Linear system identification with reverse experience replay

Jain, Kowshik, Nagaraj, Netrapalli (2021)

Geelon So, agso@eng.ucsd.edu
Time series reading group — April 21, 2022

Linear dynamical system

Consider a **linear time-invariant** (LTI) dynamical systems $(X_0, X_1, \ldots, X_T)$ where:

$$X_{t+1} = \mathbf{A}_\star X_t + \eta_t.$$

Linear dynamical system

Consider a **linear time-invariant** (LTI) dynamical systems $(X_0, X_1, \ldots, X_T)$ where:

$$X_{t+1} = \mathbf{A}_\star X_t + \eta_t.$$

- $X_t \in \mathbb{R}^d$ is the **state** at time t

Linear dynamical system

Consider a **linear time-invariant** (LTI) dynamical systems $(X_0, X_1, \ldots, X_T)$ where:

$$X_{t+1} = \mathbf{A}_\star X_t + \eta_t.$$

- $X_t \in \mathbb{R}^d$ is the **state** at time t
- $\mathbf{A}_\star \in \mathbb{R}^{d \times d}$ is the **system dynamics**

Linear dynamical system

Consider a **linear time-invariant** (LTI) dynamical systems $(X_0, X_1, \ldots, X_T)$ where:

$$X_{t+1} = \mathbf{A}_\star X_t + \eta_t.$$

- $X_t \in \mathbb{R}^d$ is the **state** at time t
- $\mathbf{A}_\star \in \mathbb{R}^{d \times d}$ is the **system dynamics**
- $\eta_t \in \mathbb{R}^d$ is an i.i.d. **noise vector** from μ a **noise distribution**

Linear system identification

Problem (System identification)

Estimate $\mathbf{A}_\star$ from a single trajectory $(X_0, \dots, X_T)$ from the linear dynamical system:

$$X_{t+1} = \mathbf{A}_\star X_t + \eta_t, \qquad \text{where } \eta_t \overset{\text{i.i.d.}}{\sim} \mu.$$

Offline setting: OLS is nearly optimal

The **ordinary least squares** (OLS) estimator is:

$$\hat{\mathbf{A}} = \arg\min_{\mathbf{A}} \sum_{t=0}^{T-1} \|X_{t+1} - \mathbf{A}X_t\|^2.$$

Offline setting: OLS is nearly optimal

The **ordinary least squares** (OLS) estimator is:

$$\hat{\mathbf{A}} = \arg\min_{\mathbf{A}} \sum_{t=0}^{T-1} \|X_{t+1} - \mathbf{A}X_t\|^2.$$

The OLS estimator is nearly optimal (Simchowitz et al., 2018; Sarkar and Rakhlin, 2019):

$$\left\|\hat{\mathbf{A}} - \mathbf{A}_\star\right\|_{\text{op}}^2 = O\left(\frac{d}{T}\right).$$

Streaming setting

In the **streaming setting**, the data X_t comes sequentially in a stream.

Streaming setting

In the **streaming setting**, the data X_t comes sequentially in a stream.

- It is possible to compute OLS in a stream (see the Woodbury-Sherman-Morrison formula to compute matrix inverses in a stream).

Streaming setting

In the **streaming setting**, the data X_t comes sequentially in a stream.

- It is possible to compute OLS in a stream (see the Woodbury-Sherman-Morrison formula to compute matrix inverses in a stream).
- However, this streaming algorithm is specific to OLS and does not extend to related settings (e.g. generalized linear models, nonlinear dynamical systems).

Streaming with a first-order gradient oracle

Question: can a SGD-style algorithm making updates achieve similarly optimal rates?

Streaming with a first-order gradient oracle

Question: can a SGD-style algorithm making updates achieve similarly optimal rates?

- Kowshik et al. (2021b) introduces *stochastic gradient descent with reverse experience replay* achieving this. The ideas are more general:

Streaming with a first-order gradient oracle

Question: can a SGD-style algorithm making updates achieve similarly optimal rates?

- Kowshik et al. (2021b) introduces *stochastic gradient descent with reverse experience replay* achieving this. The ideas are more general:
  - Extends to learning certain non-linear dynamical systems (Kowshik et al., 2021a)

Streaming with a first-order gradient oracle

Question: can a SGD-style algorithm making updates achieve similarly optimal rates?

