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 [SOF20] 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 setting, [DMR19] give concentration bounds for the covariance matrix with $T = \widetilde{\Omega}(L)$ timesteps. First, we extend this to the MIMO case in Theorem A.2 (Note that [OO19] consider the MIMO case but have extra log factors.) Then, we use Gaussian suprema arguments as in [TBPR17] 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

$$\|y - u * f\|_{[0,T-1]}^2 = \sum_{t=0}^{T-1} \left\| y(t) - u_{t:t-L}^\top f \right\|^2 = \|y_{0:T-1} - Uf\|^2 \tag{9}$$

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

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

Thus, by taking the transpose and stacking vectors,

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

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

$$M_F - M_{F^*} = (U^\top U)^{-1} U^\top M_\eta + (U^\top U)^{-1} U^\top W M_{G^*} + (U^\top U)^{-1} M_e. \tag{10}$$

A.1 Lower bounding sample covariance matrix

In this subsection we lower bound the sample covariance matrix.

Lemma A.1. *There is a constant C 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 $0 < \delta \leq \frac{1}{2}$, $T \geq C_1 L d_u \log\left(\frac{Ld_u}{\delta}\right)$,*

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

This is a corollary of the following concentration bound, which generalizes Theorem 3.4 of [DMR19] 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 C such that the following holds. Suppose $u(t)$, $0 \leq t < T$ are independent, zero-mean, and K-sub-gaussian (see Definition C.1), and let $U = \mathrm{Toep}_{T\times(L+1)}((u(t)^\top)_{t\geq 0})$. Then for $0 < \delta \leq \frac{1}{2}$, $T \geq L$,*

$$\left\|U^\top U - TI_{d_u}\right\| \leq CK^2 \left(Ld_u \log\left(\frac{T}{\delta}\right) + \sqrt{TLd_u \log\left(\frac{T}{\delta}\right)}\right)$$

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 ([BTR13]). *Let $Q(z) := \sum_{k=0}^{r-1} a_k z^k$, where $a_k \in \mathbb{C}$. For any $N \geq 4\pi r$, $\|Q\|_{\mathcal{H}_\infty} \leq \left(1 + \frac{4\pi r}{N}\right) \max_{j=0,\ldots,N-1} |Q(e^{\frac{2\pi i j}{N}})|$.*

Proof of Theorem A.2. By rescaling we may suppose $K = 1$. Decompose

$$U = U_1 + U_2 \text{ where} \tag{11}$$

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

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

Then

$$U^\top U = (T-L)I_{Ld_u} + (U_1^\top U_1 - (T-L)I_{Ld_u}) + U_1^\top U_2 + U_2^\top U_1 + U_2^\top U_2. \tag{12}$$

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

$$\begin{aligned} P_u(\omega) &= \sum_{\ell=-L}^{L} \sum_{t\in\mathbb{Z}} (\mathsf{T}^\ell u^{(1)})(t) u^{(1)}(t)^\top e^{-2\pi i \ell\omega} - T' I_{d_u} \\ &= \sum_{\substack{j,\,k\,\in\,\mathbb{Z} \\ |j-k|\,\leq\,L}} u(j)u(k)^\top e^{2\pi i (j-k)\omega} - T' I_{d_u} \\ &= \begin{pmatrix} u(0) & \cdots & u(T'-1) \end{pmatrix} M \begin{pmatrix} u(0)^\top \\ \cdots \\ u(T'-1)^\top \end{pmatrix} - T' I_{d_u} \end{aligned} \tag{13}$$

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

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

By Lemma A.3,

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

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

$$\begin{aligned}
\|P_u(\omega)\|_{\mathcal{H}_\infty} = \sup_{\omega\in[0,1]} \|P_u(\omega)\| &= \sup_{\|v\|=1} \sup_{\omega\in[0,1]} |v^\top P_u(\omega) v| \\
&= \sup_{\|v\|=1} 2 \max_{\omega\in\{0,\frac{1}{N},\cdots\}} |v^\top P_u(\omega) v| && \text{by Lemma A.4} \\
&= 2 \max_{\omega\in\{0,\frac{1}{N},\cdots\}} \left(\sup_{v\in\mathcal{N}_\varepsilon} |v^\top P_u(\omega) v| + 3\varepsilon \|P_u(\omega)\| \right) && (14)
\end{aligned}$$

