^{-1}+\beta\sqrt{n}\tau I_q)^{-1}_{ii}\right]}_{=: \ell_i(\beta,\tau)}. \tag{7.4}$$

While (7.4) is not a valid upper bound on the value of the min-max game appearing in (7.3), analyzing (7.4) is simpler and gives the correct intuition; we give a rigorous upper bound in Lemma C.9.

We start by observing that if $\beta = 0$, then regardless of the choice of τ, $\ell_i(0,\tau) = \Sigma_{ii}$, and therefore $\sum_{i=1}^q \ell_i(0,\tau_i) = \operatorname{tr}(\Sigma)$ for any $\{\tau_i\}_{i=1}^q \subset \mathbb{R}^q_{\geqslant 0}$. On the other hand, if $\beta > 0$, then $\ell_i(\beta,\tau)$ tends to $-\infty$ as $\tau \to 0^+$ and to 0 as $\tau \to \infty$. Therefore, if we can show that there exists a $v \in (0, \operatorname{tr}(\Sigma))$, such that every set of points $\{(\beta_i,\tau_i)\}_{i=1}^q \subset \mathbb{R}^2_{>0}$ satisfying:[5]

$$\frac{\partial \ell_i}{\partial \beta}(\beta_i,\tau_i) = \frac{\partial \ell_i}{\partial \tau}(\beta_i,\tau_i) = 0, \quad i = 1,\dots,q, \tag{7.5}$$

also satisfies $v = \sum_{i=1}^q \ell_i(\beta_i,\tau_i)$, then $\mathsf{SP}(\Sigma,n) = v$.

To uncover the critical points, we define the functions f and q_i, for $i = 1,\dots,q$, as:

$$f(x) := -x\sqrt{n} + x^2 \operatorname{tr}((\Sigma^{-1} + x\sqrt{n}I_q)^{-1}), \quad q_i(x) := (\Sigma^{-1} + x\sqrt{n}I_q)^{-1}_{ii}.$$

With these definitions, we can write:

$$\ell_i(\beta,\tau) = \frac{1}{\tau^2} f(\beta\tau) + q_i(\beta\tau).$$

Calculating $\frac{\partial \ell_i}{\partial \beta}(\beta,\tau) = \frac{\partial \ell_i}{\partial \tau}(\beta,\tau) = 0$ yields, for $\tau \neq 0$:

$$0 = \frac{\partial \ell_i}{\partial \tau}(\beta,\tau) = \tau^{-2} f'(\beta\tau)\beta - 2\tau^{-3} f(\beta\tau) + q_i'(\beta\tau)\beta, \tag{7.6}$$

$$0 = \frac{\partial \ell_i}{\partial \beta}(\beta,\tau) = \tau^{-2} f'(\beta\tau)\tau + q_i'(\beta\tau)\tau. \tag{7.7}$$

The second condition (7.7) implies that $q_i'(\beta\tau) = -\tau^{-2} f'(\beta\tau)$. Plugging this condition into (7.6) implies that $f(\beta\tau) = 0$, and hence $\ell_i(\beta,\tau) = q_i(\beta\tau)$ for the critical point (β,τ). We now study the positive roots of the equation $f(x) = 0$, or equivalently:

$$x\sqrt{n} = x^2 \operatorname{tr}((\Sigma^{-1} + x\sqrt{n}I_q)^{-1}).$$

Using the variable substitution $y := x\sqrt{n}$, we have, when $y > 0$, the equivalent problem:

$$\psi(y;\Sigma) := y \operatorname{tr}((\Sigma^{-1} + yI_q)^{-1}) = n.$$

Observe that $\psi(0;\Sigma) = 0$ and $\lim_{y\to\infty} \psi(y;\Sigma) = q$. Furthermore, $\psi(y;\Sigma)$ is continuous and monotonically increasing with y. Therefore, as long as $q > n$, there is exactly one $\bar{y} \in (0,\infty)$ such that $\psi(\bar{y};\Sigma) = n$, or equivalently there is exactly one $\bar{x} \in (0,\infty)$ such that $\psi(\bar{x}\sqrt{n};\Sigma) = n$. Such a quantity $\bar{x}$ supplies the curve of critical points $\mathsf{Crit}(\bar{x}) := \{(\beta,\tau) \in \mathbb{R}^2_{>0} \mid \beta\tau = \bar{x}\}$. Note that $\mathsf{Crit}(\bar{x})$ is the set of critical points for every $\ell_i(\beta,\tau)$, $i = 1,\dots,q$. Furthermore, for any $(\beta_\star,\tau_\star) \in \mathsf{Crit}(\bar{x})$ and $i \in \{1,\dots,q\}$, we have that $\ell_i(\beta_\star,\tau_\star) = q_i(\beta_\star\tau_\star) = (\Sigma^{-1} + \bar{x}\sqrt{n}I_q)^{-1}_{ii}$. Therefore:

$$\{(\beta_i,\tau_i)\}_{i=1}^T \subset \mathsf{Crit}(\bar{x}) \Longrightarrow \sum_{i=1}^q \ell_i(\beta_i,\tau_i) = \operatorname{tr}((\Sigma^{-1} + \bar{x}\sqrt{n}I_q)^{-1}) \in (0, \operatorname{tr}(\Sigma)),$$

[5] The conditions given in (7.5) are *not* in general necessary first-order optimality conditions for a nonconvex/nonconcave game [see e.g. JNJ20, Proposition 21]. However, since for every $\beta > 0$, the function $\tau \mapsto \ell_i(\beta,\tau)$ has only strictly concave stationary points (Proposition C.10), these conditions are necessary for this particular problem.

and thus:

$$\mathsf{SP}(\Sigma, n) = \frac{\sqrt{n}}{\bar{x}}, \text{ with } \bar{x} \text{ the solution to } \psi(\bar{x}\sqrt{n}; \Sigma) = n. \tag{7.8}$$

In light of (7.8), Lemma 7.1 then suggests that:

$$\mathbb{E}\operatorname{tr}((W^\mathsf{T}\Sigma W)^{-1}) \gtrapprox \frac{n}{\mathsf{SP}(\Sigma, n)} = \bar{x}\sqrt{n}, \tag{7.9}$$

where the $\gtrapprox$ notation indicates the heuristic nature of replacing the expected min-max game appearing in the bound (7.3) with the approximation (7.4).

If we briefly check (7.9) in the simple case when $\Sigma = I_q$, we see that:

$$n = \psi(\bar{x}\sqrt{n}; I_q) = \bar{x}\sqrt{n}\frac{q}{1+\bar{x}\sqrt{n}} \Longrightarrow \bar{x}\sqrt{n} = \frac{n}{q}(1+\bar{x}\sqrt{n}) \geqslant \frac{n}{q}.$$

Hence, (7.9) yields that $\mathbb{E}\operatorname{tr}((W^\mathsf{T}W)^{-1}) \gtrapprox n/q$, which is the correct scaling; the exact result is $\mathbb{E}\operatorname{tr}((W^\mathsf{T}W)^{-1}) = n/(q-n-1)$ for $q \geqslant n+2$.

7.2.3 Ideas behind Theorem 6.2

We let $X_{m,T}$ denote the data matrix associated with m iid copies of $\{x_t\}_{t=1}^T$, with $x_t \sim N(0, 2^t \cdot I_n)$ and $x_t \perp x_{t'}$ for $t \neq t'$. We also define $\Gamma_T := \frac{1}{T}\sum_{t=1}^T 2^t \cdot I_n = \frac{2}{T}(2^T-1)\cdot I_n$, and observe that $\Gamma_T \succcurlyeq \frac{2^T}{T}\cdot I_n$. By Lemma 7.1, it suffices to lower bound the quantity $\mathbb{E}\operatorname{tr}(\Gamma_T^{1/2}(X_{m,T}^\mathsf{T}X_{m,T})^{-1}\Gamma_T^{-1/2})$. Since each column of $X_{m,T}$ is independent, the matrix $X_{m,T}2^{-T/2}$ has the same distribution as $\mathsf{BDiag}(\Theta^{1/2}, m)W$, where $\Theta \in \mathbb{R}^{T\times T}$ is diagonal, $\Theta_{ii} = 2^{i-T}$ for $i \in \{1, \dots, T\}$, and $W \in \mathbb{R}^{mT\times n}$ has iid $N(0,1)$ entries. In other words, we have:

$$\mathbb{E}\operatorname{tr}(\Gamma_T^{1/2}(X_{m,T}^\mathsf{T}X_{m,T})^{-1}\Gamma_T^{-1/2}) \geqslant \frac{1}{T}\mathbb{E}\operatorname{tr}((W^\mathsf{T}\mathsf{BDiag}(\Theta, m)W)^{-1}).$$

By the arguments in Section 7.2.2, we have:

$$\mathbb{E}\operatorname{tr}((W^\mathsf{T}\mathsf{BDiag}(\Theta, m)W)^{-1}) \gtrapprox \frac{n}{\mathsf{SP}(\mathsf{BDiag}(\Theta, m), n)},$$

where the notation $\gtrapprox$ indicates the heuristic nature of the inequality as explained previously. From (7.8), we want to find $\bar{x}$ such that:

$$n = \psi(\bar{x}\sqrt{n}; \mathsf{BDiag}(\Theta, m)) = \bar{x}\sqrt{n}\cdot m\sum_{j=0}^{T-1}\frac{1}{2^j + \bar{x}\sqrt{n}}.$$

While solving this equation exactly for $\bar{x}\sqrt{n}$ is not tractable, we can estimate a lower bound on $\bar{x}\sqrt{n}$ quite easily. For any integer $T_c \in \{0, \dots, T\}$, we have the following estimate:

$$\frac{n}{m} = \bar{x}\sqrt{n}\sum_{j=0}^{T-1}\frac{1}{2^j+\bar{x}\sqrt{n}} \leqslant T_c + 2\bar{x}\sqrt{n}\cdot 2^{-T_c}.$$

Let us first assume that $\bar{x}\sqrt{n} \in [1, 2^{T-1}]$, so that $\lceil \log_2(\bar{x}\sqrt{n}) \rceil \in \{0, \ldots, T\}$. Setting $T_c = \lceil \log_2(\bar{x}\sqrt{n}) \rceil$ then yields the lower bound $\bar{x}\sqrt{n} \geqslant 2^{n/m-3}$. On the other hand, if $\bar{x}\sqrt{n} > 2^{T-1}$, then since we assume $mT \geqslant c_1 n$, we also have $\bar{x}\sqrt{n} > 2^{c_1 n/m-1}$. Finally, if $\bar{x}\sqrt{n} < 1$, we have:

$$\frac{n}{m} = \bar{x}\sqrt{n} \sum_{j=0}^{T-1} \frac{1}{2^j + \bar{x}\sqrt{n}} < \sum_{j=0}^{T-1} \frac{1}{2^j + \bar{x}\sqrt{n}} \leqslant 2 \Longrightarrow m \geqslant n/2.$$

This yields a contradiction, since by assumption $m \leqslant c_2 n$, if $c_2 < 1/2$, so we must have $\bar{x}\sqrt{n} \geqslant 2^{c' n/m-3}$ with $c' = \min\{1, c_1\}$. Now by (7.8) and (7.9):

$$\mathsf{SP}(\mathsf{BDiag}(\Theta, m), n) = \frac{n}{\bar{x}\sqrt{n}} \leqslant n 2^{-c' n/m+3} \Longrightarrow \mathbb{E}\,\mathrm{tr}(\Gamma_T^{1/2}(X_{m,T}^{\mathsf{T}} X_{m,T})^{-1}\Gamma_T^{-1/2}) \gtrapprox \frac{2^{c' n/m}}{T}.$$

We make this argument rigorous in Appendix C.5.

7.2.4 Ideas behind Theorem 6.3

We focus here on the hard instance when $A = I_n$ and $m \lesssim n$, since the cases when $A = 0_{n\times n}$ or $A = I_n$ and $m \gtrsim n$ are straightforward applications of Jensen's inequality and some basic manipulations (see Lemma C.7).

The proof used by Theorem 6.3 when $A = I_n$ and $m \lesssim n$ is actually a special case of a general proof indexed by the largest Jordan block size of the hard instance. For a maximum Jordan block size r, the hard instances are $A = \mathsf{BDiag}(J_r, n/r)$, where we assume for simplicity that r divides n; this reduces to $A = I_n$ when $r = 1$. We associate two important matrices with these hard instances. To define them, let $\mathcal{I}_r := \{1, 1+r, \ldots, 1+(T-1)r\}$, and let $E_{\mathcal{I}_r} \in \mathbb{R}^{T\times Tr}$ denote the linear operator that extracts the coordinates in $\mathcal{I}_r$. The following matrices then play a key role in our analysis:

$$\Psi_{r,T,T'} := \mathsf{BDiag}(\Gamma_{T'}^{-1/2}(J_r), T)\mathsf{BToep}(J_r, T), \quad \Theta_{r,T,T'} := E_{\mathcal{I}_r}\Psi_{r,T,T'}\Psi_{r,T,T'}^{\mathsf{T}} E_{\mathcal{I}_r}^{\mathsf{T}}. \tag{7.10}$$

The next step is to use a simple decoupling argument (see Lemma C.11) to argue that, for $A = \mathsf{BDiag}(J_r, d)$:

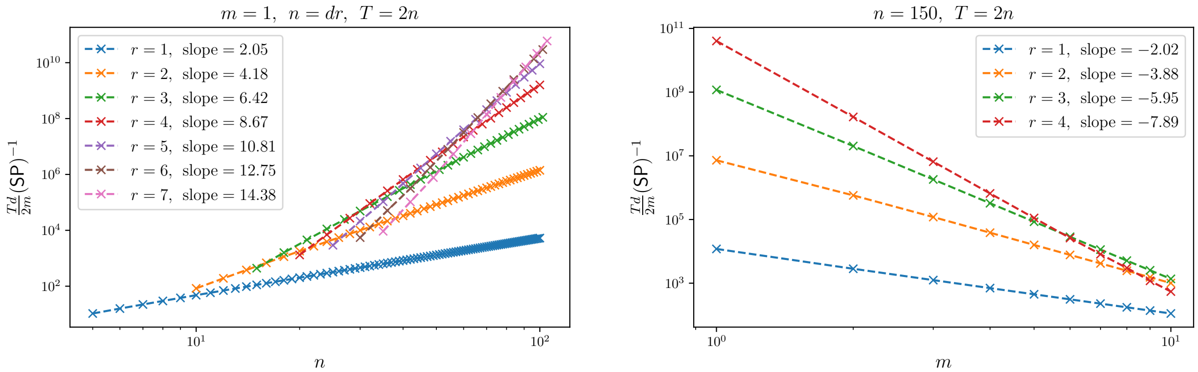
$$\mathbb{E}\,\mathrm{tr}(\Gamma_{T'}^{1/2}(X_{m,T}^{\mathsf{T}} X_{m,T})^{-1}\Gamma_{T'}^{1/2}) \geqslant \mathbb{E}\,\mathrm{tr}((W^{\mathsf{T}}\mathsf{BDiag}(\Theta_{r,T,T'}, m)W)^{-1}),$$

where $W \in \mathbb{R}^{mT\times d}$ has iid $N(0,1)$ entries. This positions us to use the arguments in Section 7.2.2 again. We first focus on the $r = 1$ case. We reduce the problem to assuming $T' = T$, by observing that since $\Gamma_t(I_n) = \frac{t+1}{2}\cdot I_n$ for any $t \in \mathbb{N}_+$, then $\Theta_{1,T,T'} = \frac{T+1}{T'+1}\cdot\Theta_{1,T,T}$. Therefore,

$$\begin{aligned}\mathbb{E}\,\mathrm{tr}((W^{\mathsf{T}}\mathsf{BDiag}(\Theta_{1,T,T'}, m)W)^{-1}) &= \frac{T'+1}{T+1}\cdot\mathbb{E}\,\mathrm{tr}((W^{\mathsf{T}}\mathsf{BDiag}(\Theta_{1,T,T}, m)W)^{-1}) \\ &\gtrapprox \frac{T'+1}{T+1}\cdot\frac{n}{\mathsf{SP}(\mathsf{BDiag}(\Theta_{1,T,T}, m), n)},\end{aligned} \tag{7.11}$$

where again the $\gtrapprox$ notation highlights the heuristic nature of the bound, used to build intuition.

To proceed, let $L_T \in \mathbb{R}^{T\times T}$ be the lower triangular matrix of all ones and define $S_T := (L_T L_T^{\mathsf{T}})^{-1}$. A computation yields that $\Theta_{1,T,T}^{-1} = \frac{T+1}{2}S_T$. Note that we can write S_T as a rank-one perturbation to a tri-diagonal matrix. Specifically, $S_T = \mathsf{Tri}(2, -1; T) - e_T e_T^{\mathsf{T}}$, where $\mathsf{Tri}(a, b; T)$



(a) Plot of $\frac{Td}{2m}(\mathsf{SP})^{-1}$ versus n.

(b) Plot of $\frac{Td}{2m}(\mathsf{SP})^{-1}$ versus m.

Figure 3: Plot of $\frac{Td}{2m}(\mathsf{SP})^{-1}$ versus n in (a) and versus m in (b), both on a log-log scale. For (a), m and p are fixed to one, d is fixed to n/r, and T is fixed to $2n$. For (b), n is fixed to 150, p is fixed to one, and T is fixed to $2n$. In the legends, the slope of the line (in log-log space) computed via linear regression is shown. Based on these plots, we conjecture that $\frac{Td}{2m}(\mathsf{SP})^{-1} \gtrsim c_r(d/m)^{2r}$, where c_r depends only on r.

denotes the symmetric $T \times T$ tri-diagonal matrix with a on the diagonal and b on the lower and upper off-diagonals. By the standard formula for the eigenvalues of a tri-diagonal matrix, we have that $\lambda_{T-k+1}(\mathsf{Tri}(2,-1;T)) = 2\left(1 - \cos\left(\frac{k\pi}{T+1}\right)\right) \asymp k^2/T^2$. In Appendix C.7, we apply the work of [KST99] to show that the rank-one perturbation is negligible: $\lambda_{T-k+1}(S_T) \asymp k^2/T^2$ as well. Therefore $\lambda_{T-k+1}(\Theta_{1,T,T}^{-1}) \asymp k^2/T$. With this bound, we have:

$$
\begin{aligned}
n = \psi(\bar{x}\sqrt{n}; \mathsf{BDiag}(\Theta_{1,T,T}, m)) &= \bar{x}\sqrt{n} \cdot m \sum_{i=1}^{T} \frac{1}{\lambda_i(\Theta_{1,T,T}^{-1}) + \bar{x}\sqrt{n}} \\
&\lesssim \bar{x}\sqrt{n} \cdot m \sum_{i=1}^{T} \frac{1}{i^2/T + \bar{x}\sqrt{n}} \leqslant \bar{x}\sqrt{n} \cdot m \int_0^T \frac{1}{x^2/T + \bar{x}\sqrt{n}}\, \mathrm{d}x \lesssim \sqrt{\bar{x}\sqrt{n}} \cdot m\sqrt{T}.
\end{aligned}
$$

This implies that $\bar{x}\sqrt{n} \gtrsim n^2/(m^2 T)$, and therefore by (7.8) and (7.9):

$$
\mathsf{SP}(\mathsf{BDiag}(\Theta_{1,T,T}, m), n) = \frac{n}{\bar{x}\sqrt{n}} \lesssim \frac{m^2 T}{n} \Longrightarrow \mathbb{E}\,\mathrm{tr}(\Gamma_{T'}^{1/2}(X_{m,T}^{\mathsf{T}} X_{m,T})^{-1}\Gamma_{T'}^{1/2}) \gtrapprox \frac{T'}{T} \cdot \frac{n^2}{m^2 T}.
$$

We make this argument rigorous in Appendix C.8.

7.2.5 Beyond diagonalizability

When $r \geqslant 2$, the analytic complexity of characterizing the solution to $n = \psi(\bar{x}\sqrt{n}; \mathsf{BDiag}(\Theta_{r,T,T'}, m))$ increases significantly. Nevertheless, we can still solve for $\bar{x}\sqrt{n}$ by numerical root finding, to look at the scaling patterns for small values of r and $T' = T$. This computation (Figure 3) leads us to conjecture a general bound of $\mathsf{R}(m, T, T; \{\mathsf{P}_x^{\mathsf{BDiag}(J_r, n/r)}\}) \gtrapprox c_r n^{2r}/(m^{2r} T)$ when $m \lesssim n$, where c_r is a constant depending only on r. A complete and precise statement is given in Appendix A.

8 Numerical simulation

We conduct a simple numerical simulation illustrating the benefits of multiple trajectories on learning. We construct a family of LDS-LS problem instances, parameterized by a scalar $\rho \in (0, \infty)$ as follows. The covariate distribution P_x is the linear dynamical system $x_{t+1} = Ax_t + w_t$ with:

$$A = U \operatorname{diag}(\underbrace{\rho, \dots, \rho}_{\lfloor n/2 \rfloor \text{ times}}, -\rho, \dots, -\rho) U^{\mathsf{T}}, \quad U \sim \mathrm{Unif}(O(n)), \quad w_t \sim N(0, I/4).$$

Here, $O(n)$ denotes the set of $n \times n$ orthonormal matrices. By construction, ρ is the spectral radius of A. The labels y_t are set as $y_t = x_{t+1}$, so that the ground truth $W_\star \in \mathbb{R}^{n \times n}$ is equal to A.

We compare the risk of the OLS estimator (3.6) on the LDS-LS problem instance, compared with its risk on the corresponding Ind-LDS-LS baseline. Specifically, we plot the ratio between OLS excess risks $\mathbb{E}[L(\cdot; T, \mathsf{P}_x)]$ on the two problem instances (P_x), respectively. We fix the covariate dimension $n = 5$ and the trajectory horizon length $T = 10n$, and vary the number of trajectories $m \in \{1, \dots, 10\}$. Figure 4 shows the result of this experiment, where we also vary $\rho \in \{0.98, 0.99, 1.0, 1.01, 1.02\}$. The error bars are plotted over 1000 trials. All computations are implemented using `jax` [BFH+18], and run with `float64` precision on a single machine.

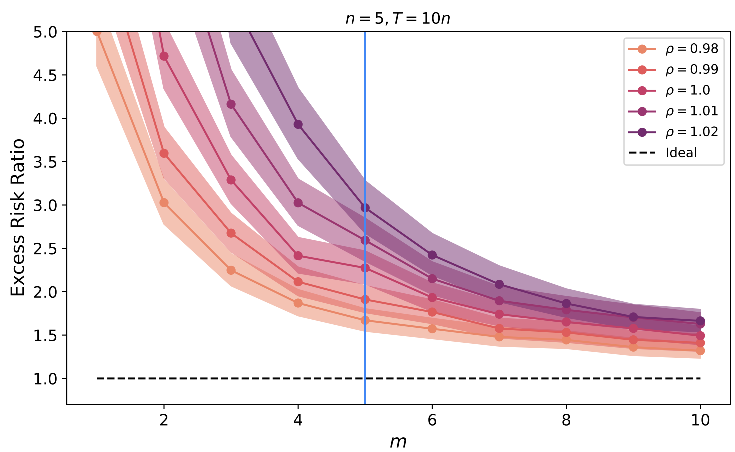


Figure 4: Plot of the ratio of excess risk for LDS-LS problem instances over its corresponding Ind-LDS-LS baseline instance, as a function of the number of trajectories m, holding both covariate dimension n and horizon length T fixed. The vertical blue line marks the transition between few trajectories ($m \leqslant n$) and many trajectories ($m \geqslant n$).

In Figure 4, we see that for the few trajectories regime ($m \leqslant n$) appearing to the left of the vertical blue line, the instability of the covariate process plays an outstanding role in determining the value of the ratio. On the other hand, for the many trajectories regime ($m \geqslant n$) appearing to the right of the blue line, the ratios quickly converge to a constant no greater than two (at $m = 10$). This behavior is consistent with Theorem 5.5. Finally, Theorem 6.3 suggests that the scaling behavior of the $\rho = 1$ curve with respect to m is on the order of $1/m$.

9 Concluding remarks

Having sharply characterized the worst-case excess risk of Seq-LS and LDS-LS, we see more precisely the trade-offs—or arguably the lack thereof—presented by resetting a system, or by simply observing parallel runs from one, where possible. After sufficient resets, one learns roughly as though examples were independent altogether (as reflected in the Ind-Seq-LS and Ind-LDS-LS baselines).

In addition to the theoretical upshot that it presents, this phenomenon seems encouraging insofar as the setup may describe reality: one does not learn to ride a bicycle by witnessing thousands of unrelated pedal strokes, nor by watching one cyclist endure the entire Tour de France, but rather by seeing and attempting many moderate rides and maneuvers.

We see a number of future directions for research, primarily in further charting out the reach of the iid-like phenomenon in learning from multiple sequences. Our work offers the trajectory small-ball criterion (Definition 4.1) as a vehicle for proving that this phenomenon occurs, or otherwise for bounding the minimax rate from above. What other notable sequential processes, outside of those covered in Section 4.1, can we capture as trajectory small-ball instances? One might look to covariate sequences generated by, say, input-to-state stable (ISS) non-linear systems, stochastic polynomial difference equations, or various Markov decision processes.

On the flip side, when must we necessarily pay a price for dependent data? One answer from our work is that a necessary gap between independent and sequentially dependent learning appears when there are insufficiently many trajectories ($m \lesssim n$). As outlined in Section 7.2.5 and Appendix A, we conjecture that this gap can be made much wider, namely by considering non-diagonalizable linear dynamical systems. That said, other pertinent problems may exhibit gaps as well. Finding them would help inform where the limits of learning from sequential data lie.

On the regression side, one might look to move beyond a well-specified linear regression model, extend to other loss functions, analyze regularized least-squares estimators in place of OLS, or consider a more adversarial analysis (e.g. measuring regret rather than risk, in an online setting).

Acknowledgements

We thank Vikas Sindhwani for organizing a lecture series on learning and control, during which this work was prompted, and for encouraging us to pursue the ensuing questions. We also thank Sumeet Singh for pointing out that the claimed first-order optimality conditions in (7.6) and (7.7) require further conditions to hold for nonconvex/nonconcave games, e.g., strictly concave stationary points of the inner maximization problem. M. Soltanolkotabi is supported by the Packard Fellowship in Science and Engineering, a Sloan Research Fellowship in Mathematics, an NSF-CAREER under award #1846369, DARPA Learning with Less Labels (LwLL) and FastNICS programs, and NSF-CIF awards #1813877 and #2008443.

References

[AYPS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. In *Conference on Learning Theory*, 2011.

[BFH+18] 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.

[BJN+20] Guy Bresler, Prateek Jain, Dheeraj Nagaraj, Praneeth Netrapalli, and Xian Wu. Least squares regression with markovian data: Fundamental limits and algorithms. In *Neural Information Processing Systems*, 2020.

[BK15] David Belanger and Sham Kakade. A linear dynamical system model for text. In *International Conference on Machine Learning*, 2015.

[BLLT20] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[BPX+07] Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

[Bre15] Guy Bresler. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, 2015.

[Cat07] Olivier Catoni. Pac-bayesian supervised classification: The thermodynamics of statistical learning. *Institute of Mathematical Statistics Lecture Notes*, 56, 2007.

[CLR+21] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Neural Information Processing Systems*, 2021.

[CP22] Yanxi Chen and H. Vincent Poor. Learning mixtures of linear dynamical systems. In *International Conference on Machine Learning*, 2022.

[CW01] Anthony Carbery and James Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical Research Letters*, 8:233–248, 2001.

[DAJJ12] John C. Duchi, Alekh Agarwal, Mikael Johansson, and Michael I. Jordan. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012.

[DDD+21] Yuval Dagan, Constantinos Daskalakis, Nishanth Dikkala, Surbhi Goel, and Anthimos Vardis Kandiros. Statistical estimation from dependent data. In *International Conference on Machine Learning*, 2021.

[DDDJ19] Yuval Dagan, Constantinos Daskalakis, Nishanth Dikkala, and Siddhartha Jayanti. Learning from weakly dependent data under dobrushin's condition. In *Conference on Learning Theory*, 2019.

[DDDK21] Yuval Dagan, Constantinos Daskalakis, Nishanth Dikkala, and Anthimos Vardis Kandiros. Learning ising models from one or multiple samples. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021.

[DDP19] Constantinos Daskalakis, Nishanth Dikkala, and Ioannis Panageas. Regression from dependent observations. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, 2019.

[DG03] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

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

[FRS20] Dylan J. Foster, Alexander Rakhlin, and Tuhin Sarkar. Learning nonlinear dynamical systems from a single trajectory. In *Learning for Dynamics & Control Conference*, 2020.

[FTM18] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.

[GLS+20] Udaya Ghai, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. No-regret prediction in marginally stable systems. In *Conference on Learning Theory*, 2020.

[GM20] Promit Ghosal and Sumit Mukherjee. Joint estimation of parameters in Ising model. *The Annals of Statistics*, 48(2):785–810, 2020.

[GR20] Rodrigo A. González and Cristian R. Rojas. A finite-sample deviation bound for stable autoregressive processes. In *Learning for Dynamics and Control*, 2020.

[GWC+20] Joshua Glaser, Matthew Whiteway, John P Cunningham, Liam Paninski, and Scott Linderman. Recurrent switching dynamical systems models for multiple interacting neural populations. In *Neural Information Processing Systems*, 2020.

[GZ01] Alexander Goldenshluger and Assaf Zeevi. Nonasymptotic bounds for autoregressive time series modeling. *The Annals of Statistics*, 29(2):417–444, 2001.

[HKZ14] Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14:569–600, 2014.

[HLS+18] Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems. In *Neural Information Processing Systems*, 2018.

[HSZ17] Elad Hazan, Karan Singh, and Cyril Zhang. Learning linear dynamical systems via spectral filtering. In *Neural Information Processing Systems*, 2017.

[JKNN21] Prateek Jain, Suhas S. Kowshik, Dheeraj Nagaraj, and Praneeth Netrapalli. Near-optimal offline and streaming algorithms for learning non-linear dynamical systems. In *Neural Information Processing Systems*, 2021.

[JLL21] Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. In *Neural Information Processing Systems*, 2021.

[JNJ20] Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, 2020.

[JP19] Yassir Jedra and Alexandre Proutiere. Sample complexity lower bounds for linear system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019.

[JP20] 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)*, 2020.

[JVS+16] Rafal Józefowicz, Oriol Vinyals, Mike Schuster, Noam M. Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

[KM15] Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008, 2015.

[KM17] Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106:93–117, 2017.

[KST99] Devadatta Kulkarni, Darrell Schmidt, and Sze-Kai Tsui. Eigenvalues of tridiagonal pseudo-toeplitz matrices. *Linear Algebra and its Applications*, 297(1–3):63–80, 1999.

[KZB11] S. Mohammad Khansari-Zadeh and Aude Billard. Learning stable nonlinear dynamical systems with gaussian mixture models. *IEEE Transactions on Robotics*, 27(5):943–957, 2011.

[LBL+18] Nevena Lazic, Craig Boutilier, Tyler Lu, Eehern Wong, Binz Roy, MK Ryu, and Greg Imwalle. Data center cooling using model-predictive control. In *Neural Information Processing Systems*, 2018.

[LJM+17] Scott Linderman, Matthew Johnson, Andrew Miller, Ryan Adams, David Blei, and Liam Paninski. Bayesian learning and inference in recurrent switching linear dynamical systems. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

[Lju98] Lennart Ljung. *System Identification: Theory for the User.* Pearson, 1998.

[LM00] Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

[LW82] Tze Leung Lai and Ching Zong Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.

[LW83] Tze Leung Lai and Ching Zong Wei. Asymptotic properties of general autoregressive models and strong consistency of least-squares estimates of their parameters. *Journal of Multivariate Analysis*, 13(1):1–23, 1983.

[Mag78] Jan R. Magnus. The moments of products of quadratic forms in normal variables. *Statistica Neerlandica*, 32(4):201–210, 1978.

[Mei00] Ron Meir. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 39:5–34, 2000.

[Men15] Shahar Mendelson. Learning without concentration. *Journal of the ACM*, 62(3):1–25, 2015.

[MFTM22] Aditya Modi, Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Joint learning of linear time-invariant dynamical systems. *arXiv preprint arXiv:2112.10955*, 2022.

[Mou22] Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics*, 50(4):2157–2178, 2022.

[MR08] Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In *Neural Information Processing Systems*, 2008.

[MR10] Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary ϕ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11(26):789–814, 2010.

[MS00] V. John Mathews and Giovanni L. Sicuranza. *Polynomial Signal Processing*. Wiley, 2000.

[MSS17] Daniel J. McDonald, Cosma Rohilla Shalizi, and Mark Schervish. Nonparametric risk bounds for time-series forecasting. *Journal of Machine Learning Research*, 18:1–40, 2017.

[MT93] Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.

[NTP11] Duy Nguyen-Tuong and Jan Peters. Model learning for robot control: a survey. *Cognitive Processing*, 12:319–340, 2011.

[Oli16] Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3):1175–1194, 2016.

[OO19] Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. In *2019 American Control Conference (ACC)*, 2019.

[OPN+18] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J. Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 7(1-2):1–179, 2018.

[PM13] Peter C.B. Phillips and Tassos Magdalinos. Inconsistent var regression with common explosive roots. *Econometric Theory*, 29:808–837, 2013.

[Pom89] Dean A. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *Neural Information Processing Systems*, 1989.

[Rug81] Wilson J. Rugh. *Nonlinear System Theory: The Volterra / Wiener Approach*. The Johns Hopkins University Press, 1981.

[Rug96] Wilson J. Rugh. *Linear system theory*. Prentice Hall, 2nd ed. edition, 1996.

[SBR19] Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. In *Conference on Learning Theory*, 2019.

[SC09] Ingo Steinwart and Andreas Christmann. Fast learning from non-i.i.d. observations. In *Neural Information Processing Systems*, 2009.

[Sha21] Cosma Rohilla Shalizi. *Advanced Data Analysis from an Elementary Point of View*. 2021. `https://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ADAfaEPoV.pdf`.

[SMT+18] 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*, 2018.

[SO20] Yahya Sattar and Samet Oymak. Non-asymptotic and accurate learning of nonlinear dynamical systems. *arXiv preprint arXiv:2002.08538*, 2020.

[SR19] Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, 2019.

[SRD21] Tuhin Sarkar, Alexander Rakhlin, and Munther A. Dahleh. Finite time lti system identification. *Journal of Machine Learning Research*, 22:1–61, 2021.

[SVL14] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Neural Information Processing Systems*, 2014.

[TB23] Stephen Tu and Ross Boczar. An elementary proof of anti-concentration for degree two non-negative gaussian polynomials. *arXiv preprint arXiv:2301.05992*, 2023.

[TOH14] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. A tight version of the gaussian min-max theorem in the presence of convexity. Technical report, Caltech, 2014.

[TP19] Anastasios Tsiamis and George J. Pappas. Finite sample analysis of stochastic system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019.

[TP21] Anastasios Tsiamis and George J. Pappas. Linear systems can be hard to learn. *arXiv preprint arXiv:2104.01120*, 2021.

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[XCS22] Lei Xin, George Chiu, and Shreyas Sundaram. Learning the dynamics of autonomous linear systems from multiple trajectories. *arXiv preprint arXiv:2203.12794*, 2022.

[XGH+21] Yu Xing, Benjamin Gravell, Xingkang He, Karl Henrik Johansson, and Tyler Summers. Identification of linear systems with multiplicative noise from multiple trajectory data. *arXiv preprint arXiv:2106.16078*, 2021.

[Yu94] Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994.

[ZL21] Yang Zheng and Na Li. Non-asymptotic identification of linear dynamical systems using multiple trajectories. *IEEE Control Systems Letters*, 5(5):1693–1698, 2021.

[ZSM22] Ingvar Ziemann, Henrik Sandberg, and Nikolai Matni. Single trajectory nonparametric learning of nonlinear dynamics. In *Conference on Learning Theory*, 2022.

[ZT22] Ingvar Ziemann and Stephen Tu. Learning with little mixing. In *Neural Information Processing Systems*, 2022.

A Beyond diagonalizability: a conjecture for the general case

Recall that various results in Section 5.1 required a diagaonalizability assumption (Assumption 5.2) on the dynamics matrix A, specifically in the many trajectories regime when $T' > T$ (Theorem 5.5), or in the few trajectories regime (Section 5.1.2). In this section, we conjecture how removing the diagonalizability assumption would affect the results. For simplicity, we focus on the few trajectories regime, and further assume that $T' = T$. Building on potential extensions of this paper's analysis, and numerical evidence detailed in Section 7, we conjecture the following extensions of Theorem 5.6 and Theorem 6.3:

Conjecture A.1 (Risk for $\mathsf{LDS\text{-}LS}$ with few trajectories under non-diagonalizable systems)**.** *There are universal positive constants c_0, c_1, c_2, c_3, and a universal mapping $\varphi : \mathbb{N}_+ \to \mathbb{R}_{>0}$ such that the following holds for any instance of $\mathsf{LDS\text{-}LS}$ satisfying Assumption 5.1 (marginal stability) and Assumption 5.3 (one-step controllability). Let $A = SJS^{-1}$ denote the Jordan normal form of the dynamics matrix A. Define $\gamma := \frac{\lambda_{\max}(S^{-1}BB^\mathsf{T}S^{-*})}{\lambda_{\min}(S^{-1}BB^\mathsf{T}S^{-*})}$, and let r be the size of the largest Jordan block in J. If $n \geqslant c_0$, $m \leqslant c_1 n$, and $mT \geqslant c_2 n$, then:*

$$\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x^{A,B})] \leqslant c_3 \sigma_\xi^2 \varphi(r) \gamma \cdot \frac{pn^{2r}}{m^{2r}T}. \tag{A.1}$$

Additionally, there exist universal positive constants c_0', c_1', c_2', c_3', and c_4' such that the following is true. Suppose $\mathcal{A} \subseteq \mathbb{R}^{n\times n}$ is any set containing all $n \times n$ matrices with Jordan blocks of size at most r. Let $T \geqslant c_0'$, $n \geqslant c_1'$, $mT \geqslant c_2' n$, and $m \leqslant c_3' n$. Then:

$$\mathsf{R}(m, T, T; \{\mathsf{P}_x^A \mid A \in \mathcal{A}\}) \geqslant c_4' \sigma_\xi^2 \varphi(r) \gamma \cdot \frac{pn^{2r}}{m^{2r}T}. \tag{A.2}$$

Lemma 5.1 provides a viable path towards proving the upper bound (A.1) from Conjecture A.1 up to logarithmic factors in the regime of constant Jordan block size r, by reducing the problem to understanding the scaling of $\underline{\lambda}(k, t; A, B) = \underline{\lambda}(\Gamma_k(A, B), \Gamma_t(A, B))$ when $k \leqslant t$. Our analysis uses diagonalizability (Assumption 5.2) of the dynamics matrices to show that $\underline{\lambda}(k, t; A, B) \gtrsim \gamma^{-1} \cdot k/t$.