- Kowshik et al. (2021b) introduces *stochastic gradient descent with reverse experience replay* achieving this. The ideas are more general:
  - Extends to learning certain non-linear dynamical systems (Kowshik et al., 2021a)
  - Extends to Q-learning tabular MDPs (Agarwal et al., 2021)

An SGD update

Let $X' = \mathbf{A}_\star X + \eta$. Then the **loss** on this example (X, X') is given by:

$$\ell(\mathbf{A}; X, X') = \left\|\mathbf{A}X - X'\right\|^2 = \left\|(\mathbf{A} - \mathbf{A}_\star)X - \eta\right\|^2$$

An SGD update

Let $X' = \mathbf{A}_\star X + \eta$. Then the **loss** on this example (X, X') is given by:

$$\ell(\mathbf{A}; X, X') = \left\|\mathbf{A}X - X'\right\|^2 = \left\|(\mathbf{A} - \mathbf{A}_\star)X - \eta\right\|^2$$

Thus, the **gradient** is:

$$\nabla\ell(\mathbf{A}; X, X') = 2\big(\mathbf{A} - \mathbf{A}_\star\big)XX^\top - 2\eta X^\top.$$

An SGD update

Let $X' = \mathbf{A}_\star X + \eta$. Then the **loss** on this example (X, X') is given by:

$$\ell(\mathbf{A}; X, X') = \left\|\mathbf{A}X - X'\right\|^2 = \left\|(\mathbf{A} - \mathbf{A}_\star)X - \eta\right\|^2$$

Thus, the **gradient** is:

$$\nabla\ell(\mathbf{A}; X, X') = 2\big(\mathbf{A} - \mathbf{A}_\star\big)XX^\top - 2\eta X^\top.$$

And the **SGD update** using the (stochastic) example (X, X') with learning rate γ is:

$$\mathbf{A}_{\text{update}} = \mathbf{A} - 2\gamma\big(\mathbf{A} - \mathbf{A}_\star\big)XX^\top - 2\gamma\eta X^\top.$$

An SGD update

Let $X' = \mathbf{A}_\star X + \eta$. Then the **loss** on this example (X, X') is given by:

$$\ell(\mathbf{A}; X, X') = \left\|\mathbf{A}X - X'\right\|^2 = \left\|(\mathbf{A} - \mathbf{A}_\star)X - \eta\right\|^2$$

Thus, the **gradient** is:

$$\nabla\ell(\mathbf{A}; X, X') = 2\big(\mathbf{A} - \mathbf{A}_\star\big)XX^\top - 2\eta X^\top.$$

And the **SGD update** using the (stochastic) example (X, X') with learning rate γ is:

$$\mathbf{A}_{\text{update}} = \mathbf{A} - 2\gamma\big(\mathbf{A} - \mathbf{A}_\star\big)XX^\top - 2\gamma\eta X^\top.$$

- Rewrite as: $\mathbf{A}_{\text{update}} - \mathbf{A}_\star = \big(\mathbf{A} - \mathbf{A}_\star\big)\underbrace{\big(\mathbf{I} - 2\gamma XX^\top\big)}_{\text{contraction factor}} - \underbrace{2\gamma\eta X^\top}_{\text{expectation zero}}$.

Uncorrelated noise and update transform

$$\mathbf{A}_{t+1} - \mathbf{A}_\star = \underbrace{\left(\mathbf{A}_t - \mathbf{A}_\star\right)\left(\mathbf{I} - 2\gamma X_t X_t^\top\right)}_{\text{contracting term}} - \underbrace{2\gamma\eta_t X_t^\top}_{\text{noise}}.$$

Uncorrelated noise and update transform

$$\mathbf{A}_{t+1} - \mathbf{A}_\star = \underbrace{\left(\mathbf{A}_t - \mathbf{A}_\star\right)\left(\mathbf{I} - 2\gamma X_t X_t^\top\right)}_{\text{contracting term}} - \underbrace{2\gamma\eta_t X_t^\top}_{\text{noise}}.$$

