Concentration Inequalities for System Identification

Stephen Tu

Google Brain Robotics

CDC 2019 Tutorial

Problem statement

Problem statement

- Given **input/output** access to $x_{t+1} = Ax_t + Bu_t + w_t$.

Problem statement

- Given **input/output** access to $x_{t+1} = Ax_t + Bu_t + w_t$.
    - Assume $w_t \sim \mathcal{N}(0, \sigma^2 I)$.

Problem statement

- Given **input/output** access to $x_{t+1} = Ax_t + Bu_t + w_t$.
    - Assume $w_t \sim \mathcal{N}(0, \sigma^2 I)$.
- Goal is to **recover** (A, B) given $\mathcal{D} = \{(x_t, u_t, x_{t+1})\}_{t=0}^{T-1}$.

Problem statement

- Given **input/output** access to $x_{t+1} = Ax_t + Bu_t + w_t$.
  - Assume $w_t \sim \mathcal{N}(0,\sigma^2 I)$.
- Goal is to **recover** (A, B) given $\mathcal{D} = \{(x_t, u_t, x_{t+1})\}_{t=0}^{T-1}$.
- Want **bounds** on estimators $(\hat{A}, \hat{B})$ of the form:

$$\mathbb{P}(\|\hat{A} - A\| \geq \varepsilon) \leq \delta,$$
$$\mathbb{P}(\|\hat{B} - B\| \geq \varepsilon) \leq \delta.$$

Least-squares estimator

Least-squares estimator

- Basic least-squares estimator:

$$(\hat{A}, \hat{B}) = \arg\min_{A,B} \frac{1}{2} \sum_{i=0}^{T-1} \|x_{i+1} - Ax_i - Bu_i\|^2.$$

Least-squares estimator

- Basic least-squares estimator:

$$(\hat{A}, \hat{B}) = \arg\min_{A,B} \frac{1}{2} \sum_{i=0}^{T-1} \|x_{i+1} - Ax_i - Bu_i\|^2.$$

- Has closed-form solution:

$$(\hat{A}, \hat{B}) = \left(\sum_{i=0}^{t-1} x_{i+1} z_i^{\mathsf{T}} \right) \left(\sum_{i=0}^{t-1} z_i z_i^{\mathsf{T}} \right)^{-1}, \quad z_i = \begin{bmatrix} x_i \\ u_i \end{bmatrix}.$$

What do we know about LS?

What do we know about LS?

- Asymptotics of the least-squares estimator are well understood (in both stable/unstable case).

What do we know about LS?

- Asymptotics of the least-squares estimator are well understood (in both stable/unstable case).
- For simplicity, let us assume scalar autonomous system:

$$x_{t+1} = ax_t + w_t$$

What do we know about LS?

- Asymptotics of the least-squares estimator are well understood (in both stable/unstable case).
- For simplicity, let us assume scalar autonomous system:
$$x_{t+1} = ax_t + w_t$$
- Then from [Mann and Wald 1943, White 1958], we have a CLT:
$$\sqrt{T}(\hat{a}_T - a) \xrightarrow{d} \mathcal{N}(0, 1 - a^2) \text{ if } |a| < 1,$$
$$T(\hat{a}_T - a) \xrightarrow{d} \Phi \text{ if } |a| = 1,$$
$$|a|^T (\hat{a}_T - a) \xrightarrow{d} (a^2 - 1)\Psi \text{ if } |a| > 1$$

What do we know about LS?

- Asymptotics of the least-squares estimator are well understood (in both stable/unstable case).
- For simplicity, let us assume scalar autonomous system:

$$x_{t+1} = ax_t + w_t$$

- Then from [Mann and Wald 1943, White 1958], we have a CLT:

$$\sqrt{T}(\hat{a}_T - a) \xrightarrow{d} \mathcal{N}(0,1 - a^2) \text{ if } |a| < 1,$$

$$T(\hat{a}_T - a) \xrightarrow{d} \Phi \text{ if } |a| = 1,$$

$$|a|^T (\hat{a}_T - a) \xrightarrow{d} (a^2 - 1)\Psi \text{ if } |a| > 1$$

  - Φ is a non-standard distribution, Ψ is standard Cauchy.

