Proceedings of Machine Learning Research vol 167:1–30, 2022 33rd International Conference on Algorithmic Learning Theory

Improved rates for prediction and identification of partially observed linear dynamical systems

Holden Lee HOLDEN.LEE@DUKE.EDU
Department of Mathematics, Duke University

Editors: Sanjoy Dasgupta and Nika Haghtalab

Abstract

Identification of a linear time-invariant dynamical system from partial observations is a fundamental problem in control theory. Particularly challenging are systems exhibiting long-term memory. A natural question is how to learn such systems with non-asymptotic statistical rates depending on the inherent dimensionality (order) d of the system, rather than on the possibly much larger memory length. We propose an algorithm that given a single trajectory of length T with gaussian observation noise, learns the system with a near-optimal rate of $\widetilde{O}\left(\sqrt{\frac{d}{T}}\right)$ in $\mathcal{H}_2$ error, with only logarithmic, rather than polynomial dependence on memory length. We also give bounds under process noise and improved bounds for learning a realization of the system. Our algorithm is based on multi-scale low-rank approximation: SVD applied to Hankel matrices of geometrically increasing sizes. Our analysis relies on careful application of concentration bounds on the Fourier domain—we give sharper concentration bounds for sample covariance of correlated inputs and for $\mathcal{H}_\infty$ norm estimation, which may be of independent interest.

Keywords: system identification, Hankel matrix, SVD, linear dynamical system

1. Introduction

We consider the problem of prediction and identification of an *unknown* partially-observed linear time-invariant (LTI) dynamical system with stochastic noise,

$$x(t) = Ax(t-1) + Bu(t-1) + \xi(t) \tag{1}$$
$$y(t) = Cx(t) + Du(t) + \eta(t), \tag{2}$$

with a single trajectory of length T, given access only to input and output data. Here, $u(t) \in \mathbb{R}^{d_u}$ are inputs, $x(t) \in \mathbb{R}^d$ are the hidden states, $y(t) \in \mathbb{R}^{d_y}$ are observations (or outputs), $\xi(t) \sim N(0, \Sigma_x)$ and $\eta(t) \sim N(0, \Sigma_y)$ are iid gaussian noise, and $A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times d_u}, C \in \mathbb{R}^{d_y\times d}, D \in \mathbb{R}^{d_y\times d_u}$ are matrices. Partial observability refers to the fact that we do not observe the state $x(t)$, but rather a noisy linear observation $y(t)$.

As a simple and tractable family of dynamical systems, LTI systems are a central object of study for control theory and time series analysis. The problem of prediction and filtering for a known system dates back to Kalman (1960). However, in many machine learning applications, the system is *unknown* and must be learned from input and output data. Identification of an unknown system is often a necessary first step for robust control (Dean et al., 2019; Boczar et al., 2018). In a long line of recent work, the interplay between machine learning and control theory has borne fruit in an improved understanding of the statistical and online learning guarantees for prediction, identification, and control for unknown systems. In machine learning, LTI systems also serve as

a simple model problem for learning from correlated data in stateful environments, and can give insight into understanding the successes of reinforcement learning (Recht, 2019; Tu and Recht, 2019) and recurrent neural networks (Hardt et al., 2018).

Partial observability poses a significant challenge to system identification: In the fully observed setting, given access to $x(t)$, there is no obstacle to learning the matrices directly through linear regression. However, in the partially observed setting, the most natural form of the optimization problem is non-convex.

Systems exhibiting *long-term memory* are particularly challenging to learn. Restricting to marginally stable systems, this occurs when the spectral radius of A, $\rho(A)$, is close to 1, and it implies that the output at a particular time cannot be accurately estimated without taking into account inputs over many previous time-steps—on the order of $O\left(\frac{1}{1-\rho(A)}\right)$ times steps. Such systems often arise in practice. A particular class of such systems are those exhibiting *multiscale* behavior, with different state variables that change on vastly different timescales (Chatterjee and Russell, 2010). For example, the body's pH level is affected both by long-term changes on a timescale of days or weeks, as well as breathing rate which changes over a timescale of seconds. For such systems, it makes sense to discretize at the scale of the fastest changing variable, which leads to a long memory for the slowest-changing variable. With few exceptions, existing guarantees for learning partially observed LTI systems degrade as the memory length increases. However, counting the number of parameters in the model (1)–(2) suggests that the right measure of statistical complexity is the intrinsic dimensionality of the system, not the memory length. This leads to the following natural question.

Question: How can we learn partially observed LTI systems with (non-asymptotic) statistical rates that depend on the *intrinsic dimensionality* of the system, rather than the memory length?

Despite the simplicity of the question, little in the way of theoretical results are known. We focus on the particular problem of learning the *impulse response (IR) function* of the system—which fully determines its input-output behavior—in $\mathcal{H}_2$ norm. This is a natural norm for prediction problems as it measures the expected prediction error under random input. Known guarantees for learning the IR depend on the memory length. One particularly undesirable consequence is that for a continuous system with time discretization Δ going to 0, the memory scales as $1/\Delta$ (while the system order stays constant), leading to suboptimal estimation by an arbitrarily large factor.

Our key contribution is an algorithm and analysis that gives statistical rates that are optimal up to logarithmic factors, in the absence of process noise. Unlike previous works, our rates depend on the system order d—the natural dimensionality of the problem—and only *logarithmically* on the memory length of the system. Our algorithm is based on taking a low-rank approximation (SVD) of the Hankel matrix, which is a widely used technique in system identification. We consider a *multiscale* version of this algorithm, where we repeat this process for a geometric sequence of sizes of the Hankel matrix. This is essential for obtaining a stronger theoretical guarantee. In the setting of zero process noise, we prove that our algorithm achieves near-optimal $\widetilde{O}\left(\sqrt{\frac{d(d_u+d_y)}{T}}\right)$ rates in $\mathcal{H}_2$ error for the learned system.

Our analysis relies on careful application of concentration bounds on the Fourier domain to give sharper concentration bounds for sample covariance and $\mathcal{H}_\infty$ norm estimation, which may be of independent interest. While we consider our algorithm in a simple setting, we hope that this is a first step to understanding and improving more complex subspace identification algorithms. Indeed,

SVD and related spectral methods are a standard step used in subspace identification algorithms such as N4SID; our analysis suggests that SVD has an important "de-noising effect".

We also give improved bounds for system identification, that is, learning the matrices A, B, C, D using the Ho-Kalman algorithm (Ho and Kalman, 1966), with $\widetilde{O}\left(\sqrt{\frac{Ld(d_u+d_y)}{T}}\right)$ rates.

1.1. Notation

Norms. We use $\|\cdot\|$ to denote the 2-norm of a vector. For a matrix A, let $\|A\| = \|A\|_2$ denote its operator norm, $\rho(A)$ denote its spectral radius (maximum absolute value of eigenvalue), and $\|A\|_{\mathsf{F}}$ denote its Frobenius norm. For a matrix-valued function $M(t) \in \mathbb{C}^{d_1 \times d_2}$, $\|M\|_{\mathsf{F}} := \sqrt{\sum_t \|M(t)\|_{\mathsf{F}}^2}$. Let $\sigma_r(A)$ denote the rth singular value of A.

Fourier transform. Given a matrix-valued function $F : \mathbb{Z} \to \mathbb{C}^{m \times n}$, define the (discrete-time) Fourier transform as the function $\widehat{F} : \mathbb{R}/\mathbb{Z} \to \mathbb{C}^{m \times n}$ given by

$$\widehat{F}(\omega) = \sum_{t=-\infty}^{\infty} F(t) e^{-2\pi i \omega t}.$$

Matrices. Given a sequence $(F(t))_{t=1}^{a+b-1}$ where each $F(t) \in \mathbb{C}^{m \times n}$, define $\mathrm{Hankel}_{a \times b}(F)$ as the $am \times bn$ block matrix such that the (i,j)th block is $[\mathrm{Hankel}_{a \times b}(F)]_{ij} = F(i+j-1)$. Given a sequence $(F(t))_{t=0}^{a-1}$ where each $F(t) \in \mathbb{C}^{m \times n}$, define the Toeplitz matrix as the block matrix such that the (i,j)th block is $[\mathrm{Toep}_{a \times b}(F)]_{ij} = F(i-j)\mathbb{1}_{i \geq j}$. For a matrix A, let $A^\top, A^H, A^\dagger$ denote its transpose, Hermitian (conjugate transpose), and pseudoinverse, respectively. For a vector-valued function $v : \{a, \ldots, b\} \to \mathbb{R}^n$, let $v_{a:b} \in \mathbb{R}^{(|a-b|+1)n}$ denote the the vertical concatenation of $v(a), \ldots, v(b)$. Let $(A_1; \ldots; A_n)$ denote the vertical concantenation of matrices $A_1, \ldots, A_n$. Let $*$ denote convolution; we define convolutions between matrix and vector-valued functions by matrix-vector multiplication: $(F * u)(t) = \sum_{s \in \mathbb{Z}} F(s) u(t-s)$.

Control theory. For a matrix $A \in \mathbb{C}^{d \times d}$, define its resolvent as $\Phi_A(z) = (zI - A)^{-1}$. For a linear dynamical system $\mathcal{D}$ given by (1)–(2), let $\Phi_{\mathcal{D}} = \Phi_{u \to y}$ denote the transfer function from u to y (response to input). Then

$$\Phi_{\mathcal{D}}(z) = \Phi_{u \to y}(z) = C\Phi_A(z)B + D = C(zI - A)^{-1}B + D$$

Let $\mathbb{T} := \{z \in \mathbb{C} : |z| = 1\}$ be the unit circle in the complex plane. For a matrix-valued function $F : \mathbb{T} \to \mathbb{C}^{d_1 \times d_2}$, define the $\mathcal{H}_2$ and $\mathcal{H}_\infty$ norms by

$$\|\Phi\|_{\mathcal{H}_2} = \sqrt{\frac{1}{2\pi} \int_{\mathbb{T}} \|\Phi(z)\|_F^2 \, dz} \qquad \|\Phi\|_{\mathcal{H}_\infty} = \sup_{z \in \mathbb{T}} \|\Phi(z)\|.$$

For a function $F : \mathbb{N}_0 \to \mathbb{C}^{d_1 \times d_2}$, define its Z-transform to be $\mathcal{Z}[F](z) = \sum_{n=0}^{\infty} F(n) z^{-n}$. Considered as a function $\mathbb{T} \to \mathbb{C}$, we can take its $\mathcal{H}_2$ and $\mathcal{H}_\infty$ norms. Overloading notation, we will let $\|F\|_{\mathcal{H}_p} := \|\mathcal{Z}F\|_{\mathcal{H}_p}$ for $p = 2, \infty$. The $\mathcal{H}_2$ and $\mathcal{H}_\infty$ norms can be interpreted as the Frobenius and operator norms of the linear operator from input to output, i.e., they measure the average power of the output signal under random or worst-case input, respectively. (Note however that there an implicit factor of d scaling between $\mathcal{H}_2$ and $\mathcal{H}_\infty$.) For background on control theory, see e.g., Zhou et al. (1996).

Constants. In proofs, K may represent different constants from line to line.

2. Main results

We consider the problem of prediction and identification for an unknown linear dynamical system (1)–(2). Our main goal is to obtain error guarantees in $\mathcal{H}_2$ norm, which determines prediction error under random input (Oymak and Ozay, 2019, Lemma 3.3).

Problem 2.1 *Consider the partially-observed LTI system $\mathcal{D}$ (1)–(2) with gaussian inputs $u(t) \sim N(0, I_{d_u})$ for $0 \le t < T$. Suppose that the system is stable, that is, $\rho(A) < 1$, and that we observe a single trajectory of length T started with $x(0) = 0$, that is, we observe $u(t) \sim N(0, I_{d_u})$ and $y(t)$ for $0 \le t < T$. Suppose also that D has rank at most d.*

The goal is to learn a LTI system $\widetilde{\mathcal{D}}$ with the aim of minimizing $\left\|\Phi_{\widetilde{\mathcal{D}}} - \Phi_{\mathcal{D}}\right\|_{\mathcal{H}_2}$. Equivalently, letting

$$F^*(t) = \begin{cases} D, & t = 0 \\ CA^{t-1}B, & t \ge 1 \end{cases}$$

denote the impulse response function (also called the Markov parameters) of the system, the goal is to learn an impulse response $\widetilde{F}$ minimizing $\left\|F^* - \widetilde{F}\right\|_{\mathcal{H}_2} = \left\|F^* - \widetilde{F}\right\|_{\mathsf{F}}$.

Note that learning F^* is sufficient to fully understand the input-output behavior of the system, but we may also ask to recover the system parameters A, B, C, D up to similarity transformation (see Theorem 2.3). Note that we require $\mathrm{rank}(D)$ to be at most the system order so that it does not take more samples to learn than A, B, C.

Previous results (Oymak and Ozay, 2019; Sarkar et al., 2019) roughly depend polynomially on the "memory" $\frac{1}{1-\rho(A)}$, which blows up as the spectral norm of A approaches 1. In the setting of zero process noise, our goal is to obtain rates that are $\widetilde{O}\left(\frac{\mathrm{poly}(d,d_u,d_y)}{\sqrt{T}}\right)$, with only poly-logarithmic dependence on $\frac{1}{1-\rho(A)}$. See Figure 1 for a comparison.

We assume that $\rho(A) < 1$ because if $\mathcal{D}$ is not stable, it is in general impossible to learn $\widetilde{\mathcal{D}}$ with finite $\mathcal{H}_2$ error, as a system with infinite response can have arbitrarily small response on any finite time interval. However, it may still be possible to learn the response up to time $L \ll T$ in this case (Simchowitz et al., 2019), or achieve other reasonable guarantees. The marginally stable case ($\rho(A) = 1$) is an important case we leave to future work.

