Non-Asymptotic System Identification

CDC'23 and https://arxiv.org/abs/2309.03873

Anastasios Tsiamis (ETH)

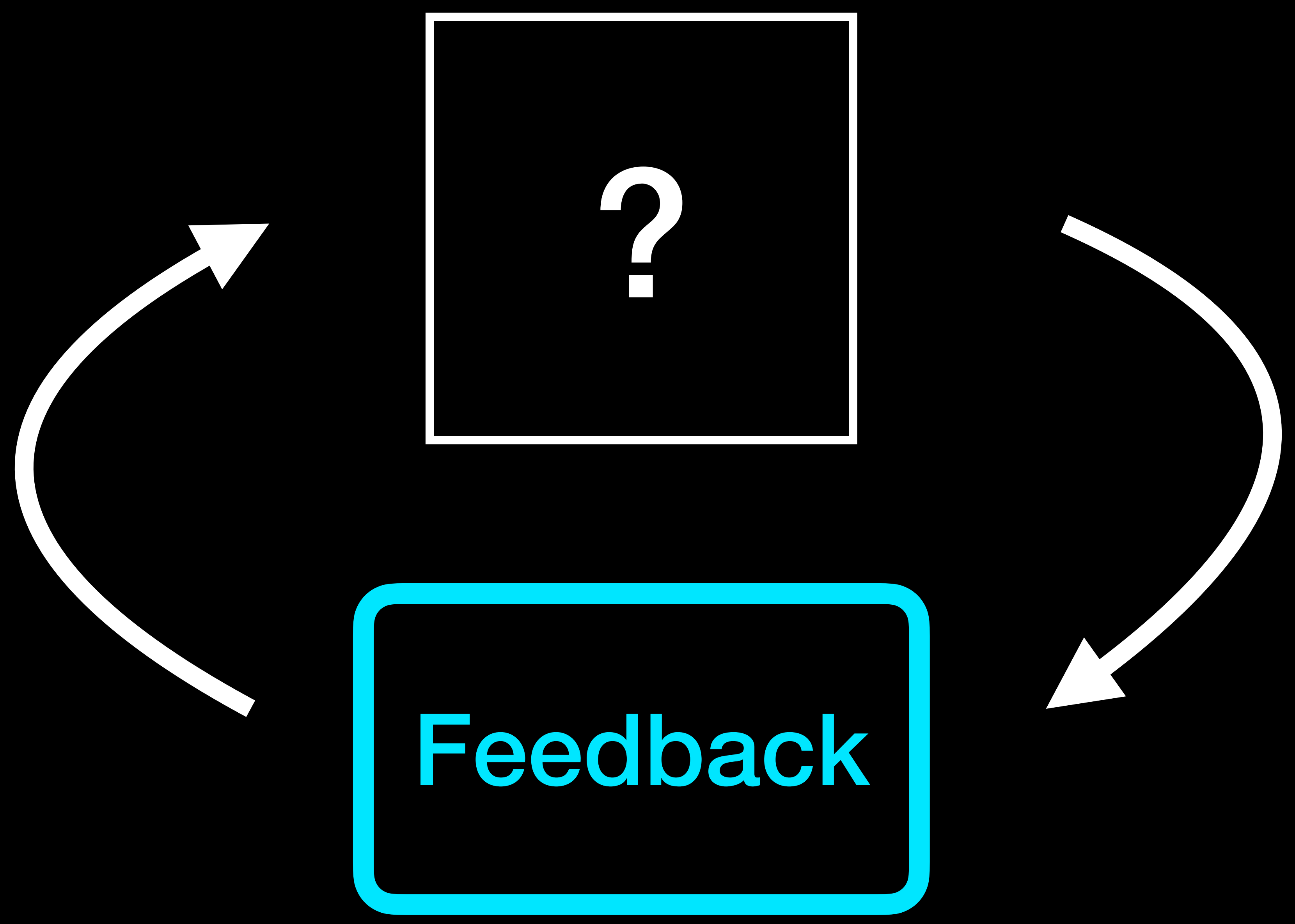
System Identification

Experiments: excite system

SysID: learn model

Today's focus: SysID

Why SysID?