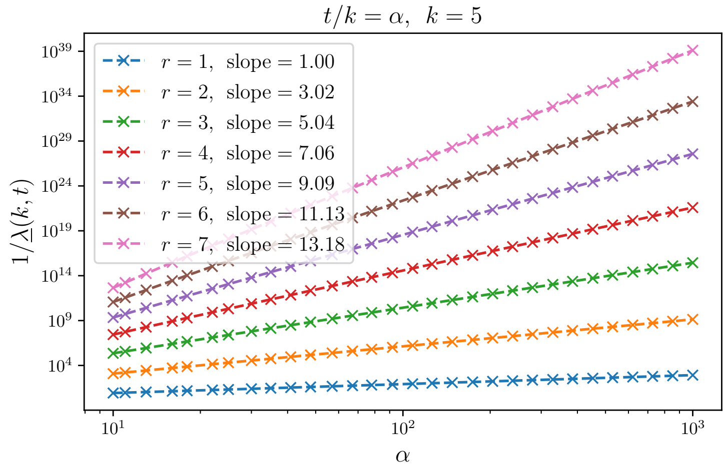


Figure 5: A plot of the ratio α versus $1/\underline{\lambda}(k,t)$ with k fixed to 5 and t fixed to $k\alpha$. Here, $\underline{\lambda}(k,t) := \underline{\lambda}(k,t;J_r,I_r)$. The slope of the line (in log-log space) computed via linear regression is reported. We conjecture that in general, $\underline{\lambda}(k,t;A,B) \gtrsim c_r \gamma^{-1} \cdot (k/t)^{2r-1}$.

Without such an assumption, analyzing $\underline{\lambda}(k,t;A,B)$ is substantially more involved. A numerical simulation (Figure 5) suggests that $\underline{\lambda}(k,t;A,B) \gtrsim c_r \gamma^{-1} \cdot (k/t)^{2r-1}$ is the general rate for dynamics matrices A with Jordan blocks at most size r, where c_r is a constant depending only on r. Assuming this scaling to be correct and plugging the rate into Lemma 5.1 yields (A.1) up to logarithmic factors. Partial progress towards analyzing $\underline{\lambda}(k,t;A,B)$ was made in [SR19, Proposition 7.6], where it is shown that $\underline{\lambda}(k,t;A,B) \gtrsim c_r \gamma^{-1} \cdot (k/t)^{r^2}$, with $1/c_r$ depending exponentially on r. We do not conjecture a form for the mapping $\varphi(r)$; $\underline{\lambda}(k,t;A,B)$ becomes numerically ill-conditioned when r is large, hindering simulation with large blocks.

On the other hand, the analytic arguments in Section 7.2.4 combined with the numerical evidence in Figure 3 suggest that the bound (A.2) holds (up to the condition number factor γ). The one caveat is that, even if we were to analytically characterize the eigenvalues of $\Theta_{r,T,T}$ for all r, our proof strategy would most likely not be able to give a sharp characterization of the leading constant $\varphi(r)$ in the lower bound. This is because our proof inherently exploits the independence between decoupled subsystems, and does not tackle the harder challenge of understanding the coupling effects within a Jordan block.

We conclude this section by noting that Conjecture A.1 does not include any logarithmic factors in the upper bound rate (A.1), and includes the condition number factor γ in the lower bound (A.2). In other words, Conjecture A.1 applied to the special case of $r = 1$ conjectures that Theorem 5.6 is loose by $\log^2(\gamma n/m)$, and that Theorem 6.3 is loose by a factor of γ.

B Analysis for upper bounds

B.1 Preliminaries

We collect various technical results which we will use in the proof of the upper bounds. The first result gives us a bound on the functional inverse of $T \mapsto T/\log T$.

Proposition B.1 ([SMT+18, Lemma A.4])**.** *For $\alpha \geqslant 1$, $T \geqslant 2\alpha \log(4\alpha)$ implies that $T \geqslant \alpha \log T$.*

The next two results study various properties of functions involving $\underline{\lambda}$.

Proposition B.2. *For $A \in \mathsf{Sym}^n_{>0}$, the map $X \mapsto \underline{\lambda}(X, A)$ is concave over symmetric matrices.*

Proof. Observe that $\underline{\lambda}(X, A) = \lambda_{\min}(A^{-1/2} X A^{-1/2}) = \inf\{\langle X, A^{-1/2} v v^\mathsf{T} A^{-1/2}\rangle \mid v \in \mathbb{S}^{n-1}\}$ is the pointwise infimum over a set of linear functions in X, and is therefore concave. □

Proposition B.3. *Fix $T \in \mathbb{N}_+$, $\{\Psi_t\}_{t=1}^T \subset \mathsf{Sym}^n_{>0}$, and $\Gamma \in \mathsf{Sym}^n_{>0}$. Suppose that $\frac{1}{T}\sum_{t=1}^T \Psi_t \preccurlyeq \Gamma$. Then $\left[\prod_{t=1}^T \underline{\lambda}(\Psi_t, \Gamma)\right]^{1/T} \leqslant 1$.*

Proof. We have that:

$$
\begin{aligned}
\left[\prod_{t=1}^T \underline{\lambda}(\Psi_t, \Gamma)\right]^{1/T} &\leqslant \frac{1}{T}\sum_{t=1}^T \underline{\lambda}(\Psi_t, \Gamma) && \text{using the AM-GM inequality} \\
&\leqslant \underline{\lambda}\left(\frac{1}{T}\sum_{t=1}^T \Psi_t, \Gamma\right) && \text{using Proposition B.2 and Jensen's inequality} \\
&\leqslant \underline{\lambda}(\Gamma, \Gamma) && \text{since } \frac{1}{T}\sum_{t=1}^T \Psi_t \preccurlyeq \Gamma \\
&= 1.
\end{aligned}
$$

□

The next result relates the anti-concentration properties of a non-negative random variable to its moment generating function on $(-\infty, 0)$.

Proposition B.4 ([Mou22, Lemma 7])**.** *Let X be a non-negative random variable. Suppose there exists an $\alpha \in (0, 1]$ and positive constant c such that:*

$$\mathbb{P}(X \leqslant t) \leqslant (ct)^\alpha \quad \forall t > 0.$$

Then:

$$\mathbb{E}[\exp(-\eta X)] \leqslant (c/\eta)^\alpha \quad \forall \eta > 0.$$

The next few results involve various properties of Gaussian and spherical distributions.

Proposition B.5 ([Mag78, Lemma 6.2])**.** *For $w \sim N(0, I)$ and symmetric matrices A, B:*

$$\mathbb{E}[w^\mathsf{T} A w w^\mathsf{T} B w] = 2\langle A, B\rangle + \operatorname{tr}(A)\operatorname{tr}(B).$$

Proposition B.6 ([DG03, Lemma 2.2])**.** *Let $n \geqslant 2$ and $v \in \mathbb{R}^n \setminus \{0\}$ be fixed. Suppose that ψ is drawn uniformly at random from the uniform measure over $\mathbb{S}^{n-1}$. We have that for all $\varepsilon > 0$:*

$$\mathbb{P}\left\{\langle v, \psi\rangle^2 \leqslant \frac{\varepsilon}{n}\|v\|_2^2\right\} \leqslant (e\varepsilon)^{1/2}.$$

Next, we state a classic result which gives us anti-concentration of arbitrary Gaussian (more generally any log-concave distribution) polynomials of bounded degree.

Theorem B.7 ([CW01, Theorem 8]). *Fix an integer $d \in \mathbb{N}_+$. There exists a universal positive constant c such that the following is true. Let $p : \mathbb{R}^n \to \mathbb{R}$ be a degree d polynomial, and let $\varepsilon > 0$. We have:*

$$\mathbb{P}\{|p(x)| \leqslant \varepsilon \cdot \mathbb{E}|p(x)|\} \leqslant c \cdot d\varepsilon^{1/d}, \quad x \sim N(0, I_n).$$

For the case when $d = 2$ and p is non-negative, we can take $c = \sqrt{e/2}$.

Theorem B.7 can be further specialized as follows. Suppose $w \sim N(0, I)$, x is fixed, and $\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12}^\mathsf{T} & Q_{22} \end{bmatrix}$ is positive semidefinite. Then:

$$\mathbb{P}\left\{ \begin{bmatrix} x \\ w \end{bmatrix}^\mathsf{T} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12}^\mathsf{T} & Q_{22} \end{bmatrix} \begin{bmatrix} x \\ w \end{bmatrix} \leqslant \varepsilon \cdot \operatorname{tr}(Q_{22}) \right\} \leqslant (e\varepsilon)^{1/2} \quad \forall \varepsilon > 0. \tag{B.1}$$

Both (B.1) and the explicit constant in Theorem B.7 for $d = 2$ and p non-negative can be derived by bounding the MGF of various Gaussian quadratic forms; see e.g. [TB23].

Next, we state a well-known result from [AYPS11], which yields an anytime bound for the size of a self-normalized martingale difference sequence (MDS).

Lemma B.8 ([AYPS11, Theorem 3]). *Fix a $\delta \in (0, 1)$ and positive definite matrix $V \in \mathbb{R}^{d \times d}$. Let $\{x_t\}_{t \geqslant 1} \subset \mathbb{R}^d$ be a stochastic process adapted to the filtration $\{\mathcal{F}_t\}_{t \geqslant 1}$. Let $\{\eta_t\}_{t \geqslant 1} \subset \mathbb{R}$ be a martingale difference sequence adapted to $\{\mathcal{F}_t\}_{t \geqslant 2}$. Suppose there exists $R > 0$ such that $\mathbb{E}[\exp(\lambda \eta_t) \mid \mathcal{F}_t] \leqslant \exp(\lambda^2 R^2/2)$ a.s. for all $\lambda \in \mathbb{R}$ and $t \geqslant 1$. Define $V_t := \sum_{k=1}^t x_k x_k^\mathsf{T}$ for $t \geqslant 1$. With probability at least $1 - \delta$,*

$$\left\| \sum_{k=1}^t \eta_k x_k \right\|_{(V_t + V)^{-1}} \leqslant \sqrt{2R^2 \log\left(\frac{\det(V_t + V)^{1/2} \det(V)^{-1/2}}{\delta} \right)} \quad \forall t \geqslant 1.$$

Lemma B.8 is generalized to vector-valued self-normalized MDS via a covering argument:

Proposition B.9 ([SR19, Proposition 8.2]). *Fix a $\delta \in (0, 1)$ and positive definite matrix $V \in \mathbb{R}^{d \times d}$. Let $\{x_t\}_{t \geqslant 1} \subset \mathbb{R}^d$ be a stochastic process adapted to the filtration $\{\mathcal{F}_t\}_{t \geqslant 1}$. Let $\{\eta_t\}_{t \geqslant 1} \subset \mathbb{R}^p$ be a stochastic process adapted to $\{\mathcal{F}_t\}_{t \geqslant 2}$. Suppose that for every fixed $v \in \mathbb{S}^{p-1}$, for every $t \geqslant 1$:*

(a) $\mathbb{E}[\langle v, \eta_t \rangle \mid \mathcal{F}_t] = 0$ *a.s.*

(b) $\mathbb{E}[\exp(\lambda \langle v, \eta_t \rangle) \mid \mathcal{F}_t] \leqslant \exp(\lambda^2 R^2/2)$ *a.s. for every $\lambda \in \mathbb{R}$.*

Define $V_t := \sum_{k=1}^t x_k x_k^\mathsf{T}$ for $t \geqslant 1$. With probability at least $1 - \delta$, for all $t \geqslant 1$:

$$\left\| \sum_{k=1}^t \eta_k x_k^\mathsf{T} (V_t + V)^{-1/2} \right\|_{\mathrm{op}} \leqslant 2\sqrt{2R^2 \log\left(\frac{5^p \det(V_t + V)^{1/2} \det(V)^{-1/2}}{\delta} \right)}.$$

The next result assumes V_t is invertible in order to simplify Proposition B.9.

Proposition B.10. *Under the same hypothesis of Proposition B.9, we have with probability at least $1-\delta$, for all $t \geqslant 1$:*

$$\mathbf{1}\{V_t \succcurlyeq V\} \left\| \sum_{k=1}^{t} \eta_k x_k^{\mathsf{T}} V_t^{-1/2} \right\|_{\mathrm{op}} \leqslant 4 \cdot \mathbf{1}\{V_t \succcurlyeq V\} \sqrt{R^2 \log \left(\frac{5^p \det(V_t + V)^{1/2} \det(V)^{-1/2}}{\delta} \right)}.$$

Proof. Observe that when $V_t \succcurlyeq V$, we have:

$$2V_t \succcurlyeq V_t + V \Longrightarrow V_t^{-1} \preccurlyeq 2(V_t + V)^{-1}.$$

For two positive definite matrices M_1 and M_2 satisfying $M_1 \preccurlyeq M_2$, and any matrix N,

$$\|NM_1^{1/2}\|_{\mathrm{op}} = \sqrt{\lambda_{\max}(NM_1N^{\mathsf{T}})} \leqslant \sqrt{\lambda_{\max}(NM_2N^{\mathsf{T}})} = \|NM_2^{1/2}\|_{\mathrm{op}}.$$

Therefore,

$$\begin{aligned} \mathbf{1}\{V_t \succcurlyeq V\} \left\| \sum_{k=1}^{t} \eta_k x_k^{\mathsf{T}} V_t^{-1/2} \right\|_{\mathrm{op}} &\leqslant 2 \cdot \mathbf{1}\{V_t \succcurlyeq V\} \left\| \sum_{k=1}^{t} \eta_k x_k^{\mathsf{T}} (V_t + V)^{-1/2} \right\|_{\mathrm{op}} \\ &\leqslant 4 \cdot \mathbf{1}\{V_t \succcurlyeq V\} \sqrt{R^2 \log \left(\frac{5^p \det(V_t + V)^{1/2} \det(V)^{-1/2}}{\delta} \right)}, \end{aligned}$$

where the last inequality holds for every t with probability at least $1-\delta$ by Proposition B.9. □

B.2 Examples of trajectory small-ball

In this section, we prove that the examples listed in Section 4.1 satisfying the trajectory small-ball condition (Definition 4.1).

Example 4.1 (Copies of a Gaussian draw)**.** *Let $\Sigma \in \mathsf{Sym}^n_{>0}$, and let P_x denote the process $x_1 \sim N(0, \Sigma)$ and $x_t = x_{t-1}$ for $t \geqslant 2$. Fix any $T \in \mathbb{N}_+$. Then P_x satisfies the $(T, T, \Sigma, e, \frac{1}{2})$-TrajSB condition.*

Proof. When $k = T$ and $\underline{\Gamma} = I_n$, the condition (4.1) simplifies to:

$$\sup_{v \in \mathbb{S}^{n-1}} \mathbb{P} \left\{ \frac{1}{T} \sum_{t=1}^{T} \langle v, x_t \rangle^2 \leqslant \varepsilon \right\} \leqslant (c_{\mathsf{sb}} \varepsilon)^{\alpha} \quad \forall \varepsilon > 0.$$

Since $x_1 = x_2 = \cdots = x_T$, this further simplifies to:

$$\sup_{v \in \mathbb{S}^{n-1}} \mathbb{P} \left\{ \langle v, x_1 \rangle^2 \leqslant \varepsilon \right\} \leqslant (c_{\mathsf{sb}} \varepsilon)^{\alpha} \quad \forall \varepsilon > 0.$$

Since $\langle v, x_1 \rangle \sim N(0, 1)$, Equation (B.1) yields that $\mathbb{P}_{X \sim N(0,1)}\{X^2 \leqslant \varepsilon\} \leqslant (e\varepsilon)^{1/2}$, so we can take $c_{\mathsf{sb}} = e$ and $\alpha = 1/2$. □

Example 4.2 (Gaussian processes)**.** *Let P_x be a Gaussian process indexed by time, i.e., for every finite index set $I \subset \mathbb{N}_+$, the collection of random variables $(x_t)_{t \in I}$ is jointly Gaussian. Let $T_{\mathsf{nd}} := \inf\{t \in \mathbb{N}_+ \mid \det(\mathbb{E}[x_t x_t^{\mathsf{T}}]) \neq 0\}$, and suppose T_{nd} is finite. Fix a $T \in \mathbb{N}_+$ satisfying $T \geqslant T_{\mathsf{nd}}$. Then P_x satisfies the $(T, T, \Gamma_T(\mathsf{P}_x), 2e, \frac{1}{2})$-TrajSB condition.*

Proof. Since the covariates $(x_1, \dots, x_T)$ are jointly Gaussian, we can write,

$$\begin{bmatrix} x_1 \\ \vdots \\ x_T \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_T \end{bmatrix} + \begin{bmatrix} M_1 \\ \vdots \\ M_T \end{bmatrix} w,$$

where $\mu_1, \dots, \mu_T \in \mathbb{R}^n$ and $M_1, \dots, M_T \in \mathbb{R}^{n \times nT}$ are fixed, and $w \sim N(0, I_{nT})$. For any $v \in \mathbb{R}^n$,

$$\frac{1}{T}\sum_{t=1}^{T} \langle v, x_t \rangle^2 = \frac{1}{T}\sum_{t=1}^{T} \langle v, \mu_t + M_t w \rangle^2.$$

This is a degree 2 non-negative polynomial in w, and therefore by Theorem B.7, for all $\varepsilon > 0$:

$$\mathbb{P}\left\{ \frac{1}{T}\sum_{t=1}^{T} \langle v, x_t \rangle^2 \leqslant \varepsilon \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \langle v, x_t \rangle^2 \right] \right\} \leqslant (2e\varepsilon)^{1/2}.$$

□

Example 4.4 (Alternating halfspaces)**.** *Suppose that $n \geqslant 4$ is even, and let $u_1, \dots, u_n$ be a fixed orthonormal basis of $\mathbb{R}^n$. Put $U_0 = \mathrm{span}(u_1, \dots, u_{n/2})$ and $U_1 = \mathrm{span}(u_{n/2+1}, \dots, u_n)$. Let $i_1 \sim \mathrm{Bern}(\frac{1}{2})$, $i_{t+1} = i_t \mod 2$ for $t \in \mathbb{N}_+$, and let P_x denote the process with conditional distribution $x_t \mid i_t$ uniform over the spherical measure on $U_{i_t} \cap \mathbb{S}^{n-1}$. For any $T \geqslant 2$, the process P_x satisfies the $(T, 2, I_n/(2n), e, \frac{1}{2})$-TrajSB condition.*

Proof. For $i \in \{0, 1\}$, let ψ_i be uniform on the uniform measure over $U_i \cap \mathbb{S}^{n-1}$, let P_{U_i} denote the orthogonal projector onto U_i, and let $v_i = P_{U_i} v$.

Fix any $v \in \mathbb{R}^n \setminus \{0\}$. We observe that for any $t \in \mathbb{N}_+$, $\langle v, x_{t+1} \rangle^2 + \langle v, x_{t+2} \rangle^2 \mid i_t$ is equal in distribution to $\langle v, \psi_0 \rangle^2 + \langle v, \psi_1 \rangle^2$, which itself is equal in distribution to $\langle v_0, \psi_0 \rangle^2 + \langle v_1, \psi_1 \rangle^2$. Suppose first that $\|v_0\|_2 \geqslant \|v_1\|_2$. Then, since $\|v\|_2^2 = \|v_0\|_2^2 + \|v_1\|_2^2 \leqslant 2\|v_0\|_2^2$:

$$\begin{aligned} \left\{ \langle v_0, \psi_0 \rangle^2 + \langle v_1, \psi_1 \rangle^2 \leqslant \frac{\varepsilon}{n}\|v\|_2^2 \right\} &\subseteq \left\{ \langle v_0, \psi_0 \rangle^2 + \langle v_1, \psi_1 \rangle^2 \leqslant \frac{2\varepsilon}{n}\|v_0\|_2^2 \right\} \\ &\subseteq \left\{ \langle v_0, \psi_0 \rangle^2 \leqslant \frac{2\varepsilon}{n}\|v_0\|_2^2 \right\}. \end{aligned}$$

Writing $\alpha_0 = (\langle u_1, v \rangle, \dots, \langle u_{n/2}, v \rangle) \in \mathbb{R}^{n/2}$, by a change of coordinates we have that $\|\alpha_0\|_2^2 = \|v_0\|_2^2$, and that $\langle v_0, \psi_0 \rangle$ is equal in distribution to $\langle \alpha_0, \zeta_0 \rangle$, where ζ_0 is uniform on $\mathbb{S}^{n/2-1}$. Since we assumed $\|v_0\|_2 \geqslant \|v_1\|_2$, we must have that $\alpha_0 \neq 0$. Hence by Proposition B.6,

$$\mathbb{P}\left\{ \langle v_0, \psi_0 \rangle^2 + \langle v_1, \psi_1 \rangle^2 \leqslant \frac{\varepsilon}{n}\|v\|_2^2 \right\} \leqslant \mathbb{P}\left\{ \langle \alpha_0, \zeta_0 \rangle^2 \leqslant \frac{2\varepsilon}{n}\|\alpha_0\|_2^2 \right\} \leqslant (e\varepsilon)^{1/2}.$$

Note that if $\|v_1\|_2 > \|v_0\|_2$, an identical argument yields the same bound. Hence, letting $\mathcal{F}_t = \sigma(x_1, \dots, x_t)$, we have shown that for all $t \geqslant 0$:

$$\mathbb{P}\left\{ \frac{1}{2}\sum_{\ell=1}^{2} \langle v, x_{t+\ell} \rangle^2 \leqslant \varepsilon \cdot v^{\mathsf{T}} \left(\frac{1}{2n} I_n \right) v \,\middle|\, \mathcal{F}_t \right\} \leqslant (e\varepsilon)^{1/2},$$

from which the claim follows. □

Example 4.5 (Normal subspaces). *Suppose that* $n \geqslant 3$. *Let* $u_1, \ldots, u_n$ *be a fixed orthonormal basis in* $\mathbb{R}^n$, *and let* $U_{\neg i} := \mathrm{span}(\{u_j\}_{j \neq i})$ *for* $i \in \{1, \ldots, n\}$. *Consider the Markov chain* $\{i_t\}_{t \geqslant 1}$ *defined by* $i_1 \sim \mathrm{Unif}(\{1, \ldots, n\})$, *and* $i_{t+1} \mid i_t \sim \mathrm{Unif}(\{1, \ldots, n\} \setminus \{i_t\})$. *Let* P_x *denote the process with conditional distribution* $x_t \mid i_t$ *uniform over the spherical measure on* $U_{\neg i_t} \cap \mathbb{S}^{n-1}$. *For any* $T \geqslant 2$, *the process* P_x *satisfies the* $(T, 2, I_n/(4n-4), e, \frac{1}{2})$*-TrajSB condition.*

Proof. Fix any $v \in \mathbb{R}^n \setminus \{0\}$, and for $i \in \{1, \ldots, n\}$, let $v_i = P_{U_{\neg i}} v$, where $P_{U_{\neg i}}$ is the orthogonal projector onto $U_{\neg i}$ Let $\{\psi_i\}_{i=1}^n$ be independent random variables, where each ψ_i is uniform on the uniform measure over $U_{\neg i} \cap \mathbb{S}^{n-1}$.

Let indices $j, k \in \{1, \ldots, n\}$ with $j \neq k$. We first observe that since $j \neq k$, we have that $U_{\neg j}^{\perp} = \mathrm{span}(u_j) \subset U_{\neg k}$. Therefore:

$$\|v\|_2^2 = \|v_j\|_2^2 + \|P_{U_{\neg j}}^{\perp} v_j\|_2^2 \leqslant \|v_j\|_2^2 + \|v_k\|_2^2.$$

Hence, assuming that $\|v_j\|_2 \geqslant \|v_k\|_2$, we have:

$$\begin{aligned}\left\{\langle v_j, \psi_j\rangle^2 + \langle v_k, \psi_k\rangle^2 \leqslant \frac{\varepsilon}{2(n-1)}\|v\|_2^2\right\} &\subseteq \left\{\langle v_j, \psi_j\rangle^2 + \langle v_k, \psi_k\rangle^2 \leqslant \frac{\varepsilon}{n-1}\|v_j\|_2^2\right\} \\ &\subseteq \left\{\langle v_j, \psi_j\rangle^2 \leqslant \frac{\varepsilon}{n-1}\|v_j\|_2^2\right\}.\end{aligned}$$

Writing $\alpha_j = (\langle u_i, v\rangle)_{i \neq j} \in \mathbb{R}^{n-1}$, by a change of coordinates we have that $\|\alpha_j\|_2^2 = \|v_j\|_2^2$, and that $\langle v_j, \psi_j\rangle$ is equal in distribution to $\langle \alpha_j, \zeta_j\rangle$, where ζ_j is uniform on $\mathbb{S}^{n-2}$. Since we assumed $\|v_j\|_2 \geqslant \|v_k\|_2$, we must have that $\alpha_j \neq 0$. Hence by Proposition B.6,

$$\mathbb{P}\left\{\langle v_j, \psi_j\rangle^2 + \langle v_k, \psi_k\rangle^2 \leqslant \frac{\varepsilon}{2(n-1)}\|v\|_2^2\right\} \leqslant \mathbb{P}\left\{\langle \alpha_j, \zeta_j\rangle^2 \leqslant \frac{\varepsilon}{n-1}\|\alpha_j\|_2^2\right\} \leqslant (e\varepsilon)^{1/2}.$$

On the other hand if $\|v_k\|_2 > \|v_j\|_2$, an identical argument yields the same bound.

Now, for any $i \in \{1, \ldots, n\}$ and $t \in \mathbb{N}_+$:

$$\begin{aligned}&\mathbb{P}\left\{\langle v, x_{t+1}\rangle^2 + \langle v, x_{t+2}\rangle^2 \leqslant \frac{\varepsilon}{2(n-1)}\|v\|_2^2 \,\middle|\, i_t = i\right\} \\ &= \sum_{j \neq i, k \neq j} \mathbb{P}\left\{\langle v, x_{t+1}\rangle^2 + \langle v, x_{t+2}\rangle^2 \leqslant \frac{\varepsilon}{2(n-1)}\|v\|_2^2 \,\middle|\, i_t = i, i_{t+1} = j, i_{t+2} = k\right\} \mathbb{P}\{i_{t+1} = j, i_{t+2} = k \mid i_t = i\} \\ &= \sum_{j \neq i, k \neq j} \mathbb{P}\left\{\langle v_j, \psi_j\rangle^2 + \langle v_k, \psi_k\rangle^2 \leqslant \frac{\varepsilon}{2(n-1)}\|v\|_2^2\right\} \mathbb{P}\{i_{t+1} = j, i_{t+2} = k \mid i_t = i\} \\ &\leqslant (e\varepsilon)^{1/2} \sum_{j \neq i, k \neq j} \mathbb{P}\{i_{t+1} = j, i_{t+2} = k \mid i_t = i\} \\ &= (e\varepsilon)^{1/2}.\end{aligned}$$

Note we also have that $\mathbb{P}\left\{\langle v, x_1\rangle^2 + \langle v, x_2\rangle^2 \leqslant \frac{\varepsilon}{2(n-1)}\|v\|_2^2\right\} \leqslant (e\varepsilon)^{1/2}$ by a nearly identical argument. Hence, letting $\mathcal{F}_t = \sigma(x_1, \ldots, x_t)$, we have shown that for all $t \geqslant 0$:

$$\mathbb{P}\left\{\frac{1}{2}\sum_{\ell=1}^{2}\langle v, x_{t+\ell}\rangle^2 \leqslant \varepsilon \cdot v^{\mathsf{T}}\left(\frac{1}{4(n-1)} I_n\right) v \,\middle|\, \mathcal{F}_t\right\} \leqslant (e\varepsilon)^{1/2},$$

from which the claim follows. ∎

For the next claim, recall that the mixing time of a Markov chain over state-space S with transition matrix P and stationary distribution π is defined as:

$$\tau_{\mathsf{mix}}(\varepsilon) = \inf\left\{ k \in \mathbb{N} \,\middle|\, \sup_{\mu \in \mathcal{P}(S)} \|\mu P^k - \pi\|_{\mathrm{tv}} \leqslant \varepsilon \right\}.$$

Here, $\mathcal{P}(S)$ denotes the set of distributions over S, and $\|\cdot\|_{\mathrm{tv}}$ is the total-variation norm over distributions.

Proposition B.11. *Let $n \geqslant 2$. Consider the Markov chain $\{i_t\}_{t \geqslant 1}$ where $i_1 \sim \mathrm{Unif}(\{1, \dots, n\})$ and $i_{t+1} \mid i_t \sim \mathrm{Unif}(\{1, \dots, n\} \setminus \{i_t\})$. We have that:*

$$\tau_{\mathsf{mix}}(\varepsilon) = \inf\left\{ k \in \mathbb{N} \,\middle|\, (n-1)^{-k} \leqslant \frac{2\varepsilon}{1 - 1/n} \right\}.$$

Proof. Let $\mathbf{1} \in \mathbb{R}^n$ denote the all ones vector. The transition matrix for this Markov chain is:

$$P = \frac{1}{n-1}(\mathbf{1}\mathbf{1}^\mathsf{T} - I_n),$$

and its stationary distribution is uniform over $\{1, \dots, n\}$. Note that for $j \geqslant 1$, $(\mathbf{1}\mathbf{1}^\mathsf{T})^j = n^{j-1}\mathbf{1}\mathbf{1}^\mathsf{T}$. Since $\mathbf{1}\mathbf{1}^\mathsf{T}$ and I_n commute, by the binomial theorem we have that:

$$\begin{aligned}
P^k &= \frac{1}{(n-1)^k} \sum_{j=0}^{k} \binom{k}{j} (\mathbf{1}\mathbf{1}^\mathsf{T})^{k-j} (-1)^j \\
&= \frac{1}{(n-1)^k} \left[\sum_{j=0}^{k-1} \binom{k}{j} n^{k-j-1} (-1)^j \mathbf{1}\mathbf{1}^\mathsf{T} + (-1)^k I_n \right] \\
&= \frac{1}{(n-1)^k} \left[\frac{1}{n} \left((n-1)^k - (-1)^k \right) \mathbf{1}\mathbf{1}^\mathsf{T} + (-1)^k I_n \right] \\
&= \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T} + \frac{(-1)^k}{(n-1)^k} \left[I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T} \right].
\end{aligned}$$

Now, let $\mu \in \mathbb{R}^n_{\geqslant 0}$ satisfy $\mu^\mathsf{T}\mathbf{1} = 1$. We have:

$$\left\| \mu^\mathsf{T} P^k - \frac{1}{n}\mathbf{1}^\mathsf{T} \right\|_1 = \frac{1}{(n-1)^k} \left\| \mu - \frac{1}{n}\mathbf{1} \right\|_1.$$

It is straightforward to check that $\sup_{\mu \in \mathbb{R}^n_{\geqslant 0}, \mu^\mathsf{T}\mathbf{1} = 1} \left\| \mu - \frac{1}{n}\mathbf{1} \right\|_1 = 1 - \frac{1}{n}$, from which the claim follows, since the TV distance between two distributions μ, ν is $\|\mu - \nu\|_{\mathrm{tv}} = \frac{1}{2}\|\mu - \nu\|_1$. ☐

Example 4.6 (Linear dynamical systems)**.** *Let (A, B) with $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times d}$ be k_c-step-controllable (Definition 4.2). Let $\mathsf{P}_x^{A,B}$ be the linear dynamical system defined in (3.8). Fix any $T, k \in \mathbb{N}_+$ satisfying $T \geqslant k \geqslant k_\mathsf{c}$. Then, $\mathsf{P}_x^{A,B}$ satisfies the $(T, k, \Gamma_k(A, B), e, \frac{1}{2})$-TrajSB condition.*

Proof. Let Γ_k be shorthand for $\Gamma_k(\mathsf{P}_x)$ and Σ_k be shorthand for $\Sigma_k(\mathsf{P}_x)$. Let $w = (w_1, \dots, w_k) \in \mathbb{R}^{nk}$ denote the vertical concatenation of the process noise variables. Let $M_t := \begin{bmatrix} A^t & \Phi_t \end{bmatrix} \in \mathbb{R}^{n \times n(k+1)}$ denote the matrix such that $x_t = M_t \begin{bmatrix} x \\ w \end{bmatrix}$. With this notation, for any $v \in \mathbb{R}^n$:

$$\frac{1}{k} \sum_{t=1}^{k} \langle v, x_t \rangle^2 = \begin{bmatrix} x \\ w \end{bmatrix}^\mathsf{T} \left(\frac{1}{k} \sum_{t=1}^{k} M_t^\mathsf{T} v v^\mathsf{T} M_t \right) \begin{bmatrix} x \\ w \end{bmatrix}.$$

By Equation (B.1), for any $\varepsilon > 0$,

$$\mathbb{P}\left\{ \begin{bmatrix} x \\ w \end{bmatrix}^\mathsf{T} \left(\frac{1}{k} \sum_{t=1}^{k} M_t^\mathsf{T} v v^\mathsf{T} M_t \right) \begin{bmatrix} x \\ w \end{bmatrix} \geqslant \varepsilon \cdot \operatorname{tr}\left(\frac{1}{k} \sum_{t=1}^{k} \Phi_t^\mathsf{T} v v^\mathsf{T} \Phi_t \right) \right\} \leqslant (e\varepsilon)^{1/2}.$$

On the other hand:

$$\operatorname{tr}\left(\frac{1}{k} \sum_{t=1}^{k} \Phi_t^\mathsf{T} v v^\mathsf{T} \Phi_t \right) = v^\mathsf{T} \left(\frac{1}{k} \sum_{t=1}^{k} \Phi_t \Phi_t^\mathsf{T} \right) v = v^\mathsf{T} \left(\frac{1}{k} \sum_{t=1}^{k} \Sigma_t \right) v = v^\mathsf{T} \Gamma_k v.$$

Because we assumed that $k \geqslant k_c$, then Γ_k is invertible. Thus, we can take $c_{\mathsf{sb}} = e$ and $\alpha = 1/2$. ☐

Proposition B.12. *Consider the scalar stochastic process $\{x_t\}_{t \geqslant 1}$ defined by:*

$$x_t = \sum_{i=0}^{t-1} \sum_{j=0}^{t-1} c_{i,j} w_{t-i-1} w_{t-j-1},$$

where $\{c_{i,j}\}_{i,j \geqslant 0}$ are the coefficients which describe the dynamics, and $\{w_t\}_{t \geqslant 0}$ are iid $N(0,1)$ random variables. Let $\{\mathcal{F}_t\}_{t \geqslant 1}$ denote the filtration defined as $\mathcal{F}_t := \sigma(w_0, \dots, w_{t-1})$, so that x_t is $\mathcal{F}_t$-measurable. Suppose that $\{c_{i,j}\}_{i,j \geqslant 0}$ is symmetric and traceless. For every $t \geqslant 1$ and $k \geqslant 0$, almost surely we have:

$$\mathbb{E}[x_{t+k}^2 \mid \mathcal{F}_k] \geqslant \mathbb{E}[x_t^2] + (\mathbb{E}[x_{t+k} \mid \mathcal{F}_k])^2.$$

Proof. For $t \in \mathbb{N}_+$, define the symmetric matrices $M_t \in \mathbb{R}^{t \times t}$ with $(M_t)_{ii} = 0$ and $(M_t)_{ij} = c_{(i-1),(j-1)}$. With this notation and with $\bar{w}_t \sim N(0, I_t)$, we can write x_t as:

$$x_t = \bar{w}_t^\mathsf{T} M_t \bar{w}_t.$$

Therefore, by Proposition B.5 and the assumption that $\operatorname{tr}(M_t) = 0$:

$$\mathbb{E}[x_t^2] = \mathbb{E}(\bar{w}_t^\mathsf{T} M_t \bar{w}_t)^2 = 2\|M_t\|_F^2 + \operatorname{tr}(M_t)^2 = 2\|M_t\|_F^2.$$

Now, partition M_{t+k} as:

$$M_{t+k} = \begin{bmatrix} M_t & D_{t,k} \\ D_{t,k}^\mathsf{T} & E_{t,k} \end{bmatrix}.$$

Let $\bar{v}_k = (w_{k-1}, \dots, w_0)$. Given $\mathcal{F}_k$, we can write x_{t+k} as:

$$x_{t+k} = \begin{bmatrix} \bar{w}_t \\ \bar{v}_k \end{bmatrix}^\mathsf{T} \begin{bmatrix} M_t & D_{t,k} \\ D_{t,k}^\mathsf{T} & E_{t,k} \end{bmatrix} \begin{bmatrix} \bar{w}_t \\ \bar{v}_k \end{bmatrix}.$$

With this notation:

$$\mathbb{E}[x_{t+k} \mid \mathcal{F}_k] = \bar{v}_k^\mathsf{T} E_{t,k}\bar{v}_k, \quad \mathbb{E}[x_{t+k}^2 \mid \mathcal{F}_k] = \mathbb{E}_{\bar{w}_t}\left(\begin{bmatrix}\bar{w}_t \\ \bar{v}_k\end{bmatrix}^\mathsf{T}\begin{bmatrix}M_t & D_{t,k} \\ D_{t,k}^\mathsf{T} & E_{t,k}\end{bmatrix}\begin{bmatrix}\bar{w}_t \\ \bar{v}_k\end{bmatrix}\right)^2.$$

Expanding the square:

$$\begin{aligned}
\left(\begin{bmatrix}\bar{w}_t \\ \bar{v}_k\end{bmatrix}^\mathsf{T}\begin{bmatrix}M_t & D_{t,k} \\ D_{t,k}^\mathsf{T} & E_{t,k}\end{bmatrix}\begin{bmatrix}\bar{w}_t \\ \bar{v}_k\end{bmatrix}\right)^2 &= (\bar{w}_t^\mathsf{T} M_t\bar{w}_t + 2\bar{w}_t^\mathsf{T} D_{t,k}\bar{v}_k + \bar{v}_k^\mathsf{T} E_{t,k}\bar{v}_k)^2 \\
&= (\bar{w}_t^\mathsf{T} M_t\bar{w}_t)^2 + 4\bar{w}_t^\mathsf{T} M_t\bar{w}_t\bar{w}_t^\mathsf{T} D_{t,k}\bar{v}_k + 2\bar{w}_t^\mathsf{T} M_t\bar{w}_t\bar{v}_k^\mathsf{T} E_{t,k}\bar{v}_k \\
&\quad + 4(\bar{w}_t^\mathsf{T} D_{t,k}\bar{v}_k)^2 + 4\bar{w}_t^\mathsf{T} D_{t,k}\bar{v}_k\bar{v}_k^\mathsf{T} E_{t,k}\bar{v}_k + (\bar{v}_k^\mathsf{T} E_{t,k}\bar{v}_k)^2.
\end{aligned}$$

Using Proposition B.5 again:

$$\begin{aligned}
\mathbb{E}[x_{t+k}^2 \mid \mathcal{F}_k] &= \mathbb{E}_{\bar{w}_t}\left(\begin{bmatrix}\bar{w}_t \\ \bar{v}_k\end{bmatrix}^\mathsf{T}\begin{bmatrix}M_t & D_{t,k} \\ D_{t,k}^\mathsf{T} & E_{t,k}\end{bmatrix}\begin{bmatrix}\bar{w}_t \\ \bar{v}_k\end{bmatrix}\right)^2 \\
&= \mathbb{E}_{\bar{w}_t}(\bar{w}_t^\mathsf{T} M_t\bar{w}_t)^2 + 2\operatorname{tr}(M_t)\bar{v}_k^\mathsf{T} E_{t,k}\bar{v}_k + 4\|D_{t,k}\bar{v}_k\|_2^2 + (\bar{v}_k^\mathsf{T} E_{t,k}\bar{v}_k)^2 \\
&= 2\|M_t\|_F^2 + 4\|D_{t,k}\bar{v}_k\|_2^2 + (\bar{v}_k^\mathsf{T} E_{t,k}\bar{v}_k)^2 \geqslant 2\|M_t\|_F^2 + (\bar{v}_k^\mathsf{T} E_{t,k}\bar{v}_k)^2.
\end{aligned}$$

To complete the proof, we recall that $\mathbb{E}[x_t^2] = 2\|M_t\|_F^2$ and $\mathbb{E}[x_{t+k} \mid \mathcal{F}_k] = \bar{v}_k^\mathsf{T} E_{t,k}\bar{v}_k$. □

Example 4.8 (Degree-2 Volterra series)**.** *Consider the following process P_x. Let $\{c_{i,j}^{(\ell)}\}_{i,j\geqslant 0}$ for $\ell \in \{1,\dots,n\}$ be symmetric, traceless, non-degenerate arrays (Definition 4.3). Let $\{w_t^{(\ell)}\}_{t\geqslant 0}$ be iid $N(0,1)$ random variables for $\ell \in \{1,\dots,n\}$. For $t \geqslant 1$, the ℓ-th coordinate of x_t, denoted $(x_t)_\ell$, is:*

$$(x_t)_\ell = \sum_{i=0}^{t-1}\sum_{j=i}^{t-1} c_{i,j}^{(\ell)} w_{t-i-1}^{(\ell)} w_{t-j-1}^{(\ell)}. \tag{4.3}$$

Let $k_{\mathsf{nd}} \in \mathbb{N}_+$ denote the smallest non-degeneracy index for all n arrays. There is a universal positive constant c such that for any T and k satisfying $T \geqslant k \geqslant k_{\mathsf{nd}}$, P_x satisfies the $(T, k, \Gamma_k(\mathsf{P}_x), c, \frac{1}{4})$-TrajSB condition.

