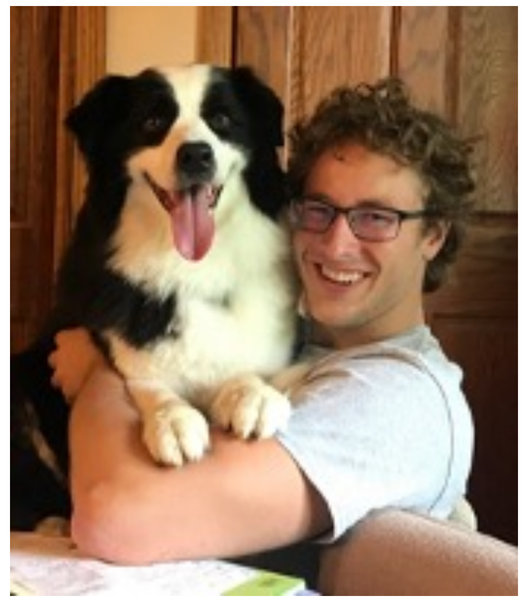
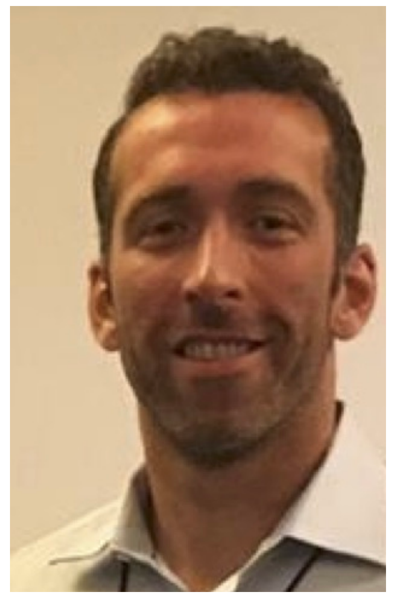
A Tutorial on the Non-Asymptotic Theory of System Identification

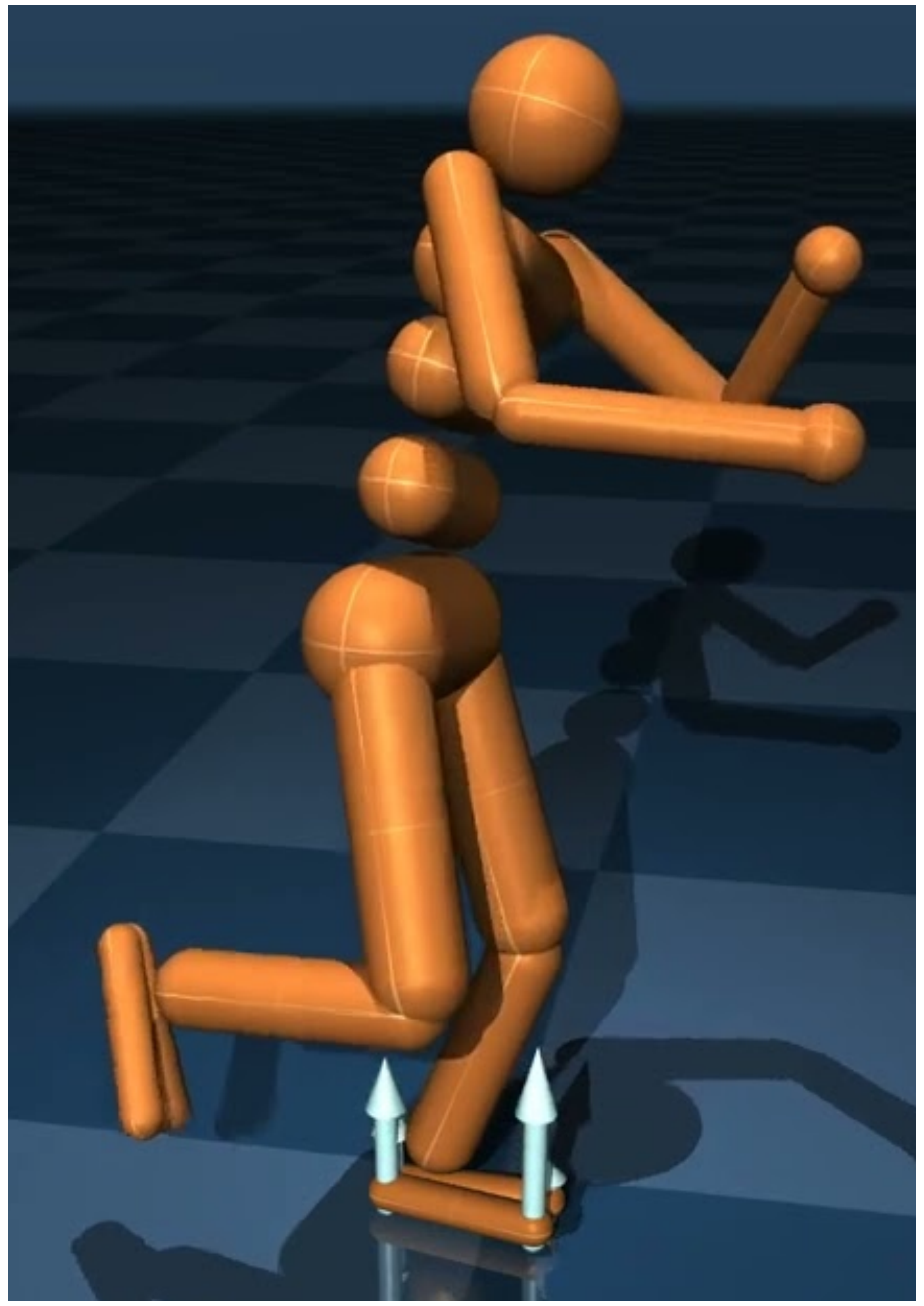
CDC’23 and https://arxiv.org/abs/2309.03873