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 C.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 C,

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

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 (13),

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

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

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

Now take $\delta_1 = \frac{\delta}{2}$. By Cor. 4.2.13 of [Ver18], there is an ε-net of size $|\mathcal{N}_\varepsilon| \le \left(1 + \frac{2}{\varepsilon}\right)^{d_u} = \exp\left(d_u \log\left(1 + \frac{2}{\varepsilon}\right)\right) = \exp\left(d_u \log\left(1 + \frac{8 d_u L T}{\delta}\right)\right)$. Letting $\delta_1 = \frac{\delta}{2|\mathcal{N}_\varepsilon|}$ and taking a union bound, with probability $1 - \delta$ we get

$$(14) \le C\left(\sqrt{TLd \log\left(\frac{T}{\delta}\right)} + Ld \log\left(\frac{T}{\delta}\right) \right)$$

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,

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

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

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

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 $1-\delta$,

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

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 $u(T'+1)^\top, \dots, u(T'+L)^\top$. 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 $C\left(\sum_{t=T-L}^{T-1} \cos^2(-2\pi t\omega)\right)^{1/2}$-sub-gaussian and the imaginary part is $C\left(\sum_{t=T-L}^{T-1} \sin^2(-2\pi t\omega)\right)^{1/2}$-sub-gaussian for some constant C. Hence

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

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

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

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

$$\begin{aligned}
\left\| U_2^\top U_2 \right\| &\le \left(\sup_{\omega\in[0,1]} \|P_{u,2}(\omega)\| \right)^2 \\
&\le \left(2 \max_{\omega\in\{0,\frac{1}{N},\dots\}} \|P_{u,2}(\omega)\| \right)^2 \\
&\le CLd_u \log\left(\frac{d_u L}{\delta}\right).
\end{aligned}$$

with probability $\ge 1-\delta$. Putting all the bounds together with (12) gives the theorem. ☐

Proof of Lemma A.1. For large enough C_2, for $T \ge C_2 L d_u \log\left(\frac{T}{\delta}\right)$, we have that by Theorem A.2 that $\left\| U^\top U - T I_{d_u} \right\| \le \frac{T}{2}$, so $\sigma_{\min}(U^\top U) \ge \frac{T}{2}$.

Finally, note that for large enough C_1, $T \ge C_1 L d_u \log\left(\frac{L d_u}{\delta}\right)$ implies $T \ge C_2 L d_u \log\left(\frac{T}{\delta}\right)$. ☐

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

Lemma A.5. *There is a constant C such that the following holds. Suppose $u(t)$, $0 \le t < T$ are independent, zero-mean, and K_u-sub-gaussian, and similarly for $w(t)$ with constant K_w. Let $U = \mathrm{Toep}_{T\times(L+1)}((u(t)^\top)_{t\ge 0})$, $W = \mathrm{Toep}_{T\times(L+1)}((w(t)^\top)_{t\ge 0})$. Then for $0 < \delta \le \frac{1}{2}$, $T \ge C_1 L d_u \log\left(\frac{Ld_u}{\delta}\right)$,*

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

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 (11). 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 C.4) gives that

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

□

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

The following Lemma A.6 generalizes the results of [TBPR17] to the MIMO setting. To get the right dimension dependence, we will use the concentration bound for covariance given by Theorem C.2.

Lemma A.6. *There is a constant C such that the following holds. Suppose that $\eta(0), \ldots, \eta(T-1) \sim N(0,\Sigma)$ are iid, $\Phi \in \mathbb{R}^{(L+1)d_u\times T}$, and $E(0), \ldots, E(L) \in \mathbb{R}^{d_y\times d_u}$ 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 $-1 \le a < L - L'$,

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

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

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

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 C.2[1],

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

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

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

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

$$\mathbb{P}\left(\max_{\omega \in \{0, \frac{1}{N}, \dots, \frac{N-1}{N}\}} \left\| M_\omega \right\| \le C\sqrt{L'} \left\| \Phi' \right\|^{\frac{1}{2}} \sqrt{d_y + d_u + \log\left(\frac{2}{\delta}\right)} \right) \ge 1 - \delta. \tag{15}$$