In our Algorithm 10, we first use linear regression to obtain a noisy estimate F of the impulse response. Next, following standard system identification procedures, we form the Hankel matrix $\mathrm{Hankel}_{L\times L}(F)$ with the entries of F on its diagonals. Because the true Hankel matrix

$$\mathrm{Hankel}_{L\times L}(F^*) = \begin{pmatrix} CB & CAB & \cdots & CA^{L-1}B \\ CAB & CA^2B & & \vdots \\ \vdots & & \ddots & \vdots \\ CA^{L-1}B & \cdots & \cdots & CA^{2L-1}B \end{pmatrix}$$

has rank d, we take a low-rank SVD R_L of the Hankel matrix to "de-noise" the impulse response. We can then read off the estimated impulse response by averaging over the corresponding diagonal

Method	Rollout type	Min # samples	IR error
Least squares (IR) Tu et al. (2017)	Multi	L	$\sigma\sqrt{\frac{L}{T}}$
Least squares (IR) Oymak and Ozay (2019)	Single	L	$\sigma\sqrt{\frac{L}{T}}$
Nuclear norm minimization	Multi	$\min\{d^2, L\}$	$\sigma\sqrt{\frac{L}{T}}$
Sun et al. (2020)	Multi	d	$\sigma\sqrt{\frac{dL}{T}}$
rank-d SVD (Theorem 2.2)	Single	L	$\sigma\sqrt{\frac{d}{T}}$

Figure 1: Here, L is the memory length for the system, which is $\widetilde{O}\left(\frac{1}{1-\rho(A)}\right)$ for well-conditioned systems. *Rollout type* refers to whether we have access to a single trajectory or multiple trajectories. *Min # samples* refers to the minimum number of samples (up to log factors) before the bounds are operational. *IR error* refers to the error in the impulse response in Frobenius/$\mathcal{H}_2$ norm. Logarithmic factors are omitted.

of R_L. For technical reasons, we need to repeat this process for a geometric sequence of sizes of the Hankel matrix: $L \times L$, $L/2 \times L/2$, $L/4 \times L/4$, and so forth. This is because the low-rank approximation objective for a $\ell \times \ell$ Hankel matrix H encourages the skew-diagonals that are $\Theta(\ell)$ (consisting of entries H_{ij} with $K_1\ell \le i + j \le c_2\ell$) to be close—as those are the diagonals with the most entries—and hence estimates $F^*(t)$ well when $t = \Theta(\ell)$. In other words, low-rank estimation for $\mathrm{Hankel}_{\ell\times\ell}(F)$ is only sensitive to the portion of the signal that is at timescale ℓ. Repeating this process ensures that we cover all scales.

Our main theorem is the following.

Theorem 2.2 *There is a constant K_1 such that following holds. In the setting of Problem 2.1, suppose that F^* is the impulse response function, G^* is the impulse response for the process noise ($G^*(t) = CA^t$, T is such that $T \ge K_1 L d_u \log\left(\frac{Ld_u}{\delta}\right)$, $\varepsilon_{\mathrm{trunc}} := \left\|F^*\mathbb{1}_{[2L,\infty)}\right\|_{\mathcal{H}_\infty}\sqrt{d_u} + \left\|G^*\mathbb{1}_{[2L,\infty)}\right\|_{\mathcal{H}_\infty}\left\|\Sigma_x^{1/2}\right\|_{\mathsf{F}}$, and $M_{x\to y} = (O, C, CA, \ldots, CA^{L-1})^\top \in \mathbb{R}^{(L+1)d\times d_y}$. Let $0 < \delta \le \frac{1}{2}$ and $\sigma = \sqrt{\|\Sigma_y\| + \|\Sigma_x\| L \log\left(\frac{Ld_u}{\delta}\right)\|M_{x\to y}\|^2}$. Then with probability at least $1-\delta$, Algorithm 10 learns an impulse response function $\widetilde{F}$ such that*

$$\left\|\widetilde{F} - F^*\right\|_{\mathsf{F}} = O\left(\sigma\sqrt{\frac{d\left(d_y + d_u + \log\left(\frac{L}{\delta}\right)\right)\log L}{T}} + \varepsilon_{\mathrm{trunc}}\sqrt{d} + \left\|F^*\mathbb{1}_{(L,\infty)}\right\|_{\mathsf{F}}\right).$$

In the absence of process noise (when $\Sigma_x = O$), when L and T are chosen large enough, the first term dominates, and ignoring log factors, the dependence is $O\left(\sqrt{\frac{d(d_y+d_u)}{T}}\right)$. We expect this to be the optimal sample complexity up to logarithmic factors. However, in the presence of process noise, there is an undesirable factor of $\sqrt{L}\,\|M_{x\to y}\|$, which (for well-conditioned matrices) is expected to be $O\left(\frac{1}{1-\rho(A)}\right)$ or $O(L)$. We leave it an open problem to improve the guarantees in this setting.

Algorithm 1: Learning impulse response through multi-scale low-rank Hankel SVD

Input: Length L (power of 2), time T.

Part 1: Linear regression to recover noisy impulse response

Let $u(t) \sim N(0, I_{d_u})$ for $0 \le t < T$, and observe the outputs $y(t) \in \mathbb{R}^{d_y}$, $0 \le t < T$.

Solve the least squares problem

$$\min_{F:\mathrm{Supp}(F)\subseteq[0,2L-1]} \sum_{t=0}^{T-1} \|y(t) - (F * u)(t)\|^2 . \tag{3}$$

to obtain the noisy impulse response $F : [0, 2L-1] \cap \mathbb{Z} \to \mathbb{R}^{d_y \times d_u}$.

Part 2: Low-rank Hankel SVD to de-noise impulse response

Let $\widetilde{F}(0)$ be the rank-d SVD of $F(0)$

for $k = 0$ *to* $\log_2 L$ **do**

  Let $\ell = 2^k$.

  Let R_ℓ be the rank-d SVD of $\mathrm{Hankel}_{\ell\times\ell}(F)$ (i.e., $\mathrm{argmin}_{\mathrm{rank}(R)\le d} \|R - \mathrm{Hankel}_{\ell\times\ell}(F)\|$).

  For $\frac{\ell}{2} < t \le \ell$, let $\widetilde{F}(t)$ be the $d_y \times d_u$ matrix given by $\widetilde{F}(t) = \frac{1}{t}\sum_{i+j=t}(R_\ell)_{ij}$, where $(\cdot)_{ij}$ denotes the (i,j)th block of the matrix.

end

Output: Estimate of impulse response $\widetilde{F}$.

Remark 1 *The L-factor dependence on the process noise is unavoidable with the current algorithm: when the process noise has covariance $\Sigma_x = I$ and decays after L steps, it can cause perturbations of size $O(\sqrt{L})$ compared to the noiseless system. Even in the case $d = 1$, when the impulse response function is $ae^{-kt/L}$ for a known k, the noise will cause the estimate of a to be off by $O(\sqrt{L})$, and hence the $\mathcal{H}_2$ norm of the impulse response to be off by $O(L)$. Our algorithm only regresses on previous inputs, but in the presence of process noise, a better approach is to regress on both the previous inputs $u(t)$ and* outputs *$y(t)$ and then take a (weighted) SVD, as in N4SID (Qin, 2006).*

Remark 2 *Note that the burn-in time—the minimum trajectory length under which our error guarantees hold—is $\Omega(L)$. A burn-in time of $\Omega(L)$ is information-theoretically required to get $\mathrm{poly}(d)$ rates. Attempting to extrapolate an impulse response function from time $o(L)$ to time L can magnify errors by $\exp(d)$, because the finite impulse response of a system of order d can approximate a polynomial of degree $d-1$ on $[0, L]$.*

Remark 3 *Commonly, one assumes that $\left\|CA^tB\right\| \le M\rho^t$ for some M and ρ (greater than or equal to the spectral radius of A). Then in the noiseless case, $\varepsilon_{trunc} \le M\frac{\rho^{2L-1}}{1-\rho}\sqrt{d_u}$. The algorithm and theorem is stated known ε_{trunc} for simplicity. Knowing ε_{trunc} allows choosing an appropriate memory length L.*

A standard technique to convert the algorithm to one that achieves the correct rates as $T \to \infty$ without knowing the memory length is to use the "doubling trick": increase the memory length by a constant every time the number of timesteps doubles, so that the memory length scales as $\log_2(T)$. See e.g., Tsiamis and Pappas (2020). This works because the impulse response decays exponentially.

We also show the following improved rates for learning the system matrices, by combining $\mathcal{H}_\infty$ bounds for the learned impulse response with stability results for the Ho-Kalman algorithm (Oymak and Ozay, 2019). Because the input-output behavior is unchanged under a similarity transformation $(A, B, C) \leftarrow\!\shortmid (W^{-1}AW, W^{-1}B, CW)$, we can only learn the parameters up to similarity transformation. We will make the standard control-theoretic assumptions that $\mathcal{D}$ is observable ($(C; CA; \ldots; CA^{d-1})$ is full-rank) and controllable ($(B, AB, \ldots, A^{d-1}B)$ is full-rank).

Theorem 2.3 *Keep the assumptions and notation of Theorem 2.2, suppose $\mathcal{D}$ is observable and controllable, and let*

$$\varepsilon' = \sigma\sqrt{\frac{L\left(d_y + d_u + \log\left(\frac{L}{\delta}\right)\right)}{T}} + \varepsilon_{\text{trunc}}.$$

Let $H^- = \text{Hankel}_{L\times(L-1)}(F^*)$. Suppose that $\varepsilon' = O(\sigma_{\min}(H^-))$. Then with probability at least $1-\delta$, the Ho-Kalman algorithm (Algorithm 9) with $T_1 = L, T_2 = L-1$ applied to the least squares solution F of (3) returns $\widehat{A}, \widehat{B}, \widehat{C}$ such that there exists a unitary matrix W satisfying

$$\max\left\{\left\|C - \widehat{C}W\right\|_{\mathsf{F}}, \left\|B - W^{-1}\widehat{B}\right\|_{\mathsf{F}}\right\} = O(\sqrt{d}\cdot\varepsilon')$$

$$\left\|A - W^{-1}\widehat{A}W\right\|_{\mathsf{F}} = O\left(\frac{1}{\sigma_{\min}(H^-)}\cdot\sqrt{d}\cdot\varepsilon'\cdot\left(\frac{\|\Phi_{\mathcal{D}}\|_{\mathcal{H}_\infty}}{\sigma_{\min}(H^-)} + 1\right)\right).$$

Algorithm 2: Ho-Kalman algorithm (from Oymak and Ozay (2019))

Input: Length T, Markov parameter matrix estimate F, system order d, Hankel shape (T_1, T_2+1) with $T_2 + T_1 + 1 = T$.

1. Form the Hankel matrix $\widehat{H} = \text{Hankel}_{T_1,T_2+1}(F)$.
2. Let $\widehat{H}^- \in \mathbb{R}^{d_yT_1\times d_uT_2}$ be the first d_uT_2 columns of $\widehat{H}$.
3. Let $\widehat{L} = U\Sigma V^\top \in \mathbb{R}^{d_yT_1\times d_uT_2}$ be the rank-n SVD of $\widehat{H}^-$.
4. Let $\widehat{O} \in \mathbb{R}^{d_yT_1\times d} = U\Sigma^{1/2}$.
5. Let $\widehat{Q} \in \mathbb{R}^{d\times d_uT_2} = \Sigma^{1/2}V^\top$.
6. Let $\widehat{C}$ be the first m rows of $\widehat{O}$.
7. Let $\widehat{B}$ be the first p columns of $\widehat{Q}$.
8. Let $\widehat{H}^+ \in \mathbb{R}^{d_yT_1\times d_uT_2}$ be the last d_uT_2 columns of $\widehat{H}$.
9. Let $\widehat{A} = \widehat{O}^\dagger\widehat{H}^+\widehat{Q}^\dagger$.

Output: $\widehat{A} \in \mathbb{R}^{d\times d}, \widehat{B} \in \mathbb{R}^{d\times d_u}, \widehat{C} \in \mathbb{R}^{d_y\times d}$.

As L can be chosen to make $\varepsilon_{\text{trunc}}$ negligible, this gives $\widetilde{O}\left(\sqrt{\frac{Ld(d_u+d_y)}{T}}\right)$ rates, however, with factors depending on the minimum eigenvalue of H. This is an improvement over the $\widetilde{O}\left(\sqrt{d}\sqrt[4]{\frac{L(d_u+d_y)}{T}}\right)$ rates in Oymak and Ozay (2019). Note that our rates still a square-root dependence on the memory L; and leave it an open question whether one can obtain logarithmic dependence similar to Theorem 2.2.

We prove Theorem 2.2 in Section 4 and Theorem 2.3 in Appendix B. We give a lower bound in Section C that shows that in the absence of process noise, the rate in Theorem 2.2 is optimal up to logarithmic factors.

3. Related work

We survey two classes of methods for learning partially observable LDS's, subspace identification and improper learning. With the exception of Rashidinejad et al. (2020), all guarantees have sample complexity depending on the memory length L, which we wish to avoid.

3.1. Subspace identification

The basic idea of subspace identification (Ljung, 1998; Qin, 2006; Van Overschee and De Moor, 2012) is to learn a certain structured matrix (such as a Hankel matrix), take a best rank-k approximation (using SVD or another linear dimensionality reduction method), and learn the system matrices A, B, C, D up to similarity transformation. Usage of spectral methods circumvents the fact that the most natural optimization problem for A, B, C, D is non-convex. However, classical guarantees for these methods are asymptotic.

Recently, various authors have given non-asymptotic guarantees for system identification algorithms. Oymak and Ozay (2019) analyzed the Ho-Kalman algorithm (Ho and Kalman, 1966) in this setting. Sarkar et al. (2019) consider the setting where system order is unknown and give an end-to-end result for prediction, while Tsiamis et al. (2020) consider the problem of online filtering, that is, recovering $x(t)$'s up to some linear transformation. Simchowitz et al. (2019) give guarantees under more general conditions of noise and marginal stability; however their main bound is for the *truncated* Markov parameters, and to capture all but ϵ of the impulse response, we would have to truncate at the memory length L, which would incur dependence on L. Moreover, they require low phase rank, a condition which we do not expect to hold generically for "random" linear dynamical systems with eigenvalues close to 1. An advantage of their approach is that they are able to achieve consistent recovery of system parameters without taking the truncation length to be as large as the memory length.