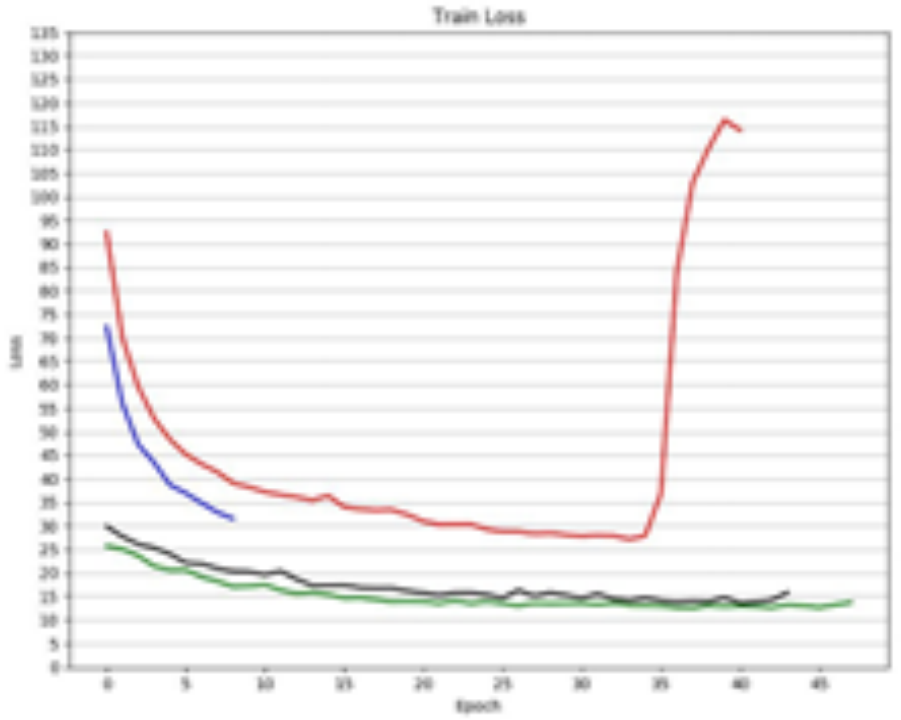
     

Ingvar Ziemann (Penn), Anastasios Tsiamis (ETH), Bruce Lee (Penn), Yassir Jedra (MIT), Nikolai Matni (Penn), George J. Pappas (Penn)