What do we know about LS?

What do we know about LS?

- CLT:

$$\sqrt{T}(\hat{a}_T - a) \xrightarrow{d} \mathcal{N}(0,1 - a^2) \text{ if } |a| < 1,$$

$$T(\hat{a}_T - a) \xrightarrow{d} \Phi \text{ if } |a| = 1,$$

$$|a|^T(\hat{a}_T - a) \xrightarrow{d} (a^2 - 1)\Psi \text{ if } |a| > 1$$

What do we know about LS?

- CLT:

$$\sqrt{T}(\hat{a}_T - a) \xrightarrow{d} \mathcal{N}(0,1 - a^2) \text{ if } |a| < 1,$$

$$T(\hat{a}_T - a) \xrightarrow{d} \Phi \text{ if } |a| = 1,$$

$$|a|^T(\hat{a}_T - a) \xrightarrow{d} (a^2 - 1)\Psi \text{ if } |a| > 1$$

- Therefore, as a becomes more “explosive”, estimation becomes easier!

Beyond asymptotics

Beyond asymptotics

- Can we prove finite-time (non-asymptotic) rates in the **scalar** case?

Beyond asymptotics

- Can we prove finite-time (non-asymptotic) rates in the **scalar** case?
- Can we generalize to the **vector** case?

Roadmap

Roadmap

- Proof sketch of autonomous scalar case.

Roadmap

- Proof sketch of autonomous scalar case.
- Discuss why vector case is a non-trivial extension.

Roadmap

- Proof sketch of autonomous scalar case.
- Discuss why vector case is a non-trivial extension.
- Discuss state-of-the-art results in the vector case.

Scalar case setup

Scalar case setup

- Dynamics are $x_{t+1} = ax_t + w_t$.

Scalar case setup

- Dynamics are $x_{t+1} = ax_t + w_t$.
- LS estimator simplifies to $\hat{a}_t = \dfrac{\sum_{i=0}^{t-1} x_i x_{i+1}}{\sum_{i=0}^{t-1} x_i^2}$.

Scalar case setup

- Dynamics are $x_{t+1} = ax_t + w_t$.
- LS estimator simplifies to $\hat{a}_t = \dfrac{\sum_{i=0}^{t-1} x_i x_{i+1}}{\sum_{i=0}^{t-1} x_i^2}$.
- Error is therefore:

$$e_t := a_t - a = \frac{\sum_{i=0}^{t-1} x_i w_i}{\sum_{i=0}^{t-1} x_i^2}.$$

Key scalar martingale

Key scalar martingale

- Define $M_t := \sum_{i=0}^{t-1} x_i w_i$ with $M_0 = 0$.

Key scalar martingale

- Define $M_t := \sum_{i=0}^{t-1} x_i w_i$ with $M_0 = 0$.
- Define the filtration $\mathcal{F}_t := \sigma(w_0, \ldots, w_{t-1})$.

Key scalar martingale

- Define $M_t := \sum_{i=0}^{t-1} x_i w_i$ with $M_0 = 0$.
- Define the filtration $\mathscr{F}_t := \sigma(w_0, \ldots, w_{t-1})$.
    - $(x_0, \ldots, x_t, M_t)$ are $\mathscr{F}_t$-measurable.

