The Lower Tail of the Empirical Covariance

Identifiability$\neq$Concentration

Ingvar Ziemann (Penn)

Identifiability

time series model:

$$Y_t = \theta^\star X_t + V_t, \quad t = 1,\ldots,T$$

Where:

- Y_t - Outputs in $\mathbb{R}^{d_Y}$
- X_t - Covariates in $\mathbb{R}^{d_X}$
- V_t - Noise in $\mathbb{R}^{d_Y}$
- $\theta^\star$ - Unknown Parameter in $\mathbb{R}^{d_Y \times d_X}$

Identifiability: Recovery of

$\theta^\star$ in a noiseless model ($V_t \equiv 0$)

Are all $\theta^\star \in \mathbb{R}^{d_Y \times d_X}$ identifiable

After T time-steps?

Yes if: Persistence of $X_{1:T}$ (span $\mathbb{R}^{d_X}$)

Equiv: $\sum_{t=1}^{T} X_t X_t^\top \succ 0$

Recall:

$$\hat{\theta} - \theta^\star = \left(\sum_{t=1}^{T} V_t X_t^\top\right)\left(\sum_{t=1}^{T} X_t X_t^\top\right)^{-1}$$

Concentration and Persistence

Let $X_{t+1} = A^{\star} X_t + W_t$ and suppose that $\rho_\star \triangleq \rho(A_\star) < 1$

Recall that we know how to control $\left\| \frac{1}{T} \sum_{t=1}^{T} X_t X_t^\top - \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top\right] \right\|_{\mathrm{op}}$

Requires order $d_X \times \mathrm{poly}\left(\frac{1}{1-\rho_\star}\right)$-many samples to guarantee persistence

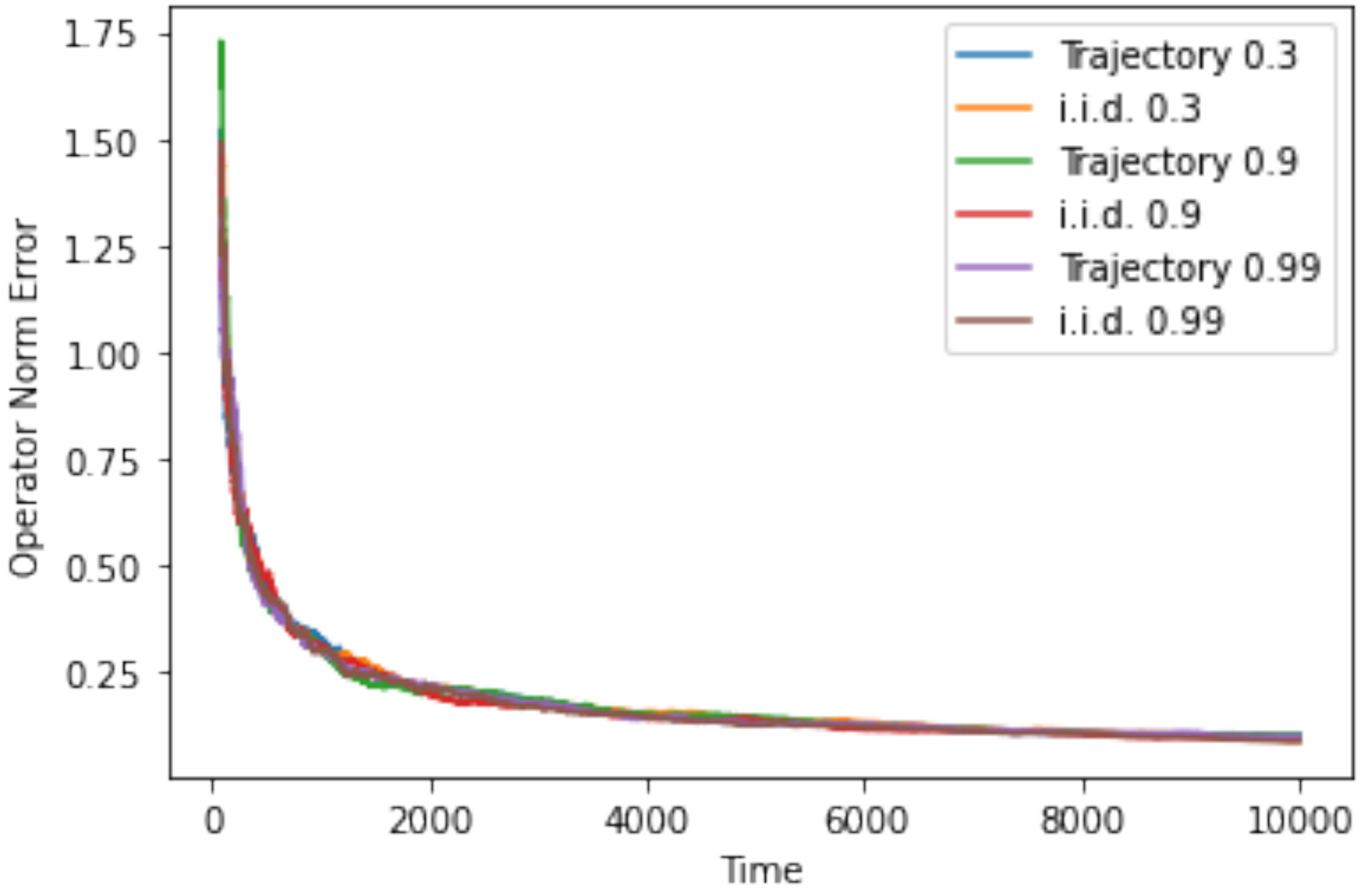
But identifiability $\approx$ linear independence $\Rightarrow$ should not depend on stability