An alternate, empirically successful approach is that of nuclear norm minimization or regularization (Fazel et al., 2013). Sun et al. (2020) (building on Cai et al. (2016)) give explicit rates of convergence, and show that the algorithm has a lower minimum sample complexity and is easier to tune.

Our algorithm is based on the classical approach of taking a low-rank approximation of the Hankel matrix, but we repeat this process with Hankel matrices of sizes $L \times L$, $L/2 \times L/2$, $L/4 \times L/4$, and so forth; this is key modification that allows us to obtain better statistical rates. Our analysis builds on the analyses given in Oymak and Ozay (2019); Sun et al. (2020). As essential part of the analysis is analyzing linear regression for correlated inputs, where we extend the work of Djehiche et al. (2019) to MIMO (multiple input multiple output) systems, as explained below.

3.1.1. LINEAR REGRESSION WITH CORRELATED INPUTS

An important step in obtaining non-asymptotic rates for system identification is analyzing linear regression for correlated inputs. The most challenging step is to lower-bound the sample covariance matrix of inputs to the linear regression. A lower bound, rather than a matrix concentration result, is sufficient (Mendelson, 2014; Simchowitz et al., 2018; Matni and Tu, 2019); however, a concentration result is obtainable in our setting.

Tu et al. (2017) give non-asymptotic bounds for learning the finite impulse response for a SISO (single input single output) system in ℓ^∞ Fourier norm; however, they require L rollouts of size

$O(L)$ and hence $\Omega(L^2)$ timesteps. Addressing the more challenging single-rollout setting, Oymak and Ozay (2019) obtain bounds for a single rollout of $\widetilde{\Omega}(L)$ timesteps, by using concentration bounds for random circulant matrices Krahmer et al. (2014) to derive concentration inequalities for the covariance matrix. These concentration inequalities for the covariance matrix were improved (by logarithmic factors) by Djehiche et al. (2019). Although Djehiche et al. (2019) give an analysis in the SISO case, as we show in Theorem A.2, the results can be extended to the MIMO case with an ε-net argument.

3.2. Improper learning using autoregressive methods

Instead of solving the statistical problem of identifying parameters, another line of work develops algorithms for regret minimization in online learning. The goal is simply to do well in predicting future observations, with small loss (regret) compared to the best predictor in hindsight; the learned predictor is allowed to be improper, that is, take a different functional form. In the stochastic case, this allows prediction almost as well as if the actual system parameters were known; however, the framework also allows for adversarial noise.

One popular strategy for improperly learning the system is to learn a linear autoregressive filter over previous inputs and observations, or ARMA model. Naturally, because we are optimizing over a larger hypothesis class, the statistical rates depend on L rather than the system order d.

(Ghai et al., 2020, Theorem 4.7) consider the problem of online prediction for a fully or partially observed LDS, and give a regret bound that depends polynomially on the memory length L. Their approach works even for marginally stable systems, that is, systems with $\rho(A) \leq 1$. See also Anava et al. (2013); Hazan et al. (2017, 2018); Kozdoba et al. (2019); Tsiamis and Pappas (2020); Rashidinejad et al. (2020) for previous work using autoregressive methods.

Of particular interest to us is Rashidinejad et al. (2020), which gives rates independent of spectral radius. Building on Hazan et al. (2017), they observe that it suffices to regress on previous inputs and outputs projected to a lower-dimensional space. Their algorithm works in the setting of process noise and competes with the Kalman filter, but only when $A - KC$ has real eigenvalues, where K is the Kalman gain.

4. Proof of main theorem

In this section, we prove Theorem 2.2. The proof hinges on the following lemma, which shows that if we observe a low-rank matrix plus noise, then taking a low-rank SVD can have a de-noising effect, producing a matrix that is closer to the true matrix.

Lemma 4.1 (De-noising effect of SVD) *There exists a constant K such that the following holds. Suppose that $A \in \mathbb{C}^{m\times n}$ is a rank-r matrix, $\widehat{A} = A + E$, and $\widehat{A}_r$ is the rank-r SVD of $\widehat{A}$. Then*

$$\left\|\widehat{A}_r - A\right\|_{\mathsf{F}} \leq K\sqrt{r}\,\|E\|\,. \tag{4}$$

Compare this with the original error $\left\|\widehat{A} - A\right\|_{\mathsf{F}} = \|E\|_{\mathsf{F}}$, which can only be bounded by $\sqrt{\min\{m,n\}}\,\|E\|$. It is an interesting question what the best constant K is. When applied to the d-SVD of the Hankel matrix, this gives a factor of $\sqrt{d}$ rather than $\sqrt{L}$ for the error.

Proof We have

$$\left\|\hat{A}_r - A\right\|_{\mathsf{F}} \leq \sqrt{2r}\left\|\hat{A}_r - A\right\|_2 \tag{5}$$

$$\leq \sqrt{2r}\left(\left\|\hat{A}_r - \hat{A}\right\|_2 + \left\|\hat{A} - A\right\|_2\right) \tag{6}$$

$$\leq 2\sqrt{2r}\,\|E\| \tag{7}$$

where (5) follows from $\hat{A}_r - A$ having rank at most $2r$, (6) follows from the triangle inequality, and (7) follows from Weyl's Theorem: $\left\|\hat{A}_r - \hat{A}\right\|_2 \leq \sigma_{r+1}(\hat{A}) \leq \sigma_{r+1}(A) + \|E\| = \|E\|$. ■

To prove Theorem 2.2, we will need to obtain bounds for $F : \{0, 1, \ldots, 2L-1\} \to \mathbb{R}^{d_y \times d_u}$ learned from linear regression in $\mathcal{H}_\infty$ norm. The following is our main technical result.

Lemma 4.2 *There are K_1, K_2 such that the following hold. Suppose $y = F^* * u + G^* * \xi + \eta$ where $u(t) \sim N(0, I_{d_u})$, $\xi(t) \sim N(0, \Sigma_x)$, $\eta(t) \sim N(0, \Sigma_y)$ for $0 \leq t < T$, and $\mathrm{Supp}(F^*), \mathrm{Supp}(G^*) \subseteq [0, \infty)$. Let $F = \mathrm{argmin}_{F \in \{0,\ldots,L\} \to \mathbb{R}^{d_y \times d_u}} \sum_{t=0}^{T-1} |y(t) - (F * u)(t)|^2$, $M_{G^*} = (G^*(0), \ldots, G^*(L))^\top \in \mathbb{R}^{(L+1)d \times d_y}$, and $\varepsilon_{\mathrm{trunc}} = \left\|F^* \mathbb{1}_{[L+1,\infty)}\right\|_{\mathcal{H}_\infty} \sqrt{d_u} + \left\|G^* \mathbb{1}_{[L+1,\infty)}\right\|_{\mathcal{H}_\infty} \left\|\Sigma_x^{1/2}\right\|_{\mathsf{F}}$. For $0 < \delta \leq \frac{1}{2}$, $T \geq K_1 L d_u \log\left(\frac{L d_u}{\delta}\right)$, $1 \leq L' \leq L$, $-1 \leq a < L - L'$,*

$$\left\|(F - F^*)\mathbb{1}_{[a+1,a+L']}\right\|_{\mathcal{H}_\infty}$$
$$\leq K_2\left[\sqrt{\frac{1}{T}}\left(\sqrt{\|\Sigma_y\|\, L'\left(d_u + d_y + \log\left(\frac{L'}{\delta}\right)\right)} + \sqrt{\|\Sigma_x\|\, L' L d_u \log\left(\frac{L d_u}{\delta}\right)}\,\|M_{G^*}\|\right) + \varepsilon_{\mathrm{trunc}}\right]$$

with probability at least $1 - \delta$.

In the case $\Sigma_x = O$, this roughly says that the error in the learned impulse response, $F - F^*$, over any interval of length L', has all Fourier coefficients bounded in spectral norm by $\widetilde{O}\left(\sqrt{\frac{L'(d_u+d_y)}{T}}\right)$, what we expect if the error from linear regression is uniformly distributed over all frequencies.

A complete proof is in Appendix A; we give a brief sketch. First, because the errors are Gaussian, the error from linear regression, $F - F^*$, follows a Gaussian distribution. To bound its covariance, we lower-bound the smallest singular value of the sample covariance of the inputs (Lemma A.1, Appendix A.1). Here, the difficulty is that the inputs are *correlated*—the input at time t is $u_{t:t-L}$. Fortunately, the translation structure means it is close to a submatrix of an infinite block Toeplitz matrix, which becomes block diagonal in the Fourier domain. This "decoupling" allows us to show concentration. Compared to the SISO setting in Djehiche et al. (2019), we require an extra ε-net argument. Once we have a bound on the covariance, we can bound any $\left\|\widehat{(F - F^*)}(\omega)\right\|$ by matrix concentration (Appendix A.2); to bound the $\mathcal{H}_\infty$ norm it suffices to bound this over a grid of ω's (Lemma A.4).

Bounding the error in $\mathcal{H}_\infty$ norm of the impulse response allows us to bound the error in operator norm of the Hankel matrix, as the following lemma shows.

Lemma 4.3 *For any $F : \mathbb{Z} \to \mathbb{C}^{m \times n}$, we have $\|\mathrm{Hankel}_{a \times b}(F)\| \leq \|F\|_{\mathcal{H}_\infty}$.*

Proof Note that when $v : \mathbb{Z} \to \mathbb{C}^n$, $\mathrm{Supp}(v) \subseteq [0, b-1]$, we have $(F*v)_{b:b+a-1} = \mathrm{Hankel}_{a\times b}(F)v_{b-1:0}$. Hence for any v set to 0 outside of $[0, b-1]$, using Parseval's Theorem and the fact that the Fourier transform of a convolution is the product of the Fourier transforms, we have

$$\|\mathrm{Hankel}_{a\times b}(F)v_{b-1:0}\|_2 \le \|F * v\|_2 = \left\|\widehat{F}\widehat{v}\right\|_2 \le \sup_{\omega\in[0,1]} \left\|\widehat{F}(\omega)\right\|_2 \|\widehat{v}\|_2 = \|F\|_{\mathcal{H}_\infty} \|v\|_2$$

This shows that $\|\mathrm{Hankel}_{a\times b}(F)\| \le \|F\|_{\mathcal{H}_\infty}$. ■

Remark 4 *In fact, something stronger is true: we can bound $\|\mathrm{Hankel}_{a\times b}(F)\|$ by the discrete Fourier transform of F over $\mathbb{Z}/(a+b-1)$ (Sun et al., 2020, Theorem 3). For consistency, we stick to using the Fourier transform over $\mathbb{Z}$; in light of Lemma A.4, this only affects the result by constant factors.*

Theorem 2.2 will follow from the following bound after an application of the triangle inequality.

Lemma 4.4 *There are K_1, K_2 such that the following holds for the setting of Problem 2.1. Suppose L is a power of 2, and $T \ge K_1 L d_u \log\left(\frac{Ld_u}{\delta}\right)$. Let $\left\|F^* \mathbb{1}_{[L+1,\infty)}\right\|_{\mathcal{H}_\infty} \sqrt{d_u} + \left\|G^* \mathbb{1}_{[L+1,\infty)}\right\|_{\mathcal{H}_\infty} \left\|\Sigma_x^{1/2}\right\|_{\mathsf{F}}$ and $M_{x\to y} = (O, C, CA, \ldots, CA^{L-1})^\top \in \mathbb{R}^{(L+1)d\times d_y}$. Then with probability at least $1-\delta$, the output $\widetilde{F}$ given by Algorithm 10 satisfies*

$$\left\|(\widetilde{F} - F^*)\mathbb{1}_{[1,L]}\right\|_{\mathsf{F}} \le K_2 \left(\sqrt{\frac{\|\Sigma_y\|\, d\left(d_y + d_u + \log\left(\frac{L}{\delta}\right)\right)\log L}{T}} + \sqrt{\frac{\|\Sigma_x\|\, L d d_u \log\left(\frac{Ld_u}{\delta}\right)}{T}}\, \|M_{G^*}\| + \varepsilon_{\mathrm{trunc}}\sqrt{d} \right)$$

Proof We are in the situation of Lemma 4.2 with $G^*(t) = CA^{t-1}\mathbb{1}_{t\ge 1}$. Let $\mathcal{H}_\ell = \mathrm{Hankel}_{\ell\times\ell}(F)$ and $\mathcal{H}_\ell^* = \mathrm{Hankel}_{\ell\times\ell}(F^*)$. Suppose $\ell \le L$ is even. Note that

$$\mathcal{H}_\ell = \underbrace{\mathrm{Hankel}_{\ell\times\ell}(F^*)}_{\mathcal{H}_\ell^*} + \mathrm{Hankel}_{\ell\times\ell}(F - F^*)$$

where $\mathcal{H}_\ell^* = \mathrm{Hankel}_{\ell\times\ell}(F^*)$ is a rank-d matrix, with error term is bounded by

$$\begin{aligned}
\|\mathrm{Hankel}_{\ell\times\ell}(F - F^*)\| &\le \left\|(F - F^*_{\mathrm{trunc}})\mathbb{1}_{[1,2\ell-1]}\right\|_{\mathcal{H}_\infty} && \text{by Lemma 4.3} \\
&< K\Bigg[\sqrt{\frac{1}{T}}\Bigg(\sqrt{\|\Sigma_y\|\,\ell\left(d_u + d_y + \log\left(\frac{L'}{\delta}\right)\right)} \\
&\quad + \sqrt{\|\Sigma_x\|\,\ell L d_u \log\left(\frac{Ld_u}{\delta}\right)}\,\|M_{G^*}\| + \varepsilon_{\mathrm{trunc}}\Bigg] && \text{by Lemma 4.2} \qquad (8)
\end{aligned}$$

with probability at least $1-\delta$. Let R_ℓ be the rank-d SVD of $\mathcal{H}_\ell$. Then by Lemma 4.1,

$$\|R_\ell - \mathcal{H}_\ell^*\|_{\mathsf{F}} = O\left(\sqrt{d}\,\|\mathrm{Hankel}_{\ell\times\ell}(F - F^*)\|\right). \qquad (9)$$