Key scalar martingale

- Define $M_t := \sum_{i=0}^{t-1} x_i w_i$ with $M_0 = 0$.
- Define the filtration $\mathcal{F}_t := \sigma(w_0, \ldots, w_{t-1})$.
    - $(x_0, \ldots, x_t, M_t)$ are $\mathcal{F}_t$-measurable.
- Furthermore, M_t is a martingale since:

$$\begin{aligned}\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] &= \mathbb{E}[M_t \mid \mathcal{F}_t] + \mathbb{E}[x_t w_t \mid \mathcal{F}_t] \\ &= M_t + x_t \mathbb{E}[w_t \mid \mathcal{F}_t] \\ &= M_t\end{aligned}$$

Key scalar martingale

Key scalar martingale

- Next, we define the **quadratic variation** $\langle M \rangle_t$ as:

$$\langle M \rangle_t := \sum_{i=0}^{t-1} \mathbb{E}[(M_{i+1} - M_i)^2 \,|\, \mathscr{F}_i].$$

Key scalar martingale

- Next, we define the **quadratic variation** $\langle M \rangle_t$ as:

$$\langle M \rangle_t := \sum_{i=0}^{t-1} \mathbb{E}[(M_{i+1} - M_i)^2 \,|\, \mathcal{F}_i].$$

- A quick computation shows that $\langle M \rangle_t = \sigma^2 \sum_{i=0}^{t-1} x_i^2$.

Key scalar martingale

- Next, we define the **quadratic variation** $\langle M \rangle_t$ as:

$$\langle M \rangle_t := \sum_{i=0}^{t-1} \mathbb{E}[(M_{i+1} - M_i)^2 \,|\, \mathcal{F}_i].$$

- A quick computation shows that $\langle M \rangle_t = \sigma^2 \sum_{i=0}^{t-1} x_i^2$.

- Therefore, we can write:

$$e_t = \frac{\sum_{i=0}^{t-1} x_i w_i}{\sum_{i=0}^{t-1} x_i^2} = \sigma^2 \frac{M_t}{\langle M \rangle_t}.$$

Self-normalized processes

Self-normalized processes

- It turns out that the quantity $\dfrac{M_t}{\langle M \rangle_t}$ is well-studied in probability theory.

Self-normalized processes

- It turns out that the quantity $\dfrac{M_t}{\langle M \rangle_t}$ is well-studied in probability theory.
- Is often referred to as a **self-normalized** process.

Self-normalized processes

- It turns out that the quantity $\dfrac{M_t}{\langle M \rangle_t}$ is well-studied in probability theory.
- Is often referred to as a **self-normalized** process.
- A rich body of concentration inequalities to draw from that are of the form:

$$\mathbb{P}(M_t \geq \alpha \langle M \rangle_t) \leq \ldots$$

Self-normalized inequality

Self-normalized inequality

- One concrete result from [Bercu and Touati 08]:

$$\mathbb{P}(M_n \geq \alpha \langle M \rangle_n) \leq \inf_{p \geq 1} \left(\mathbb{E}\left[\exp\left(-(p-1)\frac{\alpha^2}{2} \langle M \rangle_n \right) \right] \right)^{1/p}.$$

Self-normalized inequality

- One concrete result from [Bercu and Touati 08]:

$$\mathbb{P}(M_n \geq \alpha\langle M\rangle_n) \leq \inf_{p\geq 1}\left(\mathbb{E}\left[\exp\left(-(p-1)\frac{\alpha^2}{2}\langle M\rangle_n\right)\right]\right)^{1/p}.$$

- **Sanity check:** if $M_n = \sum_{i=1}^{n} w_i$ with $w_i \sim \mathcal{N}(0,\sigma^2)$, then this reduces to $\mathbb{P}\left(\sum_{i=1}^{n} w_i \geq t\right) \leq \exp(-t^2/(2n\sigma^2))$.