Proof. Fix a $v \in \mathbb{R}^n$. The relation (4.3) shows that $\langle v, x_t\rangle^2$ is a degree four polynomial in $\{w_i^{(\ell)}\}_{i=0,\ell=1}^{t-1,n}$. Let $\mathcal{F}_t = \sigma(\{w_i^{(\ell)}\}_{i=0,\ell=1}^{t-1,n})$, so that x_t is $\mathcal{F}_t$-measurable. By Theorem B.7, there exists a univeral positive constant $c > 0$ such that for any $s \geqslant 0$,

$$\mathbb{P}\left\{\frac{1}{k}\sum_{t=1}^{k}\langle v, x_{t+s}\rangle^2 \leqslant \varepsilon\mathbb{E}\left[\frac{1}{k}\sum_{t=1}^{k}\langle v, x_{t+s}\rangle^2 \,\middle|\, \mathcal{F}_s\right] \,\middle|\, \mathcal{F}_s\right\} \leqslant (c\varepsilon)^{1/4} \quad \text{a.s.}$$

To conclude the proof, we need to lower bound $\mathbb{E}\left[\frac{1}{k}\sum_{t=1}^{k}\langle v, x_{t+s}\rangle^2 \,\middle|\, \mathcal{F}_s\right]$. For any $t \geqslant 1$,

$$\mathbb{E}\left[\langle v, x_{t+s}\rangle^2 \,\middle|\, \mathcal{F}_s\right] = \mathbb{E}\left[\left(\sum_{\ell=1}^{n} v_\ell \cdot (x_{t+s})_\ell\right)^2 \,\middle|\, \mathcal{F}_s\right]$$

$$
\begin{aligned}
&\overset{(a)}{=} \sum_{\ell=1}^{n} v_\ell^2 \cdot \mathbb{E}[(x_{t+s})_\ell^2 \mid \mathcal{F}_s] + \sum_{\ell_1 \neq \ell_2}^{n} v_{\ell_1} v_{\ell_2} \cdot \mathbb{E}[(x_{t+s})_{\ell_1} \mid \mathcal{F}_s] \cdot \mathbb{E}[(x_{t+s})_{\ell_2}] \mid \mathcal{F}_s] \\
&\overset{(b)}{\geqslant} \sum_{\ell=1}^{n} v_\ell^2 \cdot \mathbb{E}[(x_t)_\ell^2] + \sum_{\ell=1}^{n} v_\ell^2 \cdot (\mathbb{E}[(x_{t+s})_\ell \mid \mathcal{F}_s])^2 \\
&\qquad + \sum_{\ell_1 \neq \ell_2}^{n} v_{\ell_1} v_{\ell_2} \cdot \mathbb{E}[(x_{t+s})_{\ell_1} \mid \mathcal{F}_s] \cdot \mathbb{E}[(x_{t+s})_{\ell_2}] \mid \mathcal{F}_s] \\
&= \sum_{\ell=1}^{n} v_\ell^2 \cdot \mathbb{E}[(x_t)_\ell^2] + \left(\sum_{\ell=1}^{n} v_\ell \cdot \mathbb{E}[(x_{t+s})_\ell \mid \mathcal{F}_s] \right)^2 \\
&\geqslant \sum_{\ell=1}^{n} v_\ell^2 \cdot \mathbb{E}[(x_t)_\ell^2] \overset{(c)}{=} v^\mathsf{T} \Sigma_t(\mathsf{P}_x) v.
\end{aligned}
$$

Above, (a) follows since each coordinate of x_t is independent by definition, (b) follows from Proposition B.12, and (c) follows since $\mathbb{E}[x_t] = 0$ and each coordinate is independent, so $\mathbb{E}[(x_t)_{\ell_1}(x_t)_{\ell_2}] = \mathbb{E}[(x_t)_{\ell_1}]\mathbb{E}[(x_t)_{\ell_2}] = 0$ for $\ell_1 \neq \ell_2$. Hence, we have shown:

$$
\mathbb{E}\left[\frac{1}{k} \sum_{t=1}^{k} \langle v, x_{t+s} \rangle^2 \,\middle|\, \mathcal{F}_s \right] \geqslant v^\mathsf{T} \left(\frac{1}{k} \sum_{t=1}^{k} \Sigma_t(\mathsf{P}_x) \right) v = v^\mathsf{T} \Gamma_k(\mathsf{P}_x) v.
$$

Note that because we assume that $k \geqslant k_{\mathsf{nd}}$, the covariances $\Sigma_t(\mathsf{P}_x)$ are all invertible (and hence so is $\Gamma_k(\mathsf{P}_x)$). The claim now follows. $\square$

Example 4.7 (Degree-D Volterra series)**.** *Fix a $D \in \mathbb{N}_+$. Let $\{c^{(d,\ell)}_{i_1,\ldots,i_d}\}_{i_1,\ldots,i_d \in \mathbb{N}}$ for $d \in \{1, \ldots, D\}$ and $\ell \in \{1, \ldots, n\}$ denote arbitrary rank-d arrays. Let $\{w_t^{(\ell)}\}_{t \geqslant 0}$ be iid $N(0, 1)$ random variables for $\ell \in \{1, \ldots, n\}$. Consider the process P_x where for $t \geqslant 1$, the ℓ-th coordinate of x_t, denoted $(x_t)_\ell$, is:*

$$
(x_t)_\ell = \sum_{d=1}^{D} \sum_{i_1,\ldots,i_d=0}^{t-1} c^{(d,\ell)}_{i_1,\ldots,i_d} \prod_{d'=1}^{d} w^{(\ell)}_{t-i_{d'}-1}. \tag{4.2}
$$

Let $T_{\mathsf{nd}} := \inf\{t \in \mathbb{N}_+ \mid \det(\Gamma_t(\mathsf{P}_x)) \neq 0\}$, and suppose T_{nd} is finite. There is a constant $c_D > 0$, depending only on D, such that for any $T \geqslant T_{\mathsf{nd}}$, P_x satisfies the $(T, T, \Gamma_T(\mathsf{P}_x), c_D, 1/(2D))$-TrajSB condition.

Proof. Fix a $v \in \mathbb{R}^n$. The definition (4.2) expresses $\langle v, x_t \rangle$ as a degree at most D polynomial in the noise variables $\{w_t^{(\ell)}\}$. By Theorem B.7, there exists a positive constant c_D, only depending on D, such that:

$$
\mathbb{P}\left\{ \frac{1}{T} \sum_{t=1}^{T} \langle v, x_t \rangle^2 \leqslant \varepsilon \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \langle v, x_t \rangle^2 \right] \right\} \leqslant (c_D \varepsilon)^{1/(2D)}.
$$

Since $T \geqslant T_{\mathsf{nd}}$, the matrix $\Gamma_T(\mathsf{P}_x)$ is invertible. The claim now follows. $\square$

B.3 Proof of Proposition 4.2

Proposition 4.2 (Average small-ball implies trajectory small-ball)**.** *Fix $T \in \mathbb{N}_+$, $k \in \{1, \dots, T\}$, $\{\Psi_j\}_{j=1}^{\lfloor T/k \rfloor} \subset \mathsf{Sym}^n_{>0}$, and $\alpha, \beta \in (0,1)$. Let P_x be a covariate distribution, with $\{x_t\}_{t \geqslant 1}$ adapted to a filtration $\{\mathcal{F}_t\}_{t \geqslant 1}$. Suppose for all $v \in \mathbb{R}^n \setminus \{0\}$ and $j \in \{1, \dots, \lfloor T/k \rfloor\}$:*

$$\frac{1}{k} \sum_{t=(j-1)k+1}^{jk} \mathbb{P}_{x_t \sim \mathsf{P}_x} \left\{ \langle v, x_t \rangle^2 \leqslant \alpha \cdot v^\mathsf{T} \Psi_j v \;\middle|\; \mathcal{F}_{(j-1)k} \right\} \leqslant \beta \;\; a.s., \tag{4.4}$$

where $\mathcal{F}_0$ is the minimal σ-algebra. Then, for all $v \in \mathbb{R}^n \setminus \{0\}$, $j \in \{1, \dots, \lfloor T/k \rfloor\}$, and $\varepsilon \in (0, \alpha)$

$$\mathbb{P}_{\{x_t\} \sim \mathsf{P}_x} \left\{ \frac{1}{k} \sum_{t=(j-1)k+1}^{jk} \langle v, x_t \rangle^2 \leqslant \varepsilon \cdot v^\mathsf{T} \Psi_j v \;\middle|\; \mathcal{F}_{(j-1)k} \right\} \leqslant \frac{\beta}{1 - \varepsilon/\alpha} \;\; a.s. \tag{4.5}$$

Proof. The following proof builds on the argument given in [SMT+18, Section E.1]. We note that a similar style of proof is used in [BLLT20, Lemma 15].

Define the shorthand notation $\mathbb{P}_t\{\cdot\} := \mathbb{P}\{\cdot \mid \mathcal{F}_t\}$, and similarly $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$. Now fix a $v \in \mathbb{R}^n \setminus \{0\}$, $j \in \{1, \dots, \lfloor T/k \rfloor\}$. Markov's inequality yields that:

$$\frac{1}{k} \sum_{t=(j-1)k+1}^{jk} \langle v, x_t \rangle^2 \geqslant \alpha v^\mathsf{T} \Psi_j v \cdot \frac{1}{k} \sum_{t=(j-1)k+1}^{jk} \mathbf{1}\{\langle v, x_t \rangle^2 > \alpha v^\mathsf{T} \Psi_j v\},$$

and therefore for all $\varepsilon > 0$:

$$\mathbb{P}_{(j-1)k} \left\{ \frac{1}{k} \sum_{t=(j-1)k+1}^{jk} \langle v, x_t \rangle^2 \leqslant \varepsilon \cdot v^\mathsf{T} \Psi_j v \right\} \leqslant \mathbb{P}_{(j-1)k} \left\{ \frac{1}{k} \sum_{t=(j-1)k+1}^{jk} \mathbf{1}\{\langle v, x_t \rangle^2 > \alpha v^\mathsf{T} \Psi_j v\} \leqslant \varepsilon/\alpha \right\}.$$

Define $Z_j := \frac{1}{k} \sum_{t=(j-1)k+1}^{jk} \mathbf{1}\{\langle v, x_t \rangle^2 > \alpha v^\mathsf{T} \Psi_j v\}$, and observe that $Z_j \in [0,1]$. By (4.4), we have:

$$\mathbb{E}_{(j-1)k}[Z_j] \geqslant 1 - \beta.$$

On the other hand:

$$\begin{aligned}
\mathbb{E}_{(j-1)k}[Z_j] &= \mathbb{E}_{(j-1)k}[Z_j \mathbf{1}\{Z_j > \varepsilon/\alpha\}] + \mathbb{E}_{(j-1)k}[Z_j \mathbf{1}\{Z_j \leqslant \varepsilon/\alpha\}] \\
&\leqslant \mathbb{P}_{(j-1)k}\{Z_j > \varepsilon/\alpha\} + \varepsilon/\alpha \cdot \mathbb{P}_{(j-1)k}\{Z_j \leqslant \varepsilon/\alpha\} && \text{since } Z_j \leqslant 1 \\
&= 1 - (1 - \varepsilon/\alpha)\mathbb{P}_{(j-1)k}\{Z_j \leqslant \varepsilon/\alpha\}.
\end{aligned}$$

Combining both these inequalities, and further restricting $\varepsilon \in (0, \alpha)$, we obtain,

$$\mathbb{P}_{(j-1)k}\{Z_j \leqslant \varepsilon/\alpha\} \leqslant \frac{\beta}{1 - \varepsilon/\alpha},$$

which implies (4.5). □

B.4 General ordinary least-squares estimator upper bound

In this section, we supply the proof of Lemma 5.1. We first start with a result which bounds the minimum eigenvalue of the empirical covariance matrix.

Lemma B.13 (Minimum eigenvalue bound via trajectory small-ball)**.** *Suppose that P_x satisfies the $(T, k, \{\Psi_j\}_{j=1}^{\lfloor T/k \rfloor}, c_{\mathsf{sb}}, \alpha)$-trajectory-small-ball condition (Definition 4.1). Put $S := \lfloor T/k \rfloor$, and $\Gamma_T := \Gamma_T(\mathsf{P}_x)$. Fix any $\underline{\Gamma} \in \mathsf{Sym}^n_{>0}$ satisfying $\frac{1}{S}\sum_{j=1}^S \Psi_j \preccurlyeq \underline{\Gamma} \preccurlyeq \Gamma_T$. Define $\tilde{X}_{m,T} := X_{m,T}\underline{\Gamma}^{-1/2}$, and:*

$$\underline{\mu}(\{\Psi_j\}_{j=1}^S, \underline{\Gamma}) := \left[\prod_{j=1}^S \underline{\lambda}(\Psi_j, \underline{\Gamma})\right]^{1/S}. \tag{B.2}$$

Suppose that:

$$n \geqslant 2, \quad \frac{mT}{kn} \geqslant \frac{32}{\alpha} \log\left(\frac{320 c_{\mathsf{sb}}}{\alpha \underline{\lambda}(\underline{\Gamma}, \Gamma_T)\underline{\mu}(\{\Psi_j\}_{j=1}^S, \underline{\Gamma})}\right). \tag{B.3}$$

For any $t \geqslant 0$, with probability at least $1 - 2e^{-t}$, the following statements simultaneously hold:

$$\operatorname{tr}(\tilde{X}_{m,T}^\mathsf{T} \tilde{X}_{m,T}) \leqslant \frac{mTne^t}{\underline{\lambda}(\underline{\Gamma}, \Gamma_T)}, \quad \lambda_{\min}(\tilde{X}_{m,T}^\mathsf{T} \tilde{X}_{m,T}) \geqslant \frac{mT\alpha \underline{\mu}(\{\Psi_j\}_{j=1}^S, \underline{\Gamma})}{8ec_{\mathsf{sb}}} \exp\left(-\frac{16kn}{mT\alpha} t\right). \tag{B.4}$$

Proof. The proof uses the PAC-Bayes argument for uniform convergence from [Mou22]. The first step is to construct a family of random variables, indexed by both $v \in \mathbb{S}^{n-1}$ and a scale parameter $\eta > 0$, such that its moment generating function is pointwise bounded by one. For notational brevity, let:

$$\underline{\lambda} := \underline{\lambda}(\underline{\Gamma}, \Gamma_T), \quad \underline{\mu} := \underline{\mu}(\{\Psi_j\}_{j=1}^S, \underline{\Gamma}).$$

Since $\underline{\Gamma} \preccurlyeq \Gamma_T$ by assumption, we have $\underline{\lambda} \in (0, 1]$. Similarly, since $\frac{1}{S}\sum_{j=1}^S \Psi_j \preccurlyeq \underline{\Gamma}$, we also have $\underline{\mu} \in (0, 1]$ by Proposition B.3.

The trajectory small-ball condition (4.1) implies for any $v \in \mathbb{S}^{n-1}$, $j \in \{1, \dots, S\}$, and $\varepsilon > 0$,

$$\mathbb{P}\left\{\frac{1}{k}\sum_{t=(j-1)k+1}^{jk} \langle v, \underline{\Gamma}^{-1/2} x_t\rangle^2 \leqslant \varepsilon \underline{\lambda}(\Psi_j, \underline{\Gamma}) \,\middle|\, \mathcal{F}_{(j-1)k}\right\} \leqslant (c_{\mathsf{sb}}\varepsilon)^\alpha.$$

Using a change of variables $\varepsilon \leftarrow \varepsilon / \underline{\lambda}(\Psi_j, \underline{\Gamma})$,

$$\mathbb{P}\left\{\frac{1}{k}\sum_{t=(j-1)k+1}^{jk} \langle v, \underline{\Gamma}^{-1/2} x_t\rangle^2 \leqslant \varepsilon \,\middle|\, \mathcal{F}_{(j-1)k}\right\} \leqslant \left(c_{\mathsf{sb}}/\underline{\lambda}(\Psi_j, \underline{\Gamma}) \cdot \varepsilon\right)^\alpha.$$

By Proposition B.4, for any $\eta > 0$,

$$\mathbb{E}\left[\exp\left(-\frac{\eta}{k}\sum_{t=(j-1)k+1}^{jk} \langle v, \underline{\Gamma}^{-1/2} x_t\rangle^2 + \alpha \log\left(\frac{\eta \underline{\lambda}(\Psi_j, \underline{\Gamma})}{c_{\mathsf{sb}}}\right)\right) \,\middle|\, \mathcal{F}_{(j-1)k}\right] \leqslant 1 \quad \text{a.s.} \tag{B.5}$$

For $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, S\}$, define the random variables $Z_j^{(i)}(v;\eta)$, $Z^{(i)}(v;\eta)$, and $Z(v;\eta)$:

$$\begin{aligned}
Z_j^{(i)}(v;\eta) &:= -\frac{\eta}{k} \sum_{t=(j-1)k+1}^{jk} \langle v, \underline{\Gamma}^{-1/2} x_t^{(i)} \rangle^2 + \alpha \log\left(\frac{\eta \underline{\lambda}(\Psi_j, \underline{\Gamma})}{c_{\mathsf{sb}}}\right), \\
Z^{(i)}(v;\eta) &:= \sum_{j=1}^{S} Z_j^{(i)}(v;\eta), \\
Z(v,\eta) &:= \sum_{i=1}^{m} Z^{(i)}(v;\eta).
\end{aligned}$$

We first claim that $\mathbb{E}[\exp(Z(v;\eta))] \leqslant 1$ for every $v \in \mathbb{S}^{n-1}$ and $\eta > 0$. Since $Z^{(i)}(v;\eta)$ is independent of $Z^{(i')}(v;\eta)$ whenever $i \neq i'$, we have that:

$$\mathbb{E}[\exp(Z(v;\eta))] = \mathbb{E}\left[\exp\left(\sum_{i=1}^{m} Z^{(i)}(v;\eta)\right)\right] = \prod_{i=1}^{m} \mathbb{E}[\exp(Z^{(i)}(v;\eta))].$$

Furthermore, by repeated applications of the tower property and (B.5), for every $i \in \{1, \ldots, m\}$,

$$\begin{aligned}
\mathbb{E}[\exp(Z^{(i)}(v;\eta))] &= \mathbb{E}\left[\exp\left(\sum_{j=1}^{S} Z_j^{(i)}(v;\eta)\right)\right] \\
&= \mathbb{E}\left[\exp\left(\sum_{j=1}^{S-1} Z_j^{(i)}(v;\eta)\right) \mathbb{E}[\exp(Z_S^{(i)}(v;\eta)) \mid \mathcal{F}_{(S-1)k}]\right] \\
&\leqslant \mathbb{E}\left[\exp\left(\sum_{j=1}^{S-1} Z_j^{(i)}(v;\eta)\right)\right] \\
&\vdots \\
&\leqslant 1.
\end{aligned}$$

Hence $\mathbb{E}[\exp(Z(v;\eta))] \leqslant 1$ for every $v \in \mathbb{S}^{n-1}$ and $\eta > 0$.

Let us now import some notation from [Mou22]. First, let π denote the spherical measure on $\mathbb{S}^{n-1}$, and let $\rho_{v,\gamma}$ denote the uniform measure over the spherical cap

$$\mathcal{C}(v,\gamma) := \{w \in \mathbb{S}^{n-1} \mid \|v - w\|_2 \leqslant \gamma\}.$$

Next, let $F_{v,\gamma}(\Sigma) := \int_{\mathcal{C}(v,\gamma)} \langle w, \Sigma w \rangle \, \mathrm{d}\rho_{v,\gamma}$ for any symmetric matrix Σ.

Fix any positive t, η. For two measures μ and ν with μ absolutely continuous w.r.t. ν, let $\mathsf{KL}(\mu,\nu) := \mathbb{E}_\mu \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)$ denote the KL-divergence between μ and ν. By the PAC-Bayes deviation bound (cf. [Cat07]), there exists an event $\mathcal{E}_{t,1}$ with probability at least $1 - e^{-t}$, such that on $\mathcal{E}_{t,1}$, we have for every $v \in \mathbb{S}^{n-1}$ and $\gamma > 0$,

$$-\frac{\eta}{k} F_{v,\gamma}\left(\underline{\Gamma}^{-1/2} \sum_{i=1}^{m} \sum_{t=1}^{kS} x_t^{(i)} (x_t^{(i)})^{\mathsf{T}} \underline{\Gamma}^{-1/2}\right) + mS\alpha \log\left(\frac{\eta \underline{\mu}}{c_{\mathsf{sb}}}\right) \leqslant \mathsf{KL}(\rho_{v,\gamma}, \pi) + t. \tag{B.6}$$

Next, by [Mou22, Sections 5.3 and 5.4], we can write $F_{v,\gamma}$ in terms of a scalar function ϕ such that:

$$F_{v,\gamma}(\Sigma) = (1-\phi(\gamma))\langle v, \Sigma v\rangle + \phi(\gamma)\frac{1}{n}\operatorname{tr}(\Sigma), \quad \phi(\gamma) \in \left[0, \frac{n}{n-1}\gamma^2\right]. \tag{B.7}$$

Furthermore, for every $v \in \mathbb{S}^{n-1}$ and $\gamma > 0$, the KL-divergence term can be upper bounded by:

$$\mathsf{KL}(\rho_{v,\gamma}, \pi) \leqslant n \log\left(1 + \frac{2}{\gamma}\right). \tag{B.8}$$

Therefore on $\mathcal{E}_{t,1}$, plugging (B.7) and (B.8) into (B.6),

$$\begin{aligned}\lambda_{\min}\left(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}\right) \geqslant\ & \frac{k}{\eta(1-\phi(\gamma))}\left[mS\alpha\log\left(\frac{\eta\underline{\mu}}{c_{\mathsf{sb}}}\right) - n\log\left(1+\frac{2}{\gamma}\right) - t\right] \\ & - \frac{\phi(\gamma)}{1-\phi(\gamma)}\frac{1}{n}\operatorname{tr}\left(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}\right).\end{aligned}$$

Restricting $\gamma \in [0, 1/2]$, we have from (B.7) that $0 \leqslant \phi(\gamma) \leqslant \frac{n}{n-1}\gamma^2 \leqslant 2\gamma^2 \leqslant 1/2$. Hence, $1 - \phi(\gamma) \in [1/2, 1]$. Furthermore, $1 + 2/\gamma \leqslant 5/(4\gamma^2)$. Therefore,

$$\lambda_{\min}\left(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}\right) \geqslant \frac{k}{\eta}\left[mS\alpha\log\left(\frac{\eta\underline{\mu}}{c_{\mathsf{sb}}}\right) - n\log\left(\frac{5}{4\gamma^2}\right) - t\right] - \frac{4\gamma^2}{n}\operatorname{tr}\left(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}\right).$$

Define the non-negative random variables $\psi_i := \sum_{t=1}^T \|\underline{\Gamma}^{-1/2}x_t^{(i)}\|_2^2$, for $i = 1, \dots, m$. It is straightforward to verify that $\operatorname{tr}(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}) = \sum_{i=1}^m \psi_i$. By Markov's inequality, for any $\beta > 0$:

$$\begin{aligned}\mathbb{P}\left(\operatorname{tr}(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}) > \beta\right) &= \mathbb{P}\left(\sum_{i=1}^m \psi_i > \beta\right) \leqslant \frac{\mathbb{E}\left[\sum_{i=1}^m \psi_i\right]}{\beta} \\ &= \frac{mT\operatorname{tr}(\underline{\Gamma}^{-1}\Gamma_T)}{\beta} \leqslant \frac{mTn\lambda_{\max}(\Gamma_T^{1/2}\underline{\Gamma}^{-1}\Gamma_T^{1/2})}{\beta} = \frac{mTn}{\underline{\lambda}\beta}.\end{aligned}$$

Therefore, setting $\beta = \frac{e^t mTn}{\underline{\lambda}}$, there exists an event $\mathcal{E}_{t,2}$ such that $\mathbb{P}(\mathcal{E}_{t,2}^c) \leqslant e^{-t}$ and on $\mathcal{E}_{t,2}$,

$$\operatorname{tr}(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}) \leqslant \frac{e^t mTn}{\underline{\lambda}}.$$

Therefore on $\mathcal{E}_{t,1} \cap \mathcal{E}_{t,2}$, which we assume holds for the remainder of the proof, we have:

$$\lambda_{\min}\left(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}\right) \geqslant \frac{k}{\eta}\left[mS\alpha\log\left(\frac{\eta\underline{\mu}}{c_{\mathsf{sb}}}\right) - n\log\left(\frac{5}{4\gamma^2}\right) - t\right] - \frac{4mTe^t}{\underline{\lambda}}\gamma^2$$

Next, we further restrict $\frac{\eta\underline{\mu}}{c_{\mathsf{sb}}} \geqslant e$ so that $\log(\eta\underline{\mu}/c_{\mathsf{sb}}) \geqslant 1$. Now consider, for positive constants A, B, C, the function $x \mapsto A\log(B/x) + Cx$ on the domain $(0, \infty)$. The derivative vanishes at $x = A/C$, and the function attains a minimum value of $A(1 + \log(BC/A))$ with this choice of x. Let us set:

$$A \leftarrow \frac{kn}{\eta}, \quad B \leftarrow \frac{5}{4}, \quad C \leftarrow \frac{4mTe^t}{\underline{\lambda}}, \quad x \leftarrow \gamma^2.$$

Then by choosing $\gamma^2 = \frac{kn\underline{\lambda}}{4\eta m T e^t}$, we have that:

$$\frac{kn}{\eta}\log\left(\frac{5}{4\gamma^2}\right) + \frac{4mTe^t}{\underline{\lambda}}\gamma^2 = \frac{kn}{\eta}\left[1+\log\left(\frac{5mTe^t\eta}{kn\underline{\lambda}}\right)\right].$$

Note that this choice of γ satisfies $\gamma \in [0, 1/2]$, since:

$$\begin{aligned}
\frac{kn\underline{\lambda}}{4\eta m T e^t} \leqslant \frac{1}{4} &\Longleftarrow \frac{kn}{\eta m T} \leqslant 1 && \text{since } t \geqslant 0 \text{ and } \underline{\lambda} \leqslant 1 \\
&\Longleftarrow \frac{kn\underline{\mu}}{ec_{\mathsf{sb}} m T} \leqslant 1 && \text{since } \eta \geqslant ec_{\mathsf{sb}}/\underline{\mu} \\
&\Longleftrightarrow \frac{n\underline{\mu}}{ec_{\mathsf{sb}}} \leqslant \frac{mT}{k} \\
&\Longleftarrow \frac{n}{ec_{\mathsf{sb}}} \leqslant \frac{mT}{k} && \text{since } \underline{\mu} \leqslant 1,
\end{aligned}$$

and the last condition holds by (B.3). With this choice of γ, we have:

$$\begin{aligned}
&\lambda_{\min}\left(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}\right) \\
&\geqslant \frac{k}{\eta}\left[mS\alpha\log\left(\frac{\eta\underline{\mu}}{c_{\mathsf{sb}}}\right) - t - n\left(1+\log\left(\frac{5mTe^t\eta}{kn\underline{\lambda}}\right)\right)\right] \\
&= \frac{k}{\eta}\left[(mS\alpha - n)\log\left(\frac{\eta\underline{\mu}}{c_{\mathsf{sb}}}\right) - t - n\left(1+\log\left(\frac{5c_{\mathsf{sb}}mTe^t}{kn\underline{\lambda}\underline{\mu}}\right)\right)\right] \\
&\geqslant \frac{k}{\eta}\left[\frac{mS\alpha}{2}\log\left(\frac{\eta\underline{\mu}}{c_{\mathsf{sb}}}\right) - t - n\left(1+\log\left(\frac{5c_{\mathsf{sb}}mTe^t}{kn\underline{\lambda}\underline{\mu}}\right)\right)\right] && \text{since } mS \geqslant 2n/\alpha \\
&= \frac{k}{\eta}\left[\frac{mS\alpha}{2}\log\left(\frac{\eta\underline{\mu}}{c_{\mathsf{sb}}}\right) - (1+n)t - n\left(1+\log\left(\frac{5c_{\mathsf{sb}}mT}{kn\underline{\lambda}\underline{\mu}}\right)\right)\right] \\
&\geqslant \frac{k}{\eta}\left[\frac{mS\alpha}{2}\log\left(\frac{\eta\underline{\mu}}{c_{\mathsf{sb}}}\right) - 2nt - n\left(1+\log\left(\frac{5c_{\mathsf{sb}}mT}{kn\underline{\lambda}\underline{\mu}}\right)\right)\right] \\
&\geqslant \frac{k}{\eta}\left[\frac{mS\alpha}{4}\log\left(\frac{\eta\underline{\mu}}{c_{\mathsf{sb}}}\right) - 2nt\right] && \text{since } \frac{mS\alpha}{4n} \geqslant 1+\log\left(\frac{5c_{\mathsf{sb}}mT}{kn\underline{\lambda}\underline{\mu}}\right) \\
&= \frac{kmS\alpha}{4c_{\mathsf{sb}}/\underline{\mu}}\left[\frac{\log(\eta\underline{\mu}/c_{\mathsf{sb}}) - \frac{8nt}{mS\alpha}}{\eta\underline{\mu}/c_{\mathsf{sb}}}\right]. && \text{(B.9)}
\end{aligned}$$

It remains to optimize over $\eta \in [ec_{\mathsf{sb}}/\underline{\mu}, \infty)$. For any $G \in \mathbb{R}$, the function $\eta' \mapsto \frac{\log\eta' - G}{\eta'}$ on $(0,\infty)$ attains a maximum of $\exp(-1-G)$ at $\eta' = \exp(1+G)$. Hence, setting $\eta = \frac{c_{\mathsf{sb}}}{\underline{\mu}}\exp(1+8nt/(mS\alpha))$, which satisfies $\eta \geqslant ec_{\mathsf{sb}}/\underline{\mu}$, we have:

$$\begin{aligned}
\lambda_{\min}\left(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}\right) &\geqslant \frac{kmS\alpha\underline{\mu}}{4ec_{\mathsf{sb}}}\exp\left(-\frac{8nt}{mS\alpha}\right) \\
&\geqslant \frac{mT\alpha\underline{\mu}}{8ec_{\mathsf{sb}}}\exp\left(-\frac{16kn}{mT\alpha}t\right) && \text{since } S \geqslant T/(2k).
\end{aligned}$$

The claim now follows by gathering the requirements on the quantity $\frac{mT}{kn}$ and simplifying as in (B.3). $\square$

Corollary B.14. *Assume the hypothesis of Lemma B.13 hold. Then, $\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}$ is invertible almost surely.*

Proof. For any $t \geqslant 0$, define the event $\mathcal{E}_t$ as:

$$\mathcal{E}_t := \left\{\lambda_{\min}\left(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}\right) < \frac{mT\alpha\underline{\mu}}{8ec_{\mathsf{sb}}}\exp\left(-\frac{16kn}{mT\alpha}t\right)\right\}.$$

The event $\{\lambda_{\min}(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}) = 0\}$ is the intersection $\bigcap_{t=1}^{\infty}\mathcal{E}_t$. By Lemma B.13, we have that $\mathbb{P}(\mathcal{E}_t) \leqslant 2e^{-t}$. Since the events $\mathcal{E}_{t'} \subseteq \mathcal{E}_t$ whenever $t' \geqslant t$, by continuity of measure from above,

$$\mathbb{P}(\lambda_{\min}(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}) = 0) = \mathbb{P}\left(\bigcap_{t=1}^{\infty}\mathcal{E}_t\right) = \lim_{t\to\infty}\mathbb{P}(\mathcal{E}_t) \leqslant \lim_{t\to\infty} 2e^{-t} = 0.$$

□

We are now ready to restate and prove Lemma 5.1.

Lemma 5.1 (General OLS upper bound)**.** *There are universal positive constants c_0 and c_1 such that the following holds. Suppose that P_x satisfies the $(T, k, \{\Psi_j\}_{j=1}^{\lfloor T/k\rfloor}, c_{\mathsf{sb}}, \alpha)$-TrajSB condition (Definition 4.1). Put $S := \lfloor T/k\rfloor$ and $\Gamma_T := \Gamma_T(\mathsf{P}_x)$. Fix any $\underline{\Gamma} \in \mathsf{Sym}^n_{>0}$ satisfying $\frac{1}{S}\sum_{j=1}^{S}\Psi_j \preccurlyeq \underline{\Gamma} \preccurlyeq \Gamma_T$, and let $\underline{\mu}(\{\Psi_j\}_{j=1}^{S}, \underline{\Gamma})$ denote the geometric mean of the minimum eigenvalues $\{\underline{\lambda}(\Psi_j, \underline{\Gamma})\}_{j=1}^{S}$, i.e.,*

$$\underline{\mu}(\{\Psi_j\}_{j=1}^{S}, \underline{\Gamma}) := \left[\prod_{j=1}^{S}\underline{\lambda}(\Psi_j, \underline{\Gamma})\right]^{1/S}. \tag{5.2}$$

Suppose that:

$$n \geqslant 2, \quad \frac{mT}{kn} \geqslant \frac{c_0}{\alpha}\log\left(\frac{\max\{e, c_{\mathsf{sb}}\}}{\alpha\underline{\lambda}(\underline{\Gamma}, \Gamma_T)\underline{\mu}(\{\Psi_j\}_{j=1}^{S}, \underline{\Gamma})}\right). \tag{5.3}$$

Then, for any $\Gamma' \in \mathsf{Sym}^n_{>0}$:

$$\mathbb{E}[\|\hat{W}_{m,T} - W_\star\|_{\Gamma'}^2] \leqslant c_1 c_{\mathsf{sb}}\sigma_\xi^2 \cdot \frac{pn}{mT\alpha\underline{\lambda}(\underline{\Gamma}, \Gamma')\underline{\mu}(\{\Psi_j\}_{j=1}^{S}, \underline{\Gamma})} \cdot \log\left(\frac{\max\{e, c_{\mathsf{sb}}\}}{\alpha\underline{\lambda}(\underline{\Gamma}, \Gamma_T)\underline{\mu}(\{\Psi_j\}_{j=1}^{S}, \underline{\Gamma})}\right). \tag{5.4}$$

Proof. For notational brevity, let:

$$\underline{\lambda} := \underline{\lambda}(\underline{\Gamma}, \Gamma_T), \quad \underline{\mu} := \underline{\mu}(\{\Psi_j\}_{j=1}^{S}, \underline{\Gamma}).$$

We choose $c_0 \geqslant 64$ such that (5.3) implies (B.3), and also so that $\frac{mT}{kn} \geqslant 64/\alpha$. By Corollary B.14, $X_{m,T}$ has full column rank almost surely, hence:

$$\hat{W}_{m,T} - W_\star = \Xi_{m,T}^{\mathsf{T}}X_{m,T}(X_{m,T}^{\mathsf{T}}X_{m,T})^{-1}.$$

Put $\tilde{X}_{m,T} := X_{m,T}\underline{\Gamma}^{-1/2}$. With this decomposition, we have:

$$\|\hat{W}_{m,T} - W_\star\|_{\Gamma'}^2 = \|\Xi_{m,T}^{\mathsf{T}}X_{m,T}(X_{m,T}^{\mathsf{T}}X_{m,T})^{-1}\|_{\Gamma'}^2$$

$$
\begin{aligned}
&= \|\Xi_{m,T}^{\mathsf{T}} X_{m,T}(X_{m,T}^{\mathsf{T}} X_{m,T})^{-1}\underline{\Gamma}^{1/2}\|^2_{\underline{\Gamma}^{-1/2}\Gamma'\underline{\Gamma}^{-1/2}} \\
&\leqslant \lambda_{\max}(\underline{\Gamma}^{-1/2}\Gamma'\underline{\Gamma}^{-1/2})\|\Xi_{m,T}^{\mathsf{T}} X_{m,T}(X_{m,T}^{\mathsf{T}} X_{m,T})^{-1}\underline{\Gamma}^{1/2}\|_F^2 \\
&= \lambda_{\max}(\underline{\Gamma}^{-1/2}\Gamma'\underline{\Gamma}^{-1/2})\|(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T})^{-1}\tilde{X}_{m,T}^{\mathsf{T}}\Xi_{m,T}\|_F^2 \\
&\leqslant \min\{n,p\}\lambda_{\max}(\underline{\Gamma}^{-1/2}\Gamma'\underline{\Gamma}^{-1/2})\|(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T})^{-1}\tilde{X}_{m,T}^{\mathsf{T}}\Xi_{m,T}\|_{\mathrm{op}}^2 \\
&\leqslant \min\{n,p\}\lambda_{\max}(\underline{\Gamma}^{-1/2}\Gamma'\underline{\Gamma}^{-1/2})\frac{\|(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T})^{-1/2}\tilde{X}_{m,T}^{\mathsf{T}}\Xi_{m,T}\|_{\mathrm{op}}^2}{\lambda_{\min}(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T})} \\
&= \min\{n,p\}\frac{\|(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T})^{-1/2}\tilde{X}_{m,T}^{\mathsf{T}}\Xi_{m,T}\|_{\mathrm{op}}^2}{\underline{\lambda}(\underline{\Gamma},\Gamma')\cdot\lambda_{\min}(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T})}.
\end{aligned}
$$

Fix any $t > 0$. By Lemma B.13, there exists an event $\mathcal{E}_{t,1}$ with probability at least $1 - 2e^{-t}$, such that on $\mathcal{E}_{t,1}$ we have:

$$
\operatorname{tr}(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}) \leqslant \frac{mnTe^t}{\underline{\lambda}}, \quad \lambda_{\min}(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}) \geqslant \frac{mT\alpha\underline{\mu}}{8ec_{\mathsf{sb}}}\exp\left(-\frac{16kn}{mT\alpha}t\right).
$$

Recall by our cohice of c_0, we have $mT/k \geqslant 64n/\alpha$. Hence, on $\mathcal{E}_{t,1}$,

$$
\lambda_{\min}(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}) \geqslant \zeta_t := \frac{mT\alpha\underline{\mu}}{8ec_{\mathsf{sb}}}\exp(-t/4).
$$