Now letting

$$\widetilde{F}(t) = \frac{1}{t}\sum_{i+j=t} (R_\ell)_{ij} \qquad \text{when } \frac{\ell}{2} < t \le \ell,$$

we have (using the fact that the mean minimizes the sum of squared errors)

$$\|R_\ell - \mathcal{H}_\ell^*\|_{\mathsf{F}}^2 \geq \sum_{t=\frac{\ell}{2}+1}^{\ell} \sum_{i+j=t} \|(R_\ell)_{ij} - F^*(t)\|_{\mathsf{F}}^2$$
$$\geq \sum_{t=\frac{\ell}{2}+1}^{\ell} \left(t \cdot \left\| \widetilde{F}(t) - F^*(t) \right\|_{\mathsf{F}}^2 \right) \geq \left(\frac{\ell}{2} + 1 \right) \sum_{t=\frac{\ell}{2}+1}^{\ell} \left(\left\| \widetilde{F}(t) - F^*(t) \right\|_{\mathsf{F}}^2 \right).$$

Note that we only get a lower bound with a factor of ℓ if we restrict to t that is $\Theta(\ell)$, i.e., restrict to diagonals that have many entries. This is the reason we will have to repeat this process for multiple sizes. Hence

$$\left\| (\widetilde{F} - F^*) \mathbb{1}_{[\frac{\ell}{2}+1,\ell]} \right\|_{\mathsf{F}}^2 \leq \frac{1}{\ell/2} \|R_\ell - \mathcal{H}_\ell^*\|_{\mathsf{F}}^2 .$$

Together with (9) and (8) this gives with probability $\geq 1 - \delta$ that

$$\left\| (\widetilde{F} - F^*) \mathbb{1}_{[\frac{\ell}{2}+1,\ell]} \right\|_{\mathsf{F}} \leq K \left(\sqrt{\frac{\|\Sigma_y\| \, d \left(d_y + d_u + \log\left(\frac{L}{\delta}\right) \right)}{T}} + \sqrt{\frac{\|\Sigma_x\| \, L d_u \log\left(\frac{L d d_u}{\delta}\right)}{T}} \|M_{G^*}\| + \frac{\varepsilon_{\text{trunc}} \sqrt{d}}{\sqrt{\ell}} \right)$$

Replacing δ by $\frac{\delta}{\log_2 L}$, using a union bound over powers of 2, and summing gives

$$\left\| (\widetilde{F} - F^*) \mathbb{1}_{[1,L]} \right\|_{\mathsf{F}} = O\left(\sqrt{\frac{\|\Sigma_y\| \, d \left(d_y + d_u + \log\left(\frac{L}{\delta}\right) \right) \log L}{T}} + \sqrt{\frac{\|\Sigma_x\| \, L d d_u \log\left(\frac{L d_u}{\delta}\right) \log L}{T}} \|M_{G^*}\| + \varepsilon_{\text{trunc}} \sqrt{d} \right).$$

■

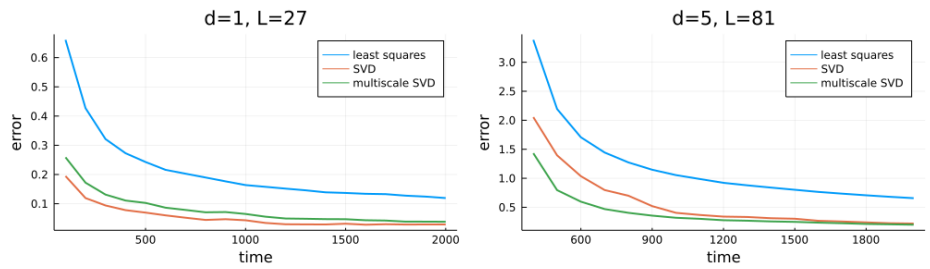
Proof [Proof of Theorem 2.2] We have the bound in Lemma 4.4, and also the same bound for $\left\| (\widetilde{F} - F^*)(0) \right\|_{\mathsf{F}}$ after applying Lemma 4.2 to $(F - F^*)\delta_0$ and then applying Lemma 4.1. Finally, note that $\left\| (\widetilde{F} - F^*) \mathbb{1}_{(L,\infty)} \right\|_{\mathsf{F}} = \left\| F^* \mathbb{1}_{(L,\infty)} \right\|_{\mathsf{F}}$ and use the triangle inequality. ■

5. Experiments

We compared three algorithms for learning the impulse response function: least-squares, and low-rank Hankel SVD with and without the multi-scale repetition. We include details of the experimental setup in Appendix E. Note that to reduce the number of scales, we consider use a slight modification of our Algorithm 10 which triples the size at each iteration instead.

The plots show the error $\left\| F^* \mathbb{1}_{[1,L]} - F \right\|_2$, where F is the estimated impulse response on $[1, L]$, averaged over 10 randomly generated LDS's, as a function of the time T elapsed. We consider systems of order $d = 1, 3, 5, 10$, and memory lengths $L = 27, 81$.

Using SVD significantly reduces the error, supporting our theory which shows that SVD has a "de-noising" effect. Additionally, multiscale SVD has better performance than naive SVD when d is moderate, L is large, and data is limited, but the performance is similar in a data-rich setting.



6. Open problems

We conclude with some open problems. It would be interesting to obtain analogous rates (depending on system order) for the nuclear norm regularized problem (Sun et al., 2020). Spectral methods also suggest the possibility of obtaining regret bounds for adaptive control of partially-observed systems with milder dependence on $\frac{1}{1-\rho(A)}$. We give several other problems below.

Process noise. A natural open question is to obtain better guarantees in the presence of process noise $\xi(t)$. Using the heuristic of parameter counting, we do not believe $\mathrm{poly}(L)$ memory dependence is necessary in this setting, and that the maximum likelihood estimator, although computationally inefficient, will attain memory-independent rates. Thus, the crux of the problem is to give a computationally efficient algorithm with mild memory dependence in this setting.

We note that in Theorem 2.2, the factor multiplying $\sqrt{\|\Sigma_x\|}$ is $\sqrt{L}\,\|M_{x\to y}\|_2$, which we expect to be on the order of L when L is the minimal sufficient memory length. This term arises because process noise can accumulate over L timesteps. In the case where $\xi(t) \sim N(0, \Sigma_x)$ is iid gaussian, the Kalman filter shows that we can rewrite the system in the predictor form (Qin, 2006)

$$x(t)^- = A_{\mathrm{KF}} x(t-1)^- + B_{\mathrm{KF}} \begin{pmatrix} u(t-1) \\ y(t-1) \end{pmatrix} \tag{10}$$

$$y(t) = Cx(t)^- + Du(t) + e(t) \tag{11}$$

where $x(t)^-$ is the maximum likelihood and least squares estimator for $x(t)$ given the values of $u(s)$ and $y(s)$ for $s < t$; A_{KF} and B_{KF} are matrices which can be calculated in terms of $A, B, C, \Sigma_x, \Sigma_y$, $e(t) \sim N(0, \Sigma_{\mathrm{KF}})$ for some covariance matrix Σ_{KF} that can be calculated in terms of $A, B, C, \Sigma_x, \Sigma_y$. This is now a *filtering* problem, where we have to regress the output on both previous inputs $u(t)$ and outputs $y(t)$. This is more challenging, because unlike previous $u(t)$, the previous $y(t)$ are highly correlated. One can perhaps treat this as a low-rank approximation in a different norm.

$\mathcal{H}_\infty$ error bounds. How can we learn the system with $\mathcal{H}_\infty$ error bounds, that is, obtain error bounds under worst case input? This is particularly useful in control. We do not expect we can achieve $\sqrt{\frac{d}{T}}$ rates under iid inputs $u(t)$. However, it may be possible to take an active learning approach, by maximally exciting the system at frequencies we wish to learn, as in (Wagenmaker and Jamieson, 2020).

Improved rates for learning system matrices. Can we learning the system matrices with rates depending logarithmically on the memory L, perhaps by incorporating the multi-scale idea into system identification?

Input design. In this work we choose iid random inputs, but can we estimate more efficiently with well-designed deterministic inputs? Can we design inputs to respect constraints such as constraints on frequencies? Sarker et al. (2020) suggests that efficient estimation is possible under general conditions on the inputs.

More general noise. Do guarantees still hold if the noise satisfies weaker conditions such as sub-gaussianity? A key difficulty is bounding the maximum Fourier coefficient (as in Lemma A.6).

Acknowledgments

The author thanks Xue Chen for useful discussions.

References

Oren Anava, Elad Hazan, Shie Mannor, and Ohad Shamir. Online learning for time series prediction. In *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, pages 172–184, 2013.

Badri Narayan Bhaskar, Gongguo Tang, and Benjamin Recht. Atomic norm denoising with applications to line spectral estimation. *IEEE Transactions on Signal Processing*, 61(23):5987–5999, 2013.

Ross Boczar, Nikolai Matni, and Benjamin Recht. Finite-data performance guarantees for the output-feedback control of an unknown system. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 2994–2999. IEEE, 2018.

Jian-Feng Cai, Xiaobo Qu, Weiyu Xu, and Gui-Bo Ye. Robust recovery of complex exponential signals from random gaussian projections via low rank hankel matrix reconstruction. *Applied and computational harmonic analysis*, 41(2):470–490, 2016.

Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57 (4):2342–2359, 2011.

Shaunak Chatterjee and Stuart Russell. Why are dbns sparse? In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 81–88. JMLR Workshop and Conference Proceedings, 2010.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pages 1–47, 2019.

Boualem Djehiche, Othmane Mazhar, and Cristian R Rojas. Finite impulse response models: A non-asymptotic analysis of the least squares estimator. *arXiv preprint arXiv:1911.12794*, 2019.

Maryam Fazel, Ting Kei Pong, Defeng Sun, and Paul Tseng. Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, 34(3):946–977, 2013.

Udaya Ghai, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. No-regret prediction in marginally stable systems. In *COLT 2020 - The 33rd Annual Conference on Learning Theory, July 9-12, 2020*, pages 1–44, 2020.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *The Journal of Machine Learning Research*, 19(1):1025–1068, 2018.

Elad Hazan, Karan Singh, and Cyril Zhang. Learning linear dynamical systems via spectral filtering. In *Advances in Neural Information Processing Systems*, pages 1–2, 2017.

Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems*, pages 4634–4643, 2018.

BL Ho and Rudolph E Kalman. Effective construction of linear state-variable models from input/output functions. *at-Automatisierungstechnik*, 14(1-12):545–548, 1966.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82.1:35–45, 1960.

Mark Kozdoba, Jakub Marecek, Tigran T. Tchrakian, and Shie Mannor. On-line learning of linear dynamical systems: Exponential forgetting in kalman filters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4098–4105, 2019.

Felix Krahmer, Shahar Mendelson, and Holger Rauhut. Suprema of chaos processes and the restricted isometry property. *Communications on Pure and Applied Mathematics*, 67(11):1877–1904, 2014.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Lennart Ljung. *System identification: Theory for the User*. Prentice Hall, Upper Saddle River, NJ, 2 edition, 1998.

Nikolai Matni and Stephen Tu. A tutorial on concentration bounds for system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3741–3749. IEEE, 2019.

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

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

S Joe Qin. An overview of subspace identification. *Computers & chemical engineering*, 30(10-12): 1502–1513, 2006.

Paria Rashidinejad, Jiantao Jiao, and Stuart Russell. Slip: Learning to predict in unknown dynamical systems with long-term memory. *arXiv preprint arXiv:2010.05899*, 2020.

Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:253–279, 2019.

Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Finite-time system identification for partially observed lti systems of unknown order. *arXiv preprint arXiv:1902.01848*, 2019.

Arnab Sarker, Joseph E Gaudio, and Anuradha M Annaswamy. Parameter estimation bounds based on the theory of spectral lines. *arXiv preprint arXiv:2006.12687*, 2020.

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

Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. *arXiv preprint arXiv:1902.00768*, 2019.

Yue Sun, Samet Oymak, and Maryam Fazel. Finite sample system identification: Improved rates and the role of regularization. In *2nd L4DC 2020 - Learning for Dynamics & Control, June 11-12, 2020*, 2020.

Anastasios Tsiamis and George Pappas. Online learning of the kalman filter with logarithmic regret. *arXiv preprint arXiv:2002.05141*, 2020.

Anastasios Tsiamis, Nikolai Matni, and George Pappas. Sample complexity of kalman filtering for unknown systems. In *Learning for Dynamics and Control*, pages 435–444. PMLR, 2020.

Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Conference on Learning Theory*, pages 3036–3083. PMLR, 2019.

Stephen Tu, Ross Boczar, Andrew Packard, and Benjamin Recht. Non-asymptotic analysis of robust control from coarse-grained identification. *arXiv preprint arXiv:1707.04791*, 2017.

Peter Van Overschee and BL De Moor. *Subspace Identification for Linear Systems*. Springer Science & Business Media, 2012.

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

Andrew Wagenmaker and Kevin Jamieson. Active learning for identification of linear dynamical systems. In *COLT 2020 - The 33rd Annual Conference on Learning Theory, July 9-12, 2020*, 2020.

Kemin Zhou, John Comstock Doyle, Keith Glover, et al. *Robust and optimal control*, volume 40. Prentice hall New Jersey, 1996.

Appendix A. Linear regression for impulse response

In this section we prove Theorem 4.2: under iid gaussian inputs, we can obtain high-probability error bounds for the transfer function of the learned impulse response in $\mathcal{H}_\infty$ norm. Moreover, these bounds kick in as soon as we have $\widetilde{\Omega}(L)$ samples from a *single* rollout. We note that analyzing the multiple-rollout setting as in Sun et al. (2020) is more straightforward, so we will not consider it here.

The main difficulty for analyzing linear regression is that the inputs are correlated. The most challenging step is to lower-bound the sample covariance matrix of inputs to the linear regression.

In the SISO (single input single output) setting, Djehiche et al. (2019) give concentration bounds for the covariance matrix with $T = \widetilde{\Omega}(L)$ timesteps. First, we extend this to the MIMO (multiple input multiple output) setting in Theorem A.2 (Note that Oymak and Ozay (2019) consider the MIMO case but have extra log factors.) Then, we use Gaussian suprema arguments as in Tu et al. (2017) to obtain bounds for the transfer function in $\mathcal{H}_\infty$ norm (Lemma A.6).