Bounding moment generating function

Bounding moment generating function

- The next step is to control:

$$\mathbb{E}\exp\left(\theta\sum_{i=0}^{T-1}x_i^2\right), \ \theta < 0.$$

Bounding moment generating function

- The next step is to control:

$$\mathbb{E}\exp\left(\theta\sum_{i=0}^{T-1}x_i^2\right),\ \theta < 0.$$

- By tower property of expectations:

$$\mathbb{E}\left[\exp\left(\theta\sum_{i=0}^{T-2}x_i^2\right)\mathbb{E}[\exp\left(\theta x_{T-1}^2\right) \mid \mathcal{F}_{T-2}]\right]$$

$$= \mathbb{E}\left[\exp\left(\theta\sum_{i=0}^{T-2}x_i^2\right)\mathbb{E}[\exp\left(\theta(ax_{T-2} + w_{T-1})^2\right) \mid \mathcal{F}_{T-2}]\right]$$

An elementary MGF bound

An elementary MGF bound

- An elementary result states that for $\theta < 0$ and μ fixed,
$$\mathbb{E} \exp(\theta(\mu + w)^2) \leq \frac{1}{\sqrt{1 - 2\sigma^2\theta}}, \; w \sim \mathcal{N}(0,\sigma^2).$$

An elementary MGF bound

- An elementary result states that for $\theta < 0$ and μ fixed,
$$\mathbb{E} \exp(\theta(\mu + w)^2) \le \frac{1}{\sqrt{1 - 2\sigma^2\theta}}, \ w \sim \mathcal{N}(0,\sigma^2).$$

- Therefore:
$$\begin{aligned}
\mathbb{E}\left[\exp\left(\theta \sum_{i=0}^{T-2} x_i^2\right) \mathbb{E}[\exp\left(\theta x_{T-1}^2\right) \mid \mathcal{F}_{T-2}]\right] &= \mathbb{E}\left[\exp\left(\theta \sum_{i=0}^{T-2} x_i^2\right) \mathbb{E}[\exp\left(\theta (a x_{T-2} + w_{T-1})^2\right) \mid \mathcal{F}_{T-2}]\right] \\
&\le \mathbb{E} \exp\left(\theta \sum_{i=0}^{T-2} x_i^2\right) \frac{1}{\sqrt{1 - 2\sigma^2\theta}} \\
&\le \dots \\
&\le \frac{1}{(1 - 2\sigma^2\theta)^{T/2}}
\end{aligned}$$

Putting it together

Putting it together

- Recall the inequality from [Bercu and Touati 08]:

$$\mathbb{P}(e_T \geq v) \leq \inf_{p \geq 1} \left(\mathbb{E}\left[\exp\left(-(p-1)\frac{v^2}{2\sigma^2} \sum_{i=0}^{T-1} x_i^2 \right) \right] \right)^{1/p}.$$

Putting it together

- Recall the inequality from [Bercu and Touati 08]:

$$\mathbb{P}(e_T \geq v) \leq \inf_{p \geq 1} \left(\mathbb{E}\left[\exp\left(-(p-1)\frac{v^2}{2\sigma^2} \sum_{i=0}^{T-1} x_i^2 \right) \right] \right)^{1/p}.$$

- Now setting $\theta = -(p-1)v^2/(2\sigma^2)$,

$$\mathbb{P}(e_T \geq v) \leq \inf_{p \geq 1} \left[\frac{1}{1+(p-1)v^2} \right]^{T/2p} \leq \left[\frac{1}{1+v^2} \right]^{T/4}.$$

Putting it together

Putting it together

- Repeating the same argument for $-e_T$, we obtain our first concentration inequality by a union bound:

$$\mathbb{P}(|e_T| \geq v) \leq 2\left[\frac{1}{1+v^2}\right]^{T/4}.$$

Putting it together

- Repeating the same argument for $-e_T$, we obtain our first concentration inequality by a union bound:

$$\mathbb{P}(|e_T| \geq v) \leq 2\left[\frac{1}{1+v^2}\right]^{T/4}.$$

- Inverting this bound for large T, this states that with probability at least $1-\delta$, we have roughly:

$$|e_T| \lesssim \sqrt{\frac{1}{T}\log(1/\delta)}.$$

Drawbacks of bound

Drawbacks of bound

- Note that this bound we derived is **not sharp!**

Drawbacks of bound

- Note that this bound we derived is **not sharp!**
- First, consider stable $|a| < 1$. From CLT we know that $\sqrt{T}e_T \xrightarrow{d} \mathcal{N}(0,1-a^2)$. Hence a more correct bound would have the form $|e_T| \asymp \sqrt{\dfrac{1-a^2}{T}}$.