We now apply Proposition B.10 with $V \leftarrow M_t := \zeta_t I_n$ and:

$$
\begin{aligned}
&x_1,\dots,x_T,x_{T+1},\dots,x_{2T},\dots,x_{(m-1)T+1},\dots,x_{mT} \leftarrow \\
&\qquad \underline{\Gamma}^{-1/2}x_1^{(1)},\dots,\underline{\Gamma}^{-1/2}x_T^{(1)},\underline{\Gamma}^{-1/2}x_1^{(2)},\dots,\underline{\Gamma}^{-1/2}x_T^{(2)},\dots,\underline{\Gamma}^{-1/2}x_1^{(m)},\dots,\underline{\Gamma}^{-1/2}x_T^{(m)},
\end{aligned}
$$

to conclude that there exists an event $\mathcal{E}_{t,2}$ with probability at least $1 - e^{-t}$ such that on $\mathcal{E}_{t,2}$:

$$
\begin{aligned}
&\mathbf{1}\{\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T} \succcurlyeq M_t\}\|(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T})^{-1/2}\tilde{X}_{m,T}^{\mathsf{T}}\Xi_{m,T}\|_{\mathrm{op}}^2 \\
&\leqslant 16\sigma_\xi^2\left[p\log 5 + \frac{1}{2}\log\det\left(I_n + \zeta_t^{-1}\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}\right) + t\right] \\
&\leqslant 32\sigma_\xi^2\left[p + \log\det\left(I_n + \zeta_t^{-1}\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T}\right) + t\right] \\
&\leqslant 32\sigma_\xi^2\left[p + n\log(1 + \zeta_t^{-1}\operatorname{tr}(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T})/n) + t\right].
\end{aligned}
$$

Above, the last inequality holds since $\log\det(X) \leqslant n\log(\operatorname{tr}(X)/n)$ for any $X \in \mathsf{Sym}^n_{\geqslant 0}$ by the AM-GM inequality. By Proposition B.1, whenever $t \geqslant 8\log 16$, we have $t \leqslant e^{t/4}$. Furthermore, for any $t \geqslant 0$ we have $1 \leqslant e^{t/4}$. Therefore, for $t \geqslant 8\log 16$, on $\mathcal{E}_{t,1}\cap\mathcal{E}_{t,2}$:

$$
\begin{aligned}
&\frac{\|(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T})^{-1/2}\tilde{X}_{m,T}^{\mathsf{T}}\Xi_{m,T}\|_{\mathrm{op}}^2}{\lambda_{\min}(\tilde{X}_{m,T}^{\mathsf{T}}\tilde{X}_{m,T})} \\
&\leqslant \frac{256ec_{\mathsf{sb}}}{mT\alpha}e^{t/4}\sigma_\xi^2\left[p + n\log\left(1 + \frac{8ec_{\mathsf{sb}}}{\alpha\underline{\lambda}\underline{\mu}}e^{(1+1/4)t}\right) + t\right] \\
&\leqslant \frac{256ec_{\mathsf{sb}}}{mT\alpha}e^{t/4}\sigma_\xi^2\left[p + n\log\left(\frac{16ec_{\mathsf{sb}}}{\alpha\underline{\lambda}\underline{\mu}}\right) + n(1+1/4)t + t\right]
\end{aligned}
$$

$$\leqslant \frac{256ec_{\mathsf{sb}}}{mT\alpha} e^{t/4}\sigma_\xi^2 \left[p + n\log\left(\frac{16ec_{\mathsf{sb}}}{\alpha\underline{\lambda}\underline{\mu}}\right) + 3nt\right]$$
$$\leqslant \frac{256ec_{\mathsf{sb}}}{mT\alpha} \sigma_\xi^2 \left[p + n\log\left(\frac{16ec_{\mathsf{sb}}}{\alpha\underline{\lambda}\underline{\mu}}\right) + 3n\right] e^{t/2}.$$

Define the random variable Z as:

$$Z := \frac{\|(\tilde{X}_{m,T}^\mathsf{T}\tilde{X}_{m,T})^{-1/2}\tilde{X}_{m,T}^\mathsf{T}\Xi_{m,T}\|_{\mathrm{op}}^2}{\lambda_{\min}(\tilde{X}_{m,T}^\mathsf{T}\tilde{X}_{m,T})}\left(\frac{256ec_{\mathsf{sb}}}{mT\alpha}\sigma_\xi^2\left[p + n\log\left(\frac{16ec_{\mathsf{sb}}}{\alpha\underline{\lambda}\underline{\mu}}\right) + 3n\right]\right)^{-1}.$$

We have shown that:

$$\mathbb{P}(Z > e^{t/2}) \leqslant 3e^{-t} \;\; \forall t \geqslant 8\log 16 \iff \mathbb{P}(Z > s) \leqslant 3s^{-2} \;\; \forall s \geqslant 16^4.$$

Hence,

$$\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z > s)\,\mathrm{d}s \leqslant 16^4 + 3\int_{16^4}^\infty s^{-2}\,\mathrm{d}s = 16^4 + 3/16^4.$$

That is, for some universal positive c_1,

$$\mathbb{E}[\|\hat{W}_{m,T} - W_\star\|_{\Gamma'}^2] \leqslant c_1\sigma_\xi^2 \min\{n,p\} c_{\mathsf{sb}}\left[\frac{p + n\log\left(\frac{\max\{e,c_{\mathsf{sb}}\}}{\alpha\underline{\lambda}\underline{\mu}}\right)}{mT\alpha\underline{\lambda}(\underline{\Gamma},\Gamma')\underline{\mu}}\right]. \tag{B.10}$$

Now, if $p \leqslant n$, (B.10) is upper bounded by:

$$c_1\sigma_\xi^2 p c_{\mathsf{sb}}\left[\frac{p + n\log\left(\frac{\max\{e,c_{\mathsf{sb}}\}}{\alpha\underline{\lambda}\underline{\mu}}\right)}{mT\alpha\underline{\lambda}(\underline{\Gamma},\Gamma')\underline{\mu}}\right] \leqslant 2c_1\sigma_\xi^2 p c_{\mathsf{sb}}\left[\frac{n\log\left(\frac{\max\{e,c_{\mathsf{sb}}\}}{\alpha\underline{\lambda}\underline{\mu}}\right)}{mT\alpha\underline{\lambda}(\underline{\Gamma},\Gamma')\underline{\mu}}\right].$$

On the other hand, if $p > n$, (B.10) is upper bounded by:

$$c_1\sigma_\xi^2 n c_{\mathsf{sb}}\left[\frac{p + n\log\left(\frac{\max\{e,c_{\mathsf{sb}}\}}{\alpha\underline{\lambda}\underline{\mu}}\right)}{mT\alpha\underline{\lambda}(\underline{\Gamma},\Gamma')\underline{\mu}}\right] < 2c_1\sigma_\xi^2 n c_{\mathsf{sb}}\left[\frac{p\log\left(\frac{\max\{e,c_{\mathsf{sb}}\}}{\alpha\underline{\lambda}\underline{\mu}}\right)}{mT\alpha\underline{\lambda}(\underline{\Gamma},\Gamma')\underline{\mu}}\right].$$

□

B.5 Proof of Theorem 5.3

Theorem 5.3 (Upper bound for Ind-Seq-LS)**.** *There are universal positive constants c_0 and c_1 such that the following holds. Fix any sequence of distributions $\{\mathsf{P}_{x,t}\}_{t\geqslant 1}$, and let $\Sigma_t := \mathbb{E}_{x_t\sim\mathsf{P}_{x,t}}[x_t x_t^\mathsf{T}]$ for $t \in \mathbb{N}_+$. Suppose there exists $c_{\mathsf{sb}} > 0$ and $\alpha \in (0,1]$ such that for all $v \in \mathbb{R}^n \setminus \{0\}$, $\varepsilon > 0$ and $t \in \mathbb{N}_+$:*

$$\mathbb{P}_{x_t\sim\mathsf{P}_{x,t}}\left\{\langle v, x_t\rangle^2 \leqslant \varepsilon\cdot v^\mathsf{T}\Sigma_t v\right\} \leqslant (c_{\mathsf{sb}}\varepsilon)^\alpha. \tag{5.7}$$

Furthermore, suppose there exists a $c_\beta \geqslant 1$ and $\beta \geqslant 0$ such that for all $s, t \in \mathbb{N}_+$ satisfying $s \leqslant t$:

$$\frac{1}{\underline{\lambda}(\Sigma_s, \Sigma_t)} \leqslant c_\beta (t/s)^\beta. \tag{5.8}$$

If:

$$n \geqslant 2, \quad mT \geqslant \frac{c_0 n}{\alpha}\left(\beta + \log\left(\frac{\max\{e, c_{\mathsf{sb}}\} c_\beta}{\alpha}\right)\right),$$

then, for $\mathsf{P}_x = \otimes_{t \geqslant 1} \mathsf{P}_{x,t}$:

$$\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x)] \leqslant c_1 c_{\mathsf{sb}} \sigma_\xi^2 c_\beta e^\beta \cdot \frac{pn}{mT\alpha} \cdot \phi\left(c_\beta(\beta+1), (T'/T)^\beta\right)\left[\beta + \log\left(\frac{\max\{e, c_{\mathsf{sb}}\} c_\beta}{\alpha}\right)\right]. \tag{5.9}$$

Proof. Equation (5.7) shows that P_x satisfies the $(T, 1, \{\Sigma_t\}_{t=1}^T, c_{\mathsf{sb}}, \alpha)$-TrajSB condition. Let $\Gamma_t := \frac{1}{t}\sum_{k=1}^t \Sigma_k$ for $t \in \mathbb{N}_+$. For any $s, t \in \mathbb{N}_+$ with $s \leqslant t$,

$$\begin{aligned}
\underline{\lambda}(\Gamma_s, \Gamma_t) &\geqslant \underline{\lambda}(\Gamma_s, \Sigma_t) && \text{since } \Gamma_t \preccurlyeq \Sigma_t \\
&\geqslant \frac{1}{s}\sum_{k=1}^s \underline{\lambda}(\Sigma_k, \Sigma_t) && \text{using Proposition B.2 and Jensen's inequality} \\
&\geqslant \frac{1}{c_\beta s}\sum_{k=1}^s (k/t)^\beta && \text{using (5.8)} \\
&\geqslant \frac{1}{c_\beta(\beta+1)}(s/t)^\beta && \text{since } x \mapsto x^\beta \text{ is increasing.}
\end{aligned}$$

Next, the growth condition (5.8) implies that:

$$\begin{aligned}
\underline{\mu}(\{\Sigma_t\}_{t=1}^T, \Gamma_T) &= \left[\prod_{t=1}^T \underline{\lambda}(\Sigma_t, \Gamma_T)\right]^{1/T} \\
&\geqslant \left[\prod_{t=1}^T \underline{\lambda}(\Sigma_t, \Sigma_T)\right]^{1/T} && \text{since } \Gamma_T \preccurlyeq \Sigma_T \\
&\geqslant \left[\prod_{t=1}^T \frac{1}{c_\beta}(t/T)^\beta\right]^{1/T} && \text{using (5.8)} \\
&= \frac{1}{c_\beta T^\beta}(T!)^{\beta/T} \\
&\geqslant \frac{1}{c_\beta e^\beta} && \text{since } T! \geqslant (T/e)^T.
\end{aligned}$$

We now apply Lemma 5.1 with $\underline{\Gamma} = \Gamma_T$. In doing so, the requirement (5.3) simplifies to:

$$n \geqslant 2, \quad \frac{mT}{n} \geqslant \frac{c_0}{\alpha}\log\left(\frac{\max\{e, c_{\mathsf{sb}}\} c_\beta e^\beta}{\alpha}\right).$$

We first assume that $T' \leqslant T$, in which case (5.4) yields:

$$\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x)] \leqslant c_1 c_{\mathsf{sb}} \sigma_\xi^2 \cdot \frac{pn}{mT\alpha} \cdot c_\beta e^\beta \cdot \log\left(\frac{\max\{e, c_{\mathsf{sb}}\} c_\beta e^\beta}{\alpha}\right).$$

On the other hand, when $T' > T$, we have $\underline{\lambda}(\Gamma_T, \Gamma_{T'}) \geqslant \frac{1}{c_b(\beta+1)}(T/T')^\beta$, and (5.4) yields:

$$\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x)] \leqslant c_1 c_{\mathsf{sb}} \sigma_\xi^2 \cdot \frac{pn}{mT\alpha} \cdot c_\beta e^\beta \cdot c_\beta(\beta+1) \left(\frac{T'}{T}\right)^\beta \cdot \log\left(\frac{\max\{e, c_{\mathsf{sb}}\} c_\beta e^\beta}{\alpha}\right).$$

□

B.6 Proofs for linear dynamical systems

B.6.1 Control of ratios of covariance matrices

Proposition B.15. *Let (A, B) be the dynamics matrices for an* LDS-LS *instance, and suppose (A, B) satisfy Assumption 5.1, Assumption 5.2, and Assumption 5.3. Put $\Sigma_t := \Sigma_t(A, B)$ for $t \in \mathbb{N}_+$ and $\gamma := \gamma(A, B)$. For any integers T_1, T_2 satisfying $1 \leqslant T_2 \leqslant T_1$,*

$$\lambda_{\min}(\Sigma_{T_1}^{-1/2} \Sigma_{T_2} \Sigma_{T_1}^{-1/2}) \geqslant \frac{1}{\gamma} \frac{T_2}{T_1}.$$

Proof. Observe that for any $t \geqslant 1$,

$$\Sigma_t = \sum_{k=0}^{t-1} A^k B B^* (A^k)^* = \sum_{k=0}^{t-1} S D^k S^{-1} B B^* S^{-*} (D^k)^* S^*.$$

By Assumption 5.3, we have that BB^* is invertible, and hence $S^{-1}BB^*S^{-*}$ is also invertible. Therefore we have the following lower and upper bound on Σ_t:

$$\lambda_{\min}(S^{-1}BB^*S^{-*}) \cdot S\left(\sum_{k=0}^{t-1} D^k (D^k)^*\right) S^* \preccurlyeq \Sigma_t \preccurlyeq \lambda_{\max}(S^{-1}BB^*S^{-*}) \cdot S\left(\sum_{k=0}^{t-1} D^k (D^k)^*\right) S^*. \tag{B.11}$$

Now recall that for two square matrices X, Y, the eigenvalues of XY coincide with the eigenvalues of YX. Letting $Q_t := \sum_{k=0}^{t-1} D^k (D^k)^*$, we have:

$$\begin{aligned}
\lambda_{\min}(\Sigma_{T_1}^{-1/2} \Sigma_{T_2} \Sigma_{T_1}^{-1/2}) &\geqslant \lambda_{\min}(S^{-1}BB^*S^{-*}) \lambda_{\min}(\Sigma_{T_1}^{-1/2} S Q_{T_2} S^* \Sigma_{T_1}^{-1/2}) \\
&= \lambda_{\min}(S^{-1}BB^*S^{-*}) \lambda_{\min}((S Q_{T_2} S^*)^{1/2} \Sigma_{T_1}^{-1} (S Q_{T_2} S^*)^{1/2}) \\
&\geqslant \frac{\lambda_{\min}(S^{-1}BB^*S^{-*})}{\lambda_{\max}(S^{-1}BB^*S^{-*})} \lambda_{\min}((S Q_{T_2} S^*)^{1/2} (S^{-*} Q_{T_1}^{-1} S^*)(S Q_{T_2} S^*)^{1/2}) \\
&= \frac{\lambda_{\min}(S^{-1}BB^*S^{-*})}{\lambda_{\max}(S^{-1}BB^*S^{-*})} \lambda_{\min}(Q_{T_2} Q_{T_1}^{-1}).
\end{aligned}$$

Let $\lambda \in \mathbb{C}$ be an eigenvalue of A. We have

$$\sum_{k=0}^{t-1} |\lambda|^{2k} = \begin{cases} \frac{1-|\lambda|^{2t}}{1-|\lambda|^2} & \text{if } |\lambda| < 1, \\ t & \text{if } |\lambda| = 1. \end{cases}$$

Therefore, $(Q_{T_2}Q_{T_1}^{-1})_{ii}$ is:

$$(Q_{T_2}Q_{T_1}^{-1})_{ii} = \begin{cases} \frac{1-|\lambda_i|^{2T_2}}{1-|\lambda_i|^{2T_1}} = \frac{1-(|\lambda_i|^{2T_1})^{T_2/T_1}}{1-|\lambda_i|^{2T_1}} & \text{if } |\lambda_i| < 1, \\ T_2/T_1 & \text{if } |\lambda_i| = 1. \end{cases}$$

Note that $\inf_{x\in(0,1)} \frac{1-x^c}{1-x} = c$ for $c \in [0,1]$. Therefore, we can lower bound:

$$\lambda_{\min}(Q_{T_2}Q_{T_1}^{-1}) \geqslant \frac{T_2}{T_1}.$$

The claim now follows. □

Proposition B.16. *Let (A,B) be the dynamics matrices for an* LDS-LS *instance, and suppose (A,B) satisfy Assumption 5.1, Assumption 5.2, and Assumption 5.3. Put $\Gamma_t := \Gamma_t(A,B)$ for $t \in \mathbb{N}_+$ and $\gamma := \gamma(A,B)$. For any integers $k,t \in \mathbb{N}_+$ satisfying $k \leqslant t$, we have:*

$$\underline{\lambda}(\Gamma_k, \Gamma_t) \geqslant \frac{1}{8\gamma}\frac{k}{t}.$$

Proof. Let $\Sigma_t := \Sigma_t(A,B)$ for $t \in \mathbb{N}_+$. We first consider the case when $k \geqslant 2$. Observe that $\Gamma_t \preccurlyeq \Sigma_t$. Furthermore, for any $k \geqslant 2$, we have:

$$\Gamma_k = \frac{1}{k}\sum_{k'=1}^{k} \Sigma_{k'} \succcurlyeq \frac{1}{k}\sum_{k'=\lfloor k/2 \rfloor}^{k} \Sigma_{\lfloor k/2 \rfloor} = \frac{k - \lfloor k/2 \rfloor + 1}{k}\Sigma_{\lfloor k/2 \rfloor} \succcurlyeq \frac{1}{2}\Sigma_{\lfloor k/2 \rfloor}.$$

Therefore,

$$\underline{\lambda}(\Gamma_k, \Gamma_t) = \lambda_{\min}(\Gamma_t^{-1/2}\Gamma_k\Gamma_t^{-1/2}) \geqslant \frac{1}{2}\lambda_{\min}(\Sigma_t^{-1/2}\Sigma_{\lfloor k/2 \rfloor}\Sigma_t^{-1/2}) \overset{(a)}{\geqslant} \frac{1}{2\gamma}\frac{\lfloor k/2 \rfloor}{t} \geqslant \frac{1}{8\gamma}\frac{k}{t}.$$

Above, (a) follows from Proposition B.15. When $k = 1$, we have $\Gamma_1 = \Sigma_1$, and therefore by Proposition B.15:

$$\underline{\lambda}(\Gamma_1, \Gamma_t) = \underline{\lambda}(\Sigma_1, \Gamma_t) \geqslant \underline{\lambda}(\Sigma_1, \Sigma_t) = \lambda_{\min}(\Sigma_t^{-1/2}\Sigma_1\Sigma_t^{-1/2}) \geqslant \frac{1}{\gamma}\frac{1}{t}.$$

The claim now follows. □

Fact B.17. *Let (A,B) be the dynamics matrices for an* LDS-LS *instance. For any $s,t \in \mathbb{N}_+$ with $s \leqslant t$:*

$$\Gamma_s(A,B) \preccurlyeq \Gamma_t(A,B).$$

Proposition B.18. *Let (A,B) be the dynamics matrices for an* LDS-LS *instance, and suppose (A,B) satisfy Assumption 5.1, Assumption 5.2, and Assumption 5.3. Put $\Gamma_t := \Gamma_t(A,B)$ for $t \in \mathbb{N}_+$, $\Sigma_t := \Sigma_t(A,B)$ for $t \in \mathbb{N}_+$, and $\gamma := \gamma(A,B)$. For any T, we have:*

$$\left[\prod_{t=1}^{T} \underline{\lambda}(\Sigma_t, \Gamma_T)\right]^{1/T} \geqslant \frac{1}{8e\gamma}.$$

Proof. By Proposition B.16, we have that $\underline{\lambda}(\Gamma_t, \Gamma_T) \geqslant \frac{1}{8\gamma}\frac{t}{T}$ for all $t \in \{1, \dots, T\}$. Therefore, since $\underline{\lambda}(\Sigma_t, \Gamma_T) \geqslant \underline{\lambda}(\Gamma_t, \Gamma_T)$, and since $n! \geqslant (n/e)^n$ for all $n \in \mathbb{N}_+$,

$$\left[\prod_{t=1}^{T} \underline{\lambda}(\Sigma_t, \Gamma_T)\right]^{1/T} \geqslant \frac{(T!)^{1/T}}{8\gamma T} \geqslant \frac{1}{8e\gamma}.$$

□

B.6.2 Many trajectory results

Lemma B.19. *There are universal positive constants c_0 and c_1 such that the following holds for any instance of* LDS-LS*. Suppose that (A, B) is k_{c}-step controllable. If $n \geqslant 2$ and $m \geqslant c_0 n$, then for any $\Gamma' \in \mathsf{Sym}^n_{>0}$:*

$$\mathbb{E}[\|\hat{W}_{m,T} - W_\star\|^2_{\Gamma'}] \leqslant c_1 \sigma_\xi^2 \cdot \frac{pn}{mT \cdot \underline{\lambda}(\Gamma_T(A, B), \Gamma')}. \tag{B.12}$$

Proof. Let $\Gamma_T := \Gamma_T(A, B)$. By Example 4.6, LDS-LS satisfies the $(T, T, \Gamma_T, e, 1/2)$-TrajSB condition. We therefore invoke Lemma 5.1 with $k = T$ and $\underline{\Gamma} = \Gamma_T$. In this case, $\underline{\mu}$ from (5.2) simplifies to $\underline{\mu} = \underline{\lambda}(\Gamma_T, \Gamma_T) = 1$, and the requirement (5.3) simplifies to $n \geqslant 2$ and $m \geqslant c_0 n$. Finally, the rate (5.4) simplifies to (B.12). □

Theorem 5.4 (Parameter recovery upper bound for LDS-LS, many trajectories)**.** *There are universal positive constants c_0 and c_1 such that the following holds for any instance of* LDS-LS*. Suppose that (A, B) is k_{c}-step controllable, If $n \geqslant 2$, $m \geqslant c_0 n$, and $T \geqslant k_{\mathsf{c}}$, then:*

$$\mathbb{E}[\|\hat{W}_{m,T} - W_\star\|^2_F] \leqslant c_1 \sigma_\xi^2 \cdot \frac{pn}{mT \cdot \lambda_{\min}(\Gamma_T(A, B))}.$$

Proof. Follows by invoking Lemma B.19 with $\Gamma' = I_n$. □

Theorem 5.5 (Risk upper bound for LDS-LS, many trajectories)**.** *There are universal positive constants c_0 and c_1 such that the following holds for any instance of* LDS-LS*. Suppose that (A, B) is k_{c}-step controllable. If $n \geqslant 2$, $m \geqslant c_0 n$, $T \geqslant k_{\mathsf{c}}$, and the evaluation horizon is strict ($T' \leqslant T$), then:*

$$\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x^{A,B})] \leqslant c_1 \sigma_\xi^2 \cdot \frac{pn}{mT}.$$

On the other hand, suppose that (A, B) satisfies Assumption 5.1, Assumption 5.2, and Assumption 5.3, with $\gamma := \gamma(A, B)$ (Definition 5.1). If $n \geqslant 2$, $m \geqslant c_0 n$, and the evaluation horizon is extended ($T' > T$), then:

$$\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x^{A,B})] \leqslant c_1 \sigma_\xi^2 \cdot \frac{pn}{mT} \cdot \gamma \frac{T'}{T}.$$

Proof. Let $\Gamma_t := \Gamma_t(A, B)$ for $t \in \mathbb{N}_+$. Invoking Lemma B.19 with $\Gamma' = \Gamma_{T'}$ yields the bound:

$$\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x^{A,B})] \leqslant c_1 \sigma_\xi^2 \cdot \frac{pn}{mT \cdot \underline{\lambda}(\Gamma_T, \Gamma_{T'})}.$$

If $T' \leqslant T$, then $\underline{\lambda}(\Gamma_T, \Gamma_{T'}) \geqslant 1$ since $\Gamma_T \succcurlyeq \Gamma_{T'}$ by Fact B.17. On the other hand, if $T' > T$, by Proposition B.16, $\underline{\lambda}(\Gamma_T, \Gamma_{T'}) \geqslant \frac{1}{8\gamma}\frac{T}{T'}$. The claim now follows. □

B.6.3 Few trajectory results

Theorem 5.6 (Risk upper bound for LDS-LS, few trajectories)**.** *There are universal positive constants c_0, c_1, and c_2 such that the following holds for any instance of* LDS-LS*. Suppose that (A, B) satisfies Assumption 5.1, Assumption 5.2, and Assumption 5.3, with $\gamma := \gamma(A, B)$ (Definition 5.1). If $n \geqslant 2$, $m \leqslant c_0 n$, and $mT \geqslant c_1 n \log(\max\{\gamma n/m, e\})$, then:*

$$\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x^{A,B})] \leqslant c_2 \sigma_\xi^2 \cdot \frac{pn \log(\max\{\gamma n/m, e\})}{mT} \cdot \phi\left(\gamma, \frac{c_1 n \log(\max\{\gamma n/m, e\})}{m} \cdot \frac{T'}{T}\right).$$

Proof. Let $\Gamma_t := \Gamma_t(A, B)$ for all $t \in \mathbb{N}_+$. By Example 4.6, for any $k \in \{1, \dots, T\}$, LDS-LS satisfies the $(T, k, \Gamma_k, e, 1/2)$-TrajSB condition. We will apply Lemma 5.1 with $\underline{\Gamma} = \Gamma_k$. The quantity $\underline{\mu}$ from (5.2) simplifies to $\underline{\mu} = \underline{\lambda}(\Gamma_k, \Gamma_k) = 1$. By Proposition B.16, we have that $\underline{\lambda}(\Gamma_k, \Gamma_T) \geqslant \frac{1}{8\gamma}\frac{k}{T}$. Hence the requirement (5.3) simplifies to $n \geqslant 2$ and

$$\frac{mT}{kn} \geqslant c \log\left(\gamma' \frac{T}{k}\right), \quad \gamma' := \max\{e, \gamma\} \tag{B.13}$$

for some universal positive constant c. Thus, for (B.13) to hold, it suffices to require:

$$\frac{T}{k} \geqslant \max\left\{\frac{2cn}{m} \log \gamma', \frac{2cn}{m} \log\left(\frac{T}{k}\right)\right\}. \tag{B.14}$$

As long as $2cn/m \geqslant 1$, then by Proposition B.1,

$$\frac{T}{k} \geqslant \frac{4cn}{m} \log\left(\frac{8cn}{m}\right) \Longrightarrow \frac{T}{k} \geqslant \frac{2cn}{m} \log\left(\frac{T}{k}\right).$$

Hence, for (B.14) to hold, it suffices to require

$$\frac{T}{k} \geqslant \frac{4cn}{m} \log\left(\frac{8c\gamma' n}{m}\right). \tag{B.15}$$

Based on (B.15), we choose k as:

$$k = \left\lfloor \frac{T}{4cn/m \cdot \log(8c\gamma' n/m)} \right\rfloor. \tag{B.16}$$

To ensure that $k \geqslant 1$, we need to ensure that:

$$mT \geqslant 4cn \log(8c\gamma' n/m). \tag{B.17}$$

On the other hand, since $2cn/m \geqslant 1$, we have that:

$$\frac{4cn}{m} \log\left(\frac{8c\gamma' n}{m}\right) \geqslant 1,$$

which ensures that $k \leqslant T$. Thus, our choice of k from (B.16) ensures that (B.13) holds. We now ready to invoke Lemma 5.1 with $\Gamma' = \Gamma_{T'}$, and conclude for a universal c':

$$\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x^{A,B})] \leqslant c' \sigma_\xi^2 \cdot \frac{pn \log(e/\underline{\lambda}(\Gamma_k, \Gamma_T))}{mT \underline{\lambda}(\Gamma_k, \Gamma_{T'})}. \tag{B.18}$$

First, we assume that $T' \leqslant k$. By Fact B.17 we have $\Gamma_k \succcurlyeq \Gamma_{T'}$, and therefore $\underline{\lambda}(\Gamma_k, \Gamma_{T'}) \geqslant 1$. Equation (B.18) yields:

$$\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x^{A,B})] \leqslant c'\sigma_\xi^2 \cdot \frac{pn \log(e/\underline{\lambda}(\Gamma_k, \Gamma_T))}{mT} \tag{B.19}$$

By Proposition B.16,

$$\underline{\lambda}(\Gamma_k, \Gamma_T) \geqslant \frac{1}{8\gamma}\frac{1}{T}\left\lfloor \frac{T}{4cn/m \cdot \log(8c\gamma' n/m)} \right\rfloor \geqslant \frac{m}{64c\gamma n \log(8c\gamma' n/m)}. \tag{B.20}$$

Plugging (B.20) into (B.19), and using the inequalities $\log x \leqslant x$ for $x > 0$ and $\phi(a, x) \geqslant 1$ for all $a \geqslant 1$ yields, for another universal c'':

$$\begin{aligned}
\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x^{A,B})] &\leqslant c'\sigma_\xi^2 \cdot \frac{pn}{mT} \cdot \log(e \cdot 64c\gamma n/m \cdot \log(8c\gamma' n/m)) \\
&\leqslant c'\sigma_\xi^2 \cdot \frac{pn \log(512e \cdot (c\gamma' n/m)^2)}{mT} \\
&\leqslant c''\sigma_\xi^2 \cdot \frac{pn \log(\max\{\gamma n/m, e\})}{mT} \\
&\leqslant c''\sigma_\xi^2 \cdot \frac{pn \log(\max\{\gamma n/m, e\})}{mT} \cdot \phi\left(\gamma, c_1 \frac{n \log(\max\{\gamma n/m, e\})}{m}\frac{T'}{T}\right).
\end{aligned}$$

On the other hand, if $T' > k$, then by Proposition B.16,

$$\underline{\lambda}(\Gamma_k, \Gamma_{T'}) \geqslant \frac{1}{8\gamma}\frac{1}{T'}\left\lfloor \frac{T}{4cn/m \cdot \log(8c\gamma' n/m)} \right\rfloor \geqslant \frac{m}{64c\gamma n \log(8c\gamma' n/m)} \cdot \frac{T}{T'}. \tag{B.21}$$

Plugging (B.20) and (B.21) into (B.18) and using again the inequality $\log x \leqslant x$ for $x > 0$ yields, for a universal c''':

$$\begin{aligned}
&\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x^{A,B})] \\
&\leqslant c'\sigma_\xi^2 \cdot \frac{pn}{mT} \cdot \log(e \cdot 64c\gamma n/m \cdot \log(8c\gamma' n/m)) \cdot 64c\gamma n/m \cdot \log(8c\gamma' n/m) \cdot \frac{T'}{T} \\
&\leqslant c'''\sigma_\xi^2 \cdot \frac{pn \log(\max\{\gamma n/m, e\})}{mT} \cdot \gamma \frac{n \log(\max\{\gamma n/m, e\})}{m} \cdot \frac{T'}{T}.
\end{aligned}$$

Furthermore, when $T' > k$, by choosing c_1 sufficiently large:

$$\begin{aligned}
8c\frac{n \log(8c\gamma' n/m)}{m}\frac{T'}{T} > 1 &\Longrightarrow c_1 \frac{n \log(\max\{\gamma n/m, e\})}{m}\frac{T'}{T} > 1 \\
&\Longrightarrow \gamma c_1 \frac{n \log(\max\{\gamma n/m, e\})}{m}\frac{T'}{T} = \phi\left(\gamma, c_1 \frac{n \log(\max\{\gamma n/m, e\})}{m}\frac{T'}{T}\right).
\end{aligned}$$

The claim now follows. □

Theorem 5.7 (Risk upper bound for Ind-LDS-LS)**.** *There are universal positive constants c_0 and c_1 such that the following holds for any instance of* Ind-LDS-LS*. Suppose that (A, B) satisfies Assumption 5.1, Assumption 5.2, and Assumption 5.3, with $\gamma := \gamma(A, B)$ (Definition 5.1). If $n \geqslant 2$ and $mT \geqslant c_0 n \log(\max\{\gamma, e\})$, then:*

$$\mathbb{E}[L(\hat{W}_{m,T}; T', \otimes_{t \geqslant 1} \mathsf{P}_{x,t}^{A,B})] \leqslant c_1 \sigma_\xi^2 \cdot \frac{pn\gamma \log(\max\{\gamma, e\})}{mT} \cdot \phi\left(\gamma, \frac{T'}{T}\right).$$

Proof. Let $\Gamma_t := \Gamma_t(A, B)$ and $\Sigma_t := \Sigma_t(A, B)$ for $t \in \mathbb{N}_+$. From Example 4.3, we have that $\mathsf{Ind\text{-}LDS\text{-}LS}$ satisfies the $(T, 1, \{\Sigma_t\}_{t=1}^T, e, 1/2)$-TrajSB condition. We will apply Lemma 5.1 with $\underline{\Gamma} = \Gamma_T$, $k = 1$, and $\Gamma' = \Gamma_{T'}$. By Proposition B.18, we have that:

$$\underline{\mu}(\{\Sigma_t\}_{t=1}^T, \Gamma_T) \geqslant \frac{1}{8e\gamma}.$$

The requirement (5.3) simplifies to $n \geqslant 2$ and $\frac{mT}{n} \geqslant c \log(\max\{\gamma, e\})$ for a universal constant c. By Lemma 5.1, for a universal c':

$$\mathbb{E}[L(\hat{W}_{m,T}; T', \mathsf{P}_x^{A,B})] \leqslant c' \sigma_\xi^2 \cdot \frac{pn \log(\max\{\gamma, e\})}{mT \cdot \underline{\lambda}(\Gamma_T, \Gamma_{T'})} \cdot 8e\gamma.$$

If $T' \leqslant T$, then $\underline{\lambda}(\Gamma_T, \Gamma_{T'}) \geqslant 1$ since $\Gamma_T \succcurlyeq \Gamma_{T'}$ by Fact B.17. On the other hand, if $T' > T$, then by Proposition B.16, $\underline{\lambda}(\Gamma_T, \gamma_{T'}) \geqslant \frac{1}{8\gamma} \frac{T}{T'}$. The claim now follows. $\square$

Theorem 5.8 (Parameter recovery upper bound for $\mathsf{LDS\text{-}LS}$, few trajectories)**.** *There are universal positive constants c_0, c_1, and c_2 such that the following holds for any instance of* $\mathsf{LDS\text{-}LS}$. *Suppose that (A, B) satisfies Assumption 5.1, Assumption 5.2, and Assumption 5.3, with $\gamma := \gamma(A, B)$ (Definition 5.1). If $n \geqslant 2$, and $mT \geqslant c_0 n \log(\max\{\gamma n/m, e\})$, then:*

$$\mathbb{E}[\|\hat{W}_{m,T} - W_\star\|_F^2] \leqslant c_1 \sigma_\xi^2 \cdot \frac{pn \log(\max\{\gamma n/m, e\})}{mT \cdot \lambda_{\min}(\Gamma_{k_\star}(A, B))}, \quad k_\star := \left\lfloor \frac{c_2 T}{n/m \cdot \log(\max\{\gamma n/m, e\})} \right\rfloor.$$

Proof. The proof is identical to that of Theorem 5.6 until (B.18), after which we set $T' = 1$ from which the result follows. $\square$

B.7 High probability upper bounds

B.7.1 Weak trajectory small ball

We first present a modified definition of trajectory small-ball (cf. Definition 4.1) which we will use to establish high probability bounds.

Definition B.1 (Weak trajectory small-ball (wTrajSB))**.** *Fix a trajectory length $T \in \mathbb{N}_+$, a parameter $k \in \{1, \ldots, T\}$, positive definite matrices $\{\Psi_j\}_{j=1}^{\lfloor T/k \rfloor} \subset \mathsf{Sym}_{>0}^n$, and constants $\alpha, \beta \in (0, 1)$. The distribution P_x satisfies the $(T, k, \{\Psi_j\}_{j=1}^{\lfloor T/k \rfloor}, \alpha, \beta)$-weak-trajectory-small-ball (wTrajSB) condition if:*

1. $\frac{1}{\lfloor T/k \rfloor} \sum_{j=1}^{\lfloor T/k \rfloor} \Psi_j \preccurlyeq \Gamma_T(\mathsf{P}_x)$,
2. *$\{x_t\}_{t \geqslant 1}$ is adapted to a filtration $\{\mathcal{F}_t\}_{t \geqslant 1}$, and*
3. *for all $v \in \mathbb{R}^n \setminus \{0\}$, $j \in \{1, \ldots, \lfloor T/k \rfloor\}$:*

$$\mathbb{P}_{\{x_t\} \sim \mathsf{P}_x} \left\{ \frac{1}{k} \sum_{t=(j-1)k+1}^{jk} \langle v, x_t \rangle^2 \leqslant \alpha \cdot v^\mathsf{T} \Psi_j v \;\middle|\; \mathcal{F}_{(j-1)k} \right\} \leqslant \beta \;\; a.s. \tag{B.22}$$

The main difference between Definition B.1 vs. Definition 4.1 is the third condition (B.22), which only needs to hold for a *fixed* resolution α and failure probability β. By contrast, in Definition 4.1, the condition must hold of for *all* resolutions—there denoted by ε—with failure probabilities that tend to zero as the resolution $\varepsilon \to 0$ (cf. (4.1)).

B.7.2 Ordinary least squares bounds

Lemma B.20 (Minimum eigenvalue bound via weak trajectory small-ball)**.** *Suppose that P_x satisfies the $(T, k, \{\Psi_j\}_{j=1}^{\lfloor T/k \rfloor}, \alpha, \beta)$-wTrajSB condition (Definition B.1). Put $S := \lfloor T/k \rfloor$ and $\Gamma_t := \Gamma_t(\mathsf{P}_x)$ for $t \in \mathbb{N}_+$. Fix any $\underline{\Gamma} \in \mathsf{Sym}^n_{>0}$ satisfying $\frac{1}{S}\sum_{j=1}^S \Psi_j \preccurlyeq \underline{\Gamma} \preccurlyeq \Gamma_T$, and define the constants:*

$$C_S := \frac{\frac{1}{S}\sum_{j=1}^S \underline{\lambda}(\Psi_j, \underline{\Gamma})^2}{\left(\frac{1}{S}\sum_{j=1}^S \underline{\lambda}(\Psi_j, \underline{\Gamma})\right)^2}, \quad \bar{\mu} := \frac{1}{S}\sum_{j=1}^S \underline{\lambda}(\Psi_j, \underline{\Gamma}). \tag{B.23}$$

(Note that $1 \leqslant C_S \leqslant S$ always). Fix $\delta \in (0, 1)$, and suppose that:

$$n \geqslant 2, \quad \frac{mT}{kn} \geqslant \frac{64 C_S}{1-\beta} \log\left(\frac{1280 C_S}{\alpha(1-\beta)\underline{\lambda}(\underline{\Gamma}, \Gamma_T)\bar{\mu}\delta}\right).$$

With probability at least $1-\delta$, the following events simulatenously hold:

$$\begin{aligned} \lambda_{\min}\left(\underline{\Gamma}^{-1/2}\sum_{i=1}^m\sum_{t=1}^T x_t^{(i)}(x_t^{(i)})^\mathsf{T}\underline{\Gamma}^{-1/2}\right) &\geqslant \frac{\alpha(1-\beta)mT\bar{\mu}}{8}, \\ \operatorname{tr}\left(\underline{\Gamma}^{-1/2}\sum_{i=1}^m\sum_{t=1}^T x_t^{(i)}(x_t^{(i)})^\mathsf{T}\underline{\Gamma}^{-1/2}\right) &\leqslant \frac{2mTn}{\underline{\lambda}(\underline{\Gamma}, \Gamma_T)\cdot\delta}. \end{aligned} \tag{B.24}$$

Proof. The proof proceeds quite similarly to the proof of Lemma B.13. Thus, we focus mostly on the parts that differ. For notational brevity, let:

$$\beta' := 1-\beta, \quad \underline{\lambda} := \underline{\lambda}(\underline{\Gamma}, \Gamma_T), \quad \underline{\lambda}_j := \underline{\lambda}(\Psi_j, \underline{\Gamma}).$$

Since $\underline{\Gamma} \preccurlyeq \Gamma_T$ by assumption, we have $\underline{\lambda} \in (0, 1]$.