Let's unroll two steps of SGD:

$$\begin{aligned}&\mathbf{A}_{t+2} - \mathbf{A}_\star \\ &= \left(\mathbf{A}_{t+1} - \mathbf{A}_\star\right)\left(\mathbf{I} - 2\gamma X_{t+1} X_{t+1}^\top\right) - 2\gamma\eta_{t+1} X_{t+1}^\top\end{aligned}$$

Uncorrelated noise and update transform

$$\mathbf{A}_{t+1} - \mathbf{A}_\star = \underbrace{\big(\mathbf{A}_t - \mathbf{A}_\star\big)\big(\mathbf{I} - 2\gamma X_t X_t^\top\big)}_{\text{contracting term}} - \underbrace{2\gamma\eta_t X_t^\top}_{\text{noise}}.$$

Let's unroll two steps of SGD:

$$\begin{aligned}
&\mathbf{A}_{t+2} - \mathbf{A}_\star \\
&= \big(\mathbf{A}_{t+1} - \mathbf{A}_\star\big)\big(\mathbf{I} - 2\gamma X_{t+1} X_{t+1}^\top\big) - 2\gamma\eta_{t+1} X_{t+1}^\top \\
&= \Big\{ \big(\mathbf{A}_t - \mathbf{A}_\star\big)\big(\mathbf{I} - 2\gamma X_t X_t^\top\big) - 2\gamma\eta_t X_t^\top \Big\}\big(\mathbf{I} - 2\gamma X_{t+1} X_{t+1}^\top\big) - 2\gamma\eta_{t+1} X_{t+1}^\top
\end{aligned}$$

Uncorrelated noise and update transform

$$\mathbf{A}_{t+1} - \mathbf{A}_\star = \underbrace{\left(\mathbf{A}_t - \mathbf{A}_\star\right)\left(\mathbf{I} - 2\gamma X_t X_t^\top\right)}_{\text{contracting term}} - \underbrace{2\gamma\eta_t X_t^\top}_{\text{noise}}.$$

Let's unroll two steps of SGD:

$$\begin{aligned}
&\mathbf{A}_{t+2} - \mathbf{A}_\star \\
&= \left(\mathbf{A}_{t+1} - \mathbf{A}_\star\right)\left(\mathbf{I} - 2\gamma X_{t+1} X_{t+1}^\top\right) - 2\gamma\eta_{t+1} X_{t+1}^\top \\
&= \left\{\left(\mathbf{A}_t - \mathbf{A}_\star\right)\left(\mathbf{I} - 2\gamma X_t X_t^\top\right) - 2\gamma\eta_t X_t^\top\right\}\left(\mathbf{I} - 2\gamma X_{t+1} X_{t+1}^\top\right) - 2\gamma\eta_{t+1} X_{t+1}^\top \\
&= \left(\mathbf{A}_t - \mathbf{A}_\star\right)\left(\mathbf{I} - 2\gamma X_t X_t^\top\right)\left(\mathbf{I} - 2\gamma X_{t+1} X_{t+1}^\top\right) - 2\gamma\eta_t X_t^\top\left(\mathbf{I} - 2\gamma X_{t+1} X_{t+1}^\top\right) - 2\gamma\eta_{t+1} X_{t+1}^\top
\end{aligned}$$

Uncorrelated noise and update transform

$$\mathbf{A}_{t+1} - \mathbf{A}_\star = \underbrace{\left(\mathbf{A}_t - \mathbf{A}_\star\right)\left(\mathbf{I} - 2\gamma X_t X_t^\top\right)}_{\text{contracting term}} - \underbrace{2\gamma\eta_t X_t^\top}_{\text{noise}}.$$