Drawbacks of bound

- Note that this bound we derived is **not sharp!**
- First, consider stable $|a| < 1$. From CLT we know that $\sqrt{T}e_T \xrightarrow{d} \mathcal{N}(0,1-a^2)$. Hence a more correct bound would have the form $|e_T| \asymp \sqrt{\dfrac{1-a^2}{T}}$.
- Situation is even worse for unstable $|a| > 1$, where we expect exponential rates: $|e_T| \asymp \dfrac{a^2-1}{|a|^T}$.

Sharpening the scalar bound

- The bound can be sharpened by a more refined MGF analysis— see Theorem B.1 from [Simchowitz et al. 19].

Difficulties of vector case

Vector case setup

Vector case setup

- Our setup is now $x_{t+1} = Ax_t + w_t$.

Vector case setup

- Our setup is now $x_{t+1} = Ax_t + w_t$.
- The error term is:

$$E_T := \left(\sum_{i=0}^{T-1} w_i x_i^\top\right)\left(\sum_{i=0}^{T-1} x_i x_i^\top\right)^{-1}.$$

Vector case setup

- Our setup is now $x_{t+1} = Ax_t + w_t$.
- The error term is:

$$E_T := \left(\sum_{i=0}^{T-1} w_i x_i^\top \right) \left(\sum_{i=0}^{T-1} x_i x_i^\top \right)^{-1}.$$

- We consider the following decomposition:

$$\|E_T\| \leq \frac{\left\| \left(\sum_{i=0}^{T-1} w_i x_i^\top \right) \left(\sum_{i=0}^{T-1} x_i x_i^\top \right)^{-1/2} \right\|}{\sqrt{\lambda_{\min} \left(\sum_{i=0}^{T-1} x_i x_i^\top \right)}}.$$

Vector case setup

Vector case setup

- The term $\left\| \left(\sum_{i=0}^{T-1} w_i x_i^\mathsf{T} \right) \left(\sum_{i=0}^{T-1} x_i x_i^\mathsf{T} \right)^{-1/2} \right\|$ is a vector-valued **self-normalized** martingale. For stable A, also not too difficult to bound [Abbasi-Yadkori et al. 11].

Vector case setup

- The term $\left\| \left(\sum_{i=0}^{T-1} w_i x_i^\top \right) \left(\sum_{i=0}^{T-1} x_i x_i^\top \right)^{-1/2} \right\|$ is a vector-valued **self-normalized** martingale. For stable A, also not too difficult to bound [Abbasi-Yadkori et al. 11].

- The tricky part is lower bounding $\lambda_{\min} \left(\sum_{i=0}^{T-1} x_i x_i^\top \right)$.

Attempt 1: Matrix Chernoff

Attempt 1: Matrix Chernoff

- Define $\Sigma_T := \sum_{i=0}^{T-1} x_i x_i^\top$. For $\theta > 0$:

$$
\begin{aligned}
\mathbb{P}\left(\lambda_{\min}(\Sigma_T) \le v\right) &= \mathbb{P}(-\theta\lambda_{\min}(\Sigma_T) \ge -\theta v) \\
&= \mathbb{P}(\exp(-\theta\lambda_{\min}(\Sigma_T)) \ge \exp(-\theta v)) \\
&\le \exp(\theta v)\mathbb{E}\exp(-\theta\lambda_{\min}(\Sigma_T)) \\
&= \exp(\theta v)\mathbb{E}\exp(\lambda_{\max}(-\theta\Sigma_T)) \\
&= \exp(\theta v)\mathbb{E}\lambda_{\max}(\exp(-\theta\Sigma_T)) \\
&\le \exp(\theta v)\mathbb{E}\operatorname{tr}\exp(-\theta\Sigma_T)
\end{aligned}
$$