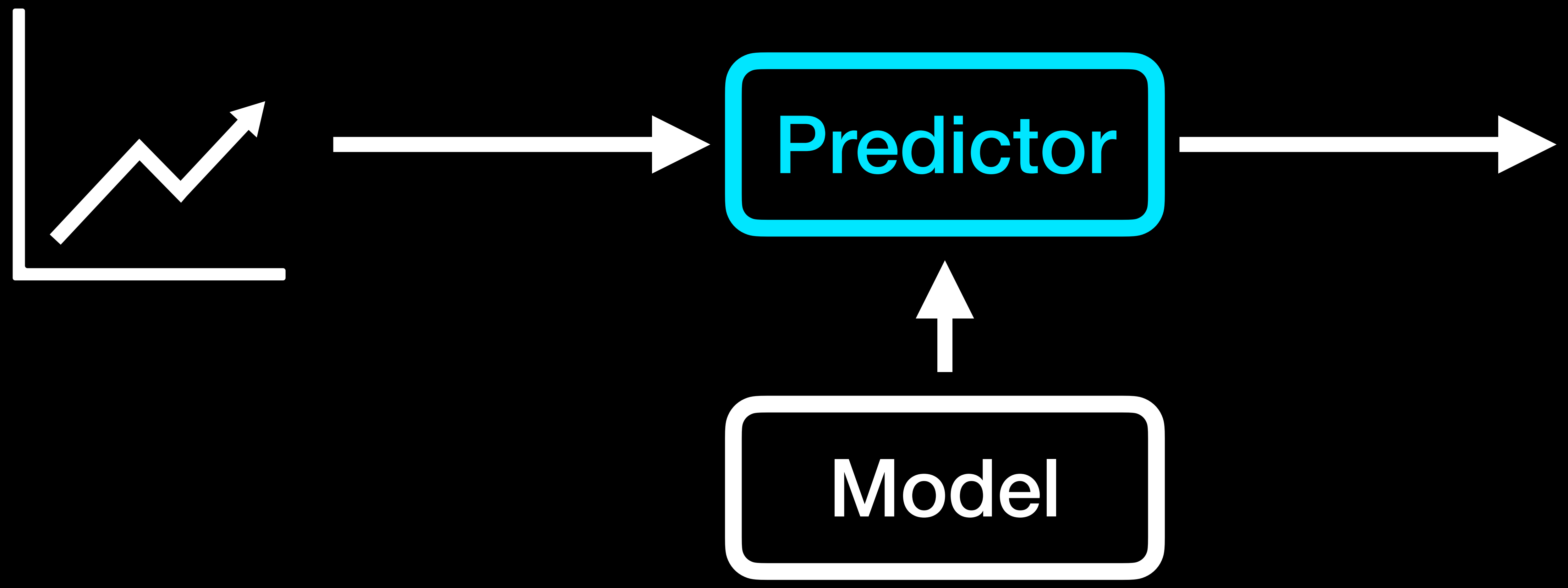


Classical **control** design: requires model

Unknown or complex system?

Learn from data

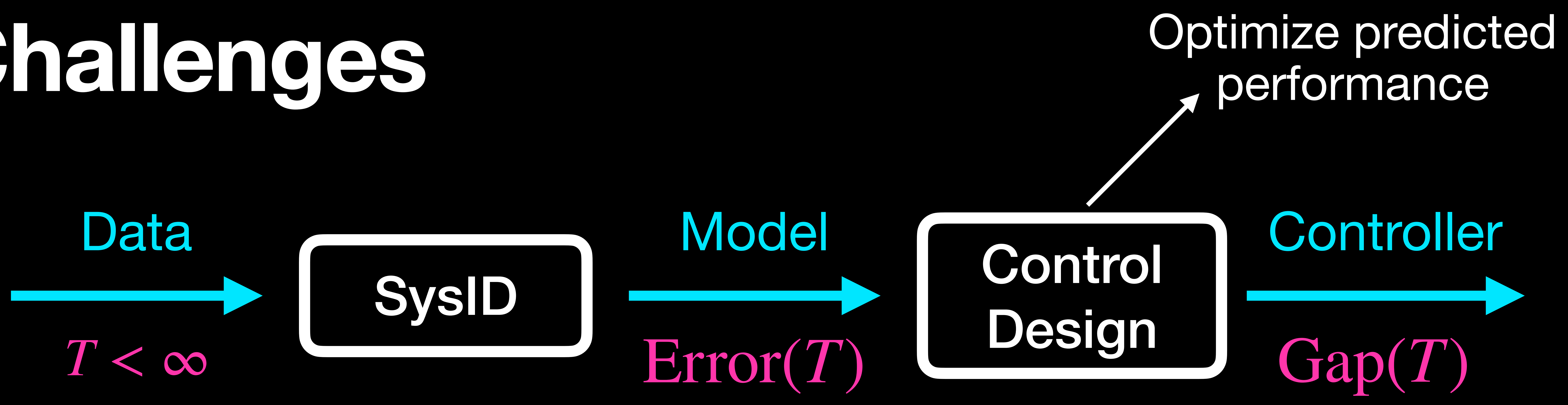
Why SysID?