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

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

Combining this with (15) gives the result. $\square$

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

Lemma A.7. *There are C_1, C_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 $F = \operatorname{argmin}_{F \in \{0,\dots,L\} \to \mathbb{R}^{d_y \times d_u}} \sum_{t=0}^{T-1} |y(t) - (F * u)(t)|^2$, $M_{G^*} = (G^*(0), \dots, G^*(L))^\top \in \mathbb{R}^{(L+1)d \times d_y}$, and $\varepsilon_{\text{trunc}} = \left\| F^* \mathbb{1}_{[L+1,\infty)} \right\|_{\mathcal{H}_\infty} \sqrt{d_u} + \left\| G^* \mathbb{1}_{[L+1,\infty)} \right\|_{\mathcal{H}_\infty} \left\| \Sigma_x^{1/2} \right\|_{\mathsf{F}}$. For $0 < \delta \le \frac{1}{2}$, $T \ge C_1 L d_u \log\left(\frac{L d_u}{\delta}\right)$, $-1 \le a < L - L'$,*

$$\begin{aligned} &\left\| (F - F^*) \mathbb{1}_{[a+1,a+L']} \right\|_{\mathcal{H}_\infty} \\ &\le C_2 \left[\sqrt{\frac{1}{T}} \left(\sqrt{\left\| \Sigma_y \right\| L' \left(d_u + d_y + \log\left(\frac{L'}{\delta}\right) \right)} + \sqrt{\left\| \Sigma_x \right\| L' L d_u \log\left(\frac{L d_u}{\delta}\right)} \left\| M_{G^*} \right\| \right) + \varepsilon_{\text{trunc}} \right] \end{aligned}$$

with probability at least $1 - \delta$.

[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.

Proof. By (10), 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 $T \ge C_1 L d_u \log\left(\frac{Ld_u}{\delta}\right)$, 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,

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

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

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

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

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

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

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

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

$$\begin{aligned}\sup_{\omega\in[0,1]}[(\phi_\omega^H \otimes I_{d_u})E_3] \le C\sqrt{\frac{1}{T}}\Bigg(&\varepsilon_{\mathrm{trunc},F}\left(\sqrt{Td_u} + \sqrt{\log\left(\frac{1}{\delta}\right)}\right) \\ &+ \varepsilon_{\mathrm{trunc},G}\left(\sqrt{T}\left\|\Sigma_x^{1/2}\right\|_{\mathsf{F}} + \sqrt{\left\|\Sigma_x\right\|\log\left(\frac{1}{\delta}\right)}\right)\Bigg)\end{aligned}$$

   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. □

B Improved rates for learning system matrices

In this section, we combine Lemma 4.2 and Lemma 4.3 with bounds in [OO19] to give improved bounds for learning the system matrices.[2]

[2]References to [OO19] are for the arXiv version `https://arxiv.org/abs/1806.05722`.

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 [OO19] more carefully, using their notation.

Lemma B.1 ([OO19, 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 [OO19]. The first line gives that the LHS is $\leq \frac{2}{\sqrt{2}-1}\frac{\left\|L-\widehat{L}\right\|_F^2}{\sigma_{\min}(L)}$. 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 [OO19].

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 $L = \operatorname{Hankel}_{L\times(L-1)}(F^*)$. 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 [OO19] for the details and just note the differences. As in [OO19], 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

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

so that (B.3)–(B.7) become

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

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

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

Proof of Theorem 2.3. Lemma 4.2 gives a bound on $\left\|H - \widehat{H}\right\|$. By [OO19, Appendix B.4],

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

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

C Concentration bounds

In this section, we collect some useful concentration results.