Attempt 1: Matrix Chernoff

- Define $\Sigma_T := \sum_{i=0}^{T-1} x_i x_i^\top$. For $\theta > 0$:

$$
\begin{aligned}
\mathbb{P}\left(\lambda_{\min}(\Sigma_T) \le v\right) &= \mathbb{P}(-\theta\lambda_{\min}(\Sigma_T) \ge -\theta v) \\
&= \mathbb{P}(\exp(-\theta\lambda_{\min}(\Sigma_T)) \ge \exp(-\theta v)) \\
&\le \exp(\theta v)\mathbb{E}\exp(-\theta\lambda_{\min}(\Sigma_T)) \\
&= \exp(\theta v)\mathbb{E}\exp(\lambda_{\max}(-\theta\Sigma_T)) \\
&= \exp(\theta v)\mathbb{E}\lambda_{\max}(\exp(-\theta\Sigma_T)) \\
&\le \exp(\theta v)\mathbb{E}\operatorname{tr}\exp(-\theta\Sigma_T)
\end{aligned}
.
$$

- Therefore:

$$
\mathbb{P}(\lambda_{\min}(\Sigma_T) \le v) \le \inf_{\theta<0} \exp(-\theta v)\mathbb{E}\operatorname{tr}\exp(\theta\Sigma_T).
$$

Attempt 1: Matrix Chernoff

Attempt 1: Matrix Chernoff

- In the scalar case, we were able to bound for $\theta < 0$,
$$\mathbb{E} \exp(\theta \sum_{i=0}^{T-1} x_i^2) \leq \frac{1}{(1 - 2\sigma^2\theta)^{T/2}}.$$

Attempt 1: Matrix Chernoff

- In the scalar case, we were able to bound for $\theta < 0$,
$$\mathbb{E} \exp(\theta \sum_{i=0}^{T-1} x_i^2) \leq \frac{1}{(1 - 2\sigma^2\theta)^{T/2}}.$$
- The matrix version is to bound $\mathbb{E} \operatorname{tr} \exp(\theta \Sigma_T)$.

Attempt 1: Matrix Chernoff

- In the scalar case, we were able to bound for $\theta < 0$,
  $$\mathbb{E} \exp(\theta \sum_{i=0}^{T-1} x_i^2) \leq \frac{1}{(1 - 2\sigma^2\theta)^{T/2}}.$$
- The matrix version is to bound $\mathbb{E} \operatorname{tr} \exp(\theta \Sigma_T)$.
- The difficulty is that $\exp(A + B) \neq \exp(A)\exp(B)$ for matrices, so the scalar proof does not go through.

Attempt 2: Scalar projections

Attempt 2: Scalar projections

- To avoid matrix issues, we can consider the scalar process $\sum_{i=0}^{T-1} \langle v, x_i \rangle^2$ for a fixed $v \in \mathcal{S}^{n-1}$.

Attempt 2: Scalar projections

- To avoid matrix issues, we can consider the scalar process $\sum_{i=0}^{T-1} \langle v, x_i \rangle^2$ for a fixed $v \in \mathcal{S}^{n-1}$.
- We can then use scalar analysis to lower bound $\sum_{i=0}^{T-1} \langle v, x_i \rangle^2$ for each fixed v.