Let's unroll two steps of SGD:

$$\begin{aligned}
&\mathbf{A}_{t+2} - \mathbf{A}_\star \\
&= \left(\mathbf{A}_{t+1} - \mathbf{A}_\star\right)\left(\mathbf{I} - 2\gamma X_{t+1} X_{t+1}^\top\right) - 2\gamma\eta_{t+1} X_{t+1}^\top \\
&= \left\{\left(\mathbf{A}_t - \mathbf{A}_\star\right)\left(\mathbf{I} - 2\gamma X_t X_t^\top\right) - 2\gamma\eta_t X_t^\top\right\}\left(\mathbf{I} - 2\gamma X_{t+1} X_{t+1}^\top\right) - 2\gamma\eta_{t+1} X_{t+1}^\top \\
&= \left(\mathbf{A}_t - \mathbf{A}_\star\right)\left(\mathbf{I} - 2\gamma X_t X_t^\top\right)\left(\mathbf{I} - 2\gamma X_{t+1} X_{t+1}^\top\right) - 2\gamma\eta_t X_t^\top\left(\mathbf{I} - 2\gamma X_{t+1} X_{t+1}^\top\right) - 2\gamma\eta_{t+1} X_{t+1}^\top
\end{aligned}$$

If η_t is independent of X_t and X_{t+1}, then the transformed noise term is mean-zero if η_t is.

Problem: independence does not hold

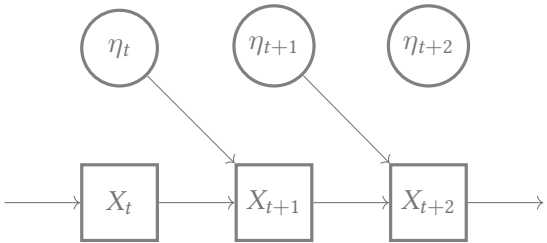

Figure 1: Recall the transformed noise term: $2\gamma\eta_t X_t^\top \left(\mathbf{I} - 2\gamma X_{t+1} X_{t+1}^\top\right)$. The past noise η_t and the SGD transformation $(\mathbf{I} - 2\gamma X_{t+1} X_{t+1}^\top)$ are not independent.

Unrolled noise term

Unrolling the whole sequence of SGD on $(X_0, X_1, \ldots, X_T)$ all the way:

$$\mathbf{A}_T - \mathbf{A}_\star = \left(\mathbf{A}_0 - \mathbf{A}_*\right) \prod_{\tau=0}^{T-1} \left(\mathbf{I} - 2\gamma X_{\tau+1} X_{\tau+1}^\top\right) + \sum_{t=0}^{T-1} N_t$$

where N_t is the noise term originating at iteration t.

Unrolled noise term

Unrolling the whole sequence of SGD on $(X_0, X_1, \ldots, X_T)$ all the way:

$$\mathbf{A}_T - \mathbf{A}_\star = \left(\mathbf{A}_0 - \mathbf{A}_*\right) \prod_{\tau=0}^{T-1} \left(\mathbf{I} - 2\gamma X_{\tau+1} X_{\tau+1}^\top\right) + \sum_{t=0}^{T-1} N_t$$

where N_t is the noise term originating at iteration t. It undergoes $T - t$ rounds updates:

$$N_t = -2\gamma\eta_t X_t^\top \prod_{\tau=t}^{T-1} \left(\mathbf{I} - 2\gamma X_{\tau+1} X_{\tau+1}^\top\right),$$

where $-2\gamma\eta_t X_t^\top$ was the original noise introduced at time t.

Unrolled noise term

Unrolling the whole sequence of SGD on $(X_0, X_1, \dots, X_T)$ all the way:

$$\mathbf{A}_T - \mathbf{A}_\star = \left(\mathbf{A}_0 - \mathbf{A}_*\right) \prod_{\tau=0}^{T-1} \left(\mathbf{I} - 2\gamma X_{\tau+1} X_{\tau+1}^\top\right) + \sum_{t=0}^{T-1} N_t$$

where N_t is the noise term originating at iteration t. It undergoes $T - t$ rounds updates:

$$N_t = \underline{-2\gamma\eta_t X_t^\top} \prod_{\tau=t}^{T-1} \left(\mathbf{I} - 2\gamma X_{\tau+1} X_{\tau+1}^\top\right),$$

where $\underline{-2\gamma\eta_t X_t^\top}$ was the original noise introduced at time t.

- **Issue:** SGD transforms the noise η_t by updates $(X_{t+1}, \dots X_T)$ that depend on η_t.