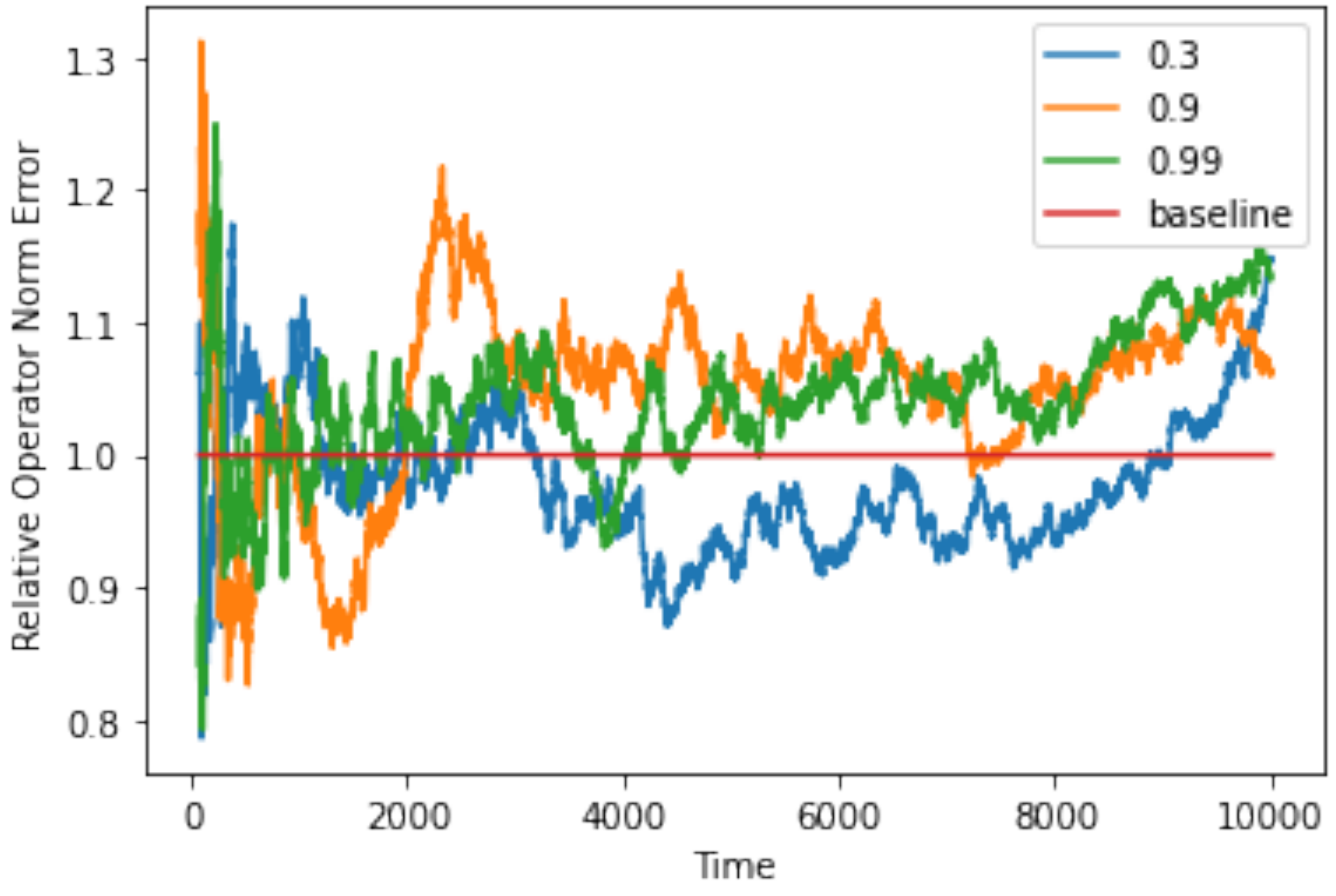
Can we remove the factor $\mathrm{poly}\left(\frac{1}{1-\rho_\star}\right)$?