Attempt 2: Scalar projections

- To avoid matrix issues, we can consider the scalar process $\sum_{i=0}^{T-1} \langle v, x_i \rangle^2$ for a fixed $v \in \mathcal{S}^{n-1}$.
- We can then use scalar analysis to lower bound $\sum_{i=0}^{T-1} \langle v, x_i \rangle^2$ for each fixed v.
- But $\lambda_{\min}(\sum_{i=0}^{T-1} x_i x_i^\top) = \inf_{\|v\|=1} \sum_{i=0}^{T-1} \langle v, x_i \rangle^2$. How do you pass to uniformly on $\mathcal{S}^{n-1}$?

Attempt 2: Scalar projections

Attempt 2: Scalar projections

- Naive covering argument:

$$\lambda_{\min}\left(\sum_{i=0}^{T-1} x_i x_i^\top\right) = \inf_{\|v\|=1} \sum_{i=0}^{T-1} \langle v, x_i \rangle^2$$

$$\geq \min_{v \in N(\varepsilon)} \sum_{i=0}^{T-1} \langle v, x_i \rangle^2 - 2\varepsilon \| \sum_{i=0}^{T-1} x_i x_i^\top \|$$

Attempt 2: Scalar projections

- Naive covering argument:

$$\lambda_{\min}\Big(\sum_{i=0}^{T-1} x_i x_i^\top\Big) = \inf_{\|v\|=1} \sum_{i=0}^{T-1} \langle v, x_i \rangle^2$$

$$\geq \min_{v \in N(\varepsilon)} \sum_{i=0}^{T-1} \langle v, x_i \rangle^2 - 2\varepsilon \Big\| \sum_{i=0}^{T-1} x_i x_i^\top \Big\|$$

- But this requires upper bound on $\| \sum_{i=0}^{T-1} x_i x_i^\top \|$, which is very unsatisfying! (nevertheless this does work in the stable case).

State of the art vector results

Stable case

Stable case

- [Simchowitz et al. 19]: If $\rho(A) < 1$, then with probability at least $1 - \delta$:

$$\|\hat{A}_T - A\| \lesssim \sqrt{\frac{n \log(n/\delta)}{T \lambda_{\min}(\Sigma_\infty)}}.$$

Stable case

- [Simchowitz et al. 19]: If $\rho(A) < 1$, then with probability at least $1 - \delta$:
$$\|\hat{A}_T - A\| \lesssim \sqrt{\frac{n \log(n/\delta)}{T\lambda_{\min}(\Sigma_\infty)}}.$$
- Here, Σ_∞ is the stationary covariance:
$$A\Sigma_\infty A^\top - \Sigma_\infty + \sigma^2 I = 0.$$

Marginally stable case

Marginally stable case

- [Simchowitz et al. 19]: In the special case when $A = O$ with O orthogonal, then with probability $1 - \delta$:
$$\|\hat{A}_T - A\| \lesssim \frac{n \log(n/\delta)}{T}.$$

"Explosive" case

- [Sarkar and Rakhlin 19]: If $|\lambda_i| > 1$ for all i, then with probability at least $1 - \delta$:
$$\|\hat{A}_T - A\| \lesssim \|A^{-T}\|/\delta.$$

References

- Simchowitz et al. *Learning Without Mixing: Towards a Sharp Analysis of Linear System Identification.* Conference on Learning Theory. 2019.
- Sarkar and Rakhlin. *Near optimal finite time identification of arbitrary linear dynamical systems*. ICML 2019.
- Rantzer. *Concentration Bounds for Single Parameter Adaptive Control.* ACC 2018.
- Abbasi-Yadkori et al. *Online Least Squares Estimation with Self-Normalized Processes: An Application to Bandit Problems.* Conference on Learning Theory 2011.
- Bercu and Touati. *Exponential inequalities for self-normalized martingales with applications.* Ann. Appl. Probab. 18(5). 2008.
- Mann and Wald. *On the Statistical Treatment of Linear Stochastic Difference Equations*. Econometrica 11(3-4). 1943.
- White. *The Limiting Distribution of the Serial Correlation Coefficient in the Explosive Case*. Ann. Math. Statist. 29(4). 1958.