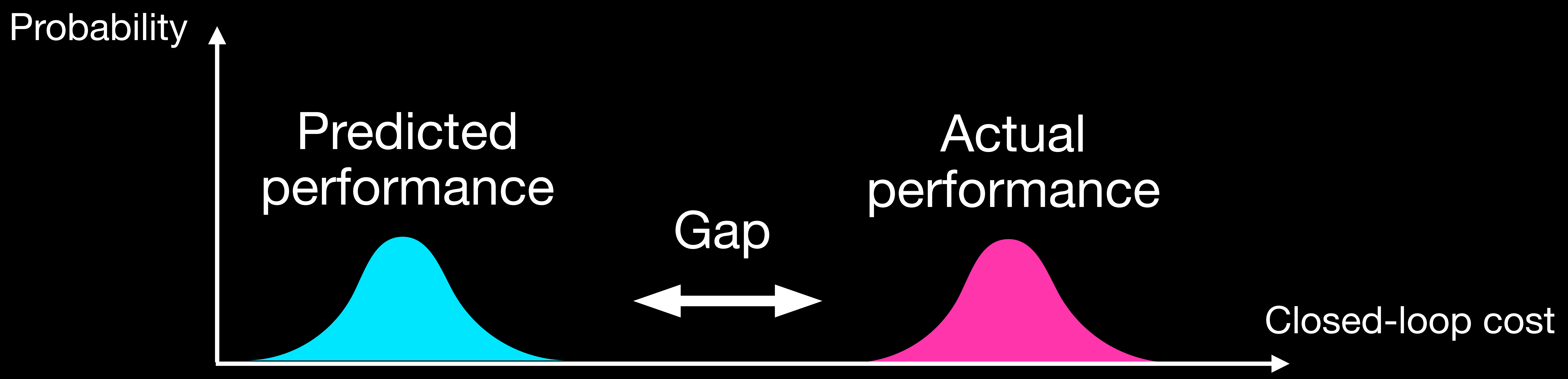


Also: **forecasting, fault detection**

Ubiquitous Applications

Challenges





Many factors (statistical, modeling bias, experiment design, sim2real, …)

Today: Finite Data $T < \infty$

Finite Sample Error of SysID

How fast does the error decay with number T of data?

How does it depend on system properties (size, structure, noise)?

Linear System Identification

Linear Systems

Unknown

- ARX: $$Y_t = \sum_{i=1}^{p} A_i^\star Y_{t-i} + \sum_{j=1}^{q} B_i^\star U_{t-j} + W_t$$

- Building block for learning state-space systems

$$Z_{t+1} = A^\star Z_t + B^\star U_t + K^\star W_t$$
$$Y_t = C^\star Z_t + W_t$$

- Linear: **simple** but **non-trivial** class

Stay tuned for non-linear!