We suppose the inputs $u(0), \ldots, u(T-1) \sim N(0, I_{d_u})$ are iid, observe $y(0), \ldots, y(T-1) \in \mathbb{R}^{d_y}$, and perform linear regression on the finite impulse response $F : \{0, 1, \ldots, L\} \to \mathbb{R}^{d_y \times d_u}$ (which we will also treat as an element of $\mathbb{R}^{(L+1)\times d_y \times d_u}$ without further comment).

Recall that given a sequence $(F(t))_{t=0}^{a-1}$ where each $F(t) \in \mathbb{C}^{m\times n}$, the Toeplitz matrix is given by

$$\mathrm{Toep}_{a\times b}(F) = \begin{bmatrix} F(0) & 0 & \cdots & 0 \\ F(1) & F(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ F(a-1) & F(a-2) & \cdots & \end{bmatrix} \in \mathbb{C}^{am\times bn}.$$

SISO setting. For simplicity, first consider the SISO setting: $d_u = d_y = 1$ and $\eta(t) \sim N(0, 1)$. In this case, we learn a finite impulse response $f \in \mathbb{R}^{L+1}$ by minimizing the loss function

$$\|y - u * f\|_{[0,T-1]}^2 = \sum_{t=0}^{T-1} \left\| y(t) - u_{t:t-L}^\top f \right\|^2 = \|y_{0:T-1} - Uf\|^2 \tag{12}$$

where we let $y_{0:T-1}$ denote the vertical concatenation of $y(0), \ldots, y(T-1)$ and similarly for $u_{t:t-L}$, and let $U = \mathrm{Toep}_{T\times(L+1)}((u(t))_{t\geq 0})$. We set $u(t) = 0$ for $t < 0$. Solving the least-squares problem gives

$$f = (U^\top U)^{-1} U^\top y_{0:T-1}.$$

Suppose that the data is generated as $y = f^* * u + \eta$ where $\eta(t) \sim N(0, 1)$ are independent and f^* is supported on $[0, L]$. Later, we will consider the effect of truncating an infinite response. We abuse notation by considering f, f^* both as functions $\mathbb{Z} \to \mathbb{R}$ and as vectors in $\mathbb{R}^{L+1}$, as they are supported in $[0, L]$. Similarly, we consider y, η as vectors in $\mathbb{R}^T$. Then as vectors in $\mathbb{R}^T$, $y = Uf^* + \eta$. Hence the error is

$$f - f^* = (U^\top U)^{-1} U^\top (Uf^* + \eta) - f^* = (U^\top U)^{-1} U^\top \eta$$

Because η has iid Gaussian entries,

$$f - f^* \sim N(0, (U^\top U)^{-1}).$$

To bound this, we need to bound $\left\|(U^\top U)^{-1}\right\|$, and hence bound the smallest singular value of $U^\top U$.

Notation for MIMO setting. For a vector or matrix-valued function $F : \{a, a+1, \ldots, b\} \to \mathbb{C}^{d_1 \times d_2}$, define

$$M_{F,a:b} = \begin{pmatrix} F(a)^\top \\ \vdots \\ F(b)^\top \end{pmatrix} \in \mathbb{C}^{(b-a+1)d_2 \times d_1}$$

with the indices omitted if they are clear from context.

MIMO setting. In the general case, we would like to learn $F = (F(t) \in \mathbb{R}^{d_y \times d_u})_{t=0}^{L} \in \mathbb{R}^{(L+1) \times d_y \times d_u}$. Now suppose the data is generated as

$$y = F^* * u + G^* * \xi + \eta$$

where F^*, G^* are supported on $[0, \infty)$ and $\eta(t) \sim N(0, \Sigma_y)$, $\xi(t) \sim N(0, \Sigma_x)$, $t \geq 0$ are independent. Let $U = \mathrm{Toep}_{T \times (L+1)}((u(t)^\top)_{t=0}^{T-1})$ as before. Truncating F^* and G^*, we have

$$\begin{aligned} y &= (F^* \mathbb{1}_{[0,L]}) * u + (G^* \mathbb{1}_{[0,L]}) * \xi + \eta + e \\ \text{where } e(t) &= (F^* \mathbb{1}_{[L+1,\infty)}) * u + (G^* \mathbb{1}_{[L+1,\infty)}) * \xi. \end{aligned}$$

Thus, by taking the transpose and stacking vectors,

$$\begin{aligned} &M_{y,0:T-1} = U M_{F^*,0:L-1} + W M_{G^*,0:L-1} + M_{\eta,0:T-1} + M_{e,0:T-1} \\ &\text{where } W = \mathrm{Toep}_{T \times (L+1)}((\xi(t))^\top). \end{aligned}$$

The least squares solution F minimizes $\|Y - U M_F\|_{\mathsf{F}}^2$, so and the error is

$$M_F - M_{F^*} = (U^\top U)^{-1} U^\top M_\eta + (U^\top U)^{-1} U^\top W M_{G^*} + (U^\top U)^{-1} M_e. \tag{13}$$

A.1. Lower bounding sample covariance matrix

In this subsection we lower bound the sample covariance matrix.

Lemma A.1 *There is a constant K such that the following holds. Let $u(t) \sim N(0, I_{d_u})$ and $U = \mathrm{Toep}_{T \times (L+1)}((u(t)^\top)_{t \geq 0})$. Then for $0 < \delta \leq \frac{1}{2}$, $T \geq K_1 L d_u \log\left(\frac{L d_u}{\delta}\right)$,*

$$\mathbb{P}\left(\sigma_{\min}(U^\top U) \geq \frac{T}{2}\right) \geq 1 - \delta.$$

This is a corollary of the following concentration bound, which generalizes Theorem 3.4 of Djehiche et al. (2019) to the MIMO setting. The main additional ingredient is an ε-net argument to reduce to the analysis of the SISO case. We also swap out the chaining argument with a use of Lemma A.4, which allows a shorter proof.

Theorem A.2 *There is K such that the following holds. Suppose $u(t)$, $0 \leq t < T$ are independent, zero-mean, and K_u-sub-gaussian (see Definition D.1), and let $U = \mathrm{Toep}_{T \times (L+1)}((u(t)^\top)_{t \geq 0})$. Then for $0 < \delta \leq \frac{1}{2}$, $T \geq L$,*

$$\left\| U^\top U - T I_{d_u} \right\| \leq K K_u^2 \left(L d_u \log\left(\frac{T}{\delta}\right) + \sqrt{T L d_u \log\left(\frac{T}{\delta}\right)} \right)$$

with probability $\geq 1 - \delta$.

We first note the fact that infinite Toeplitz matrices become diagonal in the Fourier basis.

Lemma A.3 *Consider the infinite block Toeplitz matrix* $(Z(j-k))_{j,k\in\mathbb{Z}} \in \mathbb{C}^{(\mathbb{Z}\times d_1)\times(\mathbb{Z}\times d_2)}$*, where* Z *is a function* $\mathbb{Z}\to\mathbb{C}^{d_1\times d_2}$*. In the Fourier basis, it is given by the kernel* $\widehat{Z}(\omega_1)\mathbb{1}_{\omega_1=\omega_2}$*. That is, if* $v:\mathbb{Z}\to\mathbb{R}^{d_2}$*,* $\|Z\|_1, \|v\|_1 < \infty$*, then letting*

$$w(j) = \sum_k Z(j-k)v(k),$$

we have

$$\widehat{w}(\omega) = \widehat{Z}(\omega)\widehat{v}(\omega).$$

Here, $\widehat{Z}(\omega)$ is called the **multiplication polynomial** of the matrix.
Proof Simply note that $w = Z * v$ and so $\widehat{w} = \widehat{Z}\widehat{v}$. ■

We will use the following lemma in order to bound the maximum of the Fourier transform by the maximum at a finite number of points.

Lemma A.4 (Bhaskar et al. (2013)) *Let* $Q(z) := \sum_{k=0}^{r-1} a_k z^k$*, where* $a_k \in \mathbb{C}$*. For any* $N \geq 4\pi r$*,* $\|Q\|_{\mathcal{H}_\infty} \leq \left(1+\frac{4\pi r}{N}\right)\max_{j=0,\dots,N-1}|Q(e^{\frac{2\pi i j}{N}})|$*.*

Proof [Proof of Theorem A.2] By rescaling we may suppose $K_u = 1$. Decompose

$$U = U_1 + U_2 \text{ where} \tag{14}$$

$$U_1 = \begin{pmatrix} u(0)^\top & & \mathbf{0} \\ \vdots & \ddots & u(0)^\top \\ u(T-L-1)^\top & \ddots & \vdots \\ \mathbf{0} & & u(T-L-1)^\top \end{pmatrix}$$

$$U_2 = \begin{pmatrix} \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ u(T-L)^\top & & & \vdots \\ \vdots & \ddots & & \vdots \\ u(T-1)^\top & \cdots & u(T-L)^\top & \mathbf{0} \end{pmatrix}.$$

Then

$$U^\top U = (T-L)I_{Ld_u} + (U_1^\top U_1 - (T-L)I_{Ld_u}) + U_1^\top U_2 + U_2^\top U_1 + U_2^\top U_2. \tag{15}$$

Let T be the shift operator on functions: $\mathsf{T}f(t) = f(t-1)$. Let $T' = T-L$ and let $u^{(1)} = u\mathbb{1}_{[0,T'-1]}$. Then the (j,k)th block of $U_1^\top U_1$ is

$$(U_1^\top U_1)_{jk} = \sum_{t\in\mathbb{Z}} (\mathsf{T}^j u^{(1)})(t)(\mathsf{T}^k u^{(1)})(t)^\top$$

Define the infinite block Toeplitz matrix in $\mathbb{R}^{(\mathbb{Z}\times d_u)\times(\mathbb{Z}\times d_u)}$ by

$$Z_{jk} = \sum_{t=1}^{T} (\mathsf{T}^j u^{(1)})(t)(\mathsf{T}^k u^{(1)})(t)^\top \mathbb{1}_{|j-k|\leq L} - T' I_{\mathbb{Z}\times d_u}.$$

By Lemma A.3, the multiplication polynomial of this matrix is

$$\begin{aligned} P_u(\omega) &= \sum_{\ell=-L}^{L} \sum_{t\in\mathbb{Z}} (\mathsf{T}^\ell u^{(1)})(t) u^{(1)}(t)^\top e^{-2\pi i \ell\omega} - T' I_{d_u} \\ &= \sum_{\substack{j,k\,\in\,\mathbb{Z} \\ |j-k|\,\le\,L}} u(j)u(k)^\top e^{2\pi i (j-k)\omega} - T' I_{d_u} \\ &= \begin{pmatrix} u(0) & \cdots & u(T'-1) \end{pmatrix} M \begin{pmatrix} u(0)^\top \\ \cdots \\ u(T'-1)^\top \end{pmatrix} - T' I_{d_u} \end{aligned} \tag{16}$$

where $M \in \mathbb{C}^{\mathbb{Z}\times\mathbb{Z}}$ is the matrix with $M_{jk} = e^{2\pi i(j-k)\omega}\mathbb{1}_{|j-k|\le L}$. In order to work with a scalar-valued function, we consider for $\|v\| = 1$

$$v^\top P_u(\omega) v = \sum_{j,k\in\{0,\ldots,T'-1\}} \langle v, u(j)\rangle \langle v, u(k)\rangle e^{2\pi i\omega(j-k)} \mathbb{1}_{|j-k|\le L} - T' \|v\|^2 .$$

By Lemma A.3,

$$\left\| U_1^\top U_1 - T' I_{Td_u} \right\| \le \left\| Z - T' I_{\mathbb{Z}} \right\| \le \|P_u(\omega)\|_{\mathcal{H}_\infty} .$$

Taking $N = \lceil 8\pi L \rceil$ and noting $e^{2\pi i\omega L} P(\omega)$ is a polynomial of degree at most $2L$ in $e^{2\pi i\omega}$, we have

$$\begin{aligned} \|P_u(\omega)\|_{\mathcal{H}_\infty} = \sup_{\omega\in[0,1]} \|P_u(\omega)\| &= \sup_{\|v\|=1} \sup_{\omega\in[0,1]} |v^\top P_u(\omega) v| \\ &= \sup_{\|v\|=1} 2 \max_{\omega\in\{0,\frac{1}{N},\cdots\}} |v^\top P_u(\omega) v| && \text{by Lemma A.4} \\ &= 2 \max_{\omega\in\{0,\frac{1}{N},\cdots\}} \left(\sup_{v\in\mathcal{N}_\varepsilon} |v^\top P_u(\omega) v| + 3\varepsilon \|P_u(\omega)\| \right) \end{aligned} \tag{17}$$

where $\mathcal{N}_\varepsilon$ is an ε-net of the unit sphere in $\mathbb{R}^d$. (For arbitrary v' with $\|v'\| = 1$, write $v = v + \Delta v$ where $v \in \mathcal{N}_\varepsilon$ and $\|\Delta v\| \le \varepsilon$.) We first bound $v^\top P(\omega) v$. Letting $w \in \mathbb{R}^{T'}$ be the vector with entries $w(j) = \langle v, u(j)\rangle$, we have

$$v^\top P_u(\omega) v = w^\top M w - T' \|v\|^2 .$$

Fix v. Because each $u(t)$ is independent 1-subgaussian, each entry of w is 1-subgaussian. By the Hanson-Wright inequality (Theorem D.3), for some constant $c > 0$,

$$\mathbb{P}(|v^\top P_u(\omega) v| > s) \le 2\exp\left[-c \cdot \min\left\{ \frac{s^2}{\|M\|_{\mathsf{F}}^2}, \frac{s}{\|M\|} \right\} \right] .$$

We calculate that $\|M\|_{\mathsf{F}}^2 \le (2L+1)T$ and the Fourier transform of the function $e^{2\pi i\omega t}\mathbb{1}_{|j-k|\le L}$ satisfies $\left\|\widehat{f}\right\|_\infty \le \|f\|_1 \le 2L+1$, so by Lemma A.3, $\|M\| \le 2L+1$. Then for appropriate K,

$$\mathbb{P}\left(|v^\top P_u(\omega) v| > K \left(\sqrt{TL \log\left(\frac{1}{\delta_1}\right)} + L \log\left(\frac{1}{\delta_1}\right) \right) \right) \le \delta_1 .$$