The first step, in preparation for applying the PAC-Bayes deviation inequality, is to construct a family of random variables with moment generating function upper bounded by one. To do this, we utilize the weak trajectory small-ball condition (B.22), which implies for any $v \in \mathbb{S}^{n-1}$ and $j \in \{1, \dots, S\}$:

$$\mathbb{P}\left\{\frac{1}{k}\sum_{t=(j-1)k+1}^{jk}\langle v, \underline{\Gamma}^{-1/2}x_t\rangle^2 \leqslant \alpha\underline{\lambda}(\Psi_j, \underline{\Gamma}) \,\middle|\, \mathcal{F}_{(j-1)k}\right\} \leqslant \beta.$$

Let $\tilde{x}_t := \underline{\Gamma}^{-1/2}x_t$ be the whitened vector. Define the random indicator variables for $i = 1, \dots, m$ and $j = 1, \dots, S$:

$$B_j^{(i)} := \mathbf{1}\left\{\frac{1}{k}\sum_{t=(j-1)k+1}^{jk}\langle v, \tilde{x}_t^{(i)}\rangle^2 \geqslant \alpha\underline{\lambda}(\Psi_j, \underline{\Gamma})\right\}.$$

By Markov's inequality:

$$\sum_{t=1}^T\langle v, \tilde{x}_t^{(i)}\rangle^2 \geqslant k\alpha\sum_{j=1}^S\underline{\lambda}_j\mathbf{1}\{B_j^{(i)} = 1\}.$$

Hence for any $\eta > 0$ and $v \in \mathbb{S}^{n-1}$:

$$\mathbb{E}\exp\left(-\eta\sum_{t=1}^{T}\langle v, \tilde{x}_t^{(i)}\rangle^2\right) \leqslant \mathbb{E}\exp\left(-\eta k\alpha\sum_{j=1}^{S}\underline{\lambda}_j \mathbf{1}\{B_j^{(i)} = 1\}\right).$$

Now observe:

$$\begin{aligned}
\mathbb{E}[\exp(-\eta k\alpha\underline{\lambda}_j\mathbf{1}\{B_j^{(i)} = 1\}) \mid \mathcal{F}_{(j-1)k}] &= e^{-\eta k\alpha\underline{\lambda}_j}\mathbb{P}(B_j^{(i)} = 1 \mid \mathcal{F}_{(j-1)k}) + \mathbb{P}(B_j^{(i)} = 0 \mid \mathcal{F}_{(j-1)k}) \\
&= (e^{-\eta k\alpha\underline{\lambda}_j} - 1)\mathbb{P}(B_j^{(i)} = 1 \mid \mathcal{F}_{(j-1)k}) + 1 \\
&\leqslant (e^{-\eta k\alpha\underline{\lambda}_j} - 1)\beta' + 1 \\
&\overset{(a)}{\leqslant} 1 + \left(-\eta k\alpha\underline{\lambda}_j + \frac{1}{2}\eta^2k^2\alpha^2\underline{\lambda}_j^2\right)\beta' \\
&\overset{(b)}{\leqslant} \exp\left(\left(-\eta k\alpha\underline{\lambda}_j + \frac{1}{2}\eta^2k^2\alpha^2\underline{\lambda}_j^2\right)\beta'\right).
\end{aligned}$$

Above, we used the facts (a) for $x > 0$, we have $e^{-x} - 1 \leqslant -x + \frac{x^2}{2}$, and (b) for $x \in \mathbb{R}$, we have $1 + x \leqslant e^x$. Hence by the tower property:

$$\begin{aligned}
\mathbb{E}\exp\left(-\eta\sum_{t=1}^{T}\langle v, \tilde{x}_t^{(i)}\rangle^2\right) &\leqslant \exp\left(\sum_{j=1}^{S}\left(-\eta k\alpha\underline{\lambda}_j + \frac{1}{2}\eta^2k^2\alpha^2\underline{\lambda}_j^2\right)\beta'\right) \\
&\leqslant \exp\left(-\eta k\alpha\beta'\sum_{j=1}^{S}\underline{\lambda}_j + \frac{1}{2}\eta^2k^2\alpha^2\beta'\sum_{j=1}^{S}\underline{\lambda}_j^2\right) \\
&= \exp\left(-\eta k\alpha\beta'\left(\sum_{j=1}^{S}\underline{\lambda}_j\right)\left(1 - \frac{\eta k\alpha}{2}\frac{\sum_{j=1}^{S}\underline{\lambda}_j^2}{\sum_{j=1}^{S}\underline{\lambda}_j}\right)\right).
\end{aligned}$$

Now, let us set

$$\eta = \frac{1}{k\alpha}\frac{\sum_{j=1}^{S}\underline{\lambda}_j}{\sum_{j=1}^{S}\underline{\lambda}_j^2} = \frac{1}{k\alpha}\cdot\frac{1}{\bar{\mu}}\cdot\frac{1}{C_S},$$

from which we conclude:

$$\begin{aligned}
\mathbb{E}\exp\left(-\eta\sum_{t=1}^{T}\langle v, \tilde{x}_t^{(i)}\rangle^2\right) &\leqslant \exp\left(-\frac{\beta'}{2}\frac{\left(\sum_{j=1}^{S}\underline{\lambda}_j\right)^2}{\sum_{j=1}^{S}\underline{\lambda}_j^2}\right) = \exp\left(-\frac{S\beta'}{2}\frac{\left(\frac{1}{S}\sum_{j=1}^{S}\underline{\lambda}_j\right)^2}{\frac{1}{S}\sum_{j=1}^{S}\underline{\lambda}_j^2}\right) \\
&= \exp\left(-\frac{S\beta'}{2C_S}\right).
\end{aligned}$$

By independence across the m trajectories:

$$\mathbb{E}\exp\left(-\eta\sum_{i=1}^{m}\sum_{t=1}^{T}\langle v, \tilde{x}_t^{(i)}\rangle^2 + \frac{mS\beta'}{2C_S}\right) \leqslant 1.$$

As desired, we have constructed a family of random variables indexed by $v \in \mathbb{S}^{n-1}$, with MGF bounded by one.

Using the PAC-Bayes arguments from Lemma B.13 followed by Markov's inequality, with probability at least $1 - 2e^{-t}$, for all $v \in \mathbb{S}^{n-1}$ and $\gamma \in [0, 1/2]$:

$$\sum_{i,t} \langle \tilde{x}_t^{(i)}, v \rangle^2 \geqslant \frac{1}{\eta} \left[\frac{mS\beta'}{2C_S} - n \log\left(\frac{5}{4\gamma^2} \right) - t \right] - \frac{4\gamma^2 e^t mT}{\underline{\lambda}}. \tag{B.25}$$

Choosing $\gamma^2 = \frac{n\underline{\lambda}}{4\eta m T e^t}$, we have that:

$$\frac{n}{\eta} \log\left(\frac{5}{4\gamma^2} \right) + \frac{4mTe^t}{\underline{\lambda}} \gamma^2 = \frac{n}{\eta} \left[1 + \log\left(\frac{5mTe^t \eta}{n\underline{\lambda}} \right) \right]. \tag{B.26}$$

Note that this choice of γ satisfies $\gamma \in [0, 1/2]$, since:

$$\frac{n\underline{\lambda}}{4\eta m T e^t} \leqslant \frac{1}{4} \Longleftarrow \frac{n}{\eta m T} \leqslant 1 \qquad\qquad \text{since } t \geqslant 0 \text{ and } \underline{\lambda} \leqslant 1.$$

The RHS above is ensured by:

$$\frac{mT}{kn} \geqslant \alpha \frac{\sum_{j=1}^S \underline{\lambda}_j^2}{\sum_{j=1}^S \underline{\lambda}_j} = \alpha C_S \cdot \bar{\mu} \Longleftarrow \frac{mT}{kn} \geqslant C_S \qquad\qquad \text{since } \alpha, \bar{\mu} \leqslant 1.$$

If we further enforce that:

$$\frac{mS\beta'}{4C_S} \geqslant (n+1) \left[t + \log\left(\frac{5mT\eta}{n\underline{\lambda}} \right) \right], \tag{B.27}$$

then combining (B.25) with (B.26):

$$\begin{aligned}
\sum_{i,t} \langle \tilde{x}_t^{(i)}, v \rangle^2 &\geqslant \frac{1}{\eta} \left[\frac{mS\beta'}{2C_S} - t - n - n \log\left(\frac{5mTe^t\eta}{n\underline{\lambda}} \right) \right] \\
&= \frac{1}{\eta} \left[\frac{mS\beta'}{2C_S} - (n+1)t - n - n \log\left(\frac{5mT\eta}{n\underline{\lambda}} \right) \right] \\
&\geqslant \frac{1}{\eta} \left[\frac{mS\beta'}{2C_S} - (n+1)t - (n+1) \log\left(\frac{5mT\eta}{n\underline{\lambda}} \right) \right] \\
&\geqslant \frac{mS\beta'}{4\eta C_S} = \frac{\alpha\beta' mT\bar{\mu}}{8}.
\end{aligned}$$

For (B.27), it suffices that:

$$\frac{mT}{kn} \geqslant \frac{32C_S t}{\beta'}, \quad \frac{mT}{kn} \geqslant \frac{16C_S}{\beta'} \log\left(\frac{5}{\underline{\lambda}\alpha} \frac{\sum_{j=1}^S \underline{\lambda}_j}{\sum_{j=1}^S \underline{\lambda}_j^2} \frac{mT}{kn} \right) = \frac{16C_S}{\beta'} \log\left(\frac{5}{\underline{\lambda}\alpha\bar{\mu}C_S} \right) + \frac{16C_S}{\beta'} \log\left(\frac{mT}{kn} \right).$$

For the RHS inequality, by Proposition B.1 it suffices that:

$$\frac{mT}{kn} \geqslant \frac{32C_S}{\beta'} \max\left\{ \log\left(\frac{5}{\underline{\lambda}\alpha\bar{\mu}C_S} \right), 0 \right\}, \quad \frac{mT}{kn} \geqslant \frac{64C_S}{\beta'} \log\left(\frac{128C_S}{\beta'} \right).$$

Note that $x\log(1/x) \leqslant 1/e$ for all $x > 0$, and therefore:

$$C_S \log\left(\frac{5}{\underline{\lambda}\alpha\bar{\mu}C_S}\right) = C_S \log\left(\frac{5}{\alpha\underline{\lambda}\bar{\mu}}\right) + C_S \log\left(\frac{1}{C_S}\right) \leqslant C_S \log\left(\frac{5}{\alpha\underline{\lambda}\bar{\mu}}\right) + 1 \leqslant 2C_S \log\left(\frac{5}{\alpha\underline{\lambda}\bar{\mu}}\right).$$

Hence it suffices that:

$$\frac{mT}{kn} \geqslant \frac{64C_S}{\beta'} \max\left\{\log\left(\frac{5}{\alpha\underline{\lambda}\bar{\mu}}\right), \log\left(\frac{128C_S}{\beta'}\right)\right\}.$$

The claim now follows by simplifying all the required inequalities for the quantity mT/kn. $\square$

To contrast the effects of the wTrajSB assumption from those of the TrajSB assumption, let us compare Lemma B.20 to its counterpart Lemma B.13. The minimum eigenvalue bound (B.24) from Lemma B.20 differs from the corresponding TrajSB bound (B.4) in the role of the eigenvalues of the matrices $\{\Psi_j\}_{j=1}^{\lfloor T/k\rfloor}$ from the small-ball definition. However, due to the differing requirements on the amount of data mT, neither result is necessarily sharper than the other, as detailed in the following remark:

Remark B.21. When the matrices $\{\Psi_j\}_{j=1}^{\lfloor T/k\rfloor}$ from the trajectory small-ball definition vary across j, both Lemma B.13 and Lemma B.20 yield different dependencies on the eigenvalues $\{\underline{\lambda}(\Psi_j, \underline{\Gamma})\}_{j=1}^{\lfloor T/k\rfloor}$. In particular, Lemma B.13 yields a minimum eigenvalue bound scaling as $mT\underline{\mu}$, where $\underline{\mu}$ is the *geometric mean* of the eigenvalues $\{\underline{\lambda}(\Psi_j, \underline{\Gamma})\}$, whereas Lemma B.20 yields a bound scaling as $mT\bar{\mu}$, where $\bar{\mu}$ is the *arithmetic mean* of the eigenvalues. By the AM-GM inequality, we have that $\bar{\mu} \geqslant \underline{\mu}$, so the latter bound is stronger than the former. However, Lemma B.20 has a stronger requirement on the amount of data, requiring that $mT \gtrsim knC_S$, where $C_S \in [1, S]$ is defined in (B.23), whereas Lemma B.13 has the weaker requirement that $mT \gtrsim kn$. In the worst case when $C_S \asymp S$, then the $mT \gtrsim knC_S$ requirement simplifies to the many trajectories assumption $m \gtrsim n$. Thus, the qualitative behavior of these two bounds are not necessarily comparable.

Meanwhile, although neither bound is strictly sharper than the other, if we assume polynomial growth of the $\{\Psi_j\}$ matrices, then the two bounds are roughly on par:

Remark B.22. When the matrices $\{\Psi_j\}$ exhibit low degree polynomial growth, both Lemma B.13 and Lemma B.20 yield similar qualitative behavior. Concretely, let us suppose that $k = 1$, $\Psi_j = j^p \cdot I$ for $j \in [T]$, and $\underline{\Gamma} = \frac{1}{T}\sum_{j=1}^{T} \Psi_j$. Then, $\bar{\mu} = 1$, whereas $\underline{\mu} \geqslant \frac{1+p}{e^p}$. Thus, if we consider p as constant, then $\bar{\mu} \asymp \underline{\mu}$.

We now state our general OLS upper bound under the weak trajectory small-ball condition.

Lemma B.23 (General OLS upper bound, high probability)**.** *There are universal positive constants c_0 and c_1 such that the following holds. Suppose that P_x satisfies the $(T, k, \{\Psi_j\}_{j=1}^{\lfloor T/k\rfloor}, \alpha, \beta)$-wTrajSB condition (Definition B.1). Put $S := \lfloor T/k\rfloor$ and $\Gamma_t := \Gamma_t(\mathsf{P}_x)$ for $t \in \mathbb{N}_+$. Fix any $\underline{\Gamma} \in \mathsf{Sym}^n_{>0}$ satisfying $\frac{1}{S}\sum_{j=1}^{S} \Psi_j \preccurlyeq \underline{\Gamma} \preccurlyeq \Gamma_T$, and the constants:*

$$C_S := \frac{\frac{1}{S}\sum_{j=1}^{S} \underline{\lambda}(\Psi_j, \underline{\Gamma})^2}{\left(\frac{1}{S}\sum_{j=1}^{S} \underline{\lambda}(\Psi_j, \underline{\Gamma})\right)^2}, \quad \bar{\mu} := \frac{1}{S}\sum_{j=1}^{S} \underline{\lambda}(\Psi_j, \underline{\Gamma}).$$

Fix $\delta \in (0, 1/e)$. Suppose that:

$$n \geqslant 2, \quad \frac{mT}{kn} \geqslant \frac{c_0 C_S}{1-\beta} \log\left(\frac{C_S}{\alpha(1-\beta)\underline{\lambda}(\underline{\Gamma}, \Gamma_T)\bar{\mu}\delta}\right).$$

Then, for any $\Gamma' \in \mathsf{Sym}^n_{>0}$, with probability at least $1-\delta$:

$$\|\hat{W}_{m,T} - W_\star\|^2_{\Gamma'} \leqslant c_1 \sigma_\xi^2 \left[\frac{pn \log\left(\frac{1}{\alpha(1-\beta)\underline{\lambda}(\underline{\Gamma}, \Gamma_T)\bar{\mu}\delta}\right)}{\underline{\lambda}(\underline{\Gamma}, \Gamma')\alpha(1-\beta)mT\bar{\mu}}\right].$$

Proof. Put $\beta' := 1-\beta$ and $\tilde{X}_{m,T} := X_{m,T}\underline{\Gamma}^{-1/2}$. By the arguments in the proof of Lemma 5.1,

$$\|\hat{W}_{m,T} - W_\star\|^2_{\Gamma'} \leqslant \min\{n,p\} \frac{\|(\tilde{X}_{m,T}^\mathsf{T}\tilde{X}_{m,T})^{-1/2}\tilde{X}_{m,T}^\mathsf{T}\Xi_{m,T}\|^2_{\mathrm{op}}}{\underline{\lambda}(\underline{\Gamma}, \Gamma')\cdot\lambda_{\min}(\tilde{X}_{m,T}^\mathsf{T}\tilde{X}_{m,T})}.$$

Put $M := (\alpha\beta' mT\bar{\mu}/8)\cdot I := \zeta \cdot I$. By Proposition B.10, with probability at least $1-\delta/2$:

$$\begin{aligned}
&\mathbf{1}\{\tilde{X}_{m,T}^\mathsf{T}\tilde{X}_{m,T} \succcurlyeq M\}\|(\tilde{X}_{m,T}^\mathsf{T}\tilde{X}_{m,T})^{-1/2}\tilde{X}_{m,T}^\mathsf{T}\Xi_{m,T}\|^2_{\mathrm{op}} \\
&\leqslant 16\sigma_\xi^2\left[p\log 5 + \frac{1}{2}\log\det\left(I_n + \zeta^{-1}\tilde{X}_{m,T}^\mathsf{T}\tilde{X}_{m,T}\right) + \log(2/\delta)\right] \\
&\leqslant 32\sigma_\xi^2\left[p + \log\det\left(I_n + \zeta^{-1}\tilde{X}_{m,T}^\mathsf{T}\tilde{X}_{m,T}\right) + \log(2/\delta)\right] \\
&\leqslant 32\sigma_\xi^2\left[p + n\log(1 + \zeta^{-1}\operatorname{tr}(\tilde{X}_{m,T}^\mathsf{T}\tilde{X}_{m,T})/n) + \log(2/\delta)\right].
\end{aligned}$$

Now, by Lemma B.20, with probability at least $1-\delta/2$, we also have:

$$\lambda_{\min}(\tilde{X}_{m,T}^\mathsf{T}\tilde{X}_{m,T}) \geqslant \zeta, \quad \operatorname{tr}(\tilde{X}_{m,T}^\mathsf{T}\tilde{X}_{m,T}) \leqslant \frac{4mTn}{\underline{\lambda}\delta}.$$

On both events:

$$\begin{aligned}
\|(\tilde{X}_{m,T}^\mathsf{T}\tilde{X}_{m,T})^{-1/2}\tilde{X}_{m,T}^\mathsf{T}\Xi_{m,T}\|^2_{\mathrm{op}} &\leqslant 32\sigma_\xi^2\left[p + n\log\left(1 + \frac{32}{\alpha\beta'\underline{\lambda}\bar{\mu}\delta}\right) + \log(2/\delta)\right] \\
&\leqslant 64\sigma_\xi^2\left[p + n\log\left(\frac{33}{\alpha\beta'\underline{\lambda}\bar{\mu}\delta}\right)\right].
\end{aligned}$$

Combining the inequalities:

$$\|\hat{W}_{m,T} - W_\star\|^2_{\Gamma'} \leqslant 512\sigma_\xi^2\min\{n,p\}\left[\frac{p + n\log\left(\frac{33}{\alpha\beta'\underline{\lambda}\bar{\mu}\delta}\right)}{\underline{\lambda}(\underline{\Gamma}, \Gamma')\alpha\beta' mT\bar{\mu}}\right] \leqslant 1024\sigma_\xi^2\left[\frac{pn\log\left(\frac{33}{\alpha\beta'\underline{\lambda}\bar{\mu}\delta}\right)}{\underline{\lambda}(\underline{\Gamma}, \Gamma')\alpha\beta' mT\bar{\mu}}\right].$$

By a union bound, both events hold with probability at least $1-\delta$, which concludes the proof. □

B.7.3 Mixing implies weak trajectory small-ball

One advantage of Definition B.1 is that it is implied by the standard notions of ϕ-mixing in the literature (see e.g. [MR08, DAJJ12, KM17]). In this section, we prove this reduction. First, we state the definition of ϕ-mixing.

Definition B.2 (ϕ-mixing covariate sequence)**.** *Let $\{x_t\}_{t \geqslant 1}$ be a covariate sequence which is adapted to a filtration $\{\mathcal{F}_t\}_{t \geqslant 1}$. Define the function $\phi(k)$ as:*

$$\phi(k) := \sup_{t \in \mathbb{N}_+} \sup_{B \in \mathcal{F}_t} \|\mathbb{P}_{x_{t+k}}(\cdot \mid B) - \mathbb{P}_{x_{t+k}}\|_{\mathrm{tv}}. \tag{B.28}$$

The process $\{x_t\}_{t \geqslant 1}$ is called ϕ-mixing *if $\lim_{k \to \infty} \phi(k) = 0$. We also let $\bar{\phi}(k)$ denote the* upper envelope *of $\phi(k)$, i.e., $\bar{\phi}(k) := \sup_{k' \geqslant k} \phi(k)$.*

The following result shows that a ϕ-mixing covariate sequence where each marginal distribution is weakly small-ball satisfies the weak trajectory small-ball condition.

Proposition B.24. *Fix $\alpha \in (0,1)$ and $\beta \in (0, 1/4)$. Suppose that the covariate sequence $\{x_t\}_{t \geqslant 1}$ is ϕ-mixing, and that for every $t \in \mathbb{N}_+$ and $v \in \mathbb{R}^n \setminus \{0\}$ we have:*

$$\mathbb{P}_{x_t}\{\langle v, x_t \rangle^2 \leqslant \alpha v^\mathsf{T} \Sigma_t v\} \leqslant \beta, \quad \Sigma_t := \mathbb{E}[x_t x_t^\mathsf{T}]. \tag{B.29}$$

Let $k_{\mathrm{mix}} := \inf\{k \in \mathbb{N}_+ \mid \bar{\phi}(k) \leqslant \beta\}$ and assume that $T \geqslant 2k_{\mathrm{mix}}$. Put $S := \lfloor T/(2k_{\mathrm{mix}}) \rfloor$ and suppose that $\{\Psi_j\}_{j=1}^S$ satisfies:

$$\Psi_j \preccurlyeq \frac{1}{4}\Sigma_t \quad \forall j \in [S],\, t \in [k_{\mathrm{mix}}(2j-1)+1, 2jk_{\mathrm{mix}}].$$

Then, P_x satisfies the $\left(T, 2k_{\mathrm{mix}}, \{\Psi_j\}_{j=1}^{T/(2k_{\mathrm{mix}})}, \alpha, \frac{4}{3}\left(\frac{1}{2}+\beta\right)\right)$-wTrajSB condition (cf. Definition B.1).

Proof. Fix $j \in [S]$. Since $\Psi_j \preccurlyeq \frac{1}{4}\Sigma_t$ for all $t \in [k_{\mathrm{mix}}(2j-1)+1, 2jk_{\mathrm{mix}}]$, we have:

$$\Psi_j \preccurlyeq \frac{1}{4k_{\mathrm{mix}}} \sum_{t=k_{\mathrm{mix}}(2j-1)+1}^{2jk_{\mathrm{mix}}} \Sigma_t.$$

Hence,

$$\frac{1}{S}\sum_{j=1}^S \Psi_j \preccurlyeq \frac{1}{4Sk_{\mathrm{mix}}} \sum_{j=1}^S \sum_{t=k_{\mathrm{mix}}(2j-1)+1}^{2jk_{\mathrm{mix}}} \Sigma_t \preccurlyeq \frac{1}{4Sk_{\mathrm{mix}}} \sum_{t=1}^T \Sigma_t \preccurlyeq \Gamma_T.$$

Above, the last inequality holds since $\lfloor T/(2k_{\mathrm{mix}}) \rfloor \geqslant T/(4k_{\mathrm{mix}})$. By definition of ϕ-mixing (cf. Definition B.2) and the upper envelope $\bar{\phi}$, for any $j \in [S]$ and $t \geqslant k_{\mathrm{mix}}(2j-1)+1$:

$$\begin{aligned}
\mathbb{P}_{x_t}\left\{\langle v, x_t \rangle^2 \leqslant 4\alpha \cdot v^\mathsf{T} \Psi_j v \;\middle|\; \mathcal{F}_{(j-1)2k_{\mathrm{mix}}}\right\} &\leqslant \mathbb{P}_{x_t}\left\{\langle v, x_t \rangle^2 \leqslant 4\alpha \cdot v^\mathsf{T} \Psi_j v\right\} + \beta \\
&\leqslant \mathbb{P}_{x_t}\left\{\langle v, x_t \rangle^2 \leqslant \alpha \cdot v^\mathsf{T} \Sigma_t v\right\} + \beta \\
&\leqslant 2\beta.
\end{aligned}$$

Therefore:

$$\frac{1}{2k_{\mathrm{mix}}} \sum_{t=(j-1)2k_{\mathrm{mix}}+1}^{2jk_{\mathrm{mix}}} \mathbb{P}_{x_t} \left\{ \langle v, x_t \rangle^2 \leqslant 4\alpha \cdot v^{\mathsf{T}} \Psi_j v \;\middle|\; \mathcal{F}_{(j-1)2k_{\mathrm{mix}}} \right\}$$
$$\leqslant \frac{1}{2k_{\mathrm{mix}}}[k_{\mathrm{mix}} + 2\beta k_{\mathrm{mix}}] = \frac{1}{2} + \beta.$$

The claim now follows from Proposition 4.2. □

We conclude by noting that ϕ-mixing is a stronger notion of mixing than β-mixing, where (B.28) is only required to hold in expectation. We leave to future work an analysis that only relies on the weaker β-mixing.

C Analysis for lower bounds

C.1 Preliminaries

Here, we collect the necessary auxiliary results we will use to prove the lower bound. The first result is an instance of the well-known fact that the conditional mean is the estimator which minimizes the mean squared error.

Proposition C.1. *Let $T \in \mathbb{N}_+$ and $\{\mathsf{P}_{x,t}\}_{t=1}^T$ be a sequence of distributions over $\mathbb{R}^n$ with finite second moments $\Sigma_t := \mathbb{E}_{x_t \sim \mathsf{P}_{x,t}}[x_t x_t^{\mathsf{T}}]$. Let P_W be any arbitrary distribution on $\mathbb{R}^{p \times n}$. Put $\Gamma_T := \frac{1}{T}\sum_{t=1}^T \Sigma_t$. We have:*

$$\inf_{\hat{W}} \mathbb{E}_{W \sim P_W} \left[\frac{1}{T} \sum_{t=1}^T \mathbb{E}_{x_t \sim \mathsf{P}_{x,t}} \|\hat{W}(x_t) - W x_t\|_2^2 \right] = \mathbb{E}_{W \sim P_W} \|\mathbb{E}_{W' \sim P_W}[W'] - W\|_{\Gamma_T}^2,$$

where the infimum ranges over measurable functions $\hat{W} : \mathbb{R}^n \to \mathbb{R}^p$.

Proof. Let $\mu_T := \frac{1}{T}\sum_{t=1}^T \mathsf{P}_{x,t}$ denote the uniform mixture distribution, so that

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{x_t \sim \mathsf{P}_{x,t}} \|\hat{W}(x_t) - W x_t\|_2^2 = \mathbb{E}_{\bar{x} \sim \mu_T} \|\hat{W}(\bar{x}) - W\bar{x}\|_2^2.$$

By repeated applications of Fubini's theorem,

$$\begin{aligned}
\inf_{\hat{W}} \mathbb{E}_{W \sim P_W} \mathbb{E}_{\bar{x} \sim \mu_T} \|\hat{W}(\bar{x}) - W\bar{x}\|_2^2 &= \inf_{\hat{W}} \mathbb{E}_{\bar{x} \sim \mu_T} \mathbb{E}_{W \sim P_W} \|\hat{W}(\bar{x}) - W\bar{x}\|_2^2 \\
&= \mathbb{E}_{\bar{x} \sim \mu_T} \left[\inf_{\hat{y} \in \mathbb{R}^p} \mathbb{E}_{W \sim P_W} \|\hat{y} - W\bar{x}\|_2^2 \right] \\
&= \mathbb{E}_{\bar{x} \sim \mu_T} \mathbb{E}_{W \sim P_W} \|\mathbb{E}_{W' \sim P_W}[W']\bar{x} - W\bar{x}\|_2^2 \\
&= \mathbb{E}_{W \sim P_W} \mathbb{E}_{\bar{x} \sim \mu_T} \|\mathbb{E}_{W' \sim P_W}[W']\bar{x} - W\bar{x}\|_2^2 \\
&= \mathbb{E}_{W \sim P_W} \|\mathbb{E}_{W' \sim P_W}[W'] - W\|_{\Gamma_T}^2.
\end{aligned}$$

□

The next result is a simple fact which states that if a function is strictly increasing and concave on an interval, then any root of the function is lower bounded by the root of the linear approximation at any point in the interval.

Proposition C.2. *Let $f : I \to \mathbb{R}$ be a $C^1(I)$ function that is strictly increasing and concave on an interval $I \subseteq \mathbb{R}$. Suppose that f has a (unique) root $x_0 \in I$. For any $x \in I$, we have that:*

$$x - \frac{f(x)}{f'(x)} \leqslant x_0.$$

Proof. Because f is concave on I, we have that:

$$0 = f(x_0) \leqslant f(x) + f'(x)(x_0 - x).$$

Next, because f is strictly increasing on I, we have that $f'(x) > 0$. The claim now follows by re-arranging the previous inequality. $\square$

The next result states that the trace inverse of any positive definite matrix is lower bounded by the trace inverse of any priciple submatrix. The claim is immediate from Cauchy's eigenvalue interlacing theorem, but we give a more direct proof.

Proposition C.3. *Let $M \in \mathbb{R}^{q \times n}$ have full column rank. Let $I \subseteq \{1, \ldots, n\}$ be any index set, and let $E_I : \mathbb{R}^n \to \mathbb{R}^{|I|}$ denote any linear map which extracts the coordinates associated to I. We have:*

$$\operatorname{tr}((M^\mathsf{T} M)^{-1}) \geqslant \operatorname{tr}((E_I M^\mathsf{T} M E_I^\mathsf{T})^{-1}).$$

Proof. Fix a $z \in \mathbb{R}^n$. Since M has full column rank, we have that $(M^\mathsf{T})^\dagger = M(M^\mathsf{T} M)^{-1}$. Therefore,

$$\min_{c \in \mathbb{R}^q : M^\mathsf{T} c = z} \|c\|_2^2 = \|(M^\mathsf{T})^\dagger z\|_2^2 = z^\mathsf{T} (M^\mathsf{T} M)^{-1} z.$$

Taking expectation with $z \sim N(0, I_n)$,

$$\operatorname{tr}((M^\mathsf{T} M)^{-1}) = \mathbb{E}_{z \sim N(0, I_n)} \left[\min_{c \in \mathbb{R}^q : M^\mathsf{T} c = z} \|c\|_2^2 \right].$$

On the other hand, we have that:

$$\min_{c \in \mathbb{R}^q : M^\mathsf{T} c = z} \|c\|_2^2 \geqslant \min_{c \in \mathbb{R}^q : E_I M^\mathsf{T} c = E_I z} \|c\|_2^2.$$

This is clear because for any $c \in \mathbb{R}^q$ satisfying $M^\mathsf{T} c = z$, the equality $E_I M^\mathsf{T} c = E_I z$ trivially holds. This means we have the following set inclusion:

$$\{c \in \mathbb{R}^q \mid M^\mathsf{T} c = z\} \subseteq \{c \in \mathbb{R}^q \mid E_I M^\mathsf{T} c = E_I z\}.$$

Therefore, minimizing any function over the first set will be lower bounded by minimizing the same function over the second set. From this inclusion, we conclude for any index set I:

$$\begin{aligned} \operatorname{tr}((M^\mathsf{T} M)^{-1}) &= \mathbb{E}_{z \sim N(0, I_n)} \left[\min_{c \in \mathbb{R}^q : M^\mathsf{T} c = z} \|c\|_2^2 \right] \geqslant \mathbb{E}_{z \sim N(0, I_n)} \left[\min_{c \in \mathbb{R}^q : E_I M^\mathsf{T} c = E_I z} \|c\|_2^2 \right] \\ &= \mathbb{E}_{z \sim N(0, I_{|I|})} \left[\min_{c \in \mathbb{R}^q : E_I M^\mathsf{T} c = z} \|c\|_2^2 \right] = \operatorname{tr}((E_I M^\mathsf{T} M E_I^\mathsf{T})^{-1}). \end{aligned}$$

$\square$

Next, we state well-known upper and lower tail bounds for chi-squared random variables.

Lemma C.4 ([LM00, Lemma 1])**.** *Let $g_1, \dots, g_D$ be iid $N(0,1)$ random variables, and let $a_1, \dots, a_D$ be non-negative scalars. For any $t > 0$, we have:*

$$\mathbb{P}\left\{ \sum_{i=1}^D a_i(g_i^2 - 1) \geqslant 2\sqrt{t}\sqrt{\sum_{i=1}^D a_i^2} + 2t \max_{i=1,\dots,D} a_i \right\} \leqslant e^{-t},$$

$$\mathbb{P}\left\{ \sum_{i=1}^D a_i(g_i^2 - 1) \leqslant -2\sqrt{t}\sqrt{\sum_{i=1}^D a_i^2} \right\} \leqslant e^{-t}.$$

Finally, we conclude with a convex extension of Gordon's min-max theorem.

Theorem C.5 ([TOH14, Theorem II.1])**.** *Let $A \in \mathbb{R}^{m\times n}$, $g \in \mathbb{R}^m$, and $h \in \mathbb{R}^n$ have iid $N(0,1)$ entires and be independent of each other. Suppose that $S_1 \subset \mathbb{R}^n$ and $S_2 \subset \mathbb{R}^m$ are non-empty compact convex sets, and let $\psi : S_1 \times S_2 \to \mathbb{R}$ be a continuous, convex-concave function. For every $t \in \mathbb{R}$, we have:*

$$\mathbb{P}\left\{ \min_{x\in S_1} \max_{y \in S_2} \left[y^\mathsf{T} A x + \psi(x,y) \right] \geqslant t \right\} \leqslant 2\mathbb{P}\left\{ \min_{x\in S_1} \max_{y\in S_2} \left[\|x\|_2 g^\mathsf{T} y + \|y\|_2 h^\mathsf{T} x + \psi(x,y) \right] \geqslant t \right\}.$$

C.2 Proof of Lemma 6.1

We first prove the following intermediate result, which holds under the Gaussian observation noise model (Definition 7.1).

Lemma C.6. *Let $T \in \mathbb{N}_+$, $\{\mathsf{P}_{x,t}\}_{t=1}^T$ be a sequence of distributions over $\mathbb{R}^n$ with finite second moments $\Sigma_t := \mathbb{E}_{x_t \sim \mathsf{P}_{x,t}}[x_t x_t^\mathsf{T}]$, and $\sigma_\xi > 0$. Let P_X be a distribution on $\mathbb{R}^{q\times n}$ with $q \geqslant n$ such that for $X \sim P_X$, $X^\mathsf{T} X$ is invertible almost surely. For $W \in \mathbb{R}^{p\times n}$, let P_W be the distribution over $\mathbb{R}^{q\times n} \times \mathbb{R}^{q\times p}$ with $(X,Y) \sim P_W$ satisfying $X \sim P_X$ and $Y \mid X = XW^\mathsf{T} + \Xi$, where $\Xi \in \mathbb{R}^{q\times p}$ has iid $N(0, \sigma_\xi^2)$ entries (and is independent of everything else). Put $\Gamma_T := \frac{1}{T}\sum_{t=1}^T \Sigma_t$. We have that:*

$$\begin{aligned} &\inf_{\hat{W}} \sup_{W\in\mathbb{R}^{p\times n}} \mathbb{E}_{(X,Y)\sim P_W} \left[\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{x_t\sim\mathsf{P}_{x,t}} \|\hat{W}(X,Y,x_t) - Wx_t\|_2^2 \right] \\ &\qquad \geqslant \sigma_\xi^2 p \cdot \mathbb{E}_{X\sim P_X} \operatorname{tr}(\Gamma_T^{1/2}(X^\mathsf{T} X)^{-1}\Gamma_T^{1/2}), \end{aligned}$$

where the infimum ranges over all measurable functions $\hat{W} : \mathbb{R}^{q\times n} \times \mathbb{R}^{q\times p} \times \mathbb{R}^n \to \mathbb{R}^p$.