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

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

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

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

Theorem C.2 ([Ver18, Ex. 4.7.3])**.** *There is a constant C 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,*

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

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

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

Theorem C.3 (Hanson-Wright inequality, [RV$^+$13, 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-sub-gaussian entries. Then for every $s \ge 0$,*

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

Corollary C.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 $s \ge 0$,*

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

Proof. Apply Theorem C.3 for $v \hookleftarrow \begin{pmatrix} v \\ w \end{pmatrix}$ and $A \hookleftarrow \begin{pmatrix} O & A \\ O & O \end{pmatrix}$. $\square$

Theorem C.5 (Sub-gaussian concentration, [RV$^+$13, 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-sub-gaussian entries. Then for every $s \ge 0$,*

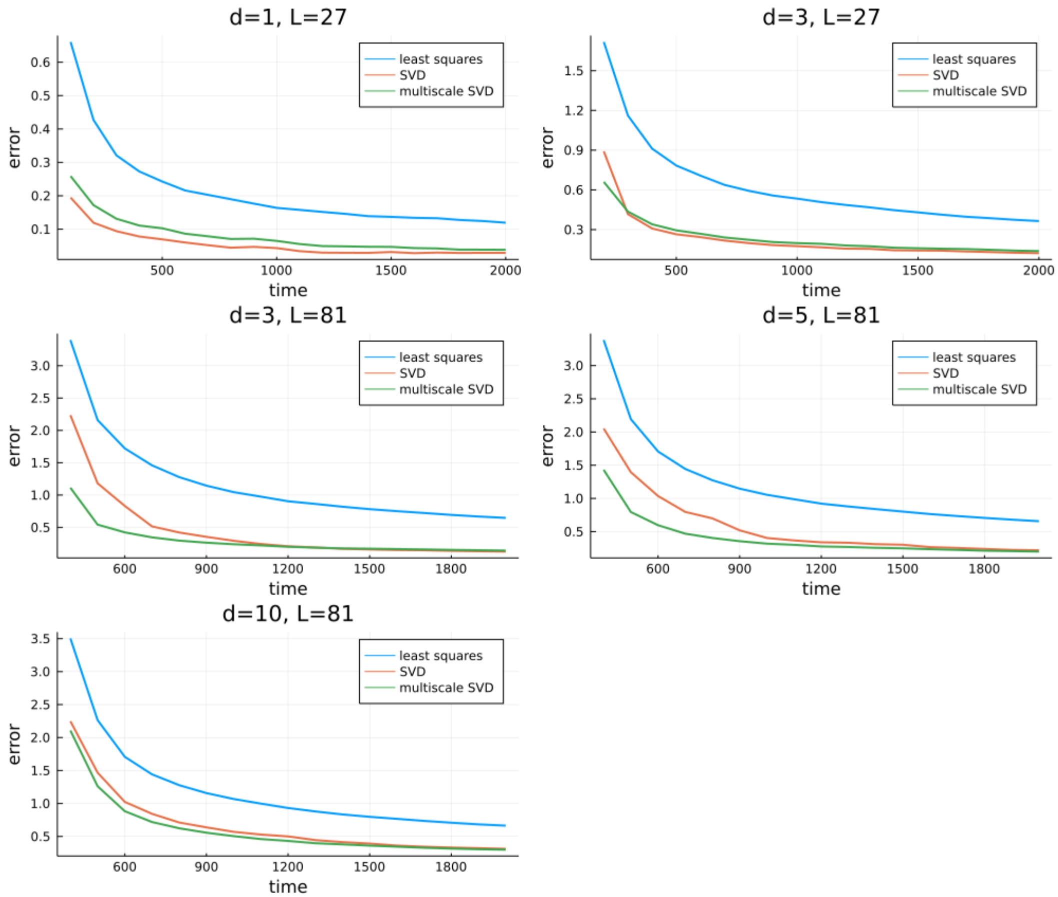
$$\mathbb{P}\left[|\|Av\|_2 - \|A\|_{\mathsf{F}}| > s\right] \le 2\exp\left(-\frac{cs^2}{K^4\|A\|^2}\right).$$

D 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 1 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 ($k \ge 1$), 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 $3^{k-1} < t \le 2^k$. 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.

E 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 [SOF20]. 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 $\xi(t)$. 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 [Qin06]

$$x(t)^- = A_{\mathrm{KF}} x(t-1)^- + B_{\mathrm{KF}} \begin{pmatrix} u(t-1) \\ y(t-1) \end{pmatrix} \tag{16}$$

$$y(t) = Cx(t)^- + Du(t) + e(t) \tag{17}$$

where $e(t) \sim N(0, \Sigma_{\text{KF}})$ 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 [WJ20].

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? [SGA20] 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).