Does poly $\left(1/(1-\rho_\star)\right)$ matter?

Doesn't seem so!

Let $X_{t+1} = A^\star X_t + W_t$ and suppose that $\rho_\star \to 1$





Basically no loss in performance

Persistence of Causal Processes

Want to guarantee $\sum_{t=1}^{T} X_t X_t^\top \succ 0$ for "reasonable" linear models (e.g. ARX)

Fix p-dim i.i.d. K^2-subG source of randomness $W_{1:T}$ with $\mathbf{E} W_{1:T} W_{1:T}^\top = I_{pT}$

Causality: $X_{1:T}$ is $k-$causal w/ subG incr if $\exists$ a block-lower triangular matrix $\mathbf{L}$ w/ form

$$(k \mid T) \;\; \mathbf{L} = \begin{bmatrix} \mathbf{L}_{1,1} & 0 & 0 & 0 & 0 \\ \mathbf{L}_{2,1} & \mathbf{L}_{2,2} & 0 & 0 & 0 \\ \mathbf{L}_{3,1} & \mathbf{L}_{3,2} & \mathbf{L}_{3,3} & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{L}_{T/k,1} & \dots & \dots & \dots & \dots \mathbf{L}_{T/k,T/k} \end{bmatrix} = \begin{bmatrix} \mathbf{L}_1 \\ \mathbf{L}_2 \\ \mathbf{L}_3 \\ \vdots \\ \mathbf{L}_{T/k} \end{bmatrix} \quad \text{and} \quad X_{1:T} = \mathbf{L} W_{1:T}$$

Decoupling Causal Processes

$X_{1:T} = \mathbf{L}W_{1:T}$ is (typically) a highly dependent process

We will relate $X_{1:T} = \mathbf{L}W_{1:T}$ to $\tilde{X}_{1:T} = \tilde{\mathbf{L}}W_{1:T}$ where

$$\tilde{\mathbf{L}} \triangleq \begin{bmatrix} \mathbf{L}_{1,1} & 0 & 0 & 0 \\ 0 & \mathbf{L}_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \mathbf{L}_{T/k,T/k} \end{bmatrix}$$