A snapshot of what lead us here

Ambition:





Reality:

Not just in sim:

A snapshot of what lead us here

Journals & Magazines > IEEE Control Systems Magazine > Volume: 43 Issue: 5

Data-Driven Control: Part One of Two

appears in part 2/2

Statistical Learning Theory for Control

A FINITE SAMPLE PERSPECTIVE

Anastasios Tsiamis*, Ingvar Ziemann*, Nikolai Matni, and George J. Pappas

Unknown System

$$X_{t+1} = A^\star X_t + B^\star U_t + W_t, \quad U_t = KX_t$$

Find K such that

$$\sum_{t=1}^{T} \mathbf{E}\left\{X_t^\top Q X_t + U_t^\top R U_t\right\} = !\min$$

today!

experiments $\xrightarrow[(X, U)_{1:T}]{\text{data}}$ system identification $\xrightarrow[(\widehat{A}, \widehat{B})]{\text{model estimates}}$ control

$U_t \sim N(0, I)$

$U_t = \widehat{K} X_t$

https://arxiv.org/abs/2209.05423

Overview

Introduction and Roadmap (20 minutes)

Concentration Inequalities (30 minutes)

- Hanson-Wright Inequality and Self-Normalized Martingales

Lower Tail of the Empirical Covariance (20 min)

- Concentration =/= Persistence of Excitation

System Identification (30 min)

- ARX-identification

An Alternate Approach: The Offset Basic Inequality (20 min)

- Extension to nonlinear problems

Statistical Setup

Consider a time series model

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

benign noise

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 $\mathbb{R}^{d_Y \times d_X}$

Example ARX(p,q):

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

In other words…

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

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

$$V_t = W_t .$$

Least Squares Estimation (LSE)

Consider a time-series model:

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

Least Squares Estimator:

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

$\Rightarrow$

$$\widehat{\theta} \triangleq \left(\sum_{t=1}^{T} Y_t X_t^\top \right) \left(\sum_{t=1}^{T} X_t X_t^\top \right)^{-1}$$

Interested in:

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

Today: Modern perspective on LSE

Draw on tools from:

Machine Learning Theory

High-Dimensional Statistics

High-Dimensional Probability

Problem

Fix:

accuracy $\epsilon > 0$

failure probability $\delta \in (0,1)$

a norm $\| \cdot \|$

and a 'reasonable' estimator $\widehat{\theta}$

Establish finite sample guarantees:

$\| \widehat{\theta} - \theta^\star \| \leq \epsilon$ wpal. $1 - \delta$

Typically we can prove:

$$\epsilon \propto (\text{noise scale}) \times \sqrt{\frac{\text{dimension} + \log(1/\delta)}{\text{sample size}}}$$

As long as:

$$\text{sample size} \gtrsim \text{dimension} \ + \ \log(1/\delta)$$

Persistence of Excitation

The Path Ahead

Covariance Concentration

At LLN Scale ~ 1

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

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

Random Matrix!

Random Walk at CLT Scale $\sim \sqrt{T}$

Todo:

CLT-analogue: Self-Normalized Martingale Ineq.