SGD: noise is transformed by correlated data

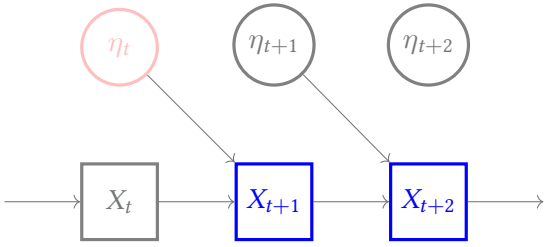

Figure 2: After running SGD, the noise term η_t is transformed updates generated by $X_{t+1}, \ldots, X_T$, leading to noise that is not mean zero.

Reverse experience replay (RER)

Idea: apply SGD on the reversed stream $(X_T, X_T - 1, \dots, X_0)$.

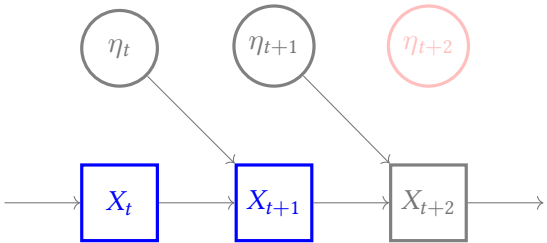

Figure 3: In SGD with *reverse experience replay*, the noise term is transformed by updates generated by past data it is independent of.

Reverse experience replay (RER) update

SGD update:

$$\mathbf{A}_{t+1} - \mathbf{A}_\star = \left(\mathbf{A}_t - \mathbf{A}_\star\right)\left(\mathbf{I} - 2\gamma X_t X_t^\top\right) - 2\gamma\eta_t X_t^\top.$$

Reverse experience replay (RER) update

SGD update:

$$\mathbf{A}_{t+1} - \mathbf{A}_\star = \big(\mathbf{A}_t - \mathbf{A}_\star\big)\big(\mathbf{I} - 2\gamma X_t X_t^\top\big) - 2\gamma\eta_t X_t^\top.$$

SGD with RER update:

$$\mathbf{A}_{t+1} - \mathbf{A}_\star = \big(\mathbf{A}_t - \mathbf{A}_\star\big)\big(\mathbf{I} - 2\gamma X_{-t} X_{-t}^\top\big) - 2\gamma\eta_{-t} X_{-t}^\top,$$

where $X_{-t} := X_{(T-1)-t}$ and $\eta_{-t} = \eta_{(T-1)-t}$.

Empirical comparison

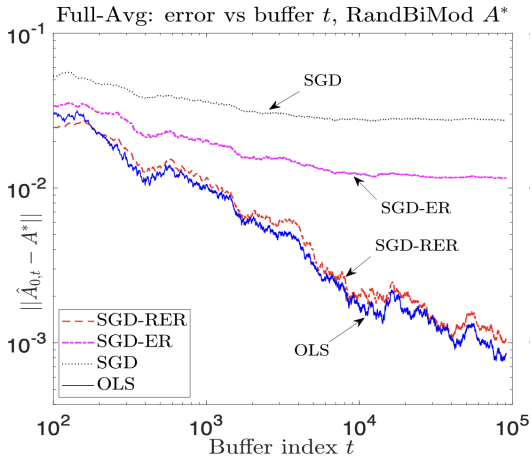

Figure 4: Comparison between SGD (forward traversal), SGD–RER (reverse traversal), SGD with experience replay (random traversal), and OLS (full gradient descent).

Streaming algorithm

Reverse traversal requires the whole stream to be stored.

Streaming algorithm

Reverse traversal requires the whole stream to be stored.

- This intuitive description of SGD–RER is not a **streaming algorithm**, which generally assumes we don't have enough memory to store the whole stream.

Streaming algorithm

Reverse traversal requires the whole stream to be stored.

- This intuitive description of SGD–RER is not a **streaming algorithm**, which generally assumes we don't have enough memory to store the whole stream.

Idea: given memory buffer size B, we can process the stream in pieces, where we replay the small batch of data in reverse order.

Streaming algorithm

Reverse traversal requires the whole stream to be stored.

- This intuitive description of SGD–RER is not a **streaming algorithm**, which generally assumes we don't have enough memory to store the whole stream.

Idea: given memory buffer size B, we can process the stream in pieces, where we replay the small batch of data in reverse order.