Proof. The proof extends the Bayesian argument from [Mou22, Theorem 1]. Let p_λ be any prior distribution over $\mathbb{R}^{p\times n}$. Let $\mu_T := \frac{1}{T}\sum_{t=1}^T \mathsf{P}_{x,t}$ denote the uniform mixture. Bounding the minimax risk from below by the Bayes risk:

$$\begin{aligned} &\inf_{\hat{W}} \sup_{W\in\mathbb{R}^{p\times n}} \mathbb{E}_{(X,Y)\sim P_W} \mathbb{E}_{\bar{x}\sim\mu_T} \|\hat{W}(X,Y,\bar{x}) - W\bar{x}\|_2^2 \\ &\geqslant \inf_{\hat{W}} \mathbb{E}_{W_\lambda\sim p_\lambda} \mathbb{E}_{(X,Y)\sim P_{W_\lambda}} \mathbb{E}_{\bar{x}\sim\mu_T} \|\hat{W}(X,Y,\bar{x}) - W_\lambda \bar{x}\|_2^2 \\ &= \inf_{\hat{W}} \mathbb{E}_{(X,Y)} \mathbb{E}_{W_\lambda|(X,Y)} \mathbb{E}_{\bar{x}\sim\mu_T} \|\hat{W}(X,Y,\bar{x}) - W_\lambda\bar{x}\|_2^2 \qquad \text{using Fubini's theorem} \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{(X,Y)} \inf_{\hat{W}_{X,Y}} \mathbb{E}_{W_\lambda|(X,Y)} \mathbb{E}_{\bar{x}\sim\mu_T} \|\hat{W}_{X,Y}(\bar{x}) - W_\lambda \bar{x}\|_2^2 && \text{where } \hat{W}_{X,Y} \text{ maps } \mathbb{R}^n \to \mathbb{R}^p \\
&= \mathbb{E}_{(X,Y)} \mathbb{E}_{W_\lambda|(X,Y)} \|\mathbb{E}[W_\lambda \mid X, Y] - W_\lambda\|_{\Gamma_T}^2 && \text{using Proposition C.1.}
\end{aligned}
$$

Now let $W_\lambda \sim p_\lambda$ have iid $N(0, 1/\lambda)$ entries for $\lambda > 0$. Noting that

$$
\mathrm{vec}(Y) = (I_p \otimes X)\mathrm{vec}(W_\lambda^\mathsf{T}) + \mathrm{vec}(\Xi),
$$

we see that the vector $\begin{bmatrix} \mathrm{vec}(W_\lambda^\mathsf{T}) \\ \mathrm{vec}(Y) \end{bmatrix}$ is jointly Gaussian conditioned on X:

$$
\begin{bmatrix} \mathrm{vec}(W_\lambda^\mathsf{T}) \\ \mathrm{vec}(Y) \end{bmatrix} \mid X \sim N\left(0, \begin{bmatrix} \frac{1}{\lambda} I_{pn} & \frac{1}{\lambda}(I_p \otimes X^\mathsf{T}) \\ * & \frac{1}{\lambda}(I_p \otimes XX^\mathsf{T}) + \sigma_\xi^2 I_{qp} \end{bmatrix}\right).
$$

Therefore, the distribution of $\mathrm{vec}(W_\lambda^\mathsf{T}) \mid X, Y$ is:

$$
\begin{aligned}
\mathrm{vec}(W_\lambda^\mathsf{T}) \mid X, Y &\sim N(\mu_\lambda, \Sigma_\lambda), \\
\mu_\lambda &:= \frac{1}{\lambda}(I_p \otimes X^\mathsf{T}) \left[\frac{1}{\lambda}(I_p \otimes XX^\mathsf{T}) + \sigma_\xi^2 I_{qp}\right]^{-1} \mathrm{vec}(Y), \\
\Sigma_\lambda &:= \frac{1}{\lambda} I_{pn} - \frac{1}{\lambda^2}(I_p \otimes X^\mathsf{T}) \left[\frac{1}{\lambda}(I_p \otimes XX^\mathsf{T}) + \sigma_\xi^2 I_{qp}\right]^{-1} (I_p \otimes X).
\end{aligned}
$$

A generalization of the identity $X^\mathsf{T}(\frac{1}{\lambda} XX^\mathsf{T} + \sigma_\xi^2 I_q)^{-1} = (\frac{1}{\lambda} X^\mathsf{T} X + \sigma_\xi^2 I_n)^{-1} X^\mathsf{T}$ yields:

$$
(I_p \otimes X^\mathsf{T}) \left[\frac{1}{\lambda}(I_p \otimes XX^\mathsf{T}) + \sigma_\xi^2 I_{qp}\right]^{-1} = \left[\frac{1}{\lambda}(I_p \otimes X^\mathsf{T} X) + \sigma_\xi^2 I_{np}\right]^{-1} (I_p \otimes X^\mathsf{T}).
$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}[\mathrm{vec}(W_\lambda^\mathsf{T}) \mid X, Y] - \mathrm{vec}(W_\lambda^\mathsf{T}) \\
&= \mu_\lambda - \mathrm{vec}(W_\lambda^\mathsf{T}) \\
&= \left[(I_p \otimes X^\mathsf{T} X) + \sigma_\xi^2 \lambda I_{np}\right]^{-1} (I_p \otimes X^\mathsf{T})\mathrm{vec}(Y) - \mathrm{vec}(W_\lambda^\mathsf{T}) \\
&= \left[\left[(I_p \otimes X^\mathsf{T} X) + \sigma_\xi^2 \lambda I_{np}\right]^{-1} (I_p \otimes X^\mathsf{T} X) - I_{np}\right] \mathrm{vec}(W_\lambda^\mathsf{T}) \\
&\qquad + \left[(I_p \otimes X^\mathsf{T} X) + \sigma_\xi^2 \lambda I_{np}\right]^{-1} (I_p \otimes X^\mathsf{T})\mathrm{vec}(\Xi).
\end{aligned}
$$

Observing that

$$
\|\mathbb{E}[W_\lambda \mid X, Y] - W_\lambda\|_{\Gamma_T}^2 = \|\mathbb{E}[\mathrm{vec}(W_\lambda^\mathsf{T}) \mid X, Y] - \mathrm{vec}(W_\lambda^\mathsf{T})\|_{I_p \otimes \Gamma_T}^2,
$$

and defining $M_X(\lambda) := (I_p \otimes X^\mathsf{T} X) + \sigma_\xi^2 \lambda I_{np}$, we have the following bias-variance decomposition:

$$
\begin{aligned}
&\mathbb{E}_{X,\Xi,W_\lambda} \|\mathbb{E}[W_\lambda \mid X, Y] - W_\lambda\|_{\Gamma_T}^2 \\
&= \mathbb{E}_{X,\Xi,W_\lambda} \left\| \left[M_X^{-1}(\lambda)(I_p \otimes X^\mathsf{T} X) - I_{np}\right] \mathrm{vec}(W_\lambda^\mathsf{T}) \right\|_{I_p \otimes \Gamma_T}^2
\end{aligned}
$$

$$
\begin{aligned}
&+ \sigma_\xi^2 \mathbb{E}_X \operatorname{tr}\left((I_p \otimes \Gamma_T^{1/2}) M_X^{-1}(\lambda)(I_p \otimes X^\mathsf{T} X) M_X^{-1}(\lambda)(I_p \otimes \Gamma_T^{1/2})\right) \\
&\geqslant \sigma_\xi^2 \mathbb{E}_X \operatorname{tr}\left((I_p \otimes \Gamma_T^{1/2}) M_X^{-1}(\lambda)(I_p \otimes X^\mathsf{T} X) M_X^{-1}(\lambda)(I_p \otimes \Gamma_T^{1/2})\right).
\end{aligned}
$$

Since $\lambda \mapsto \operatorname{tr}\left((I_p \otimes \Gamma_T^{1/2}) M_X^{-1}(\lambda)(I_p \otimes X^\mathsf{T} X) M_X^{-1}(\lambda)(I_p \otimes \Gamma_T^{1/2})\right)$ is non-negative and decreasing in λ for $\lambda > 0$, by the monotone convergence theorem:

$$
\begin{aligned}
&\lim_{\lambda \to 0^+} \mathbb{E}_{X,\Xi,W_\lambda} \|\mathbb{E}[W_\lambda \mid X, Y] - W_\lambda\|_{\Gamma_T}^2 \\
&\geqslant \sigma_\xi^2 \lim_{\lambda \to 0^+} \mathbb{E}_X \operatorname{tr}\left((I_p \otimes \Gamma_T^{1/2}) M_X^{-1}(\lambda)(I_p \otimes X^\mathsf{T} X) M_X^{-1}(\lambda)(I_p \otimes \Gamma_T^{1/2})\right) \\
&= \sigma_\xi^2 \mathbb{E}_X \operatorname{tr}\left((I_p \otimes \Gamma_T^{1/2}) M_X^{-1}(0)(I_p \otimes X^\mathsf{T} X) M_X^{-1}(0)(I_p \otimes \Gamma_T^{1/2})\right) \\
&= \sigma_\xi^2 \mathbb{E}_X \operatorname{tr}((I_p \otimes \Gamma_T^{1/2}(X^\mathsf{T} X)^{-1} \Gamma_T^{1/2})) \\
&= \sigma_\xi^2 p \cdot \mathbb{E}_X \operatorname{tr}(\Gamma_T^{1/2}(X^\mathsf{T} X)^{-1} \Gamma_T^{1/2}).
\end{aligned}
$$

Since the first expression above lower bounds the minimax risk, this concludes the proof. □

We now restate and prove Lemma 6.1.

Lemma 6.1 (Expected trace of inverse covariance bounds risk from below)**.** *Fix $m, T \in \mathbb{N}_+$ and a set of covariate distributions $\mathcal{P}_x$. Suppose that for every $\mathsf{P}_x \in \mathcal{P}_x$, the data matrix $X_{m,T} \in \mathbb{R}^{mT \times n}$ drawn from $\otimes_{i=1}^m \mathsf{P}_x$ has full column rank almost surely. The minimax risk $\mathsf{R}(m, T, T'; \mathcal{P}_x)$ satisfies:*

$$
\mathsf{R}(m, T, T'; \mathcal{P}_x) \geqslant \sigma_\xi^2 p \cdot \sup_{\mathsf{P}_x \in \mathcal{P}_x} \mathbb{E}_{\otimes_{i=1}^m \mathsf{P}_x} \left[\operatorname{tr}\left(\Gamma_{T'}^{1/2}(\mathsf{P}_x)(X_{m,T}^\mathsf{T} X_{m,T})^{-1} \Gamma_{T'}^{1/2}(\mathsf{P}_x)\right)\right].
$$

Proof. Fix a $\mathsf{P}_x \in \mathcal{P}_x$, and let $\{\mathsf{P}_{x,t}\}_{t=1}^{T'}$ denote its marginal distributions up to time T'. Let $\mathsf{P}_\xi^{\mathsf{g}}$ denote the σ_ξ-MDS corresponding to the Gaussian observation noise model (Definition 7.1). Note that for any hypothesis $f : \mathbb{R}^n \to \mathbb{R}^p$, we have from (3.2):

$$
L(\hat{f}; T', \mathsf{P}_x) = \mathbb{E}_{\mathsf{P}_x}\left[\frac{1}{T'} \sum_{t=1}^{T'} \|\hat{f}(x_t) - W_\star x_t\|_2^2\right] = \frac{1}{T'} \sum_{t=1}^{T'} \mathbb{E}_{x_t \sim \mathsf{P}_{x,t}} \|\hat{f}(x_t) - W_\star x_t\|_2^2.
$$

By the definition of $\mathsf{R}(m, T, T'; \mathcal{P}_x)$ from (3.5) and Lemma C.6:

$$
\begin{aligned}
\mathsf{R}(m, T, T'; \mathcal{P}_x) &\geqslant \inf_{\mathsf{Alg}} \sup_{W_\star} \mathbb{E}_{\otimes_{i=1}^m \mathsf{P}_{x,y}^{W_\star}[\mathsf{P}_x, \mathsf{P}_\xi^{\mathsf{g}}]}\left[L\left(\mathsf{Alg}(\{(x_t^{(i)}, y_t^{(i)})\}_{i=1,t=1}^{m,T}); T', \mathsf{P}_x\right)\right] \\
&\geqslant \sigma_\xi^2 p \cdot \mathbb{E}_{\otimes_{i=1}^m \mathsf{P}_x}\left[\operatorname{tr}\left(\Gamma_{T'}^{1/2}(\mathsf{P}_x)(X_{m,T}^\mathsf{T} X_{m,T})^{-1} \Gamma_{T'}^{1/2}(\mathsf{P}_x)\right)\right].
\end{aligned}
$$

Since the bound above holds for any $\mathsf{P}_x \in \mathcal{P}_x$, we can take the supremum over $\mathsf{P}_x \in \mathcal{P}_x$, from the which the claim follows. □

C.3 A general risk lower bound

We now state a lower bound which applies with an arbitrary number of trajectories.

Lemma C.7. *Suppose that $\mathcal{P}_x$ is any set containing $\mathsf{P}_x^{0_{n\times n}}$ and $\mathsf{P}_x^{I_n}$. Let $mT \geqslant n$. Then:*

$$\mathsf{R}(m,T,T';\mathcal{P}_x) \geqslant \frac{\sigma_\xi^2}{2} \cdot \frac{pn}{mT} \cdot \max\left\{\frac{T'}{T}, 1\right\}.$$

Proof. Define $\zeta(A) := \mathbb{E}_{\otimes_{i=1}^m \mathsf{P}_x^A}\left[\operatorname{tr}\left(\Gamma_{T'}^{1/2}(A)(X_{m,T}^\mathsf{T} X_{m,T})^{-1}\Gamma_{T'}^{1/2}(A)\right)\right]$. By Lemma 6.1:

$$\mathsf{R}(m,T,T';\mathcal{P}_x) \geqslant \sigma_\xi^2 p \cdot \max\{\zeta(0_{n\times n}), \zeta(I_n)\}. \tag{C.1}$$

Next, for any $M \in \mathsf{Sym}^n_{\geqslant 0}$, the function $X \mapsto \operatorname{tr}(M^{1/2}X^{-1}M^{1/2})$ is convex on the domain $\mathsf{Sym}^n_{>0}$. To see this, we define $f(X;v) := v^\mathsf{T} X^{-1} v$ for $X \in \mathsf{Sym}^n_{>0}$. We can write $f(X;v)$ as $f(X;v) = \sup\left\{-z^\mathsf{T} X z + 2v^\mathsf{T} z \mid z \in \mathbb{R}^n\right\}$; therefore $X \mapsto f(X;v)$ is convex on $\mathsf{Sym}^n_{>0}$, since it is the pointwise supremum of an affine function in X. Now we see that $X \mapsto \operatorname{tr}(M^{1/2}X^{-1}M^{1/2})$ is convex, since $\operatorname{tr}(M^{1/2}X^{-1}M^{1/2}) = \sum_{i=1}^n f(X; M^{1/2}e_i)$, which is the sum of convex functions. Therefore by Jensen's inequality, whenever $X_{m,T}^\mathsf{T} X_{m,T}$ is invertible almost surely,

$$\begin{aligned}
\zeta(A) &= \mathbb{E}_{\otimes_{i=1}^m \mathsf{P}_x^A}\left[\operatorname{tr}\left(\Gamma_{T'}^{1/2}(A)(X_{m,T}^\mathsf{T} X_{m,T})^{-1}\Gamma_{T'}^{1/2}(A)\right)\right] \\
&\geqslant \operatorname{tr}\left(\Gamma_{T'}^{1/2}(A)(\mathbb{E}_{\otimes_{i=1}^m \mathsf{P}_x^A}[X_{m,T}^\mathsf{T} X_{m,T}])^{-1}\Gamma_{T'}^{1/2}(A)\right) \\
&= \operatorname{tr}\left(\Gamma_{T'}^{1/2}(A)(mT \cdot \Gamma_T(A))^{-1}\Gamma_{T'}^{1/2}(A)\right) \\
&= \frac{\operatorname{tr}(\Gamma_{T'}(A)\Gamma_T^{-1}(A))}{mT}.
\end{aligned}$$

We first consider the case when $A = 0_{n\times n}$. Under these dynamics, it is a standard fact that when $mT \geqslant n$, then $X_{m,T}^\mathsf{T} X_{m,T}$ is invertible almost surely. Furthermore, $\Gamma_t(0_{n\times n}) = I_n$ for all t, Hence, $\zeta(0_{n\times n}) \geqslant \frac{n}{mT}$.

Next, we consider the case when $A = I_n$. We first argue that as long as $mT \geqslant n$, the matrix $X_{m,T}^\mathsf{T} X_{m,T}$ is invertible almost surely. We write $x_t^{(i)} = \sum_{k=1}^t w_k^{(i)}$, where $\{w_t^{(i)}\}_{i=1,t=1}^{m,T}$ are all iid $N(0, I_n)$ vectors. Let $p : \mathbb{R}^{mTn} \to \mathbb{R}$ be the polynomial $p(\{w_t^{(i)}\}) = \det(X_{m,T}^\mathsf{T} X_{m,T})$. The zero-set of p is either all of $\mathbb{R}^{mTn}$, or Lebesgue measure zero. We will select $\{w_t^{(i)}\}$ so that $p(\{w_t^{(i)}\}) \neq 0$, which shows that the zeros of this polynomial are not all of $\mathbb{R}^{mTn}$, and hence Lebesgue measure zero. Since the Gaussian measure on $\mathbb{R}^{mTn}$ is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^{mTn}$, this implies that $\det(X_{m,T}^\mathsf{T} X_{m,T}) \neq 0$ almost surely.

To select $\{w_t^{(i)}\}$, we introduce some notation. Let $e_i \in \mathbb{R}^n$ denote the i-th standard basis vector. For any positive integer k, let $U(k) \in \mathbb{R}^{k\times k}$ be the upper triangular matrix with ones for all its non-zero entries. Let $S(k) = U(k)U(k)^\mathsf{T}$. By construction, $S(k)$ is invertible since $U(k)$ is invertible. We put $w_t^{(i)} = e_{(i-1)T+t} \cdot \mathbf{1}\{(i-1)T + t \leqslant n\}$. We now claim that with this choice of $\{w_t^{(i)}\}$, the matrix $X_{m,T}^\mathsf{T} X_{m,T}$ is invertible.

Suppose first that $T \geqslant n$. Then we have that $X_{m,T}^\mathsf{T} X_{m,T} = S(n)$, and therefore $\det(X_{m,T}^\mathsf{T} X_{m,T}) \neq 0$. On the other hand, suppose that $T < n$. Because $mT \geqslant n$, then we have that:

$$X_{m,T}^\mathsf{T} X_{m,T} = \mathsf{BDiag}(\underbrace{S(T), \dots, S(T)}_{\lfloor n/T \rfloor \text{ times}}, S(n - T\lfloor n/T \rfloor)),$$

where $\mathsf{BDiag}(M_1, \ldots, M_k)$ denotes the block diagonal matrices with block diagonals $M_1, \ldots, M_k$. Since $S(T)$ and $S(n - T\lfloor n/T \rfloor)$ are both invertible, so is $X_{m,T}^\mathsf{T} X_{m,T}$ and therefore $\det(X_{m,T}^\mathsf{T} X_{m,T}) \neq 0$. Thus, $X_{m,T}^\mathsf{T} X_{m,T}$ is invertible almost surely.

Next, we note that $\Sigma_t(I_n) = t \cdot I_n$ and $\Gamma_t(I_n) = \left(\frac{1}{t}\sum_{k=1}^t k\right) \cdot I_n = \frac{t+1}{2} \cdot I_n$. Hence we have $\Gamma_{T'}(I_n)\Gamma_T^{-1}(I_n) = \frac{T'+1}{T+1} \cdot I_n \succcurlyeq \frac{T'}{2T} \cdot I_n$, and therefore $\zeta(I_n) \geqslant \frac{n}{2mT}\frac{T'}{T}$.

Combining our bounds on $\zeta(0_{n\times n})$ and $\zeta(I_n)$, we have the desired claim:

$$\mathsf{R}(m, T, T'; \mathcal{P}_x) \geqslant \sigma_\xi^2 p \cdot \max\left\{\frac{n}{mT}, \frac{n}{2mT}\frac{T'}{T}\right\} \geqslant \frac{\sigma_\xi^2}{2} \cdot \frac{pn}{mT} \cdot \max\left\{\frac{T'}{T}, 1\right\}.$$

□

C.4 Non-isotropic random gramian matrices

The goal of this subsection is to prove Lemma 7.1, which gives a bound on the expected trace inverse of a non-isotropic random gramian matrix. We first prove an auxiliary lemma, which will be used as a building block in the proof.

Lemma C.8. *Fix any $x \in \mathbb{R}^q$. Let $g \in \mathbb{R}^q$ and $h \in \mathbb{R}^n$ be random vectors with iid $N(0, 1)$ entries, and let $W \in \mathbb{R}^{q\times n}$ be a random matrix with iid $N(0, 1)$ entries. Let $\Sigma \in \mathbb{R}^{q\times q}$ be positive definite. We have that:*

$$\mathbb{E} \min_{\alpha\in\mathbb{R}^n} \|\Sigma^{1/2}W\alpha - x\|_2^2 \leqslant \mathbb{E} \min_{\beta\geqslant 0} \max_{\tau\geqslant 0} \left[-\frac{\beta\|h\|_2}{\tau} + \|\beta g - \Sigma^{-1/2}x\|^2_{(\Sigma^{-1}+\beta\|h\|_2\tau I_q)^{-1}}\right].$$

Proof. The proof invokes the convex Gaussian min-max lemma (Theorem C.5) via a limiting argument. In what follows, let $\{\alpha_k\}_{k\geqslant 1}$ and $\{v_k\}_{k\geqslant 1}$ be any two positive, increasing sequences of scalars tending to $+\infty$. It is clear that for every W,

$$\lim_{k\to\infty} \min_{\|\alpha\|_2\leqslant\alpha_k} \|\Sigma^{1/2}W\alpha - x\|_2^2 = \min_{\alpha\in\mathbb{R}^n} \|\Sigma^{1/2}W\alpha - x\|_2^2.$$

Since $\alpha = 0$ is always a feasible solution to $\min_{\|\alpha\|_2\leqslant\alpha_k} \|\Sigma^{1/2}W\alpha - x\|_2^2$, we have for every $k \geqslant 1$:

$$0 \leqslant \min_{\|\alpha\|_2\leqslant\alpha_k} \|\Sigma^{1/2}W\alpha - x\|_2^2 \leqslant \|x\|_2^2.$$

Therefore, by the dominated convergence theorem,

$$\mathbb{E} \min_{\alpha\in\mathbb{R}^n} \|\Sigma^{1/2}W\alpha - x\|_2^2 = \mathbb{E} \lim_{k\to\infty} \min_{\|\alpha\|_2\leqslant\alpha_k} \|\Sigma^{1/2}W\alpha - x\|_2^2 = \lim_{k\to\infty} \mathbb{E} \min_{\|\alpha\|_2\leqslant\alpha_k} \|\Sigma^{1/2}W\alpha - x\|_2^2. \quad \text{(C.2)}$$

We next state two variational forms which we will use:

$$\frac{1}{2}\|x\|_2^2 = \max_{v\in\mathbb{R}^q}\left\{v^\mathsf{T} x - \frac{\|v\|_2^2}{2}\right\}, \quad \text{(C.3)}$$

$$\|x\|_2 = \min_{\tau\geqslant 0}\left\{\frac{\|x\|_2^2\tau}{2} + \frac{1}{2\tau}\right\}. \quad \text{(C.4)}$$

Using the first variational form (C.3), we have for every W and $k_1 \geqslant 1$,

$$\begin{aligned}
\min_{\|\alpha\|_2 \leqslant \alpha_{k_1}} \frac{1}{2}\|\Sigma^{1/2}W\alpha - x\|_2^2 &= \min_{\|\alpha\|_2 \leqslant \alpha_{k_1}} \max_{v \in \mathbb{R}^q} \left[v^\mathsf{T}(\Sigma^{1/2}W\alpha - x) - \frac{\|v\|_2^2}{2} \right] \\
&= \min_{\|\alpha\|_2 \leqslant \alpha_{k_1}} \max_{v \in \mathbb{R}^q} \left[v^\mathsf{T}W\alpha - v^\mathsf{T}\Sigma^{-1/2}x - \frac{v^\mathsf{T}\Sigma^{-1}v}{2} \right] \\
&= \min_{\|\alpha\|_2 \leqslant \alpha_{k_1}} \max_{\|v\|_2 \leqslant \|\Sigma W\|_{\mathrm{op}}\alpha_{k_1} + \|\Sigma^{1/2}x\|_2} \left[v^\mathsf{T}W\alpha - v^\mathsf{T}\Sigma^{-1/2}x - \frac{v^\mathsf{T}\Sigma^{-1}v}{2} \right] \\
&= \lim_{k_2 \to \infty} \min_{\|\alpha\|_2 \leqslant \alpha_{k_1}} \max_{\|v\|_2 \leqslant v_{k_2}} \left[v^\mathsf{T}W\alpha - v^\mathsf{T}\Sigma^{-1/2}x - \frac{v^\mathsf{T}\Sigma^{-1}v}{2} \right].
\end{aligned}$$

Observe that for every $k_2 \geqslant 1$,

$$\begin{aligned}
0 &\leqslant \min_{\|\alpha\|_2 \leqslant \alpha_{k_1}} \max_{\|v\|_2 \leqslant v_{k_2}} \left[v^\mathsf{T}W\alpha - v^\mathsf{T}\Sigma^{-1/2}x - \frac{v^\mathsf{T}\Sigma^{-1}v}{2} \right] \\
&\leqslant \max_{\|v\|_2 \leqslant v_{k_2}} \left[-v^\mathsf{T}\Sigma^{-1/2}x - \frac{v^\mathsf{T}\Sigma^{-1}v}{2} \right] \\
&\leqslant \max_{v \in \mathbb{R}^q} \left[-v^\mathsf{T}\Sigma^{-1/2}x - \frac{v^\mathsf{T}\Sigma^{-1}v}{2} \right] = \frac{1}{2}\|x\|_2^2.
\end{aligned}$$

Therefore, by (C.2) and another application of the dominated convergence theorem:

$$\begin{aligned}
\mathbb{E} \min_{\alpha \in \mathbb{R}^n} \|\Sigma^{1/2}W\alpha - x\|_2^2 &= \lim_{k_1 \to \infty} \mathbb{E} \min_{\|\alpha\|_2 \leqslant \alpha_{k_1}} \|\Sigma^{1/2}W\alpha - x\|_2^2 \\
&= \lim_{k_1 \to \infty} \mathbb{E} \lim_{k_2 \to \infty} \min_{\|\alpha\|_2 \leqslant \alpha_{k_1}} \max_{\|v\|_2 \leqslant v_{k_2}} \left[v^\mathsf{T}W\alpha - v^\mathsf{T}\Sigma^{-1/2}x - \frac{v^\mathsf{T}\Sigma^{-1}v}{2} \right] \\
&= \lim_{k_1 \to \infty} \lim_{k_2 \to \infty} \mathbb{E} \min_{\|\alpha\|_2 \leqslant \alpha_{k_1}} \max_{\|v\|_2 \leqslant v_{k_2}} \left[v^\mathsf{T}W\alpha - v^\mathsf{T}\Sigma^{-1/2}x - \frac{v^\mathsf{T}\Sigma^{-1}v}{2} \right]. \qquad \text{(C.5)}
\end{aligned}$$

We now apply Theorem C.5 to the expectation on the RHS of (C.5):

$$\begin{aligned}
&\mathbb{E} \min_{\|\alpha\|_2 \leqslant \alpha_{k_1}} \max_{\|v\|_2 \leqslant v_{k_2}} \left[v^\mathsf{T}W\alpha - v^\mathsf{T}\Sigma^{-1/2}x - \frac{v^\mathsf{T}\Sigma^{-1}v}{2} \right] \\
&= \int_0^\infty \mathbb{P}\left\{ \min_{\|\alpha\|_2 \leqslant \alpha_{k_1}} \max_{\|v\|_2 \leqslant v_{k_2}} \left[v^\mathsf{T}W\alpha - v^\mathsf{T}\Sigma^{-1/2}x - \frac{v^\mathsf{T}\Sigma^{-1}v}{2} \right] \geqslant t \right\} \mathrm{d}t \\
&\overset{(a)}{\leqslant} 2\int_0^\infty \mathbb{P}\left\{ \min_{\|\alpha\|_2 \leqslant \alpha_{k_1}} \max_{\|v\|_2 \leqslant v_{k_2}} \left[\|\alpha\|_2 g^\mathsf{T}v + \|v\|_2 h^\mathsf{T}\alpha - v^\mathsf{T}\Sigma^{-1/2}x - \frac{v^\mathsf{T}\Sigma^{-1}v}{2} \right] \geqslant t \right\} \mathrm{d}t \\
&= 2\mathbb{E} \min_{\|\alpha\|_2 \leqslant \alpha_{k_1}} \max_{\|v\|_2 \leqslant v_{k_2}} \left[\|\alpha\|_2 g^\mathsf{T}v + \|v\|_2 h^\mathsf{T}\alpha - v^\mathsf{T}\Sigma^{-1/2}x - \frac{v^\mathsf{T}\Sigma^{-1}v}{2} \right] \\
&\leqslant 2\mathbb{E} \min_{\|\alpha\|_2 \leqslant \alpha_{k_1}} \max_{v \in \mathbb{R}^q} \left[\|\alpha\|_2 g^\mathsf{T}v + \|v\|_2 h^\mathsf{T}\alpha - v^\mathsf{T}\Sigma^{-1/2}x - \frac{v^\mathsf{T}\Sigma^{-1}v}{2} \right]. \qquad \text{(C.6)}
\end{aligned}$$

Above, inequality (a) is an application of Theorem C.5. Now for every k_1, g, and h, define

$$\psi_{k_1}(g,h) := \min_{\|\alpha\|_2 \leqslant \alpha_{k_1}} \max_{v \in \mathbb{R}^q} \left[\|\alpha\|_2 g^\mathsf{T} v + \|v\|_2 h^\mathsf{T} \alpha - v^\mathsf{T} \Sigma^{-1/2} x - \frac{v^\mathsf{T} \Sigma^{-1} v}{2} \right].$$

For every k_1, g, and h, we have

$$0 \leqslant \psi_{k_1}(g,h) \leqslant \max_{v \in \mathbb{R}^q} \left[-v^\mathsf{T} \Sigma^{-1/2} x - \frac{v^\mathsf{T} \Sigma^{-1} v}{2} \right] = \frac{\|x\|_2^2}{2}.$$

Furthermore, since $\{\alpha_k\}$ is an increasing sequence, the sequence $\{\psi_k(g,h)\}_{k \geqslant 1}$ is montonically decreasing. Therefore, by the monotone convergence theorem,

$$\begin{aligned} \lim_{k \to \infty} \psi_k(g,h) &= \inf\{\psi_k(g,h) \mid k \in \mathbb{N}_+\} \\ &= \min_{\alpha \in \mathbb{R}^n} \max_{v \in \mathbb{R}^q} \left[\|\alpha\|_2 g^\mathsf{T} v + \|v\|_2 h^\mathsf{T} \alpha - v^\mathsf{T} \Sigma^{-1/2} x - \frac{v^\mathsf{T} \Sigma^{-1} v}{2} \right]. \end{aligned}$$

Therefore by another application of the dominated convergence theorem, we have that:

$$\begin{aligned} &\lim_{k_1 \to \infty} \mathbb{E} \min_{\|\alpha\|_2 \leqslant \alpha_{k_1}} \max_{v \in \mathbb{R}^q} \left[\|\alpha\|_2 g^\mathsf{T} v + \|v\|_2 h^\mathsf{T} \alpha - v^\mathsf{T} \Sigma^{-1/2} x - \frac{v^\mathsf{T} \Sigma^{-1} v}{2} \right] \\ &= \mathbb{E} \lim_{k_1 \to \infty} \min_{\|\alpha\|_2 \leqslant \alpha_{k_1}} \max_{v \in \mathbb{R}^q} \left[\|\alpha\|_2 g^\mathsf{T} v + \|v\|_2 h^\mathsf{T} \alpha - v^\mathsf{T} \Sigma^{-1/2} x - \frac{v^\mathsf{T} \Sigma^{-1} v}{2} \right] \\ &= \mathbb{E} \min_{\alpha \in \mathbb{R}^n} \max_{v \in \mathbb{R}^q} \left[\|\alpha\|_2 g^\mathsf{T} v + \|v\|_2 h^\mathsf{T} \alpha - v^\mathsf{T} \Sigma^{-1/2} x - \frac{v^\mathsf{T} \Sigma^{-1} v}{2} \right]. \end{aligned} \tag{C.7}$$

Chaining together inequalities (C.5), (C.6), (C.7), we have:

$$\mathbb{E} \min_{\alpha \in \mathbb{R}^n} \|\Sigma^{1/2} W \alpha - x\|_2^2 \leqslant 2\mathbb{E} \min_{\alpha \in \mathbb{R}^n} \max_{v \in \mathbb{R}^q} \left[\|\alpha\|_2 g^\mathsf{T} v + \|v\|_2 h^\mathsf{T} \alpha - v^\mathsf{T} \Sigma^{-1/2} x - \frac{v^\mathsf{T} \Sigma^{-1} v}{2} \right]. \tag{C.8}$$

We now proceed to study the RHS of (C.8), which we denote by (AO) (the *auxiliary optimization* problem):

$$\begin{aligned} (\mathsf{AO}) &:= \min_{\alpha \in \mathbb{R}^n} \max_{v \in \mathbb{R}^q} \left[\|\alpha\|_2 g^\mathsf{T} v + \|v\|_2 h^\mathsf{T} \alpha - v^\mathsf{T} \Sigma^{-1/2} x - \frac{v^\mathsf{T} \Sigma^{-1} v}{2} \right] \\ &= \min_{\beta \geqslant 0} \min_{\theta \in [-1,1]} \max_{v \in \mathbb{R}^q} \left[\beta g^\mathsf{T} v + \beta \|v\|_2 \|h\|_2 \theta - v^\mathsf{T} \Sigma^{-1/2} x - \frac{v^\mathsf{T} \Sigma^{-1} v}{2} \right] \\ &\overset{(a)}{=} \min_{\beta \geqslant 0} \max_{v \in \mathbb{R}^q} \left[\beta g^\mathsf{T} v - \beta \|v\|_2 \|h\|_2 - v^\mathsf{T} \Sigma^{-1/2} x - \frac{v^\mathsf{T} \Sigma^{-1} v}{2} \right] \\ &\overset{(b)}{=} \min_{\beta \geqslant 0} \max_{v \in \mathbb{R}^q} \max_{\tau \geqslant 0} \left[\beta g^\mathsf{T} v - \beta \|h\|_2 \|v\|_2^2 \frac{\tau}{2} - \frac{\beta \|h\|_2}{2\tau} - v^\mathsf{T} \Sigma^{-1/2} x - \frac{v^\mathsf{T} \Sigma^{-1} v}{2} \right] \\ &= \min_{\beta \geqslant 0} \max_{\tau \geqslant 0} \max_{v \in \mathbb{R}^q} \left[\beta g^\mathsf{T} v - \beta \|h\|_2 \|v\|_2^2 \frac{\tau}{2} - \frac{\beta \|h\|_2}{2\tau} - v^\mathsf{T} \Sigma^{-1/2} x - \frac{v^\mathsf{T} \Sigma^{-1} v}{2} \right] \\ &= \min_{\beta \geqslant 0} \max_{\tau \geqslant 0} \left[-\frac{\beta \|h\|_2}{2\tau} + \frac{1}{2} \|\beta g - \Sigma^{-1/2} x\|^2_{(\Sigma^{-1} + \beta \|h\|_2 \tau I_q)^{-1}} \right]. \end{aligned}$$

Above, (b) holds by the variational form (C.4). The proof is now finished after justifying (a). First, let $h_\beta(\theta, v)$ denote the term in the bracket, so that

$$(\mathsf{AO}) = \min_{\beta \geqslant 0} \min_{\theta \in [-1,1]} \max_{v \in \mathbb{R}^q} h_\beta(\theta, v).$$

Fix a $\beta \geqslant 0$. By weak duality,

$$\min_{\theta \in [-1,1]} \max_{v \in \mathbb{R}^q} h_\beta(\theta, v) \geqslant \max_{v \in \mathbb{R}^q} \min_{\theta \in [-1,1]} h_\beta(\theta, v) = \max_{v \in \mathbb{R}^q} h_\beta(-1, v).$$

On the other hand,

$$\min_{\theta \in [-1,1]} \max_{v \in \mathbb{R}^q} h_\beta(\theta, v) \leqslant \min_{\theta \in [-1,0]} \max_{v \in \mathbb{R}^q} h_\beta(\theta, v) = \max_{v \in \mathbb{R}^q} \min_{\theta \in [-1,0]} h_\beta(\theta, v) = \max_{v \in \mathbb{R}^q} h_\beta(-1, v).$$

The first equality above is Sion's minimax theorem, since the function $\theta \mapsto h_\beta(\theta, v)$ is affine for every v and the function $v \mapsto h_\beta(\theta, v)$ is concave for $\theta \in [-1, 0]$. Therefore,

$$\min_{\theta \in [-1,1]} \max_{v \in \mathbb{R}^q} h_\beta(\theta, v) = \max_{v \in \mathbb{R}^q} h_\beta(-1, v).$$

$\square$

With Lemma C.8 in hand, we can now restate and prove Lemma 7.1.

Lemma 7.1. *Let q, n be positive integers with $q \geqslant n$ and $n \geqslant 2$. Let $W \in \mathbb{R}^{q \times n}$ have iid $N(0,1)$ entries, and let $\Sigma \in \mathbb{R}^{q \times q}$ be positive definite. Let $g \sim N(0, I_q)$ and $h \sim N(0, I_{n-1})$, with g and h independent. Also, let $\{e_i\}_{i=1}^q$ be the standard basis vectors in $\mathbb{R}^q$. We have:*

$$\mathbb{E}\operatorname{tr}((W^\mathsf{T}\Sigma W)^{-1}) \geqslant \frac{n}{\sum_{i=1}^q \mathbb{E}\min_{\beta \geqslant 0}\max_{\tau \geqslant 0}\left[-\frac{\beta\|h\|_2}{\tau} + \|\beta g - e_i\|^2_{(\Sigma^{-1} + \beta\|h\|_2 \tau I_q)^{-1}}\right]}. \tag{7.3}$$

Proof. We rewrite $\mathbb{E}\operatorname{tr}((W^\mathsf{T}\Sigma W)^{-1})$ in a way that is amenable to Lemma C.8. Let $w_1 \in \mathbb{R}^q$ denote the first column of W, so that $W = \begin{bmatrix} w_1 & W_2 \end{bmatrix}$ with $W_2 \in \mathbb{R}^{q \times (n-1)}$. We write:

$$W^\mathsf{T}\Sigma W = \begin{bmatrix} \|w_1\|_\Sigma^2 & w_1^\mathsf{T}\Sigma W_2 \\ W_2^\mathsf{T}\Sigma w_1 & W_2^\mathsf{T}\Sigma W_2 \end{bmatrix}.$$

Using the block matrix inversion formula to compute the $(1,1)$ entry of $(W^\mathsf{T}\Sigma W)^{-1}$:

$$\begin{aligned} ((W^\mathsf{T}\Sigma W)^{-1})_{11} &= (w_1^\mathsf{T}(\Sigma - \Sigma W_2^\mathsf{T}(W_2^\mathsf{T}\Sigma W_2)^{-1}W_2^\mathsf{T}\Sigma)w_1)^{-1} \\ &= (w_1^\mathsf{T}\Sigma^{1/2}(I - P_{\Sigma^{1/2}W_2})\Sigma^{1/2}w_1)^{-1} \\ &= (w_1^\mathsf{T}\Sigma^{1/2}P^\perp_{\Sigma^{1/2}W_2}\Sigma^{1/2}w_1)^{-1}. \end{aligned}$$

Since the columns of W are all independent and identically distributed, this calculation shows that the law of $((W^\mathsf{T}\Sigma W)^{-1})_{ii}$ is the same as the law of $((W^\mathsf{T}\Sigma W)^{-1})_{11}$ for all $i = 1, \ldots, n$. Therefore:

$$\mathbb{E}\operatorname{tr}((W^\mathsf{T}\Sigma W)^{-1}) = \sum_{i=1}^n \mathbb{E}((W^\mathsf{T}\Sigma W)^{-1})_{ii} = n \cdot \mathbb{E}(w_1^\mathsf{T}\Sigma^{1/2}P^\perp_{\Sigma^{1/2}W_2}\Sigma^{1/2}w_1)^{-1}$$

$$\geqslant \frac{n}{\mathbb{E}\,\mathrm{tr}(\Sigma^{1/2} P^{\perp}_{\Sigma^{1/2} W_2} \Sigma^{1/2})}.$$

The last inequality follows from Jensen's inequality combined with the independence of w_1 and W_2. By decomposing $\mathrm{tr}(\Sigma^{1/2} P^{\perp}_{\Sigma^{1/2} W_2} \Sigma^{1/2}) = \sum_{i=1}^q \|P^{\perp}_{\Sigma^{1/2} W_2} \Sigma^{1/2} e_i\|_2^2$ and observing that

$$\|P^{\perp}_{\Sigma^{1/2} W_2} x\|_2^2 = \min_{\alpha \in \mathbb{R}^{n-1}} \|\Sigma^{1/2} W_2 \alpha - x\|_2^2 \quad \forall x \in \mathbb{R}^q,$$

we have the following identity:

$$\mathbb{E}\,\mathrm{tr}(\Sigma^{1/2} P^{\perp}_{\Sigma^{1/2} W_2} \Sigma^{1/2}) = \sum_{i=1}^q \mathbb{E} \min_{\alpha_i \in \mathbb{R}^{n-1}} \|\Sigma^{1/2} W_2 \alpha_i - \Sigma^{1/2} e_i\|_2^2.$$

Invoking Lemma C.8 with $x = \Sigma^{1/2} e_i$ for $i = 1, \ldots, q$ yields

$$\mathbb{E}\,\mathrm{tr}(\Sigma^{1/2} P^{\perp}_{\Sigma^{1/2} W_2} \Sigma^{1/2}) \leqslant \sum_{i=1}^q \mathbb{E} \min_{\beta \geqslant 0} \max_{\tau \geqslant 0} \left[-\frac{\beta \|h\|_2}{\tau} + \|\beta g - e_i\|^2_{(\Sigma^{-1} + \beta \|h\|_2 \tau I_q)^{-1}} \right],$$

where $g \sim N(0, I_q)$ and $h \sim N(0, I_{n-1})$. The claim now follows. $\square$

We conclude this section with the following technical result which we will use in the sequel.