LLN-analogue: Covariance (anti-)Concentration

Finite sample aspects and comparison with asymptotic

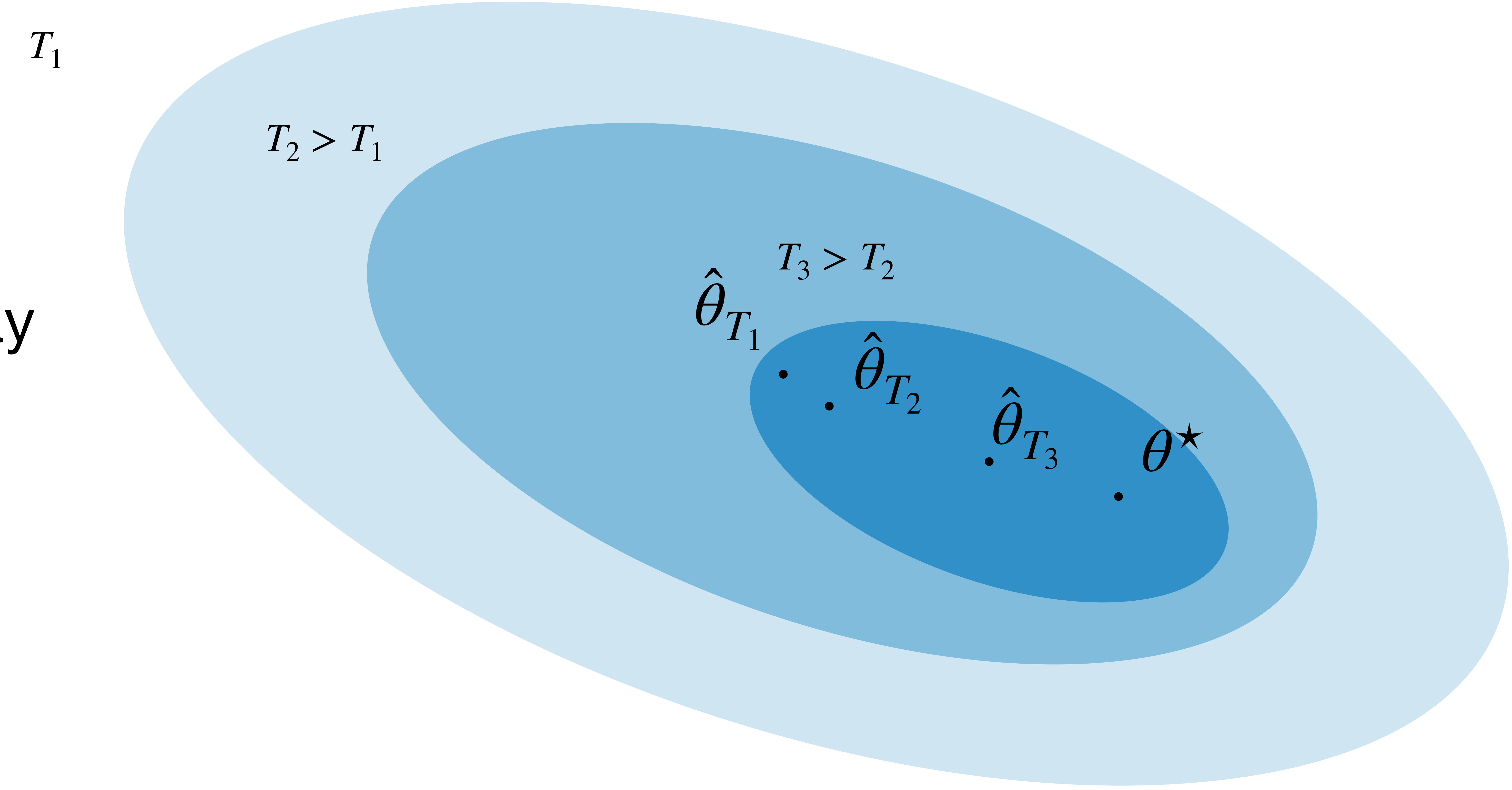
Finite Sample Guarantees

What is the error $\hat{\theta}_T - \theta^\star$?

Finite Sample Guarantees

$\theta^\star \in \mathscr{B}(T)$

Q1: Error decay rate?