Next we bound $\|P_u(\omega)\|$ and choose ε appropriately. A crude bound with Markov's inequality suffices to bound $\|P_u(\omega)\|$. We have (because the second moment is at most the sub-gaussian constant)

$$\mathbb{E}\left\|\begin{pmatrix} u(0) & \cdots & u(T'-1)\end{pmatrix}\right\|_{\mathsf{F}}^2 \le \mathbb{E}\sum_{j=1}^{d_u}\sum_{t=0}^{T'-1}\langle e_j, u(t)\rangle^2 \le d_u T'$$

so with probability $\ge 1-\delta_2$, $\left\|\begin{pmatrix} u(0) & \cdots & u(T'-1)\end{pmatrix}\right\|_{\mathsf{F}}^2 \le \frac{d_u T'}{\delta_2}$. Hence, for every $\omega \in [0,1]$, by (16),

$$\left\|P_u(\omega) + T' I_{d_u}\right\| \le \left\|\begin{pmatrix} u(0) & \cdots & u(T'-1)\end{pmatrix}\right\|_{\mathsf{F}}^2 \|M\| \le \frac{d_u T'}{\delta_2} 2L.$$

Choose $\varepsilon = \frac{\delta_2}{2d_u LT}$. Then with probability $\ge 1-\delta_2$, we have

$$\sup_{\omega\in[0,1]} 3\varepsilon \|P_u(\omega)\| \le 3\cdot\frac{\delta_2}{2d_u LT}\cdot\left(\frac{d_u T'}{\delta_2}2L + T'\right) \le 4.5.$$

Now take $\delta_1 = \frac{\delta}{2}$. By Cor. 4.2.13 of Vershynin (2018), there is an ε-net of size $|\mathcal{N}_\varepsilon| \le \left(1+\frac{2}{\varepsilon}\right)^{d_u} = \exp\left(d_u \log\left(1+\frac{2}{\varepsilon}\right)\right) = \exp\left(d_u \log\left(1+\frac{8d_u LT}{\delta}\right)\right)$. Letting $\delta_1 = \frac{\delta}{2|\mathcal{N}_\varepsilon|}$ and taking a union bound, with probability $1-\delta$ we get

$$(17) \le K\left(\sqrt{TLd\log\left(\frac{T}{\delta}\right)} + Ld\log\left(\frac{T}{\delta}\right)\right)$$

Next consider the term $U_1^\top U_2$. Let $u^{(1)} = u\mathbb{1}_{[0,T'-1]}$, $u^{(2)} = u\mathbb{1}_{[T',T-1]}$. This is part of the infinite Toeplitz matrix with $Z_{jk} = \sum_{t\in\mathbb{Z}}(\mathsf{T}^j u^{(1)})(t)(T^k u^{(2)})(t)^\top \mathbb{1}_{|j-k|\le L-1}$. In the Fourier basis,

$$\begin{aligned} P_{u,12}(\omega) &= \sum_{\substack{j,k\in\mathbb{Z}\\ |j-k|\le L}}\sum_{t\in\mathbb{Z}} e^{-2\pi i j\omega}(\mathsf{T}^j u^{(1)})(t)(\mathsf{T}^k u^{(2)})(t)^\top e^{2\pi i k\omega} \\ &\quad \sum_{\substack{j,k\in\mathbb{Z}\\ |j-k|\le L}} \mathbb{1}_{[0,T'-1]}(j)\mathbb{1}_{[T-L,T-1]}(k) e^{2\pi i(j-k)\omega} u(j)u(k) \\ &= \begin{pmatrix} u(0) & \cdots & u(T'-1)\end{pmatrix} M \begin{pmatrix} u(0)^\top \\ \cdots \\ u(T'-1)^\top\end{pmatrix} \end{aligned}$$

where $M_{jk} = \mathbb{1}_{[0,T'-1]}(j)\mathbb{1}_{[T-L,T-1]}(k)\mathbb{1}_{|j-k|\le L}e^{2\pi i(j-k)\omega}$. As before, we have

$$\left\|U_1^\top U_2\right\| \le 2\max_{\omega\in\{0,\frac{1}{N},\cdots\}}\sup_{v\in\mathcal{N}_\varepsilon}|v^\top P_{u,12}(\omega)v| + 3\varepsilon\|P_{u,12}(\omega)\|.$$

We calculate $\|M\|_{\mathsf{F}}^2 \le (T-L)(2L+1)$ and each block in M is part of a Toeplitz matrix, so similarly to before $\|M\| \le 2L+1$. Hence, with probability at least $1-\delta$,

$$\left\|U_1^\top U_2\right\| \le K\left(\sqrt{TL\log\left(\frac{1}{\delta}\right)} + L\log\left(\frac{1}{\delta}\right)\right)$$

Note $\left\|U_1^\top U_2\right\| = \left\|U_2^\top U_1\right\|$. Finally, we bound $U_2^\top U_2$. Note U_2 is part of an infinite Hankel matrix with entries $u(T'+1)^\top, \ldots, u(T'+L)^\top$. The multiplication polynomial is

$$P_{u,2}(\omega) = e^{-2\pi i(T-L)\omega} \sum_{t=0}^{L-1} u(T-L+t)^\top e^{-2\pi i t\omega}.$$

The real part is $K\left(\sum_{t=T-L}^{T-1} \cos^2(-2\pi t\omega)\right)^{1/2}$-sub-gaussian and the imaginary part is $K\left(\sum_{t=T-L}^{T-1} \sin^2(-2\pi t\omega)\right)^{1/2}$-sub-gaussian for some constant K. Hence

$$\mathbb{P}\left(\langle e_j, P_{u,2}(\omega)\rangle^2 \leq KL\log\left(\frac{1}{\delta}\right)\right) \geq 1-\delta.$$

Using this for $j = 1, \ldots, d_u$, replacing $\delta \leftarrow\!\shortmid \frac{\delta}{d_u}$, and using a union bound gives

$$\mathbb{P}\left(\|P_{u,2}(\omega)\|^2 \leq KLd_u\log\left(\frac{d}{\delta}\right)\right) \geq 1-\delta.$$

Now for $N \geq 4\pi L$, using another union bound gives

$$\left\|U_2^\top U_2\right\| \leq \left(\sup_{\omega\in[0,1]} \|P_{u,2}(\omega)\|\right)^2 \leq \left(2\max_{\omega\in\{0,\frac{1}{N},\ldots\}} \|P_{u,2}(\omega)\|\right)^2 \leq KLd_u\log\left(\frac{d_uL}{\delta}\right).$$

with probability $\geq 1-\delta$. Putting all the bounds together with (15) gives the theorem. ■

Proof [Proof of Lemma A.1] For large enough K_2, for $T \geq K_2Ld_u\log\left(\frac{T}{\delta}\right)$, we have that by Theorem A.2 that $\left\|U^\top U - TI_{d_u}\right\| \leq \frac{T}{2}$, so $\sigma_{\min}(U^\top U) \geq \frac{T}{2}$.

Finally, note that for large enough K_1, $T \geq K_1Ld_u\log\left(\frac{Ld_u}{\delta}\right)$ implies $T \geq K_2Ld_u\log\left(\frac{T}{\delta}\right)$. ■

We show here a bound similar to Theorem A.2 that will be useful to us later.

Lemma A.5 *There is a constant K such that the following holds. Suppose $u(t)$, $0 \leq t < T$ are independent, zero-mean, and K_u-sub-gaussian, and similarly for $w(t)$ with constant K_w. Let $U = \mathrm{Toep}_{T\times(L+1)}((u(t)^\top)_{t\geq 0})$, $W = \mathrm{Toep}_{T\times(L+1)}((w(t)^\top)_{t\geq 0})$. Then for $0 < \delta \leq \frac{1}{2}$, $T \geq K_1Ld_u\log\left(\frac{Ld_u}{\delta}\right)$,*

$$\left\|U^\top W\right\| \leq KK_uK_w\left(Ld_u\log\left(\frac{T}{\delta}\right) + \sqrt{TLd_u\log\left(\frac{T}{\delta}\right)}\right)$$

with probability at least $1-\delta$.

Proof By scaling we may assume $K_u = K_w = 1$. Decompose $U = U_1 + U_2$ and $W = W_1 + W_2$ as in (14). Let $S_a = \{0, \ldots, T-L-1\}$ and $S_b = \{T-L, \ldots, T-1\}$. We have

$$U^\top W = \sum_{a,b\in\{0,1\}} U_a^\top W_b.$$

Let $u^{(a)} = u\mathbb{1}_{S_a}$ and similarly define $w^{(b)} = w\mathbb{1}_{S_b}$. Then the (j,k)th block of $U_a^\top W_b$ is

$$(U_a^\top W_b)_{jk} = \sum_{t\in\mathbb{Z}} (\mathsf{T}^j u^{(a)})(t)(\mathsf{T}^k w^{(b)})(t)^\top.$$

This is part of the infinite block Toeplitz matrix in $\mathbb{R}^{(\mathbb{Z}\times d_u)\times(\mathbb{Z}\times d_u)}$ defined by

$$Z_{jk} = \sum_{t\in\mathbb{Z}} (\mathsf{T}^j u^{(a)})(t)(\mathsf{T}^k w^{(b)})(t)^\top \mathbb{1}_{|j-k|\le L},$$

with multiplication polynomial

$$\begin{aligned} P_{u,ab}(\omega) &= \sum_{|j-k|\le L} u^{(a)}(j) w^{(b)}(k) e^{2\pi i(j-k)\omega} \\ &= \begin{pmatrix} u(0) & \cdots & u(T-1) \end{pmatrix} M \begin{pmatrix} w(0)^\top \\ \cdots \\ w(T-1)^\top \end{pmatrix} \\ \text{where } M_{jk} &= e^{2\pi i(j-k)} \mathbb{1}_{j\in S_a} \mathbb{1}_{k\in S_b} \mathbb{1}_{|j-k|\le L}. \end{aligned}$$

We calculate that $\|M\|_F^2 \le T(2L+1)$ and $\|M\| \le 2L+1$ so the same argument as in Theorem A.2 (but using the version of Hanson-Wright given by Corollary D.4) gives that

$$\left\|U^\top W\right\| \le K K_u K_w \left(L d_u \log\left(\frac{T}{\delta}\right) + \sqrt{T L d_u \log\left(\frac{T}{\delta}\right)} \right).$$

■

A.2. Upper bound in $\mathcal{H}_\infty$ norm

The following Lemma A.6 generalizes the results of Tu et al. (2017) to the MIMO setting. To get the right dimension dependence, we will use the concentration bound for covariance given by Theorem D.2.

Lemma A.6 *There is a constant K such that the following holds. Suppose that $\eta(0),\dots,\eta(T-1) \sim N(0,\Sigma)$ are iid, $\Phi \in \mathbb{R}^{(L+1)d_u\times T}$, and $E(0),\dots,E(L) \in \mathbb{R}^{d_y\times d_u}$ are such that*

$$\begin{pmatrix} E(0)^\top \\ \vdots \\ E(L)^\top \end{pmatrix} = M_E = \Phi M_\eta \in \mathbb{R}^{(L+1)d_u\times d_y}$$

For any $0 < \delta \le \frac{1}{2}$ and $-1 \le a < L - L'$,

$$\mathbb{P}\left(\left\|E\mathbb{1}_{[a+1,a+L']}\right\|_{\mathcal{H}_\infty} \le K\sqrt{L'}\,\|\Sigma\|^{\frac{1}{2}}\,\|\Phi\| \sqrt{d_y + d_u + \log\left(\frac{L'}{\delta}\right)} \right) \ge 1-\delta.$$

Proof First, by considering

$$M_E\Sigma^{-1/2} = \Phi(M_\eta\Sigma^{-1/2}),$$

we may reduce to the case where $\eta(t) \sim N(0, I_{d_u})$ are iid, i.e., all entries of M_η are iid standard gaussian.

Let $M_\omega = (E\mathbb{1}_{[a+1,a+L']})^\wedge(\omega) \in \mathbb{C}^{d_y\times d_u}$. Note that

$$M_\omega = (\phi_\omega^H \otimes I_{d_u})M_E$$
$$\text{where } \phi_\omega = (e^{2\pi i k\omega}\mathbb{1}_{a+1\le k\le a+L'})_{0\le k\le L} \in \mathbb{R}^{L+1}$$

as a column vector. Because the columns of M_η are independent and distributed as $N(0, I_T)$, the columns m_j of M_ω are independent. To bound M_ω, it suffices to bound $M_\omega M_\omega^H = \sum_{j=1}^{d_y} m_j m_j^H$. Note that

$$\left\|\mathbb{E}m_j m_j^H\right\| = \left\|(\phi_\omega^H \otimes I_{d_u})\Phi\Phi^\top(\phi_\omega \otimes I_{d_u})\right\| \le L' \left\|\Phi\right\|^2.$$

Let $\Phi' = (\phi_\omega^H \otimes I_{d_u})\Phi\Phi^\top(\phi_\omega \otimes I_{d_u})$. By Theorem D.2[1],

$$\mathbb{P}\left(\left\|\frac{1}{d_y}\sum_{j=1}^{d_y} m_j m_j^H - \Phi'\right\| \le KL'\left\|\Phi'\right\|^2\left(\sqrt{\frac{d_u+s}{d_y}} + \frac{d_u+s}{d_y}\right)\right) \ge 1-2e^{-s}$$
$$\implies \mathbb{P}\left(\left\|\frac{1}{d_y}\sum_{j=1}^{d_y} m_j m_j^H\right\| \le KL'\left\|\Phi'\right\|^2\left(1+\frac{d_u+\log\left(\frac{2}{\delta}\right)}{d_y}\right)\right) \ge 1-\delta$$

by taking $u = \log\left(\frac{2}{\delta}\right)$. Multiplying by d_y gives

$$\mathbb{P}\left(\left\|M_\omega M_\omega^H\right\| \le KL'\left\|\Phi'\right\|^2\left(d_y + d_u + \log\left(\frac{2}{\delta}\right)\right)\right) \ge 1-\delta.$$

Replacing δ by $\frac{\delta}{N}$, taking the square root, and taking a union bound gives

$$\mathbb{P}\left(\max_{\omega\in\{0,\frac{1}{N},\dots,\frac{N-1}{N}\}}\left\|M_\omega\right\| \le K\sqrt{L'}\left\|\Phi'\right\|^{\frac{1}{2}}\sqrt{d_y + d_u + \log\left(\frac{2}{\delta}\right)}\right) \ge 1-\delta. \tag{18}$$

Finally, we note that by Lemma A.4, for $N = \lceil 4\pi L'\rceil$,

$$\begin{aligned}\left\|E\mathbb{1}_{[a+1,a+L']}\right\|_{\mathcal{H}_\infty} &= \sup_{\omega\in[0,1]}\left\|E\widehat{\mathbb{1}_{[a+1,a+L']}}(\omega)\right\| = \sup_{\|v\|_2=1}\sup_{\omega\in[0,1]}\left\|E\widehat{\mathbb{1}_{[a+1,a+L']}}(\omega)v\right\| \\ &\le \sup_{\|v\|_2\le 1}\max_{\omega\in\{0,\frac{1}{N},\dots,\frac{N-1}{N}\}} 2\left\|E\widehat{\mathbb{1}_{[a+1,a+L']}}(\omega)v\right\| \\ &\le 2\max_{\omega\in\{0,\frac{1}{N},\dots,\frac{N-1}{N}\}}\left\|E\widehat{\mathbb{1}_{[a+1,a+L']}}(\omega)\right\|.\end{aligned}$$

Combining this with (18) gives the result. ■

Finally we can put everything together to obtain a $\mathcal{H}_\infty$ error bound for linear regression.