- To maintain near-independence across buffers, we can throw away a small amount of data in between pieces.

SGD – RER with memory buffer

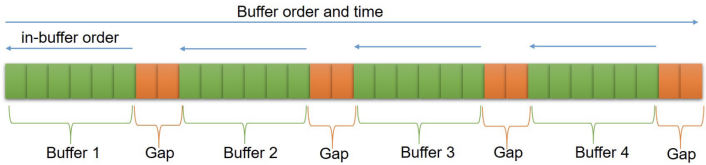

Figure 5: Given a memory buffer, we can store pieces of the stream, and perform SGD on each piece in reverse order.

Convergence result

Theorem (Informal, Kowshik et al. (2021b))

Assume that $\|\mathbf{A}_\star\|_{\mathrm{op}} < 1$ (stable dynamics), the noise η is ν^2-sub-Gaussian, and X_0 is drawn from the stationary distribution of the LTI. Let $\Sigma = \mathbb{E}[X_0 X_0^\top]$. With an appropriate γ, w.h.p.,

$$\left\|\hat{\mathbf{A}} - \mathbf{A}_\star\right\|_{\mathrm{op}}^2 = O\left(\frac{\nu^2}{\sigma_{\min}(\Sigma)} \frac{d + \log T}{T}\right).$$

- Read $\sigma_{\min}(\Sigma)/\nu^2$ as the signal-to-noise ratio.

Nearly matching lower bound

Theorem (Informal, Simchowitz et al. (2018))

Let $\rho < 1$ and $\mu = \mathcal{N}(0, \nu^2 I)$. For any estimator, there exists $\mathbb{A}_\star$ such that $\|\mathbf{A}_\star\|_{\mathrm{op}} = \rho$ such that with probability at least δ,

$$\left\|\hat{\mathbf{A}} - \mathbf{A}_\star\right\|_{\mathrm{op}}^2 = \Omega\left((1-\rho) \cdot \frac{d + \log 1/\delta}{T}\right).$$

Nearly matching lower bound

Theorem (Informal, Simchowitz et al. (2018))

Let $\rho < 1$ and $\mu = \mathcal{N}(0, \nu^2 I)$. For any estimator, there exists $\mathbb{A}_\star$ such that $\|\mathbf{A}_\star\|_{\mathrm{op}} = \rho$ such that with probability at least δ,

$$\left\|\hat{\mathbf{A}} - \mathbf{A}_\star\right\|_{\mathrm{op}}^2 = \Omega\left((1-\rho) \cdot \frac{d + \log 1/\delta}{T}\right).$$

- It turns out that $\mathbf{A}_\star$ can be chosen so that the covariance Σ of the associated stationary distribution is $\nu^2/(1-\rho^2)$, so we have:

$$\frac{\nu^2}{\sigma_{\min}(\Sigma)} \sim 1 - \rho.$$

Nearly matching lower bound

Theorem (Informal, Simchowitz et al. (2018))

Let $\rho < 1$ and $\mu = \mathcal{N}(0, \nu^2 I)$. For any estimator, there exists $\mathbb{A}_\star$ such that $\|\mathbf{A}_\star\|_{\mathrm{op}} = \rho$ such that with probability at least δ,

$$\left\|\hat{\mathbf{A}} - \mathbf{A}_\star\right\|_{\mathrm{op}}^2 = \Omega\left((1-\rho) \cdot \frac{d + \log 1/\delta}{T}\right).$$

- It turns out that $\mathbf{A}_\star$ can be chosen so that the covariance Σ of the associated stationary distribution is $\nu^2/(1-\rho^2)$, so we have:

$$\frac{\nu^2}{\sigma_{\min}(\Sigma)} \sim 1 - \rho.$$

  - Thus, SGD – RER convergence matches the minimax bound up to a factor of $\log T$.

References

Naman Agarwal, Syomantak Chaudhuri, Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Online target Q-learning with reverse experience replay: Efficiently finding the optimal policy for linear MDPs. *arXiv preprint arXiv:2110.08440*, 2021.

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

Suhas Kowshik, Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. Streaming linear system identification with reverse experience replay. *Advances in Neural Information Processing Systems*, 34, 2021b.

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

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