Key Idea: Lower bound

$\sum_{t=1}^{T} X_t X_t^\top$ by $\sum_{t=1}^{T} \mathbf{E}\tilde{X}_t\tilde{X}_t^\top$

Obtain $\tilde{\mathbf{L}}$ by discarding sub-diagonal of $\mathbf{L}$

For an LTI system this amounts to “restarting” the process every k steps

$\Rightarrow$ instead of one long trajectory work with T/k independent trajectories

Example: AR(1)

Let $X_{t+1} = A^\star X_t + W_t$, then:

$$\mathbf{L} = \begin{bmatrix} I & 0 & 0 & \dots & 0 \\ A^\star & I & 0 & \dots & 0 \\ A^{2,\star} & A^\star & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ A^{T-1,\star} & \dots & \dots & A^\star & I \end{bmatrix}$$

general k:

$$\tilde{\mathbf{L}} = \mathrm{blkdiag}\left(\begin{bmatrix} I & 0 & 0 & \dots & 0 \\ A^\star & I & 0 & \dots & 0 \\ A^{2,\star} & A^\star & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ A^{k-1,\star} & \dots & \dots & A^\star & I \end{bmatrix} \right)$$

With $k = 1$:

$$\tilde{\mathbf{L}} = \begin{bmatrix} I & 0 & 0 & \dots & 0 \\ 0 & I & 0 & \dots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & I \end{bmatrix}$$

With $k = 2$:

$$\tilde{\mathbf{L}} = \begin{bmatrix} I & 0 & 0 & \dots & 0 \\ A^\star & I & 0 & \dots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & I & 0 \\ 0 & \dots & 0 & A^\star & I \end{bmatrix}$$

Key Decoupling Inequality

$Q \succeq 0$, x arbitrary, W isotropic K^2-subG, mean zero indep. entries (Prop. 3.1)

$$\lambda \in \left[0, \frac{1}{8\sqrt{2}K^2\|Q\|_{\mathrm{op}}}\right] \quad \Rightarrow \quad \mathbf{E}\exp\left(-\lambda \begin{bmatrix} x \\ W \end{bmatrix}^\top \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} x \\ W \end{bmatrix}\right)$$

$$\leq \exp\left(-\lambda \mathrm{tr} Q_{22} + 36K^4\lambda^2 \mathrm{tr} Q_{22}^2\right)$$

Use case? $v \in$ unit sphere, $J_v = \mathrm{blkdiag}(vv^\top)$

$$\sum_{t=1}^{T} \langle X_t, v\rangle^2 = \sum_{t=1}^{T-k} \langle X_t, v\rangle^2 + \sum_{t=T-k+1}^{T} \langle X_t, v\rangle^2 = \sum_{t=1}^{T-k} \langle X_t, v\rangle^2 + W_{0:T-1}^\top \mathbf{L}_{T/k}^\top J_v \mathbf{L}_{T/k} W_{0:T-1}$$

$x = W_{1:T-k}$

$W = W_{T-k+1:T}$

$$= W_{1:T-k}^\top [\,*\,] W_{1:T-k} + \begin{bmatrix} W_{1:T-k} \\ W_{T-k+1:T} \end{bmatrix} \begin{bmatrix} * & * \\ * & \mathbf{L}_{T/k,T/k} vv^\top \mathbf{L}_{T/k,T/k} \end{bmatrix} \begin{bmatrix} W_{1:T-k} \\ W_{T-k+1:T} \end{bmatrix}^\top$$

The Lower Spectrum of the Empirical Covariance

(Theorem 3.1)