1. The theorem is stated for real matrices, but we can view the matrix as acting on a real vector space of twice the dimension.

Lemma A.7 *There are K_1, K_2 such that the following hold. Suppose $y = F^* * u + G^* * \xi + \eta$ where $u(t) \sim N(0, I_{d_u})$, $\xi(t) \sim N(0, \Sigma_x)$, $\eta(t) \sim N(0, \Sigma_y)$ for $0 \le t < T$, and $\mathrm{Supp}(F^*), \mathrm{Supp}(G^*) \subseteq [0, \infty)$. Let $F = \mathrm{argmin}_{F \in \{0,\dots,L\} \to \mathbb{R}^{d_y \times d_u}} \sum_{t=0}^{T-1} |y(t) - (F * u)(t)|^2$, $M_{G^*} = (G^*(0), \dots, G^*(L))^\top \in \mathbb{R}^{(L+1)d \times d_y}$, and $\varepsilon_{\mathrm{trunc}} = \left\|F^* \mathbb{1}_{[L+1,\infty)}\right\|_{\mathcal{H}_\infty} \sqrt{d_u} + \left\|G^* \mathbb{1}_{[L+1,\infty)}\right\|_{\mathcal{H}_\infty} \left\|\Sigma_x^{1/2}\right\|_{\mathsf{F}}$. For $0 < \delta \le \frac{1}{2}$, $T \ge K_1 L d_u \log\left(\frac{L d_u}{\delta}\right)$, $1 \le L' \le L$, $-1 \le a < L - L'$,*

$$\left\|(F - F^*)\mathbb{1}_{[a+1,a+L']}\right\|_{\mathcal{H}_\infty}$$
$$\le K_2 \left[\sqrt{\frac{1}{T}} \left(\sqrt{\|\Sigma_y\| L' \left(d_u + d_y + \log\left(\frac{L'}{\delta}\right)\right)} + \sqrt{\|\Sigma_x\| L' L d_u \log\left(\frac{L d_u}{\delta}\right)} \|M_{G^*}\|\right) + \varepsilon_{\mathrm{trunc}}\right]$$

with probability at least $1 - \delta$.

Proof By (13), using the notation defined there,

$$M_F - M_{F^*} = \underbrace{(U^\top U)^{-1} U^\top M_\eta}_{=:E_1} + \underbrace{(U^\top U)^{-1} U^\top W M_{G^*}}_{=:E_2} + \underbrace{(U^\top U)^{-1} M_e}_{=:E_3}.$$

We wish to bound $\left\|(F - F^*)\mathbb{1}_{[a+1,a+L']}\right\|_{\mathcal{H}_\infty} = \sup_{\omega \in [0,1]} [(\phi_\omega^H \otimes I_{d_u})(M_F - M_{F^*})]$.

We bound the contributions from E_1, E_2, E_3. First note that $\left\|(U^\top U)^{-1} U^\top\right\|_2 = \left\|(U^\top U)^{-1}\right\|_2^{1/2}$ and by Lemma A.1, for $T \ge K_1 L d_u \log\left(\frac{L d_u}{\delta}\right)$, with probability $1 - \delta$, $\left\|(U^\top U)^{-1}\right\| \le \frac{2}{T}$. Call this event $\mathcal{A}$.

1. Under the event $\mathcal{A}$, by Lemma A.6,

$$\mathbb{P}\left(\sup_{\omega \in [0,1]} [(\phi_\omega^H \otimes I_{d_u}) E_1] \le K \sqrt{L' \|\Sigma_y\|} \sqrt{\frac{1}{T}} \sqrt{d_y + d_u + \log\left(\frac{L'}{\delta}\right)}\right) \ge 1 - \delta.$$

2. By Lemma A.5 and the condition on T,

$$\left\|U^\top W\right\| \le \left\|U^\top (W \Sigma_x^{-1/2})\right\| \|\Sigma_x\|^{1/2} \le \sqrt{T L d_u \log\left(\frac{L d_u}{\delta}\right) \|\Sigma_x\|}$$

   Under $\mathcal{A}$, we bound the spectral norm (for all ω)

$$\begin{aligned}\left\|(\phi_\omega^H \otimes I_{d_u})(U^\top U)^{-1} U^\top W M_{G^*}\right\| &\le \|\phi_\omega^H \otimes I_{d_u}\| \left\|(U^\top U)^{-1}\right\| \left\|U^\top W\right\| \|M_{G^*}\| \\ &\le \sqrt{L'} \frac{K}{\sqrt{T}} \sqrt{T L d_u \log\left(\frac{L d_u}{\delta}\right) \|\Sigma_x\|} \|M_{G^*}\| \\ &\le K \sqrt{\|\Sigma_x\| L' L d_u \log\left(\frac{L d_u}{\delta}\right)} \|M_{G^*}\|.\end{aligned}$$

3. Let $\varepsilon_{\mathrm{trunc},F} = \left\|F^* \mathbb{1}_{[L+1,\infty)}\right\|_{\mathcal{H}_\infty}$ and similarly define $\varepsilon_{\mathrm{trunc},G}$. We bound the last term by noting

$$\left\|(F^* \mathbb{1}_{[L+1,\infty)}) * u\right\|_2 \le \left\|(F^* \mathbb{1}_{[L+1,\infty)})^\wedge \cdot \widehat{u}\right\|_2 \le \varepsilon_{\mathrm{trunc},F} \|u\|_2$$

and similarly $\left\|(G^* \mathbb{1}_{[L+1,\infty)}) * \xi\right\|_2 \leq \varepsilon_{\text{trunc},G} \|\xi\|_2$. We have $\mathbb{P}\left(\|\eta_{0:T-1}\| > \sqrt{Td_u} + K\sqrt{\log\left(\frac{1}{\delta}\right)}\right) \leq \delta$ and $\mathbb{P}\left(\|\xi_{0:T-1}\| > \sqrt{T}\left\|\Sigma_x^{1/2}\right\|_{\mathsf{F}} + K\sqrt{\|\Sigma_x\| \log\left(\frac{1}{\delta}\right)}\right) \leq \delta$ by Theorem D.5, so conditioned on event $\mathcal{A}$,

$$\sup_{\omega\in[0,1]} [(\phi_\omega^H \otimes I_{d_u})E_3] \leq K\sqrt{\frac{1}{T}}\left(\varepsilon_{\text{trunc},F}\left(\sqrt{Td_u} + \sqrt{\log\left(\frac{1}{\delta}\right)}\right)\right.$$
$$\left. + \varepsilon_{\text{trunc},G}\left(\sqrt{T}\left\|\Sigma_x^{1/2}\right\|_{\mathsf{F}} + \sqrt{\|\Sigma_x\| \log\left(\frac{1}{\delta}\right)}\right)\right)$$

with probability at least $1-\delta$. By the condition on T, the first terms are dominant.

Finish by replacing δ by $\frac{\delta}{4}$ and using the triangle inequality and a union bound. ■

Appendix B. Improved rates for learning system matrices

In this section, we combine Lemma 4.2 and Lemma 4.3 with bounds in Oymak and Ozay (2019) to give improved bounds for learning the system matrices.[2]

As L can be chosen to make $\varepsilon_{\text{trunc}}$ negligible, this gives $\sqrt{\frac{Ld}{T}}$ rates, however, with factors depending on the minimum eigenvalue of H.

We first re-do some of the bounds in Oymak and Ozay (2019) more carefully, using their notation.

Lemma B.1 ((Oymak and Ozay, 2019, Lemma B.1)) *Suppose $\sigma_{\min}(L) \geq 2\left\|L - \widehat{L}\right\|$ where $\sigma_{\min}(L)$ is the smallest nonzero singular value of L. Let rank-d matrices $L, \widehat{L}$ have singular value decompositions $U\Sigma V^*$ and $\widehat{U}\widehat{\Sigma}\widehat{V}^*$. There exists a $n \times n$ unitary matrix W so that*

$$\left\|U\Sigma^{1/2} - \widehat{U}\widehat{\Sigma}^{1/2}W\right\|_F^2 + \left\|V\Sigma^{1/2} - \widehat{V}\widehat{\Sigma}^{1/2}W\right\|_F^2 \leq \frac{4(\sqrt{2}+1)d\left\|L - \widehat{L}\right\|^2}{\sigma_{\min}(L)}.$$

Proof This inequality is given as an intermediate inequality in the proof of Lemma B.1 in Oymak and Ozay (2019). The first line gives that the LHS is $\leq \frac{2}{\sqrt{2}-1}\frac{\|L-\widehat{L}\|_F^2}{\sigma_{\min}(L)}$. Then use the fact that $\operatorname{rank}(L - \widehat{L}) \leq 2d$, so $\left\|L - \widehat{L}\right\|_F^2 \leq 2d\left\|L - \widehat{L}\right\|^2$. ■

Using this instead of Lemma B.1 gives the following for Theorem 4.3 of Oymak and Ozay (2019).

Lemma B.2 *Let $\widehat{A}, \widehat{B}, \widehat{C}$ be the state-space realization corresponding to the output of Ho-Kalman with input $\widehat{G}$. Suppose the system is observable and controllable. Let $L = \text{Hankel}_{L\times(L-1)}(F^*)$.*

2. References to Oymak and Ozay (2019) are for the arXiv version https://arxiv.org/abs/1806.05722.

Suppose $\sigma_{\min}(L) > 0$ and the low-rank approximation from Ho-Kalman satisfies $\left\|L - \widehat{L}\right\| \leq \sigma_{\min}(L)/2$. Then there exists a unitary matrix $W \in \mathbb{R}^{d\times d}$ such that

$$\left\|B - W^{-1}\widehat{B}\right\|_F, \left\|C - \widehat{C}W\right\|_F \leq \frac{2\sqrt{(\sqrt{2}+1)d}\left\|L - \widehat{L}\right\|}{\sqrt{\sigma_{\min}(L)}}$$

$$\left\|A - W^{-1}\widehat{A}W\right\|_F \leq \frac{2\sqrt{d}}{\sigma_{\min}(L)}\left(\frac{2\sqrt{\sqrt{2}+1}\left\|L - \widehat{L}\right\|}{\sigma_{\min}(L)}\left(2\left\|H^+\right\| + \left\|H^+ - \widehat{H}^+\right\|\right) + \left\|H^+ - \widehat{H}^+\right\|\right).$$

Proof We refer the reader to Oymak and Ozay (2019) for the details and just note the differences. As in Oymak and Ozay (2019), the first inequality follows from taking the square root in Lemma B.1.

For the second inequality, using Lemma B.1, the inequality for $\left\|O^\dagger - X^\dagger\right\|_{\mathsf{F}}$ becomes instead

$$\begin{aligned}\left\|O^\dagger - X^\dagger\right\|_{\mathsf{F}} &\leq \|O - X\|_{\mathsf{F}} \max\left\{\left\|X^\dagger\right\|^2, \left\|O^\dagger\right\|^2\right\} \\ &\leq \frac{2\sqrt{(\sqrt{2}+1)d}\left\|L - \widehat{L}\right\|_{\mathsf{F}}}{\sigma_{\min}(L)^{1/2}} \cdot \frac{2}{\sigma_{\min}(L)} \leq \frac{4\sqrt{(\sqrt{2}+1)d}\left\|L - \widehat{L}\right\|_F}{\sigma_{\min}(L)^{3/2}}\end{aligned}$$

so that (B.3)–(B.7) become

$$\begin{aligned}\left\|(O^\dagger - X^\dagger)H^+Q^\dagger\right\|_{\mathsf{F}} &\leq \frac{4\sqrt{(2+\sqrt{2})d}\left\|L - \widehat{L}\right\|_{\mathsf{F}}}{\sigma_{\min}(L)^2}\left\|H^+\right\| \\ \left\|X^\dagger\widehat{H}^+(Q^\dagger - Y^\dagger)\right\|_{\mathsf{F}} &\leq \frac{4\sqrt{(2+\sqrt{2})d}\left\|L - \widehat{L}\right\|_{\mathsf{F}}}{\sigma_{\min}(L)^2}\left(\left\|H^+\right\| + \left\|H^+ - \widehat{H}^+\right\|\right).\end{aligned}$$

Substituting in (B.2) then gives the theorem. ■

Proof [Proof of Theorem 2.3] Lemma 4.2 gives a bound on $\left\|H - \widehat{H}\right\|$. By (Oymak and Ozay, 2019, Appendix B.4),

$$\left\|H^+ - \widehat{H}^+\right\| \leq \left\|H - \widehat{H}\right\|, \qquad \left\|H^+\right\| \leq \|H\|, \qquad \left\|L - \widehat{L}\right\| \leq 2\left\|H - \widehat{H}\right\|.$$

By Lemma 4.3, $\|H\| \leq \|F^*\|_{\mathcal{H}_\infty} = \|\Phi_{\mathcal{D}}\|_{\mathcal{H}_\infty}$. Plugging this into Lemma B.2 gives the theorem. ■

Appendix C. Lower bound

We prove a minimax lower bound which shows that the rates in Theorem 2.2 are optimal up to logarithmic factors. In fact, the lower bound already holds in the simple case where there is no time dependency, in which case the problem reduces to a low-rank linear regression problem.