Assumptions

$$Y_t = \sum_{i=1}^{p} A_i^{\star} Y_{t-i} + \sum_{j=1}^{q} B_i^{\star} U_{t-j} + W_t$$

1. Marginally Stable system (poles on or inside unit circle)
2. Known p, q (upper bounds)
3. σ^2-sub-Gaussian noise, zero-mean, independent
4. White-noise inputs
5. Single trajectory data: $Y_0, U_0, \ldots, Y_T, U_T$

Problem

Unknown

$$Y_t = \sum_{i=1}^{p} A_i^\star Y_{t-i} + \sum_{j=1}^{q} B_i^\star U_{t-j} + W_t$$

Given data: $Y_0, U_0, \ldots, Y_T, U_T$

Return $\hat{A}_1^\star, \ldots, \hat{A}_p^\star, \hat{B}_1^\star, \ldots, \hat{B}_q^\star$ and finite-sample bounds for the error

ARX in the Least Squares Framework

$$Y_t = \theta^\star X_t + W_t$$

$$Y_t = \sum_{i=1}^{p} A_i^\star Y_{t-i} + \sum_{j=1}^{q} B_i^\star U_{t-j} + W_t$$

Instance of least squares setup

$$\begin{bmatrix} Y_{t-1} \\ \vdots \\ Y_{t-p} \end{bmatrix} \begin{bmatrix} U_{t-1} \\ \vdots \\ U_{t-q} \end{bmatrix} \longrightarrow X_t = \begin{bmatrix} Y_{t-1:t-p}^\top & U_{t-1:t-q}^\top \end{bmatrix}^\top$$

$$\begin{bmatrix} A_{1:p}^\star & B_{1:q}^\star \end{bmatrix} \longrightarrow \theta^\star$$

$$\hat{\theta} \in \operatorname{argmin}_{\theta \in \mathbb{R}^{d_Y \times d_X}} \left\{ \frac{1}{T} \sum_{t=1}^{T} \| Y_t - \theta X_t \|_2^2 \right\}$$

ARX Least Squares error

$$Y_t = \theta^\star X_t + W_t$$

$$\hat{\theta} \in \operatorname{argmin}_{\theta \in \mathbb{R}^{d_Y \times d_X}} \left\{ \frac{1}{T} \sum_{t=1}^{T} \| Y_t - \theta X_t \|_2^2 \right\}$$

Sample Covariance
Lower Tail Methods

$$\hat{\theta} - \theta^\star = \left(\sum_{t=1}^{T} W_t X_t^\top \right) \left(\sum_{t=1}^{T} X_t X_t^\top \right)^{-1/2} \left(\sum_{t=1}^{T} X_t X_t^\top \right)^{-1/2}$$

Self-normalized
martingale methods

Bounding the terms

$$\Sigma_t \triangleq \mathbf{E} X_t X_t^\top$$

$$\frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top \geq c\Sigma_\tau, \quad \text{for } T \geq T_{\text{burn}}(\tau, \delta)$$

Empirical cov. Covariance

$$\left(\sum_{t=1}^{T} W_t X_t^\top\right)\left(\sum_{t=1}^{T} X_t X_t^\top\right)^{-1/2} \leq c'\sigma^2(d_X + \log\det \Sigma_T \Sigma_\tau^{-1} + \log\frac{1}{\delta})$$

With prob. $1 - \delta$

Mismatch? Empirical Cov. at T
Covariance at some $\tau < T$

$$\hat{\theta} - \theta^\star = \left(\sum_{t=1}^{T} W_t X_t^\top\right)\left(\sum_{t=1}^{T} X_t X_t^\top\right)^{-1/2}\left(\sum_{t=1}^{T} X_t X_t^\top\right)^{-1/2}$$

τ: tunable to reduce mismatch

But: It increases burn-in time

Finite-sample bound

$$\Sigma_t \triangleq \mathbf{E} X_t X_t^\top$$

size of $\theta^\star$

Confidence level

$$\|\widehat{\theta_T} - \theta^\star\|_{\mathrm{op}}^2 \lesssim \frac{c}{T} \frac{\sigma^2}{\lambda_{\min}(\Sigma_\tau)} \left(p d_Y + q d_U + \log\det \Sigma_T \Sigma_\tau^{-1} + \log 1/\delta\right)$$

Hides log factors

SNR^{-1}: Signal-to-noise ratio (inverted)