Lemma C.9. *Let $q, n \in \mathbb{N}_+$ with $q \geqslant n$ and $n \geqslant 6$, and let $\Sigma \in \mathsf{Sym}^q_{>0}$. Let $g \sim N(0, I_q)$ and $h \sim N(0, I_{n-1})$ with g and h independent. Define the random variables Z_i for $i \in \{1, \ldots, q\}$ as:*

$$Z_i := \min_{\beta \geqslant 0} \max_{\tau \geqslant 0} \left[-\frac{\beta \|h\|_2}{2\tau} + \beta^2 \|g\|^2_{(\Sigma^{-1} + \beta \|h\|_2 \tau I_q)^{-1}} + (\Sigma^{-1} + \beta \|h\|_2 \tau I_q)^{-1}_{ii} \right]. \tag{C.9}$$

Let $\{\lambda_i\}_{i=1}^q$ denote the eigenvalues of Σ^{-1} listed in decreasing order. Define n_1 and the random function $p(y)$ as:

$$n_1 := \frac{n}{64}, \quad p(y) := \sum_{i=1}^q \frac{y}{\lambda_i + y} g_i^2 - \frac{n_1}{2}. \tag{C.10}$$

There exists an event $\mathcal{E}$ (over the probability of g and h) such that the following statements hold:

(a) $\mathbb{P}(\mathcal{E}^c) \leqslant e^{-n/128} + e^{-q/16}$.

(b) On $\mathcal{E}$, there exists a unique root $y^* \in (0, \infty)$ such that $p(y^*) = 0$.

(c) The following bounds hold for $i \in \{1, \ldots, q\}$:

$$Z_i \leqslant \Sigma_{ii}, \quad \mathbf{1}\{\mathcal{E}\} Z_i \leqslant \mathbf{1}\{\mathcal{E}\} (\Sigma^{-1} + y^* I_q)^{-1}_{ii}. \tag{C.11}$$

Proof. First, we observe that we can trivially upper bound the value of Z_i by setting $\beta = 0$ and obtaining the bound $Z_i \leqslant \Sigma_{ii}$. Furthermore, by the rotational invariance of g and the fact that g and h are independent, we have that Z_i is equal in distribution to:

$$Z_i = \min_{\beta \geqslant 0} \max_{\tau \geqslant 0} \left[-\frac{\beta \|h\|_2}{2\tau} + \beta^2 \sum_{i=1}^q \frac{g_i^2}{\lambda_i + \beta \|h\|_2 \tau} + (\Sigma^{-1} + \beta \|h\|_2 \tau I_q)^{-1}_{ii} \right].$$

Define the following events:

$$\mathcal{E}_h := \left\{\|h\|_2 \geqslant \sqrt{n}/8\right\}, \quad \mathcal{E}_g := \left\{\sum_{i=1}^{q} g_i^2 \geqslant q/2\right\},$$

and put $\mathcal{E} := \mathcal{E}_h \cap \mathcal{E}_g$. Since $n \geqslant 6$, by a standard computation we have that $\mathbb{E}\|h\|_2 \geqslant \sqrt{n}/4$. Therefore, by Gaussian concentration of Lipschitz functions [cf. Wai19, Chapter 2], $\mathbb{P}(\mathcal{E}_h^c) \leqslant e^{-n/128}$. Furthermore, Lemma C.4 yields that $\mathbb{P}(\mathcal{E}_g^c) \leqslant e^{-q/16}$. By a union bound, $\mathbb{P}(\mathcal{E}^c) \leqslant e^{-n/128} + e^{-q/16}$.

We now focus on upper bounding the quantity:

$$\mathbf{1}\{\mathcal{E}\}Z_i \leqslant \mathbf{1}\{\mathcal{E}\} \min_{\beta \geqslant 0} \max_{\tau \geqslant 0} \underbrace{\left[-\frac{\beta\sqrt{n}}{16\tau} + \beta^2 \sum_{i=1}^{q} \frac{g_i^2}{\lambda_i + \beta\sqrt{n}\tau/8} + (\Sigma^{-1} + \beta\sqrt{n}\tau/8 I_q)^{-1}_{ii}\right]}_{=:\ell_i(\beta,\tau)}.$$

Let us bracket the value of the game $\min_{\beta \geqslant 0} \max_{\tau \geqslant 0} \ell_i(\beta, \tau)$. We previously noted that $\ell_i(0, \tau) = \Sigma_{ii}$ for all $\tau \in [0, \infty)$. Next, for any $\beta > 0$, $\lim_{\tau \to \infty} \ell_i(\beta, \tau) = 0$. Hence,

$$\min_{\beta \geqslant 0} \max_{\tau \geqslant 0} \ell_i(\beta, \tau) \in [0, \Sigma_{ii}].$$

Recalling from (C.10) that $n_1 = n/64$ (so that $\sqrt{n_1} = \sqrt{n}/8$) and defining f, q_i as:

$$\begin{aligned} f(x) &:= -\frac{x\sqrt{n_1}}{2} + x^2 \sum_{i=1}^{q} \frac{g_i^2}{\lambda_i + x\sqrt{n_1}}, \\ q_i(x) &:= (\Sigma^{-1} + x\sqrt{n_1})^{-1}_{ii}, \end{aligned}$$

we have that $\ell_i(\beta, \tau) = \frac{1}{\tau^2} f(\beta\tau) + q_i(\beta\tau)$.

In order to sharpen our estimate for the value of the game, we will study the positive critical points $(\beta, \tau) \in \mathbb{R}^2_{>0}$ of the game $\min_\beta \max_\tau \ell_i(\beta, \tau)$, i.e., the points $(\beta, \tau) \in \mathbb{R}^2_{>0}$ satisfying $\frac{\partial \ell_i}{\partial \beta}(\beta, \tau) = 0$ and $\frac{\partial \ell_i}{\partial \tau}(\beta, \tau) = 0$. Note that in general for a nonconvex/nonconcave game, this is *not* a necessary first order optimality condition for the global min/max value [see e.g. JNJ20, Proposition 21]. However, for every fixed $\beta > 0$, stationary points of the function $\tau \mapsto \ell_i(\beta, \tau)$ on $\mathbb{R}_{>0}$ are strictly concave by Proposition C.10. Hence, by the implicit function theorem (or alternatively [JNJ20, Theorem 23]), the first order stationarity conditions $\frac{\partial \ell_i}{\partial \beta}(\beta, \tau) = 0$ and $\frac{\partial \ell_i}{\partial \tau}(\beta, \tau) = 0$ are necessary for global min/max optimality. For $\tau \neq 0$, this yields:

$$\begin{aligned} 0 &= \tau^{-2} f'(\beta\tau)\beta - 2\tau^{-3} f(\beta\tau) + q_i'(\beta\tau)\beta, \\ 0 &= \tau^{-2} f'(\beta\tau)\tau + q_i'(\beta\tau)\tau. \end{aligned}$$

Together, these conditions imply that $f(\beta\tau) = 0$, and that the value of the game at such a critical point is $q_i(\beta\tau)$. Thus, we are interested in the positive roots of $f(x) = 0$. To proceed, recall the definition of $p(y)$ from (C.10):

$$p(y) = \sum_{i=1}^{q} \frac{y}{\lambda_i + y} g_i^2 - \frac{n_1}{2}.$$

Note that y^* is a positive root of p iff $y^*/\sqrt{n_1}$ is a positive root of f. Since $q \geqslant n$ by assumption, observe that on $\mathcal{E}$:

$$\lim_{y\to\infty} p(y) = \sum_{i=1}^{q} g_i^2 - n_1/2 \geqslant q/2 - n_1/2 \geqslant n/2 - n/64 > 0.$$

On the other hand, $p(0) = -n_1/2 < 0$. Since $p(y)$ is continuous and strictly increasing, on $\mathcal{E}$ there exists a unique $y^* \in (0, \infty)$ such that $p(y^*) = 0$. Thus,

$$\mathbf{1}\{\mathcal{E}\}Z_i \leqslant \mathbf{1}\{\mathcal{E}\}(\Sigma^{-1} + y^* I_q)^{-1}_{ii}.$$

$\square$

Proposition C.10. *Let M, A be $n \times n$ positive definite matrices, and let α, β be positive numbers. Consider the function:*

$$f(\tau) := -\frac{\alpha}{\tau} + \langle (A + \beta\tau I)^{-1}, M\rangle.$$

Suppose there exists a $\tau \in (0, \infty)$ satisfying $f'(\tau) = 0$. Then, $f''(\tau) < 0$.

Proof. A straightforward computation yields the following expressions for $f'(\tau)$ and $f''(\tau)$:

$$\begin{aligned}
f'(\tau) &= \alpha\tau^{-2} - \beta\langle (A + \beta\tau I)^{-2}, M\rangle, \\
f''(\tau) &= -2\alpha\tau^{-3} + 2\beta^2\langle (A + \beta\tau I)^{-3}, M\rangle.
\end{aligned}$$

The assumed condition $f'(\tau) = 0$ implies that:

$$\alpha\tau^{-2} = \beta\langle (A + \beta\tau I)^{-2}, M\rangle \Longrightarrow -2\alpha\tau^{-3} = -2\beta\tau^{-1}\langle (A + \beta\tau I)^{-2}, M\rangle.$$

Next, let $A = Q\Lambda Q^\mathsf{T}$ be the eigendecomposition of A, with $\Lambda = \mathrm{diag}(\{\lambda_i\}_{i=1}^n)$. For any integer k:

$$\langle (A + \beta\tau I)^{-k}, M\rangle = \mathrm{tr}(MQ(\Lambda + \beta\tau I)^{-k}Q^\mathsf{T}) = \langle QMQ^\mathsf{T}, (\Lambda + \beta\tau I)^{-k}\rangle = \sum_{i=1}^{n} \frac{(QMQ^\mathsf{T})_{ii}}{(\lambda_i + \beta\tau)^k}.$$

Now, since M is positive definite, $(QMQ^\mathsf{T})_{ii} > 0$ for all $i \in [n]$. Furthermore, since A is positive definite, $\lambda_i > 0$ for all $i \in [n]$. Hence plugging these expressions into the expression of $f''(\tau)$:

$$\begin{aligned}
f''(\tau) &= -2\beta^2 \sum_{i=1}^{n} \frac{(QMQ^\mathsf{T})_{ii}}{(\lambda_i + \beta\tau)^2\beta\tau} + 2\beta^2 \sum_{i=1}^{n} \frac{(QMQ^\mathsf{T})_{ii}}{(\lambda_i + \beta\tau)^2(\lambda_i + \beta\tau)} \\
&< -2\beta^2 \sum_{i=1}^{n} \frac{(QMQ^\mathsf{T})_{ii}}{(\lambda_i + \beta\tau)^2\beta\tau} + 2\beta^2 \sum_{i=1}^{n} \frac{(QMQ^\mathsf{T})_{ii}}{(\lambda_i + \beta\tau)^2\beta\tau} \\
&= 0.
\end{aligned}$$

$\square$

C.5 Proof of Theorem 6.2

Theorem 6.2 (Need for growth assumptions in $\mathsf{Ind\text{-}Seq\text{-}LS}$ when $m \lesssim n$)**.** *There exists universal constant c_0, c_1, and c_2 such that the following holds. Suppose that $\mathsf{P}_x = \otimes_{t \geqslant 1} N(0, 2^t \cdot I_n)$, $n \geqslant 6$, $mT \geqslant n$, and $m \leqslant c_0 n$. Then:*

$$\mathsf{R}(m, T, T; \{\mathsf{P}_x\}) \geqslant c_1 \sigma_\xi^2 \cdot \frac{p \cdot 2^{c_2 n/m}}{T}.$$

Proof. Let $\Gamma_T := \Gamma_T(\mathsf{P}_x)$. We have that $\Gamma_T = \frac{2}{T}(2^T - 1) I_n \succcurlyeq \frac{2^T}{T} I_n$. By Lemma 6.1:

$$\mathsf{R}(m, T, T; \{\mathsf{P}_x\}) \geqslant \sigma_\xi^2 p \cdot \mathbb{E} \operatorname{tr}(\Gamma_T^{1/2} (X_{m,T}^\mathsf{T} X_{m,T})^{-1} \Gamma_T^{1/2}) \geqslant \frac{\sigma_\xi^2 p}{T} \cdot \mathbb{E} \operatorname{tr}((2^{-T/2} X_{m,T}^\mathsf{T} X_{m,T} 2^{-T/2})^{-1}).$$

Since each column of $X_{m,T}$ is independent, the matrix $X_{m,T} 2^{-T/2}$ has the same distribution as $\mathsf{BDiag}(\Theta^{1/2}, m) W$, where $\Theta \in \mathbb{R}^{T \times T}$ is diagonal, $\Theta_{ii} = 2^{i-T}$ for $i \in \{1, \ldots, T\}$, and $W \in \mathbb{R}^{mT \times n}$ has iid $N(0,1)$ entries. Let $\lambda_t = 2^{T-t}$ for $t \in \{1, \ldots, T\}$. With this notation:

$$\mathbb{E} \operatorname{tr}((2^{-T/2} X_{m,T}^\mathsf{T} X_{m,T} 2^{-T/2})^{-1}) = \mathbb{E} \operatorname{tr}((W^\mathsf{T} \mathsf{BDiag}(\Theta, m) W)^{-1}).$$

Let $\{g_j\}_{j=1}^m$ be independent isotropic Gaussian random vectors in $\mathbb{R}^T$, and let $h \sim N(0, I_{n-1})$ be independent from $\{g_j\}$. Define the random variables $\{Z_i\}_{i=1}^T$ as:

$$Z_i := \min_{\beta \geqslant 0} \max_{\tau \geqslant 0} \left[-\frac{\beta \|h\|_2}{2\tau} + \beta^2 \sum_{j=1}^m \sum_{t=1}^T \frac{g_{j,t}^2}{\lambda_t + \beta \|h\|_2 \tau} + \frac{1}{\lambda_i + \beta \|h\|_2 \tau} \right]. \tag{C.12}$$

By Lemma 7.1,

$$\mathbb{E} \operatorname{tr}((W^\mathsf{T} \mathsf{BDiag}(\Theta, m) W)^{-1}) \geqslant \frac{n}{2m} \left[\sum_{i=1}^T \mathbb{E}[Z_i] \right]^{-1}.$$

Next, define

$$n_1 := \frac{n}{64}, \quad p(y) := \sum_{j=1}^m \sum_{t=1}^T \frac{y}{\lambda_t + y} g_{j,t}^2 - \frac{n_1}{2}.$$

Since $n \geqslant 6$ and $mT \geqslant n$, we can invoke Lemma C.9 to conclude there exists an event $\mathcal{E}_1$ (over the probability of $\{g_j\}$ and h) such that:

(a) on $\mathcal{E}_1$, there exists a unique root $y^* \in (0, \infty)$ such that $p(y^*) = 0$,

(b) the following inequalities holds:

$$Z_i \leqslant \frac{1}{\lambda_i}, \quad \mathbf{1}\{\mathcal{E}_1\} Z_i \leqslant \mathbf{1}\{\mathcal{E}_1\} \frac{1}{\lambda_i + y^*}, \tag{C.13}$$

(c) the following estimate holds:

$$\mathbb{P}(\mathcal{E}_1^c) \leqslant e^{-n/128} + e^{-mT/16}.$$

Now, let $c = 1/20$, and assume that $cn_1/m \geqslant 4$. We can check easily that $\lceil cn_1/m \rceil \leqslant T$. Fix a $\delta \in (0, e^{-2}]$ to be chosen later. Define the integer $T_c := \lceil cn_1/m \rceil \in \{4, \dots, T\}$, and the events (over the probability of $\{g_j\}$ and h):

$$\mathcal{E}_2^{g,T_c} := \left\{ \sum_{j=1}^m \sum_{t=1}^{T_c} g_{j,t}^2 \leqslant 5mT_c \right\}, \quad \mathcal{E}_2^{g,+} := \left\{ \max_{t=1,\dots,T} \sum_{j=1}^m g_{j,t}^2 \leqslant 2m + 4 \log\left(\frac{t^2\pi^2}{6\delta}\right) \right\}.$$

By Lemma C.4, $\mathbb{P}((\mathcal{E}_2^{g,T_c})^c) \leqslant e^{-mT_c}$. Next, Gaussian concentration for Lipschitz functions [cf Wai19, Chapter 2] yields, for any $\eta \in (0,1)$:

$$\max_{t=1,\dots,T} \mathbb{P}\left\{ \sqrt{\sum_{j=1}^m g_{j,t}^2} \geqslant \sqrt{m} + \sqrt{2\log(1/\eta)} \right\} \leqslant \eta.$$

Hence by a union bound, and the fact that $6\delta/\pi^2 \sum_{t=1}^T t^{-2} \leqslant 6\delta/\pi^2 \sum_{t=1}^\infty t^{-2} = \delta$, we have that $\mathbb{P}((\mathcal{E}_2^{g,+})^c) \leqslant \delta$. Putting $\mathcal{E} := \mathcal{E}_1 \cup \mathcal{E}_2^{g,T_c} \cup \mathcal{E}_2^{g,+}$, we have:

$$\begin{aligned} \mathbb{P}(\mathcal{E}^c) &\leqslant e^{-n/128} + e^{-mT/16} + e^{-mT_c} + \delta \\ &\leqslant e^{-n/128} + e^{-mT/16} + e^{-cn_1} + \delta. \end{aligned} \tag{C.14}$$

Next, noting that $t/2 \geqslant \log_2 \log((t+1)^2\pi^2/6)$ for all $t \geqslant 4$:

$$\begin{aligned} &\sum_{t=T_c}^{T-1} 2^{-t} \log((t+1)^2\pi^2/(6\delta)) \\ &= \sum_{t=T_c}^{T-1} 2^{-t+\log_2 \log((t+1)^2\pi^2/6)} + \log(1/\delta) \sum_{t=T_c}^{T-1} 2^{-t} \\ &\leqslant \sum_{t=T_c}^{T-1} 2^{-t/2} + \log(1/\delta) \sum_{t=T_c}^{T-1} 2^{-t} && \text{since } T_c \geqslant 4 \\ &= \sqrt{2}/(\sqrt{2}-1)(2^{-T_c/2} - 2^{-T/2}) + 2\log(1/\delta)(2^{-T_c} - 2^{-T}) \\ &\leqslant (4 + 2\log(1/\delta))2^{-T_c/2} \\ &\leqslant 4\log(1/\delta)2^{-T_c/2} && \text{since } \delta \in (0, e^{-2}). \end{aligned} \tag{C.15}$$

Now, on $\mathcal{E}$:

$$\begin{aligned} \frac{n_1}{2} &= \sum_{j=1}^m \sum_{t=1}^T \frac{y^*}{\lambda_t + y^*} g_{j,t}^2 && \text{since } p(y^*) = 0 \\ &\leqslant \sum_{j=1}^m \sum_{t=1}^{T_c} g_{j,t}^2 + y^* \sum_{j=1}^m \sum_{t=T_c}^{T-1} 2^{-t} g_{j,t+1}^2 \\ &= \sum_{j=1}^m \sum_{t=1}^{T_c} g_{j,t}^2 + y^* \sum_{t=T_c}^{T-1} 2^{-t} \left[\sum_{j=1}^m g_{j,t+1}^2 \right] \end{aligned}$$

$$
\begin{aligned}
&\leqslant 5mT_c + y^* \sum_{t=T_c}^{T-1} 2^{-t}\left[2m + 4\log((t+1)^2\pi^2/(6\delta))\right] && \text{using } \mathcal{E} \\
&\leqslant 5mT_c + 4my^* 2^{-T_c} + 16y^*\log(1/\delta)2^{-T_c/2} && \text{using (C.15)} \\
&\leqslant 5mT_c + 18my^*\log(1/\delta)2^{-T_c/2} && \text{since } \delta \in (0, e^{-2}).
\end{aligned}
$$

This inequality implies the following lower bound on y^*:

$$
\begin{aligned}
y^* &\geqslant \frac{2^{cn_1/(2m)}}{18\log(1/\delta)}\left[\frac{n_1}{2m} - 5c\frac{n_1}{m} - 5\right] \\
&= \frac{2^{cn_1/(2m)}}{18\log(1/\delta)}\left[\frac{n_1}{4m} - 5\right] && \text{since } c = 1/20 \\
&\geqslant \frac{2^{cn_1/(2m)}}{144\log(1/\delta)}\frac{n_1}{m} && \text{since } cn_1/m \geqslant 4 \implies n_1/(8m) \geqslant 5 \\
&=: \underline{y}^*.
\end{aligned}
$$

We now bound,

$$
\begin{aligned}
\sum_{i=1}^{T}\mathbb{E}[Z_i] &= \sum_{i=1}^{T}\left[\mathbb{E}[\mathbf{1}\{\mathcal{E}\}Z_i] + \mathbb{E}[\mathbf{1}\{\mathcal{E}^c\}Z_i]\right] \\
&\leqslant \sum_{i=1}^{T}\left(\mathbb{E}\left[\mathbf{1}\{\mathcal{E}\}\frac{1}{\lambda_i + y^*}\right] + \mathbb{P}(\mathcal{E}^c)\frac{1}{\lambda_i}\right) && \text{using (C.13)} \\
&\leqslant \sum_{i=1}^{T}\frac{1}{\lambda_i + \underline{y}^*} + \mathbb{P}(\mathcal{E}^c)\sum_{t=1}^{T}\frac{1}{\lambda_i} && \text{since } y^* \geqslant \underline{y}^* \text{ on } \mathcal{E} \\
&\leqslant \sum_{t=0}^{T-1}\frac{1}{2^t + \underline{y}^*} + 2\left(e^{-n/128} + e^{-mT/16} + e^{-cn_1} + \delta\right) && \text{using (C.14)} \\
&\leqslant \frac{T_c}{\underline{y}^*} + 2\cdot 2^{-T_c} + 2\left(e^{-n/128} + e^{-mT/16} + e^{-cn_1} + \delta\right) \\
&\leqslant 288c\log(1/\delta)2^{-cn_1/(2m)} + 2\cdot 2^{-cn_1/m} \\
&\qquad + 2\left(e^{-n/128} + e^{-mT/16} + e^{-cn_1} + \delta\right).
\end{aligned}
$$

Since $cn_1/m \geqslant 4$, we can choose $\delta = e^{-cn_1/(2m)} \in (0, e^{-2}]$ and obtain:

$$
\sum_{i=1}^{T}\mathbb{E}[Z_i] \leqslant 144c^2\frac{n_1}{m}2^{-cn_1/(2m)} + 2\cdot 2^{-cn_1/m} + 2\left(e^{-n/128} + e^{-mT/16} + e^{-cn_1} + e^{-cn_1/(2m)}\right).
$$

Since $1 \leqslant cn_1/(4m)$, $mT \geqslant n$, and $m \geqslant 1$, this inequality implies there exists universal positive constants c_1, c_2 such that:

$$
\sum_{i=1}^{T}\mathbb{E}[Z_i] \leqslant \frac{c_1 n}{m}2^{-c_2 n/m}.
$$

Hence:

$$\mathsf{R}(m, T, T; \{\mathsf{P}_x\}) \geqslant \frac{\sigma_\xi^2 p}{T} \frac{n}{2m} \left[\sum_{i=1}^{T} \mathbb{E}[Z_i]\right]^{-1} \geqslant \frac{\sigma_\xi^2 p}{T} \frac{n}{2m} \frac{m}{c_1 n} 2^{c_2 n/m} = \frac{\sigma_\xi^2 p 2^{c_2 n/m}}{2c_1 T}.$$

□

C.6 Block decoupling

We now use a block decoupling argument to study lower bounds on the risk. The first step is the following result, which bounds the risk from below by a particular random gramian matrix.

Lemma C.11. *Let $n = dr$ with both d, r positive integers. Define $\mathcal{I}_r := \{1, 1+r, \dots, 1+(T-1)r\}$, and let $E_{\mathcal{I}_r} \in \mathbb{R}^{T \times Tr}$ denote the linear operator which extracts the coordinates in $\mathcal{I}_r$, so that $(E_{\mathcal{I}_r} x)_i = x_{1+(i-1)r}$ for $i = 1, \dots, T$. Recall the following definitions from Equation (7.10):*

$$\begin{aligned}
\Psi_{r,T,T'} &= \mathsf{BDiag}(\Gamma_{T'}^{-1/2}(J_r), T)\mathsf{BToep}(J_r, T) \in \mathbb{R}^{Tr \times Tr}, \\
\Theta_{r,T,T'} &= E_{\mathcal{I}_r} \Psi_{r,T,T'} \Psi_{r,T,T'}^{\mathsf{T}} E_{\mathcal{I}_r}^{\mathsf{T}} \in \mathbb{R}^{T \times T}.
\end{aligned}$$

Then, for $A = \mathsf{BDiag}(J_r, d)$ we have:

$$\mathbb{E}_{\otimes_{i=1}^m \mathsf{P}_x^A}\left[\operatorname{tr}\left(\Gamma_{T'}^{1/2}(A)(X_{m,T}^{\mathsf{T}} X_{m,T})^{-1}\Gamma_{T'}^{1/2}(A)\right)\right] \geqslant \mathbb{E}\operatorname{tr}((W^{\mathsf{T}}\mathsf{BDiag}(\Theta_{r,T,T'}, m)W)^{-1}),$$

where $W \in \mathbb{R}^{mT \times d}$ is a matrix with independent $N(0,1)$ entries.

Proof. We apply Proposition C.3 with:

$$M = X_{m,T}\Gamma_{T'}^{-1/2}, \quad I = \{1, 1+r, 1+2r, \dots, 1+(d-1)r\}, \quad |I| = d.$$

Note that the block diagonal structure of A yields the same block diagonal structure on $\Gamma_{T'}$ and its inverse square root, specifically $\Gamma_{T'}(A) = \mathsf{BDiag}(\Gamma_{T'}(J_r), d)$ and $\Gamma_{T'}^{-1/2}(A) = \mathsf{BDiag}(\Gamma_{T'}^{-1/2}(J_r), d)$. Hence, it is not hard to see that the columns of ME_I^{T} are not only independent, but also identically distributed. Furthermore, the distribution of each column obeys a multivariate Gaussian in $\mathbb{R}^{mT}$. Hence, ME_I^{T} is equal in distribution to $Q_{m,T}^{1/2}W$, where $W \in \mathbb{R}^{mT \times d}$ is a matrix of iid Gaussians and $Q_{m,T} \in \mathsf{Sym}_{>0}^{mT}$ is a positive definite covariance matrix to be determined. Furthermore, because ME_I^{T} contains the vertical concatenation of m independent trajectories, $Q_{m,T}$ itself is block diagonal:

$$Q_{m,T} = \mathsf{BDiag}(Q_T, m), \quad Q_T \in \mathsf{Sym}_{>0}^{T}.$$

Let us now compute an expression for Q_T. Consider the dynamics:

$$x_{t+1}^r = J_r x_t^r + w_t^r, \quad x_0^r = 0, \quad w_t^r \sim N(0, \sigma_w^2 I_r).$$

It is not hard to see that, with $w_{0:T-1}^r = (w_0, \dots, w_{T-1}) \in \mathbb{R}^{rT}$,

$$\begin{bmatrix} \Gamma_{T'}^{-1/2}(J_r)x_1^r \\ \vdots \\ \Gamma_{T'}^{-1/2}(J_r)x_T^r \end{bmatrix} = \Psi_{r,T,T'} w_{0:T-1}^r.$$

From this, we see that every column of ME_I^{T} is equal in distribution to $E_{\mathcal{I}_r}\Psi_{r,T,T'}w^r_{0:T-1}$, and therefore has distribution $N(0, E_{\mathcal{I}_r}\Psi_{r,T,T'}\Psi_{r,T,T'}^{\mathsf{T}}E_{\mathcal{I}_r}^{\mathsf{T}})$. Therefore:

$$Q_T = E_{\mathcal{I}_r}\Psi_{r,T,T'}\Psi_{r,T,T'}^{\mathsf{T}}E_{\mathcal{I}_r}^{\mathsf{T}} = \Theta_{r,T,T'}.$$

The claim now follows. □

C.7 Eigenvalue analysis of a tridiagonal matrix

For any $T \in \mathbb{N}_+$, recall that L_T denotes the $T \times T$ lower triangle matrix with ones in the lower triangle, and $\mathsf{Tri}(a, b; T)$ denotes the symmetric $T \times T$ tri-diagonal matrix with a on the digonal and b on the lower and upper off-diagonals. In this section, we study the eigenvalues of $(L_T L_T^{\mathsf{T}})^{-1}$, which we denote by S_T:

$$S_T = (L_T L_T^{\mathsf{T}})^{-1} = \mathsf{Tri}(2, -1; T) - e_T e_T^{\mathsf{T}}. \tag{C.16}$$

Understanding the eigenvalues of this matrix will be necessary in the proof of Lemma C.15. The following result sharply characterizes the spectrum of S_T up to constant factors.

Lemma C.12. *Suppose $T \geqslant 8$. For all $k = 1, \ldots, T$, we have that:*

$$0.02\frac{k^2}{T^2} \leqslant \lambda_{T-k+1}(S_T) \leqslant \pi^2\frac{k^2}{T^2}.$$

Proof. We prove the upper bound in Proposition C.13, and the lower bound in Proposition C.14. □

The next result gives the necessary upper bounds on the eigenvalues of S_T.

Proposition C.13. *We have that:*

$$\lambda_{T-k+1}(S_T) \leqslant \pi^2\frac{k^2}{T^2}, \quad k = 1, \ldots, T.$$

Proof. By (C.16), we immediately produce a semidefinite upper bound on S_T:

$$S_T = \mathsf{Tri}(2, -1; T) - e_T e_T^{\mathsf{T}} \preccurlyeq \mathsf{Tri}(2, -1; T).$$

Therefore by the Courant min-max theorem, followed by the closed-form expression for the eigenvalues of $\mathsf{Tri}(2, -1; T)$, we have:

$$\lambda_{T-k+1}(S_T) \leqslant \lambda_{T-k+1}(\mathsf{Tri}(2, -1; T)) = 2\left(1 - \cos\left(\frac{k\pi}{T+1}\right)\right), \quad k = 1, \ldots, T.$$

Next, we have the following elementary lower bounds for $\cos(x)$ on $x \in [0, \pi]$:

$$\cos(x) \geqslant \begin{cases} 1 - x^2/2 & \text{if } x \in [0, 2\pi/3], \\ (x-\pi)^2/4 - 1 & \text{if } x \in [2\pi/3, \pi]. \end{cases}$$

Therefore, when $k \in \left\{1, \ldots, \left\lfloor \frac{2(T+1)}{3} \right\rfloor\right\}$, we immediately have that:

$$\lambda_{T-k+1}(S_T) \leqslant \pi^2\frac{k^2}{(T+1)^2}.$$

For the case when $k \in \{\lfloor \frac{2(T+1)}{3} \rfloor + 1, \dots, T\}$, we use the cosine lower bounds to bound:

$$\begin{aligned}
\lambda_{T-k+1}(S_T) &\leqslant 4 - \frac{\pi^2}{2}\left(1 - \frac{k}{T+1}\right)^2 \\
&\leqslant 4\left[1 - \left(1 - \frac{k}{T+1}\right)^2\right] \\
&= 4\left(\frac{k}{T+1}\right)\left(2 - \frac{k}{T+1}\right) \\
&= 4\left(\frac{k}{T+1}\right)\left(\frac{2(T+1)-k}{T+1}\right) \\
&\leqslant 4\left(\frac{k}{T+1}\right)\left(\frac{3k-k}{T+1}\right) && \text{since } k \geqslant 2(T+1)/3 \\
&= 8\frac{k^2}{(T+1)^2}.
\end{aligned}$$

The claim now follows by taking the maximum over the upper bounds. □

We now move to the lower bound on $\lambda_{T-k+1}(S_T)$. At this point, it would be tempting to use Weyl's inequalities, which imply that $\lambda_i(S_T) \geqslant \lambda_i(\mathsf{Tri}(2,-1;T)) - 1$. However, this bound becomes vacuous, since $\lambda_T(\mathsf{Tri}(2,-1;T)) \lesssim 1/T^2$. To get finer grained control, we need to use the eigenvalue interlacing result of [KST99]. This is done in the following result:

Proposition C.14. *Suppose that $T \geqslant 8$. We have that*

$$\lambda_{T-k+1}(S_T) \geqslant 0.02\frac{k^2}{T^2}, \quad k = 1, \dots, T.$$

Proof. The proof relies on the interlacing result from [KST99, Theorem 4.1]. However, the interlacing result does not cover the minimum eigenvalue of S_T, so we first explicitly derive a lower bound for $\lambda_{\min}(S_T)$. To do this, we note that:

$$\lambda_{\min}(S_T) = \lambda_{\min}((L_T L_T^{\mathsf{T}})^{-1}) = \frac{1}{\|L_T\|_{\mathrm{op}}^2}.$$

Letting $l_i \in \mathbb{R}^T$ denote the i-th column of L_T, by the variational form of the operator norm followed by Cauchy-Schwarz,

$$\|L_T\|_{\mathrm{op}} = \max_{\|v\|_2=1} \|L_T v\|_2 \leqslant \max_{\|v\|_2=1} \sum_{i=1}^T \|l_i\|_2 |v_i| \leqslant \sqrt{\sum_{i=1}^T \|l_i\|_2^2} = \sqrt{\sum_{i=1}^T i} = \sqrt{T(T+1)/2}.$$

Hence:

$$\lambda_{\min}(S_T) \geqslant \frac{2}{T(T+1)} \geqslant \frac{1}{T^2}.$$

Now we may proceed with the remaining eigenvalues. We can write S_T as the following block matrix, with $e_{T-1} \in \mathbb{R}^{T-1}$ denoting the $(T-1)$-th standard basis vector:

$$S_T = \begin{bmatrix} \mathsf{Tri}(2,-1;T-1) & -e_{T-1} \\ -e_{T-1}^{\mathsf{T}} & 1 \end{bmatrix}.$$

This matrix is of the form studied in [KST99, Theorem 4.1]; for what follows we will borrow their notation. Let $U_T(x)$ denote the T-th degree Chebyshev polynomial of the 2nd kind. We know that the eigenvalues of S_T are given by $\lambda = 2(1-x)$, where x are the roots of the polynomial $p_T(x)$ defined as:

$$p_T(x) := (1+2x)U_{T-1}(x) - U_{T-2}(x). \tag{C.17}$$

Therefore, letting $\psi_1 \leqslant \ldots \leqslant \psi_T$ denote the roots of (C.17) listed in increasing order, we have:

$$\lambda_i(S_T) = 2(1-\psi_i), \quad i = 1, \ldots, T.$$

Let $\eta_1 < \cdots < \eta_{T-2}$ denote the $T-2$ roots of $U_{T-2}(x)$ listed in increasing order. Put $\eta_0 := -\infty$ and $\eta_{T-1} := +\infty$. Because the roots of $U_{T-2}(x)$ are given by $x = \cos(\frac{k\pi}{T-1})$, $k = 1, \ldots, T-2$, we have that:

$$\eta_i = \cos\left(\frac{(T-1-i)\pi}{T-1}\right), \quad i = 1, \ldots, T-2.$$

[KST99, Theorem 4.1] states that there is exactly one root of $p_T(x)$ in each of the intervals (η_j, η_{j+1}) for $j \in \{0, \ldots, T-2\} \setminus \{i_\star\}$, with $i_\star$ satisfying:

$$i_\star \in \begin{cases} \{\lfloor \frac{2(T-1)}{3} \rfloor\} & \text{if } 2(T-1) \mod 3 \neq 0, \\ \{\frac{2(T-1)}{3}, \frac{2(T-1)}{3} + 1\} & \text{otherwise,} \end{cases}$$

and furthermore $(\eta_{i_\star}, \eta_{i_\star+1})$ contains exactly two roots of $p_T(x)$. Therefore, for $i \in \{i_\star+3, \ldots, T-1\}$:

$$\psi_i \leqslant \eta_{i-1} \Longrightarrow \lambda_i(S_T) \geqslant 2(1-\eta_{i-1}) = 2\left(1 - \cos\left(\frac{(T-i)\pi}{T-1}\right)\right).$$

For $i \in \{i_\star + 3, \ldots, T-1\}$, we have:

$$\frac{T-i}{T-1} \leqslant \frac{T-i_\star-3}{T-1} = \frac{T - (\frac{2(T-1)}{3} - 1) - 3}{T-1} = \frac{1}{3} - \frac{1}{T-1} \leqslant \frac{1}{3}.$$

It is elementary to check that:

$$2(1-\cos(x)) \geqslant \frac{x^2}{2} \quad \forall x \in [0, \pi/3].$$

Therefore for $i \in \{i_\star + 3, \ldots, T-1\}$,

$$\lambda_i(S_T) \geqslant \frac{\pi^2}{2}\left(\frac{T-i}{T-1}\right)^2.$$

Furthermore, for $i \in \{1, \ldots, i_\star + 2\}$,

$$\psi_i \leqslant \eta_{i_\star+1} \Longrightarrow \lambda_i(S_T) \geqslant 2(1-\eta_{i_\star+1}) = 2\left(1 - \cos\left(\frac{(T-1-i_\star-1)\pi}{T-1}\right)\right) \geqslant 2(1-\cos(\pi/21)).$$

The last inequality holds by:

$$\cos\left(\frac{(T-1-i_\star-1)\pi}{T-1}\right) \leqslant \cos\left(\frac{(T-1)-(2(T-1)/3+2)}{T-1}\pi\right) \qquad \text{since } i_\star \leqslant \frac{2(T-1)}{3} + 1$$

$$
\begin{aligned}
&= \cos\left(\left(\frac{1}{3} - \frac{2}{T-1}\right)\pi\right) \\
&\leqslant \cos(\pi/21) && \text{since } T \geqslant 8.
\end{aligned}
$$

Summarizing, we have shown that:

$$
\lambda_{T-k+1}(S_T) \geqslant \begin{cases} \frac{1}{T^2} & \text{if } k = 1, \\ \frac{\pi^2}{2}\left(\frac{k-1}{T-1}\right)^2 & \text{if } k \in \{2, \dots, T - i_\star - 2\}, \\ 2(1-\cos(\pi/21)) & \text{if } k \in \{T - i_\star - 1, \dots, T\}. \end{cases}
$$

Since $\frac{k-1}{T-1} \geqslant \frac{k}{2T}$ when $k \geqslant 2$, and since $2(1-\cos(\pi/21)) \geqslant 2(1-\cos(\pi/21))\frac{k^2}{T^2}$ trivially, we have shown the desired conclusion:

$$
\lambda_{T-k+1}(S_T) \geqslant \min\left\{1, \frac{\pi^2}{8}, 2(1-\cos(\pi/21))\right\}\frac{k^2}{T^2} \geqslant 0.02\frac{k^2}{T^2}, \quad k = 1, \dots, T.
$$