Theorem C.1 *Given the setting in Problem 2.1 with $\Sigma_\xi = \sigma^2 I_{d_y}$ and $\Sigma_\eta = O$, if $\max\{d_u, d_y\} \geq d$, then with probability $\geq 1-\delta$ over the randomness of the $u(t)$, the minimax error for estimating F^* (even when $A = O$, $B = O$, and $C = O$) in squared Frobenius norm is at least $\frac{\max\{d_u,d_y\}d\sigma^2}{T+2\left(\sqrt{T\log\left(\frac{d}{\delta}\right)}+\log\left(\frac{d}{\delta}\right)\right)}$, i.e., any estimator $\widehat{F}(y)$ satisfies*

$$\max_{\mathcal{D}:A=O,B=O,C=O} \left\|\widehat{F}(y) - F^*\right\|_{\mathsf{F}}^2 \geq \frac{\max\{d_u, d_y\}d\sigma^2}{T + 2\left(\sqrt{T\log\left(\frac{d}{\delta}\right)} + \log\left(\frac{d}{\delta}\right)\right)}.$$

Note alternatively we can restrict to $D = O$ and $A = O$; the only difference is that the impulse response is now nonzero only at $t = 1$ rather than $t = 0$.

We will use the following theorem.

Theorem C.2 ((Candes and Plan, 2011, Theorem 2.5)) *Suppose that $\mathcal{A} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ is a linear transformation such that*

$$\|\mathcal{A}(X)\|_2^2 \leq K\|X\|_{\mathsf{F}}^2$$

for all matrices X of rank at most r. Suppose that $n = \max\{n_1, n_2\} \geq r$. Given the observation $y = \mathcal{A}(X) + z$ where $z \sim N(0, \sigma^2 I_m)$, the minimax error in squared Frobenius norm over $\{X \in \mathbb{R}^{n_1\times n_2} : \operatorname{rank}(X) \leq r\}$ is at least $\frac{nr\sigma^2}{K}$, i.e., any estimator $\widehat{M}(y)$ satisfies

$$\sup_{M:\operatorname{rank}(M)\leq r} \mathbb{E}\left\|\widehat{M}(y) - M\right\|_{\mathsf{F}}^2 \geq \frac{nr\sigma^2}{K}.$$

Note that (Candes and Plan, 2011, Theorem 2.5) assumed a stronger condition—a matrix RIP (restricted isometry property)—but the lower bound coming from RIP is not used in the proof.

Proof [Proof of Theorem C.1] We are in the setting of Theorem C.2 where $\mathcal{A}(D) = (Du(0); \ldots; Du(T-1))$. Note that by change of basis, we may assume D is diagonal. If $\sigma_1, \ldots, \sigma_d$ are the singular values of D, then $\|\mathcal{A}(D)\|_2^2 = \sum_{t=0}^{T-1}\|Du(t)\|^2$ is distributed as $\sum_{i=1}^d \sigma_i^2 X_i$ where each $X_i \sim \chi_T^2$ is a χ^2-random variable with T degrees of freedom. By the tail bound in Laurent and Massart (2000), $\mathbb{P}(X_i \geq T + 2(\sqrt{Tu} + u)) \leq e^{-u}$. Letting $u = \log\left(\frac{d}{\delta}\right)$ and $Q = T + 2(\sqrt{Tu} + u)$, we get that

$$\mathbb{P}\left(\max_{1\leq i\leq d} X_i \geq Q\right) \leq d\mathbb{P}(X_1 \geq Q) \leq \delta.$$

Then with probability $\geq 1-\delta$, the event $\max_{1\leq i\leq d} X_i < Q$ holds, so

$$\|\mathcal{A}(D)\|_2^2 = \sum_{i=1}^d \sigma_i^2 X_i \leq \left(\sum_{i=1}^d \sigma_i^2\right)\max X_i \leq \|D\|_{\mathsf{F}}^2 Q.$$

Then by Theorem C.2, for any estimate $\widehat{D}(y)$, $\sup_{D:\operatorname{rank}(D)\leq d}\mathbb{E}\left\|\widehat{D}(y) - D\right\|_{\mathsf{F}}^2 \geq \frac{\max\{d_u,d_y\}d\sigma^2}{Q}$, as desired. ■

Appendix D. Concentration bounds

In this section, we collect some useful concentration results.

Definition D.1 *A $\mathbb{R}$-valued random variable X is sub-gaussian with constant K_X if*

$$\|X\|_{\psi_2} := \inf\left\{s > 0 : \mathbb{E}[\exp((x/s)^2) - 1] \le 1\right\} \le K_X,$$

and a $\mathbb{R}^n$-valued random variable X is sub-gaussian with constant K_X if

$$\|X\|_{\psi_2} := \sup_{v \in \mathbb{S}^{n-1}} \|\langle X, x\rangle\|_{\psi_2} \le K_X.$$

Theorem D.2 ((Vershynin, 2018, Ex. 4.7.3)) *There is a constant K such that the following holds. Let $X_1, \ldots, X_m$ be iid copies of a random vector X in $\mathbb{R}^n$ satisfying the sub-gaussian bound for any x,*

$$\|\langle X, x\rangle\|_{\psi_2} \le K_X \mathbb{E}[\langle X, x\rangle^2].$$

Let $\Sigma_m = \frac{1}{m}\sum_{i=1}^m X_i X_i^\top$. Then for any $s \ge 0$,

$$\mathbb{P}\left(\|\Sigma_m - \Sigma\| \le K K_X^2 \left(\sqrt{\frac{n+s}{m}} + \frac{n+s}{m}\right)\|\Sigma\|\right) \ge 1 - 2e^{-s}.$$

Theorem D.3 (Hanson-Wright inequality, (Rudelson et al., 2013, Theorem 1.1)) *There is a constant $c > 0$ such that the following holds. Let $A \in \mathbb{C}^{n\times n}$ be a matrix, and let $v \in \mathbb{R}^n$ be a random vector with independent, mean-0, K_v-sub-gaussian entries. Then for every $s \ge 0$,*

$$\mathbb{P}(|v^\top A v - \mathbb{E} v^\top A v| > s) \le 2\exp\left[-c \cdot \min\left\{\frac{s^2}{K_v^4 \|A\|_{\mathsf{F}}^2}, \frac{s}{K_v^2 \|A\|}\right\}\right].$$

Corollary D.4 *There is a constant $c > 0$ such that the following holds. Let $A \in \mathbb{C}^{m\times n}$ be a matrix, and let $v \in \mathbb{R}^m, w \in \mathbb{R}^n$ be random vectors with independent, mean-0, K_v and K_w sub-gaussian entries, respectively. Then for every $s \ge 0$,*

$$\mathbb{P}(|v^\top A w| > s) \le 2\exp\left[-c \cdot \min\left\{\frac{s^2}{K_v^2 K_w^2 \|A\|_{\mathsf{F}}^2}, \frac{s}{K_v K_w \|A\|}\right\}\right].$$

Proof Apply Theorem D.3 for $v \leftarrow\!\shortmid \begin{pmatrix} v \\ w \end{pmatrix}$ and $A \leftarrow\!\shortmid \begin{pmatrix} O & A \\ O & O \end{pmatrix}$. ■

Theorem D.5 (Sub-gaussian concentration, (Rudelson et al., 2013, Theorem 2.1)) *There is a constant $c > 0$ such that the following holds. Let $A \in \mathbb{C}^{m\times n}$ be a matrix, and let $v \in \mathbb{R}^n$ be a random vector with independent, mean-0, K_v-sub-gaussian entries. Then for every $s \ge 0$,*

$$\mathbb{P}\left[|\|Av\|_2 - \|A\|_{\mathsf{F}}| > s\right] \le 2\exp\left(-\frac{cs^2}{K_v^4 \|A\|^2}\right).$$

Appendix E. Experimental details

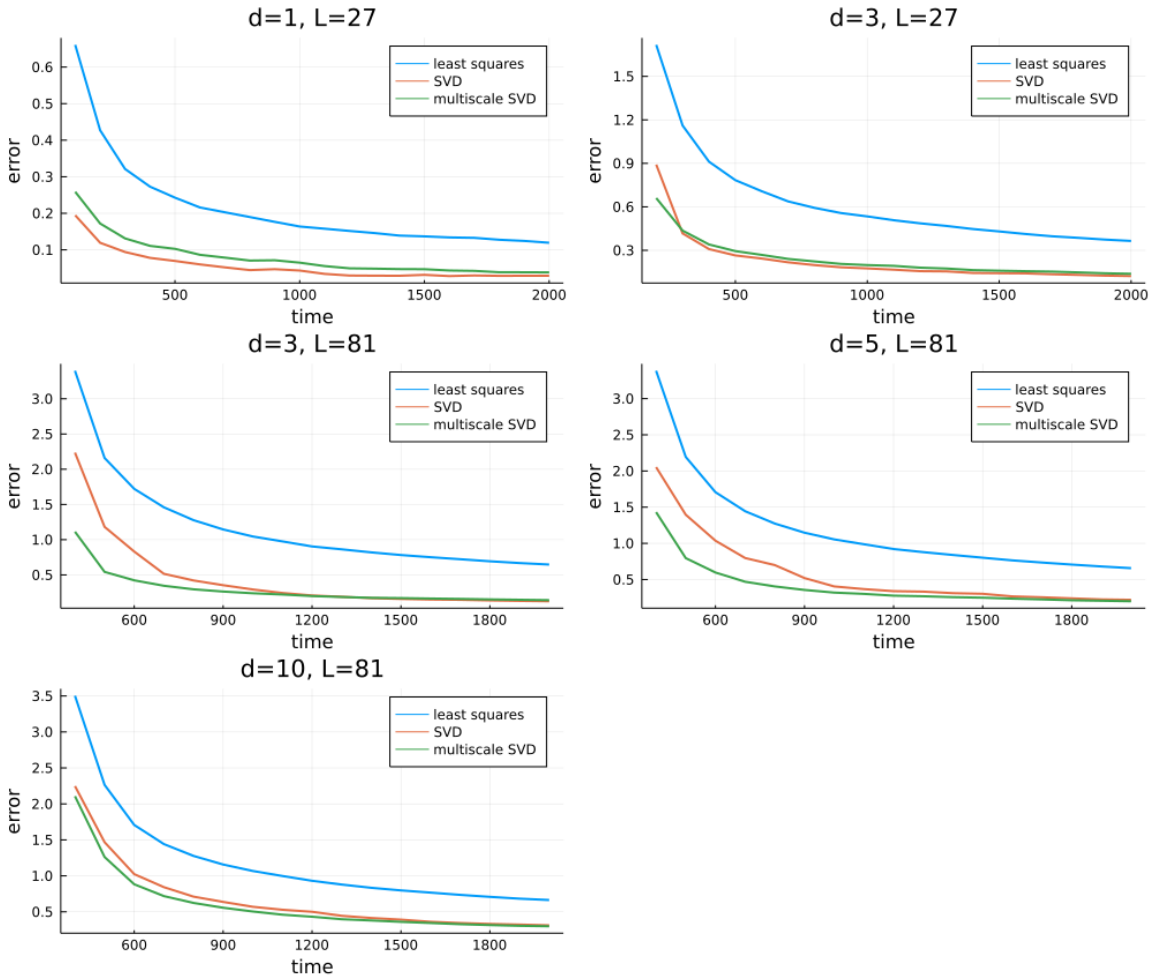
We generate random LDS's as follows. For B and C, the rows or columns are chosen to be a random set of orthonormal vectors (depending on whether they have more rows or columns). For A, the entries are first chosen to be iid standard gaussians, and then A is re-normalized so that its maximum eigenvalue has absolute value $\lambda_{\max}$. For simplicity, we take $D = O$.

We make a slight modification of Algorithm 10 which triples the size at each iteration instead. For $L = 3^a$, we estimate a finite impulse response of length $4 \cdot 3^{a-1} - 1$. Then, for the multiscale SVD algorithm, at the kth scale ($k \geq 1$), we consider the rank-d SVD of $\mathrm{Hankel}_{\ell \times \ell}(F)$, where $\ell = 2 \cdot 3^{k-1}$, and use this SVD to estimate the $F(t)$ for $3^{k-1} < t \leq 2^k$. For the single-scale SVD, we estimate all $F(t)$ from the rank-d SVD of $\mathrm{Hankel}_{\ell \times \ell}(F)$, where $\ell = 2 \cdot 3^{a-1}$.

The plots show the error $\left\| F^* \mathbb{1}_{[1,L]} - F \right\|_2$, where F is the estimated impulse response on $[1, L]$, averaged over 10 randomly generated LDS's, as a function of the time T elapsed, for the following settings of parameters:

1. $d = d_u = d_y = 1$, $L = 27$, $\lambda_{\max} = 0.9$.
2. $d = d_u = d_y = 3$, $L = 27$, $\lambda_{\max} = 0.9$.
3. $d = d_u = d_y = 3$, $L = 81$, $\lambda_{\max} = 0.95$.
4. $d = 5$, $d_u = d_y = 3$, $L = 81$, $\lambda_{\max} = 0.95$.
5. $d = 10$, $d_u = d_y = 3$, $L = 81$, $\lambda_{\max} = 0.95$.

The code was written in Julia, and is available at https://github.com/holdenlee/hankel-svd.


d=1, L=27
d=3, L=27
d=3, L=81
d=5, L=81
d=10, L=81
least squares
SVD
multiscale SVD
error
time