"Mismatched" condition number of covariances

$\sim \log T$: marginally Stable

~ 1: Stable

Non-asymptotic rate of $1/T$

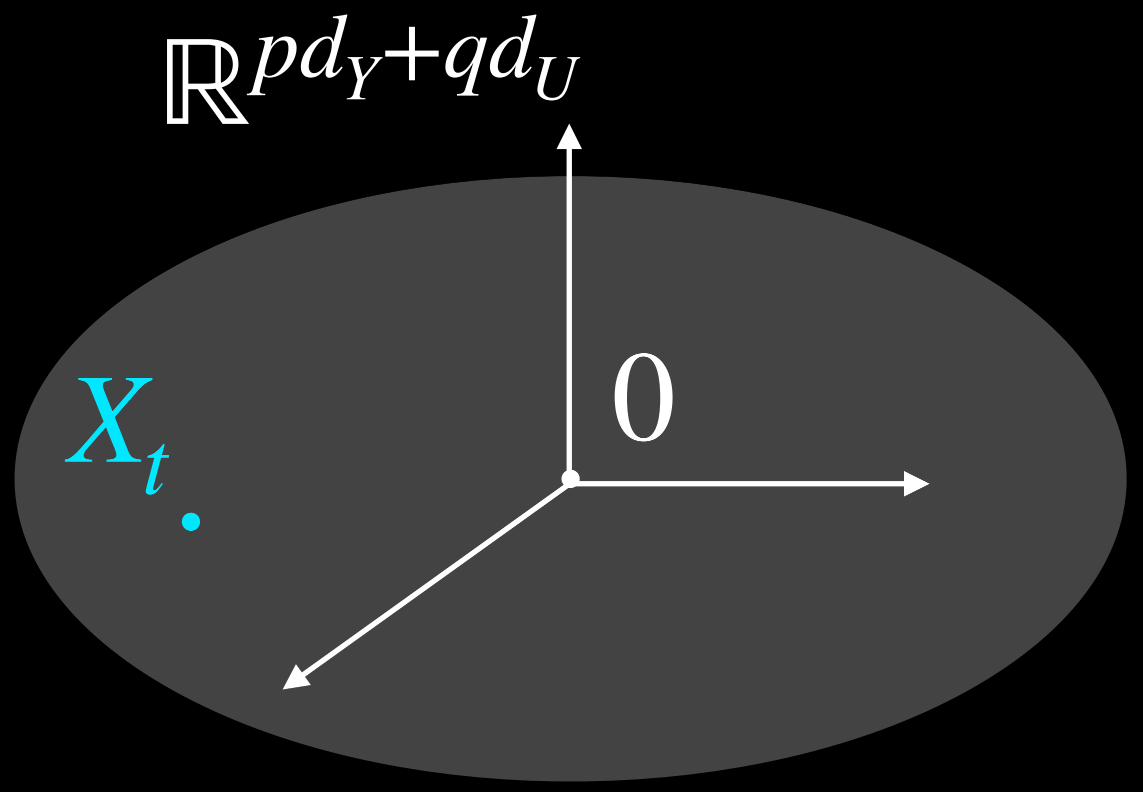
Agrees with asymptotics (Central Limit Theorem)

Persistency of excitation

$\Sigma_t \triangleq \mathbf{E} X_t X_t^\top$

$$\frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top \geq c\Sigma_\tau, \quad \text{for } T \geq T_{\text{burn}}(\tau, \delta)$$

Before T_{burn} matrix can be singular



After T_{burn} matrix invertible & persistently away from zero

$$T_{\text{burn}} \simeq c'\tau\left(\log 1/\delta + (pd_Y + qd_U)\cdots\right)$$

Trades excitation with larger Burn-in

Dimension of X_t

Tuning the bounds

$\Sigma_t \triangleq \mathbf{E} X_t X_t^\top$

$$\frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top \geq c\Sigma_\tau, \quad \text{for } T \geq T_{\text{burn}}(\tau, \delta)$$

$$T_{\text{burn}} \simeq c\tau \left(\log 1/\delta + (pd_Y + qd_U)\cdots\right)$$

Tune τ?

$\tau = 1 < p, q$ ✗

$$X_1 = \begin{bmatrix} Y_0^\top & 0 & \cdots & U_0^\top & 0 & \cdots & 0 \end{bmatrix}^\top$$

Singular Σ_τ

$\tau = \max\{p, q\}$ ✓

$$X_\tau = \begin{bmatrix} Y_{\tau-1:\tau-p}^\top & U_{\tau-1:\tau-q}^\top \end{bmatrix}^\top$$

Invertible Σ_τ (details in paper)

τ arbitrary large ?

only if $T \geq T_{\text{burn}}(\tau, \delta) \simeq c\tau(\cdots)$

Feasible if $\tau = o(T)$, e.g. $\tau = \sqrt{T}$

But large burn-in!