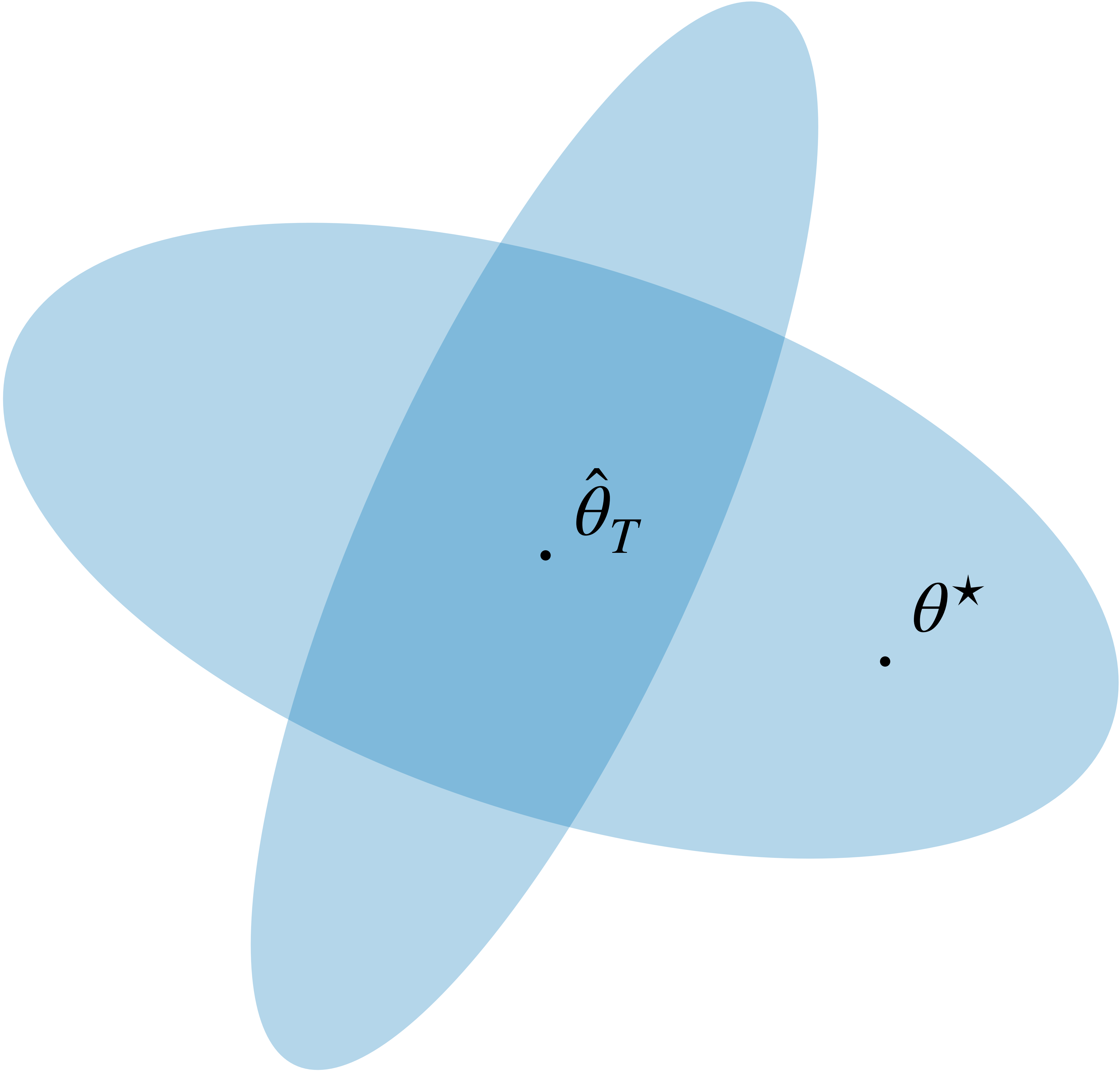


Finite Sample Guarantees

$$\theta^\star \in \mathscr{B}(T)$$

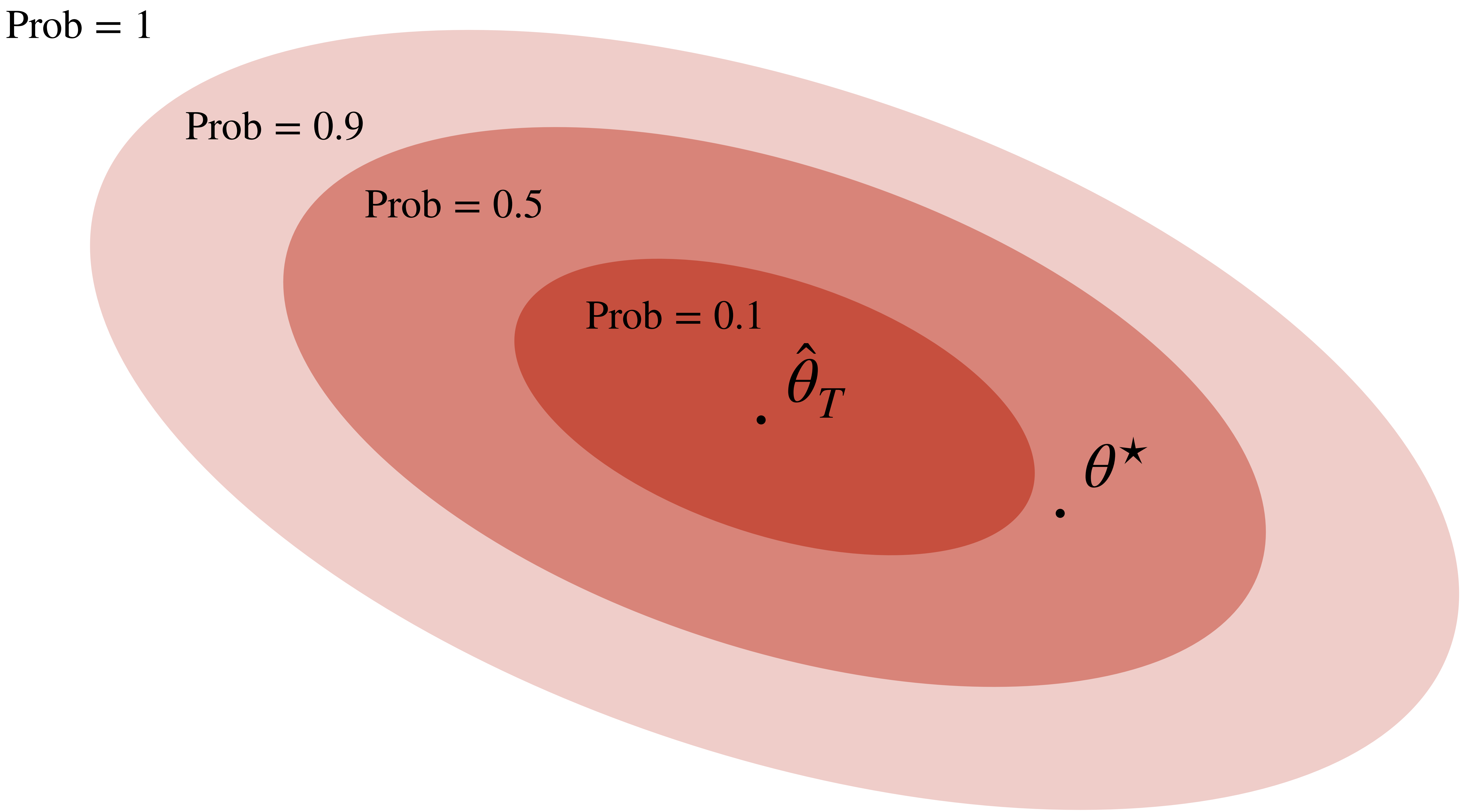
Q1: Error decay rate?

Q2: Shape of uncertainty?



Finite Sample Guarantees

$$\mathrm{Prob}(\theta^\star \in \mathcal{B}(T, \delta)) \geq 1 - \delta$$



Q1: Error decay rate?

Q2: Shape of uncertainty?

Q3: Confidence?