Fix $k, T \in \mathbb{N}$ with $k \mid T$

Let $X_{1:T}$ be k-causal with K^2-subG incr. $(X_{1:T} = \mathbf{L} W_{1:T})$

Let $\mathbf{L}_{1,1} = \mathbf{L}_{2,2} = \ldots$ (diag. stationarity)

$$\sum_{t=1}^{T} \mathbf{E} \tilde{X}_t \tilde{X}_t^\top \succ 0$$

(k-step controllability)

Then w.p $1 - \delta$:

$$\frac{1}{T} \sum_{t=1}^{T} X_t X_t^\top \succeq \frac{1}{8T} \sum_{t=1}^{T} \mathbf{E} \tilde{X}_t \tilde{X}_t^\top$$

As long as: $T/k \gtrsim K^2 d(\log C^*_{\text{sys}} + \log(1/\delta))$

$$C_{\text{sys}} = O\left(\text{poly}\left(T, \lambda_{\max}\left(\sum_{t=1}^{T} \mathbf{E} X_t X_t^\top\right), \lambda_{\min}^{-1}\left(\sum_{t=1}^{T} \mathbf{E} \tilde{X}_t \tilde{X}_t^\top\right)\right)\right)$$

*terms and conditions apply

Simchowitz, Max, et al. "Learning without mixing: Towards a sharp analysis of linear system identification." *Conference On Learning Theory*. PMLR, 2018.
Ziemann, Ingvar. "A note on the smallest eigenvalue of the empirical covariance of causal Gaussian processes." *IEEE Transactions on Automatic Control* (2023).

Example AR(1)

Let $X_{t+1} = A^\star X_t + W_t$, then:

$$\mathbf{L} = \begin{bmatrix} I & 0 & 0 & \dots & 0 \\ A^\star & I & 0 & \dots & 0 \\ A^{2,\star} & A^\star & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ A^{T-1,\star} & \dots & \dots & A^\star & I \end{bmatrix}$$

$$\tilde{\mathbf{L}} = \text{blkdiag} \begin{bmatrix} I & 0 & 0 & \dots & 0 \\ A^\star & I & 0 & \dots & 0 \\ A^{2,\star} & A^\star & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ A^{k-1,\star} & \dots & \dots & A^\star & I \end{bmatrix}$$

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{E}\tilde{X}_t\tilde{X}_t^\top = \frac{1}{k}\sum_{t=1}^{k} \mathbf{E}X_tX_t^\top = \frac{1}{k}\sum_{t=1}^{k}\Gamma_t$$

$$\Gamma_t = \sum_{j=0}^{t-1} A^{\star,j}A^{\star,j,\top}$$

Hence Theorem 3.1 informs us that:

$\frac{1}{T}\sum_{t=1}^{T} X_tX_t^\top \succeq \frac{1}{8k}\sum_{t=1}^{k}\Gamma_t$ with probability $1-\delta$ as long as $T/k \gtrsim K^2(d\log C_{\text{sys}} + \log(1/\delta))$

Takeway: Persistence does not require stability

$$X_{t+1} = A^{\star} X_t + W_t \qquad \Gamma_l = \sum_{j=0}^{l-1} A^{\star,j} A^{\star,j,\top}$$

Requires k-step controllability of $(A_{\star}, I)$

Theorem 3.1 informs us that:

$\frac{1}{T}\sum_{t=1}^{T} X_t X_t^{\top} \succeq \frac{1}{8k}\sum_{t=1}^{k} \Gamma_t$ with probability $1-\delta$ as long as $T/k \gtrsim K^2(d \log C_{\text{sys}} + \log(1/\delta))$

Grow polynomially with T unless $\rho(A_{\star}) < 1$

Saved by the logarithm

Results from the previous presentation showed:

$\frac{1}{T}\sum_{t=1}^{T} X_t X_t^{\top} \succeq \frac{1}{8T}\sum_{t=1}^{T} \Gamma_t$ with probability $1-\delta$ as long as $T \gtrsim K^2 C'_{\text{sys}}(\log(1/\delta) + d)$

Requires strict stability