Tuning for Stable Systems

$\Sigma_t \triangleq \mathbf{E} X_t X_t^\top$

$$\|\widehat{\theta_T} - \theta^\star\|_{\text{op}}^2 \lesssim \frac{c}{T} \frac{\sigma^2}{\lambda_{\min}(\Sigma_\tau)} \left(pd_Y + qd_U + \log\det \Sigma_T \Sigma_\tau^{-1} + \log 1/\delta\right)$$

For stable plants we have $\Sigma_t \to \Sigma$

$$\|\widehat{\theta_T} - \theta^\star\|_{\text{op}}^2 \lesssim \frac{c'}{T} \frac{\sigma^2}{\lambda_{\min}(\Sigma)} \left(pd_Y + qd_U + \log 1/\delta\right)$$

where $\tau = \tau_{\text{mix}}$: mixing time (settling time of dominant pole)

Remarks

Guarantees valid even if $\tau = \max\{p, q\} < \tau_{\text{mix}}$

- Stability or mixing is not required
- Such tradeoffs not captured by asymptotics

System theoretic constants

- $\|\widehat{\theta}_T - \theta^\star\|_{\text{op}}^2 \lesssim \frac{c}{T} \frac{\sigma^2}{\lambda_{\min}(\Sigma_\tau)} (p d_Y + q d_U + \cdots)$

Model size

Identifiability/ excitability

- SNR depends on system's controllability structure
- Both size and structure affect difficulty

Remarks

$$\|\widehat{\theta_T} - \theta^\star\|_{\text{op}}^2 \lesssim \frac{c}{T}\frac{\sigma^2}{\lambda_{\min}(\Sigma_\tau)}(\cdots)$$

operator norm: worst-case error vs worst-case excitation

Alternative statements are possible, e.g.

$$(\widehat{\theta_T} - \theta^\star)\Sigma_T(\widehat{\theta_T} - \theta^\star)^\top \preceq \frac{c}{T}\sigma^2(\cdots)$$

Generalization to State Space

$$Z_t = A^\star Z_{t-1} + B^\star U_{t-1} + K^\star W_{t-1}$$
$$Y_{t-1} = C^\star Z_{t-1} + W_{t-1}$$

Innovation Form

$\updownarrow$ Under conditions, e.g. Gaussianity

$$Z_t = A^\star Z_{t-1} + B^\star U_{t-1} + W_{t-1}$$
$$Y_{t-1} = C^\star Z_{t-1} + V_{t-1}$$

Standard form

1. (Marginally) Stable system
2. σ^2-sub-Gaussian noise, zero-mean, independent
3. White-noise inputs
4. Minimum-phase: $A^\star - K^\star C^\star$ is stable (linked to observability)

Reformulation

$$Z_t = A^\star Z_{t-1} + B^\star U_{t-1} + K^\star W_{t-1}$$

$$Y_{t-1} = C^\star Z_{t-1} + W_{t-1}$$

$$A_{\mathrm{cl}}^\star \triangleq A^\star - K^\star C^\star$$

$$A_i^\star \triangleq C^\star (A_{\mathrm{cl}}^\star)^{i-1} K^\star$$

$$B_i^\star \triangleq C^\star (A_{\mathrm{cl}}^\star)^{i-1} B^\star$$

Can we approximate it as ARX?

$$\begin{aligned}
Y_t &= C^\star Z_t + W_t \\
&= C^\star A^\star Z_{t-1} + C^\star B^\star U_{t-1} + C^\star K^\star W_{t-1} + W_t \\
&= C^\star A_{\mathrm{cl}}^\star Z_{t-1} + C^\star B^\star U_{t-1} + C^\star K^\star Y_{t-1} + W_t \\
&= C^\star (A_{\mathrm{cl}}^\star)^2 Z_{t-2} + C^\star A_{\mathrm{cl}}^\star B^\star U_{t-2} + C^\star A_{\mathrm{cl}}^\star K^\star Y_{t-2} \\
&\quad + C^\star B^\star U_{t-1} + C^\star K^\star Y_{t-1} + W_t \\
&= \sum_{i=1}^{p} A_i^\star Y_{t-i} + \sum_{j=1}^{p} B_i^\star U_{t-j} + W_t + C^\star (A_{\mathrm{cl}}^\star)^p Z_{t-p}
\end{aligned}$$