□

C.8 A risk lower bound in the few trajectories regime

Lemma C.15. *There exist universal positive constants c_0, c_1, c_2, and c_3 such that the following is true. Suppose $\mathcal{A} \subseteq \mathbb{R}^{n \times n}$ is any set containing I_n. Let $T \geqslant c_0$, $n \geqslant c_1$, $mT \geqslant n$, and $m \leqslant c_2 n$. We have that:*

$$
\mathsf{R}(m, T, T'; \{\mathsf{P}_x^A \mid A \in \mathcal{A}\}) \geqslant c_3 \sigma_\xi^2 p \cdot \frac{n^2}{m^2 T} \cdot \frac{T'}{T}.
$$

Proof. Let $\{g_j\}_{j=1}^m$ be independent $N(0, I_T)$ random vectors, and let $h \sim N(0, I_{n-1})$ be independent from $\{g_j\}$. Let $\{\lambda_t\}_{t=1}^T$ denote the eigenvalues of $\Theta_{1,T,T}^{-1}$ listed in decreasing order. Define the random variables $\{Z_i\}_{i=1}^T$ as:

$$
Z_i := \min_{\beta \geqslant 0} \max_{\tau \geqslant 0} \left[-\frac{\beta\|h\|_2}{2\tau} + \beta^2 \sum_{j=1}^m \sum_{t=1}^T \frac{g_{j,t}^2}{\lambda_t + \beta\|h\|_2 \tau} + (\Theta_{1,T,T}^{-1} + \beta\|h\|_2 \tau I_T)_{ii}^{-1} \right]. \tag{C.18}
$$

We now lower bound the minimax risk as follows:

$$
\begin{aligned}
&\mathsf{R}(m, T, T'; \{\mathsf{P}_x^{I_n}\}) \\
&\geqslant \sigma_\xi^2 p \cdot \mathbb{E}_{\otimes_{i=1}^m \mathsf{P}_x^{I_n}}\left[\operatorname{tr}\left(\Gamma_{T'}(I_n)^{1/2}(X_{m,T}^\mathsf{T} X_{m,T})^{-1}\Gamma_{T'}(I_n)^{1/2}\right)\right] && \text{by Lemma 6.1} \\
&\geqslant \sigma_\xi^2 p \cdot \mathbb{E}\operatorname{tr}((W^\mathsf{T} \mathsf{BDiag}(\Theta_{1,T,T'}, m) W)^{-1}) && \text{by Lemma C.11} \\
&= \sigma_\xi^2 p \cdot \frac{T'+1}{T+1} \cdot \mathbb{E}\operatorname{tr}((W^\mathsf{T} \mathsf{BDiag}(\Theta_{1,T,T}, m) W)^{-1}) && \text{using (7.11)} \\
&\geqslant \sigma_\xi^2 p \cdot \frac{T'}{2T} \cdot \mathbb{E}\operatorname{tr}((W^\mathsf{T} \mathsf{BDiag}(\Theta_{1,T,T}, m) W)^{-1}) \\
&\geqslant \sigma_\xi^2 p \cdot \frac{T'}{2T} \cdot \frac{n}{2m} \cdot \left[\sum_{i=1}^T \mathbb{E}[Z_i]\right]^{-1} && \text{by Lemma 7.1.}
\end{aligned} \tag{C.19}
$$

Next, define:

$$n_1 := \frac{n}{64}, \quad p(y) := \sum_{j=1}^{m} \sum_{t=1}^{T} \frac{y}{\lambda_t + y} g_{j,t}^2 - \frac{n_1}{2}.$$

Assuming that $c_1 \geqslant 6$ so that $n \geqslant 6$ and $mT \geqslant n$, we can invoke Lemma C.9 to conclude there exists an event $\mathcal{E}_1$ (over the probability of $\{g_j\}$ and h) such that:

(a) on $\mathcal{E}_1$, there exists a unique root $y^* \in (0, \infty)$ such that $p(y^*) = 0$,

(b) the following inequalities holds:

$$Z_i \leqslant (\Theta_{1,T,T})_{ii}, \quad \mathbf{1}\{\mathcal{E}_1\} Z_i \leqslant \mathbf{1}\{\mathcal{E}_1\} \frac{1}{\lambda_i + y^*}, \tag{C.20}$$

(c) the following estimate holds:

$$\mathbb{P}(\mathcal{E}_1^c) \leqslant e^{-n/128} + e^{-mT/16}.$$

The remainder of the proof is to estimate a lower bound on y^*. Towards this goal, we define an auxiliary function:

$$\tilde{p}(y) := \mathbb{E}[p_1(y)] = m \sum_{t=1}^{T} \frac{y}{\lambda_t + y} - \frac{n_1}{2}.$$

Let $\bar{y}^*$ be the unique solution to $\tilde{p}(y) = 0$. A unique root exists because $\tilde{p}(0) < 0$, $\lim_{y\to\infty} \tilde{p}(y) = mT - n_1/2 \geqslant n - n/64 > 0$, and $\tilde{p}$ is continuous and strictly increasing. We derive a lower bound on y^* through a lower bound on $\bar{y}^*$. For any fixed $\alpha > 0$, the function $x \mapsto \frac{x}{\alpha + x}$ is monotonically increasing and concave on $\mathbb{R}_{>0}$. Therefore, the function $p(y)$ is monotonically increasing and concave on $\mathbb{R}_{>0}$. By Proposition C.2, the root of the linear approximation to $p(y)$ at $\bar{y}^*$ is a lower bound to y^*:

$$\mathbf{1}\{\mathcal{E}_1\} y^* \geqslant \mathbf{1}\{\mathcal{E}_1\} \left[\bar{y}^* - \frac{p(\bar{y}^*)}{p'(\bar{y}^*)} \right]. \tag{C.21}$$

Equation (C.21) is a crucial step for the proof, because it turns analyzing y^*, which is the root of a random function, into analyzing the pointwise evaluation of a random function on a deterministic quantity. To lower bound the RHS, we need a upper bound on $p(\bar{y}^*)$ and lower bounds on both $\bar{y}^*$ and $p'(\bar{y}^*)$.

Upper and lower bounds on $\bar{y}^*$. We first derive a crude upper bound by Jensen's inequality. Observe that $\tilde{p}(\bar{y}^*) = 0$ implies that:

$$mT - \frac{n_1}{2} = m \sum_{t=1}^{T} \frac{\lambda_t}{\lambda_t + \bar{y}^*}.$$

The function $x \mapsto x/(x+\bar{y}^*)$ is concave on $\mathbb{R}_{>0}$. Let $\bar{\lambda} := \frac{1}{T}\sum_{t=1}^T \lambda_t$. Jensen's inequality states that $T\frac{\bar{\lambda}}{\bar{\lambda}+\bar{y}^*} \geqslant \sum_{t=1}^T \frac{\lambda_t}{\lambda_t+\bar{y}^*}$. Therefore:

$$1 - \frac{n_1}{2mT} \leqslant \frac{\bar{\lambda}}{\bar{\lambda}+\bar{y}^*} \Longrightarrow \bar{y}^* \leqslant \bar{\lambda}\frac{n_1}{2mT}\frac{1}{1-n_1/(2mT)}.$$

Recalling the definition of S_T from (C.16), we can immediately bound

$$\bar{\lambda} = \frac{1}{T}\sum_{t=1}^T \lambda_t = \frac{1}{T}\operatorname{tr}(\Theta_{1,T,T}^{-1}) = \frac{1}{T}\operatorname{tr}\left(\frac{T+1}{2}S_T\right) \leqslant \operatorname{tr}(S_T) \leqslant 2T.$$

Therefore, since $mT \geqslant n$,

$$\bar{y}^* \leqslant \frac{n_1}{m}\frac{1}{1-n_1/(2mT)} \leqslant \frac{2n_1}{m}.$$

Now for the lower bound on $\bar{y}^*$. Noting that $\lambda_{T-k+1} = \lambda_{T-k+1}(\Theta_{1,T,T}^{-1}) = \frac{T+1}{2}\lambda_{T-k+1}(S_T)$, Corollary C.12 implies (assuming that $c_0 \geqslant 8$ so $T \geqslant 8$) that

$$0.01\frac{k^2}{T} \leqslant \lambda_{T-k+1} \leqslant \pi^2\frac{k^2}{T}, \quad k = 1, \ldots, T. \tag{C.22}$$

Therefore, $\tilde{p}(\bar{y}^*) = 0$ implies that:

$$\begin{aligned}\frac{1}{\bar{y}^*} &= \frac{2m}{n_1}\sum_{t=1}^T \frac{1}{\lambda_t+\bar{y}^*} \leqslant \frac{2m}{n_1}\sum_{t=1}^T \frac{1}{0.01t^2/T+\bar{y}^*} \\ &\leqslant \frac{2m}{n_1}\int_0^T \frac{1}{0.01x^2/T+\bar{y}^*}\,\mathrm{d}x = \frac{20m\sqrt{T}}{n_1\sqrt{\bar{y}^*}}\tan^{-1}\left(\frac{\sqrt{T}}{10\sqrt{\bar{y}^*}}\right) \leqslant \frac{10\pi m\sqrt{T}}{n_1\sqrt{\bar{y}^*}}.\end{aligned}$$

Solving for $\bar{y}^*$ yields:

$$\bar{y}^* \geqslant \frac{1}{100\pi^2}\frac{n_1^2}{m^2T}.$$

Next, we use this lower bound on $\bar{y}^*$ to bootstrap our upper bound $\bar{y}^* \leqslant 2n_1/m$ into something stronger. Using the upper bounds on λ_t from (C.22),

$$\begin{aligned}\frac{1}{\bar{y}^*} &= \frac{2m}{n_1}\sum_{t=1}^T \frac{1}{\lambda_t+\bar{y}^*} \geqslant \frac{2m}{n_1}\sum_{t=1}^T \frac{1}{\pi^2t^2/T+\bar{y}^*} \geqslant \frac{2m}{n_1}\int_1^{T+1} \frac{1}{\pi^2x^2/T+\bar{y}^*}\,\mathrm{d}x \\ &= \frac{2m\sqrt{T}}{\pi n_1\sqrt{\bar{y}^*}}\left[\tan^{-1}\left(\frac{(T+1)\pi}{\sqrt{T\bar{y}^*}}\right) - \tan^{-1}\left(\frac{\pi}{\sqrt{T\bar{y}^*}}\right)\right].\end{aligned}$$

The function $\tan^{-1}(x)$ is increasing. Using the $\bar{y}^* \leqslant 2n_1/m$ upper bound and the assumption that $mT \geqslant n$,

$$\frac{(T+1)\pi}{\sqrt{T\bar{y}^*}} \geqslant \pi\sqrt{\frac{mT}{2n_1}} \geqslant \pi\sqrt{32} \Longrightarrow \tan^{-1}\left(\frac{(T+1)\pi}{\sqrt{T\bar{y}^*}}\right) \geqslant \tan^{-1}(\pi\sqrt{32}).$$

On the other handing, using the bound $\bar{y}^* \geqslant \frac{1}{100\pi^2}\frac{n_1^2}{m^2 T}$ and the assumption that $m \leqslant \sqrt{2}n/320$,

$$\frac{\pi}{\sqrt{T\bar{y}^*}} \leqslant 10\pi\frac{m}{n_1} \leqslant \pi\sqrt{32}/2 \Longrightarrow \tan^{-1}\left(\frac{\pi}{\sqrt{T\bar{y}^*}}\right) \leqslant \tan^{-1}(\pi\sqrt{32}/2).$$

Combining these inequalities:

$$\frac{1}{\bar{y}^*} \geqslant \frac{2m\sqrt{T}}{\pi n_1\sqrt{\bar{y}^*}}\left[\tan^{-1}(\pi\sqrt{32}) - \tan^{-1}(\pi\sqrt{32}/2)\right] \geqslant \frac{2 \cdot 0.05}{\pi}\frac{m\sqrt{T}}{n_1\sqrt{\bar{y}^*}} \Longrightarrow \bar{y}^* \leqslant 791\pi^2\frac{n_1^2}{m^2 T}.$$

Therefore we have the following upper and lower bounds on $\bar{y}^*$:

$$\frac{1}{100\pi^2}\frac{n_1^2}{m^2 T} \leqslant \bar{y}^* \leqslant \min\left\{791\pi^2\frac{n_1^2}{m^2 T}, 2\frac{n_1}{m}\right\}. \tag{C.23}$$

For the remainder of the proof, in order to avoid precisely tracking constants, we let c_0, c_1, c_2, c_3 be any positive universal constants such that:

$$c_0\frac{k^2}{T} \leqslant \lambda_{T-k+1} \leqslant c_1\frac{k^2}{T}, \quad k = 1, \ldots, T, \tag{C.24}$$

$$c_2\frac{n_1^2}{m^2 T} \leqslant \bar{y}^* \leqslant c_3\frac{n_1^2}{m^2 T}. \tag{C.25}$$

Equations (C.22) and (C.23) give one valid setting of these constants.

Upper bound on $p(\bar{y}^*)$. To upper bound $p(\bar{y}^*)$, we note that:

$$\begin{aligned}
p(\bar{y}^*) &= \sum_{j=1}^m\sum_{t=1}^T \frac{\bar{y}^*}{\lambda_t + \bar{y}^*}g_{j,t}^2 - \frac{n_1}{2} \\
&= \sum_{j=1}^m\sum_{t=1}^T \frac{\bar{y}^*}{\lambda_t + \bar{y}^*}(g_{i,j}^2 - 1) + \sum_{j=1}^m\sum_{t=1}^T \frac{\bar{y}^*}{\lambda_t + \bar{y}^*} - \frac{n_1}{2} \\
&= \sum_{j=1}^m\sum_{t=1}^T \frac{\bar{y}^*}{\lambda_t + \bar{y}^*}(g_{i,j}^2 - 1) && \text{since } \tilde{p}(\bar{y}^*) = 0.
\end{aligned}$$

Therefore, by Lemma C.4,

$$\mathbb{P}\left(p(\bar{y}^*) > 2\sqrt{t}\sqrt{m\sum_{t=1}^T\left(\frac{\bar{y}^*}{\lambda_t + \bar{y}^*}\right)^2} + 2t\max_{t=1,\ldots,T}\frac{\bar{y}^*}{\lambda_t + \bar{y}^*}\right) \leqslant e^{-t} \quad \forall t > 0. \tag{C.26}$$

We upper bound:

$$\begin{aligned}
m\sum_{t=1}^T\left(\frac{\bar{y}^*}{\lambda_t + \bar{y}^*}\right)^2 &\leqslant m\sum_{t=1}^T\left(\frac{\bar{y}^*}{c_0 t^2/T + \bar{y}^*}\right)^2 && \text{using (C.24)} \\
&\leqslant m\int_0^T\left(\frac{\bar{y}^*}{c_0 x^2/T + \bar{y}^*}\right)^2 \mathrm{d}x
\end{aligned}$$

$$
\begin{aligned}
&= \frac{m(\bar{y}^*)^2 T}{2c_0 T\bar{y}^* + 2(\bar{y}^*)^2} + \frac{\sqrt{T\bar{y}^*}}{2\sqrt{c_0}} \tan^{-1}\left(\sqrt{\frac{c_0 T}{\bar{y}^*}}\right) \\
&\leqslant \frac{m\bar{y}^*}{2c_0} + \frac{\pi\sqrt{T\bar{y}^*}}{4\sqrt{c_0}} \\
&\leqslant \frac{c_3}{2c_0}\frac{n_1^2}{mT} + \frac{\pi}{4}\sqrt{\frac{c_3}{c_0}\frac{n_1}{m}} && \text{using (C.25)} \\
&= \left[\frac{c_3}{128c_0} + \frac{\pi}{4}\sqrt{\frac{c_3}{c_0}}\right] n_1 && \text{since } mT \geqslant n \text{ and } m \geqslant 1 \\
&=: c_4 n_1. && \text{(C.27)}
\end{aligned}
$$

Next, we immediately have:

$$
\max_{t=1,\ldots,T} \frac{\bar{y}^*}{\lambda_t + \bar{y}^*} \leqslant 1. \tag{C.28}
$$

Thus, combining (C.26), (C.27), and (C.28), we have:

$$
\mathbb{P}\left(p(\bar{y}^*) > 2\sqrt{t_u}\sqrt{c_4 n_1} + 2t_u\right) \leqslant e^{-t_u} \; \forall t_u > 0. \tag{C.29}
$$

Lower bound on $p'(\bar{y}^*)$. Differentiating $p(y)$ yields:

$$
p'(y) = \sum_{j=1}^{m}\sum_{t=1}^{T} \frac{\lambda_t}{(\lambda_t + y)^2} g_{j,t}^2.
$$

Applying Lemma C.4 yields,

$$
\mathbb{P}\left(p'(\bar{y}^*) < m\sum_{t=1}^{T}\frac{\lambda_t}{(\lambda_t+\bar{y}^*)^2} - 2\sqrt{t}\sqrt{m\sum_{t=1}^{T}\frac{\lambda_t^2}{(\lambda_t+\bar{y}^*)^4}}\right) \leqslant e^{-t} \quad \forall t > 0. \tag{C.30}
$$

Our first goal is to lower bound $m\sum_{t=1}^{T}\frac{\lambda_t}{(\lambda_t+\bar{y}^*)^2}$. The function $x \mapsto x/(x+\bar{y}^*)^2$ is increasing when $x \in [0, \bar{y}^*]$ and decreasing when $x \in (\bar{y}^*, \infty)$. Let $t^* \in \{0, \ldots, T\}$ be such that $c_1 t^2/T \leqslant \bar{y}^*$ for $t \in \{1, \ldots, t^*\}$ and $c_1 t^2/T > \bar{y}^*$ for $t \in \{t^*+1, \ldots, T\}$ ($t^* = 0$ if $c_1/T > \bar{y}^*$). We write:

$$
\begin{aligned}
m\sum_{t=1}^{T}\frac{\lambda_t}{(\lambda_t+\bar{y}^*)^2} &\geqslant \frac{c_0}{c_1} m \sum_{t=1}^{T} \frac{c_1 t^2/T}{(c_1 t^2/T + \bar{y}^*)^2} && \text{using (C.24)} \\
&= \frac{c_0}{c_1} m \left[\sum_{t=1}^{t^*}\frac{c_1 t^2/T}{(c_1 t^2/T + \bar{y}^*)^2} + \sum_{t=t^*+1}^{T}\frac{c_1 t^2/T}{(c_1 t^2/T + \bar{y}^*)^2}\right] \\
&\geqslant \frac{c_0}{c_1} m \left[\int_0^{t^*}\frac{c_1 x^2/T}{(c_1 x^2/T + \bar{y}^*)^2}\,\mathrm{d}x + \int_{t^*+1}^{T+1}\frac{c_1 x^2/T}{(c_1 x^2/T + \bar{y}^*)^2}\,\mathrm{d}x\right] \\
&= \frac{c_0}{c_1} m \left[\int_0^{T+1}\frac{c_1 x^2/T}{(c_1 x^2/T + \bar{y}^*)^2}\,\mathrm{d}x - \int_{t^*}^{t^*+1}\frac{c_1 x^2/T}{(c_1 x^2/T + \bar{y}^*)^2}\,\mathrm{d}x\right].
\end{aligned}
$$

The function $z \mapsto \frac{z}{(z+\bar{y}^*)^2}$ is upper bounded by $\frac{1}{4\bar{y}^*}$. Therefore,

$$\int_{t^*}^{t^*+1} \frac{c_1 x^2/T}{(c_1 x^2/T + \bar{y}^*)^2} \, \mathrm{d}x \leqslant \frac{1}{4\bar{y}^*} \leqslant \frac{1}{4c_2} \frac{m^2 T}{n_1^2}.$$

Next,

$$\begin{aligned}
\int_0^{T+1} \frac{c_1 x^2/T}{(c_1 x^2/T + \bar{y}^*)^2} \, \mathrm{d}x &= c_1 T \left[\frac{1}{2c_1^{3/2}\sqrt{T\bar{y}^*}} \tan^{-1}\left(\frac{(T+1)\sqrt{c_1}}{\sqrt{T\bar{y}^*}} \right) - \frac{T+1}{2c_1^2 (T+1)^2 + 2c_1 T \bar{y}^*} \right] \\
&\geqslant c_1 T \left[\frac{m}{2c_1^{3/2}\sqrt{c_3} n_1} \tan^{-1}\left(\frac{(T+1)\sqrt{c_1}}{\sqrt{T\bar{y}^*}} \right) - \frac{1}{2c_1^2 T} \right] \\
&\geqslant c_1 T \left[\frac{m}{2c_1^{3/2}\sqrt{c_3} n_1} \tan^{-1}\left(64\sqrt{\frac{c_1}{c_3}} \right) - \frac{1}{2c_1^2 T} \right].
\end{aligned}$$

The last inequality holds because:

$$\frac{(T+1)\sqrt{c_1}}{\sqrt{T\bar{y}^*}} \geqslant (T+1)\sqrt{\frac{c_1}{c_3}\frac{m}{n_1}} \geqslant \sqrt{\frac{c_1}{c_3}\frac{mT}{n_1}} \geqslant 64\sqrt{\frac{c_1}{c_3}}.$$

Above, the first inequality holds using (C.25) and the last inequality holds since $mT \geqslant n$. Therefore, assuming that $mT \geqslant 2\sqrt{\frac{c_3}{c_1}} \frac{1}{\tan^{-1}(64\sqrt{c_1/c_3})} n_1$,

$$\int_0^{T+1} \frac{c_1 x^2/T}{(c_1 x^2/T + \bar{y}^*)^2} \, \mathrm{d}x \geqslant \frac{\tan^{-1}(64\sqrt{c_1/c_3})}{4\sqrt{c_1 c_3}} \frac{mT}{n_1}.$$

Combining these inequalities, assuming that $m \leqslant \frac{c_2}{2\sqrt{c_1 c_3}} \tan^{-1}(64\sqrt{c_1/c_3}) n_1$, we have:

$$m \sum_{t=1}^{T} \frac{\lambda_t}{(\lambda_t + \bar{y}^*)^2} \geqslant \frac{c_0}{c_1} m \left[\frac{\tan^{-1}(64\sqrt{c_1/c_3})}{4\sqrt{c_1 c_3}} \frac{mT}{n_1} - \frac{m^2 T}{4c_2 n_1^2} \right] \geqslant \frac{c_0 \tan^{-1}(64\sqrt{c_1/c_3})}{8c_1^{3/2}\sqrt{c_3}} =: c_5 \frac{m^2 T}{n_1}. \tag{C.31}$$

Next, we turn to upper bounding $m\sum_{t=1}^{T} \frac{\lambda_t^2}{(\lambda_t+\bar{y}^*)^4}$. Again the function $x \mapsto x^2/(x+\bar{y}^*)^4$ is increasing when $x \in [0, \bar{y}^*]$ and decreasing when $x \in (\bar{y}^*, \infty)$, and therefore $x^2/(x+\bar{y}^*)^4 \leqslant \frac{1}{16(\bar{y}^*)^2}$ for all $x \geqslant 0$. Let $t^* \in \{0, \ldots, T\}$ be such that $c_0 t^2/T \leqslant \bar{y}^*$ for $t \in \{1, \ldots, t^*\}$ and $c_0 t^2/T > \bar{y}^*$ for $t \in \{t^*+1, \ldots, T\}$. In the case when $c_0/T > \bar{y}^*$, we set $t^* = 0$. We have:

$$\begin{aligned}
& m \sum_{t=1}^{T} \frac{\lambda_t^2}{(\lambda_t + \bar{y}^*)^4} \\
&\leqslant \frac{c_1^2}{c_0^2} m \sum_{t=1}^{T} \frac{(c_0 t^2/T)^2}{(c_0 t^2/T + \bar{y}^*)^4} && \text{using (C.24)} \\
&= \frac{c_1^2}{c_0^2} m \left[\sum_{t=1}^{t^*-1} \frac{(c_0 t^2/T)^2}{(c_0 t^2/T + \bar{y}^*)^4} + \sum_{t=t^*+2}^{T} \frac{(c_0 t^2/T)^2}{(c_0 t^2/T + \bar{y}^*)^4} + \frac{(c_0 (t^*)^2/T)^2}{(c_0 (t^*)^2/T + \bar{y}^*)^4} + \frac{(c_0 (t^*+1)^2/T)^2}{(c_0 (t^*+1)^2/T + \bar{y}^*)^4} \right]
\end{aligned}$$

$$\leqslant \frac{c_1^2}{c_0^2} m \left[\int_1^{t^*} \frac{(c_0x^2/T)^2}{(c_0x^2/T+\bar{y}^*)^4}\,\mathrm{d}x + \int_{t^*+1}^{T} \frac{(c_0x^2/T)^2}{(c_0x^2/T+\bar{y}^*)^4}\,\mathrm{d}x + \frac{(c_0(t^*)^2/T)^2}{(c_0(t^*)^2/T+\bar{y}^*)^4} + \frac{(c_0(t^*+1)^2/T)^2}{(c_0(t^*+1)^2/T+\bar{y}^*)^4}\right]$$
$$\leqslant \frac{c_1^2}{c_0^2} m \left[\int_0^{T} \frac{(c_0x^2/T)^2}{(c_0x^2/T+\bar{y}^*)^4}\,\mathrm{d}x + \frac{(c_0(t^*)^2/T)^2}{(c_0(t^*)^2/T+\bar{y}^*)^4} + \frac{(c_0(t^*+1)^2/T)^2}{(c_0(t^*+1)^2/T+\bar{y}^*)^4}\right]$$
$$\leqslant \frac{c_1^2}{c_0^2} m \left[\int_0^{T} \frac{(c_0x^2/T)^2}{(c_0x^2/T+\bar{y}^*)^4}\,\mathrm{d}x + \frac{1}{8(\bar{y}^*)^2}\right] \qquad \text{since } \max_{x>0} \frac{x^2}{(x+\bar{y}^*)^4} \leqslant \frac{1}{16(\bar{y}^*)^2}$$
$$\leqslant \frac{c_1^2}{c_0^2} m \left[\int_0^{T} \frac{(c_0x^2/T)^2}{(c_0x^2/T+\bar{y}^*)^4}\,\mathrm{d}x + \frac{1}{8c_2^2}\frac{m^4T^2}{n_1^4}\right].$$

We now bound:

$$\int_0^{T} \frac{(c_0x^2/T)^2}{(c_0x^2/T+\bar{y}^*)^4}\,\mathrm{d}x = c_0^2T^2\left[\frac{(3c_0T+\bar{y}^*)(c_0T-3\bar{y}^*)}{48c_0^2T\bar{y}_*(c_0T+\bar{y}^*)^3} + \frac{\tan^{-1}\left(\sqrt{\frac{c_0T}{\bar{y}^*}}\right)}{16c_0^{5/2}T^{3/2}(\bar{y}^*)^{3/2}}\right]$$
$$\leqslant c_0^2T^2\left[\frac{1}{16c_0\bar{y}^*(c_0T+\bar{y}^*)^2} + \frac{\pi}{32c_0^{5/2}T^{3/2}(\bar{y}^*)^{3/2}}\right]$$
$$\leqslant c_0^2T^2\left[\frac{1}{16c_0^3\bar{y}^*T^2} + \frac{\pi}{32c_0^{5/2}T^{3/2}(\bar{y}^*)^{3/2}}\right]$$
$$\leqslant c_0^2T^2\left[\frac{1}{16c_0^3c_2}\frac{m^2}{n_1^2T} + \frac{\pi}{32c_0^{5/2}c_2^{3/2}}\frac{m^3}{n_1^3}\right] \qquad \text{using (C.25)}$$
$$\leqslant \left[\frac{1}{1024c_0c_2} + \frac{\pi}{32c_0^{1/2}c_2^{3/2}}\right]\frac{m^3T^2}{n_1^3} \qquad \text{since } mT \geqslant n.$$

Combining these inequalities, assuming that $m \leqslant n_1$:

$$m\sum_{t=1}^{T}\frac{\lambda_t^2}{(\lambda_t+\bar{y}^*)^4} \leqslant \frac{c_1^2}{c_0^2}\left[\left[\frac{1}{1024c_0c_2} + \frac{\pi}{32c_0^{1/2}c_2^{3/2}}\right]\frac{m^4T^2}{n_1^3} + \frac{1}{8c_2^2}\frac{m^5T^2}{n_1^4}\right]$$
$$\leqslant \frac{c_1^2}{c_0^2}\left[\frac{1}{1024c_0c_2} + \frac{\pi}{32c_0^{1/2}c_2^{3/2}} + \frac{1}{8c_2^2}\right]\frac{m^4T^2}{n_1^3} \qquad \text{since } m \leqslant n_1$$
$$=: c_6\frac{m^4T^2}{n_1^3}. \tag{C.32}$$

Combining (C.30), (C.31), and (C.32) yields

$$\mathbb{P}\left(p'(\bar{y}^*) < c_5\frac{m^2T}{n_1} - 2\sqrt{t_\ell}\sqrt{c_6}\frac{m^2T}{n_1^{3/2}}\right) \leqslant e^{-t_\ell}\ \forall t_\ell > 0. \tag{C.33}$$

Lower bounds on y^*. We now combine (C.29) with (C.33) to established a lower bound on y^*. Equations (C.21) and (C.25) imply that:

$$y^* \geqslant \bar{y}^* - \frac{p(\bar{y}^*)}{p'(\bar{y}^*)} \geqslant \frac{c_2n_1^2}{m^2T} - \frac{p(\bar{y}^*)}{p'(\bar{y}^*)}.$$

We first set $t_\ell = \frac{c_5^2}{16c_6} n_1$, so that by (C.33),

$$\mathbb{P}\left(p'(\bar{y}^*) < \frac{c_5}{2}\frac{m^2T}{n_1}\right) \leqslant e^{-\frac{c_5^2}{16c_6}n_1}.$$

We next set $t_u = \beta n_1$ for a $\beta > 0$ to be specified. By (C.29),

$$\mathbb{P}\left(p(\bar{y}^*) > 2(\sqrt{c_4\beta} + \beta)n_1\right) \leqslant e^{-\beta n_1}.$$

Let $\mathcal{E}_2$ denote the event:

$$\mathcal{E}_2 := \left\{p'(\bar{y}^*) \geqslant \frac{c_5}{2}\frac{m^2T}{n_1}\right\} \cap \left\{p(\bar{y}^*) \leqslant 2(\sqrt{c_4\beta} + \beta)n_1\right\}.$$

By a union bound, $\mathbb{P}(\mathcal{E}_2^c) \leqslant e^{-\frac{c_5^2}{16c_6}n_1} + e^{-\beta n_1}$. Furthermore,

$$\mathbf{1}\{\mathcal{E}_2\}\left[\frac{c_2 n_1^2}{m^2T} - \frac{p(\bar{y}^*)}{p'(\bar{y}^*)}\right] \geqslant \mathbf{1}\{\mathcal{E}_2\}\left[c_2 - \frac{4(\sqrt{c_4\beta}+\beta)}{c_5}\right]\frac{n_1^2}{m^2T}.$$

Setting $\beta = c_7 := \min\left\{\frac{c_2c_5}{16}, \frac{c_2^2c_5^2}{16^2c_4}\right\}$, we have that $c_2 - \frac{4(\sqrt{c_4\beta}+\beta)}{c_5} \geqslant c_2/2$, and therefore from (C.21),

$$\mathbf{1}\{\mathcal{E}_1\}y^* \geqslant \mathbf{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\}\left[\frac{c_2 n_1^2}{m^2T} - \frac{p(\bar{y}^*)}{p'(\bar{y}^*)}\right] \geqslant \mathbf{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\}\frac{c_2}{2}\frac{n_1^2}{m^2T}. \tag{C.34}$$

Finishing the proof. Define $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2$ and define $\underline{y}^* := \frac{c_2}{2}\frac{n_1^2}{m^2T}$. By a union bound,

$$\begin{aligned}
\mathbb{P}(\mathcal{E}) &\leqslant e^{-n/128} + e^{-mT/16} + e^{-\frac{c_5^2}{16c_6}n_1} + e^{-c_7 n_1} \\
&\leqslant e^{-n/128} + e^{-n/16} + e^{-\frac{c_5^2}{1024c_6}n} + e^{-\frac{c_7}{64}n} && \text{since } mT \geqslant n \\
&\leqslant 4\exp\left(-\min\left\{\frac{1}{128}, \frac{1}{16}, \frac{c_5^2}{1024c_6}, \frac{c_7}{64}\right\}n\right) =: 4e^{-c_8 n}. && (C.35)
\end{aligned}$$

From (C.20), since $y^* \geqslant \underline{y}^*$ on $\mathcal{E}$ by (C.34),

$$\mathbf{1}\{\mathcal{E}\}Z_i \leqslant \mathbf{1}\{\mathcal{E}\}\frac{1}{\lambda_i + y^*} \leqslant \frac{1}{\lambda_i + \underline{y}^*}. \tag{C.36}$$

Next, by Proposition B.1, if $n \geqslant 2\max\{1, c_8^{-1}\}\log(4\max\{1, c_8^{-1}\})$, then we have

$$n \geqslant c_8^{-1}\log n \Longleftrightarrow ne^{-c_8 n} \leqslant 1.$$

We now bound,

$$\sum_{i=1}^T \mathbb{E}[Z_i] = \sum_{i=1}^T [\mathbb{E}[\mathbf{1}\{\mathcal{E}\}Z_i] + \mathbb{E}[\mathbf{1}\{\mathcal{E}^c\}Z_i]]$$

$$
\begin{aligned}
&\leqslant \sum_{t=1}^{T}\left[\frac{1}{\lambda_t+\underline{y}^*}+\mathbb{P}(\mathcal{E}^c)(\Theta_{1,T,T})_{tt}\right] && \text{using (C.36) and } Z_i \leqslant (\Theta_{1,T,T})_{ii}\\
&= \sum_{t=1}^{T}\frac{1}{\lambda_t+\underline{y}^*}+\mathbb{P}(\mathcal{E}^c)T && \text{since } \operatorname{tr}(\Theta_{1,T,T})=T\\
&\leqslant \sum_{t=1}^{T}\frac{1}{c_0t^2/T+\underline{y}^*}+4Te^{-c_8n} && \text{using (C.24) and (C.35)}\\
&\leqslant \int_0^T \frac{1}{c_0x^2/T+\underline{y}^*}\,\mathrm{d}x+4Te^{-c_8n}\\
&\leqslant \frac{\pi}{2}\sqrt{\frac{T}{c_0\underline{y}^*}}+4Te^{-c_8n}=\frac{\sqrt{2}\pi}{2\sqrt{c_0c_2}}\frac{mT}{n_1}+4Te^{-c_8n}\\
&\leqslant \left[\frac{\sqrt{2}\pi}{2\sqrt{c_0c_2}}+\frac{1}{16}\right]\frac{mT}{n_1}=: c_8\frac{mT}{n_1} && \text{since } ne^{-c_8n}\leqslant 1 \text{ and } m\geqslant 1.
\end{aligned}
$$

Plugging this upper bound into (C.19):

$$
\mathsf{R}(m,T,T';\{\mathsf{P}_x^{I_n}\}) \geqslant \sigma_\xi^2 p\cdot\frac{T'}{2T}\cdot\frac{n}{2m}\cdot\frac{1}{c_8}\frac{n_1}{mT}=\frac{1}{256c_8}\sigma_\xi^2\cdot\frac{pn^2}{m^2T}\cdot\frac{T'}{T}.
$$

The claim now follows. □

C.9 Proof of Theorem 6.3

Theorem 6.3 (Risk lower bound)**.** *There are universal positive constants c_0, c_1, and c_2 such that the following holds. Recall that $\mathsf{P}_x^{I_n}$ (resp. $\mathsf{P}_x^{0_{n\times n}}$) denotes the covariate distribution for a linear dynamical system with $A = I_n$ and $B = I_n$ (resp. $A = 0_{n\times n}$ and $B = I_n$). If $T \geqslant c_0$, $n \geqslant c_1$, and $mT \geqslant n$, then:*

$$
\mathsf{R}(m,T,T';\{\mathsf{P}_x^{0_{n\times n}},\mathsf{P}_x^{I_n}\}) \geqslant c_2\sigma_\xi^2\cdot\frac{pn}{mT}\cdot\max\left\{\frac{nT'}{mT},\frac{T'}{T},1\right\}.
$$

Proof. Let $\mathcal{P}_x := \{\mathsf{P}_x^{0_{n\times n}},\mathsf{P}_x^{I_n}\}$. We let c_0', c_1', c_2', and c_3' denote the universal positive constants in the statement of Lemma C.15. We first invoke Lemma C.7 to conclude that:

$$
\mathsf{R}(m,T,T';\mathcal{P}_x) \geqslant \frac{\sigma_\xi^2}{2}\cdot\frac{pn}{mT}\cdot\max\left\{\frac{T'}{T},1\right\}. \tag{C.37}
$$

The proof now proceeds in three cases:

Case $nT'/(mT) \leqslant 1$. In this case, we trivially have $\max\left\{\frac{T'}{T},1\right\} = \max\left\{\frac{nT'}{mT},\frac{T'}{T},1\right\}$. Therefore, (C.37) yields:

$$
\mathsf{R}(m,T,T';\mathcal{P}_x) \geqslant \frac{\sigma_\xi^2}{2}\cdot\frac{pn}{mT}\cdot\max\left\{\frac{nT'}{mT},\frac{T'}{T},1\right\}.
$$

Case $nT'/(mT) > 1$ and $m \leqslant c_2' n$. In this case, we can invoke Lemma C.15 to conclude that:

$$\mathsf{R}(m, T, T'; \mathcal{P}_x) \geqslant c_3' \sigma_\xi^2 \cdot \frac{pn}{mT} \cdot \frac{nT'}{mT} = c_3' \sigma_\xi^2 \cdot \frac{pn}{mT} \cdot \max\left\{\frac{nT'}{mT}, 1\right\}. \tag{C.38}$$

Since $n/m \geqslant 1/c_2'$, we have that $nT'/(mT) \geqslant T'/(c_2' T)$. Therefore:

$$\max\left\{\frac{nT'}{mT}, 1\right\} = \max\left\{\frac{nT'}{mT}, \frac{T'}{c_2' T}, 1\right\} \geqslant \min\{1, 1/c_2'\} \max\left\{\frac{nT'}{mT}, \frac{T'}{T}, 1\right\}.$$

Hence, from (C.38),

$$\mathsf{R}(m, T, T'; \mathcal{P}_x) \geqslant \min\{c_3', c_3'/c_2'\} \sigma_\xi^2 \cdot \frac{pn}{mT} \cdot \max\left\{\frac{nT'}{mT}, \frac{T'}{T}, 1\right\}.$$

Case $nT'/(mT) > 1$ and $m > c_2' n$. In this case, we have $T'/T > c_2' nT'/(mT)$. Therefore, we have:

$$\max\left\{\frac{T'}{T}, 1\right\} = \max\left\{c_2' \frac{nT'}{mT}, \frac{T'}{T}, 1\right\} \geqslant \min\{1, c_2'\} \max\left\{\frac{nT'}{mT}, \frac{T'}{T}, 1\right\}.$$

Hence, from (C.37),

$$\mathsf{R}(m, T, T'; \mathcal{P}_x) \geqslant \min\{1/2, c_2'/2\} \sigma_\xi^2 \cdot \frac{pn}{mT} \cdot \max\left\{\frac{nT'}{mT}, \frac{T'}{T}, 1\right\}.$$

The claim now follows taking $c_0 = c_0'$, $c_1 = c_1'$, and $c_2 = \min\{1/2, c_3', c_3'/c_2', c_2'/2\}$. $\square$