ARX Approximation

$$Y_t = \sum_{i=1}^{p} A_i^\star Y_{t-i} + \sum_{j=1}^{p} B_i^\star U_{t-j} + W_t + C^\star (A_{\text{cl}}^\star)^p Z_{t-p}$$ → Bias term

$$A_{\text{cl}}^\star \triangleq A^\star - K^\star C^\star$$
$$A_i^\star \triangleq C^\star (A_{\text{cl}}^\star)^{i-1} K^\star$$
$$B_i^\star \triangleq C^\star (A_{\text{cl}}^\star)^{i-1} B^\star$$

$$X_t = \begin{bmatrix} Y_{t-1:t-p}^\top & U_{t-1:t-p}^\top \end{bmatrix}^\top$$

$$\theta_p^\star = \begin{bmatrix} A_{1:p}^\star & B_{1:p}^\star \end{bmatrix}$$

Minimum-phase

$$\|(A_{\text{cl}}^\star)^p\| \leq c\rho^p$$

For $p = c' \log T$

$$Y_t \approx \theta_p^\star X_t + W_t$$

Learn Markov Parameters $C^\star A_{\text{cl}}^i B^\star$
instead of original $A^\star, B^\star, C^\star, K^\star$

State-Space bound

$\Sigma_{p,t} \triangleq \mathbf{E} X_t X_t^\top$

Size of $\theta_p^\star$

Confidence level

$$\|\widehat{\theta}_T - \theta_p^\star\|_{op}^2 \lesssim \frac{c}{T} \frac{\sigma^2}{\lambda_{\min}(\Sigma_{p,\tau})} \left(p(d_Y + d_U) + \log\det \Sigma_{p,T}\Sigma_{p,\tau}^{-1} + \log 1/\delta\right)$$

Hides log factors

SNR^{-1}: Signal-to-noise ratio (inverted)

"Mismatched" condition number of covariances

$\sim \log T$: marginally Stable

~ 1: Stable

Non-asymptotic rate of $1/T$

But $p = c \log T$

Persistency of excitation

$$\Sigma_{p,t} \triangleq \mathbf{E} X_t X_t^\top$$

$$\frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top \geq c\Sigma_{p,\tau}, \quad \text{for } T \geq T_{\text{burn}}(\tau, \delta) \qquad p = c' \log T$$

$$T_{\text{burn}} \simeq c\tau \left(\log 1/\delta + (p(d_Y + d_U))\cdots\right)$$

$$= c''\tau \left(\log 1/\delta + (\log T(d_Y + d_U))\cdots\right)$$

$$= c'''\log T \left(\log 1/\delta + (\log T(d_Y + d_U))\cdots\right)$$

$\tau = p$

$$X_\tau = \begin{bmatrix} Y_{\tau-1:\tau-p}^\top & U_{\tau-1:\tau-p}^\top \end{bmatrix}^\top$$

Invertible $\Sigma_{p,\tau}$ (details in paper)

Comments

Guarantees valid even if $\tau = \max\{p, q\} < \tau_{\text{mix}}$

- (Strict) Stability or mixing (of $A^\star$) is not required

However we need minimum phase

- Stability of $A^\star - K^\star C^\star$ (linked to observability, not mixing)

Improper learning

- Learn Markov Parameters $C^\star A_{\text{cl}}^i B^\star$ instead of original $A^\star, B^\star, C^\star, K^\star$
- Recover original using realization theory

Moving forward

Improve universal constants c

Rates are upper bounds. Are they optimal? What about lower bounds?

Statistical Learning for Control
A finite sample perspective

Nonlinear systems

Thank you!

Statistical Learning for Control
A finite sample perspective

A Tutorial on the Non-Asymptotic Theory of System Identification