Finite Sample Guarantees

$$\mathrm{Prob}(\theta^\star \in c \cdot \mathcal{B}(T, \delta)) \geq 1 - \delta$$

System specific

& Universal constants

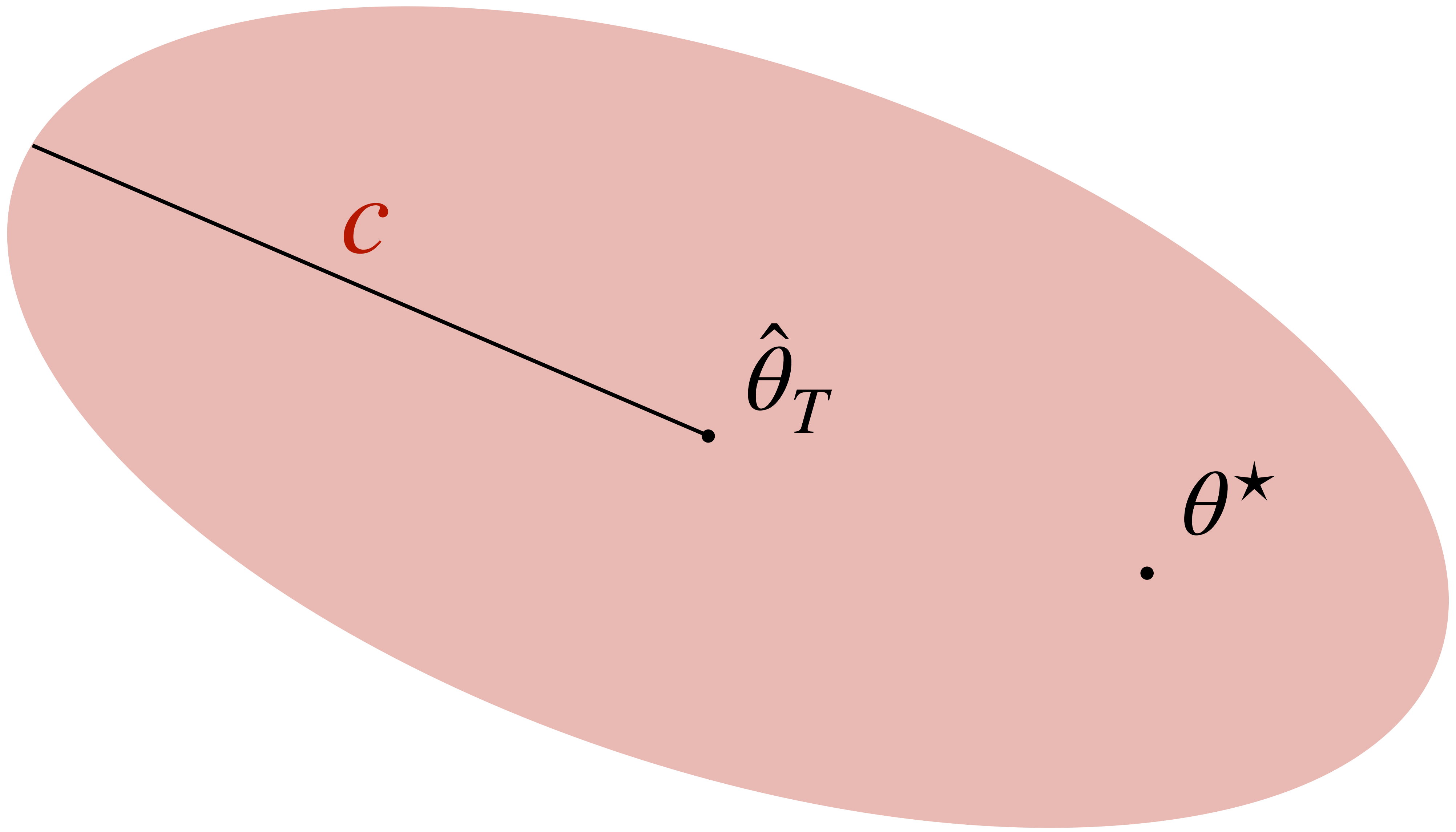
Relative scaling

Q1: Error decay rate?

Q2: Shape of uncertainty?

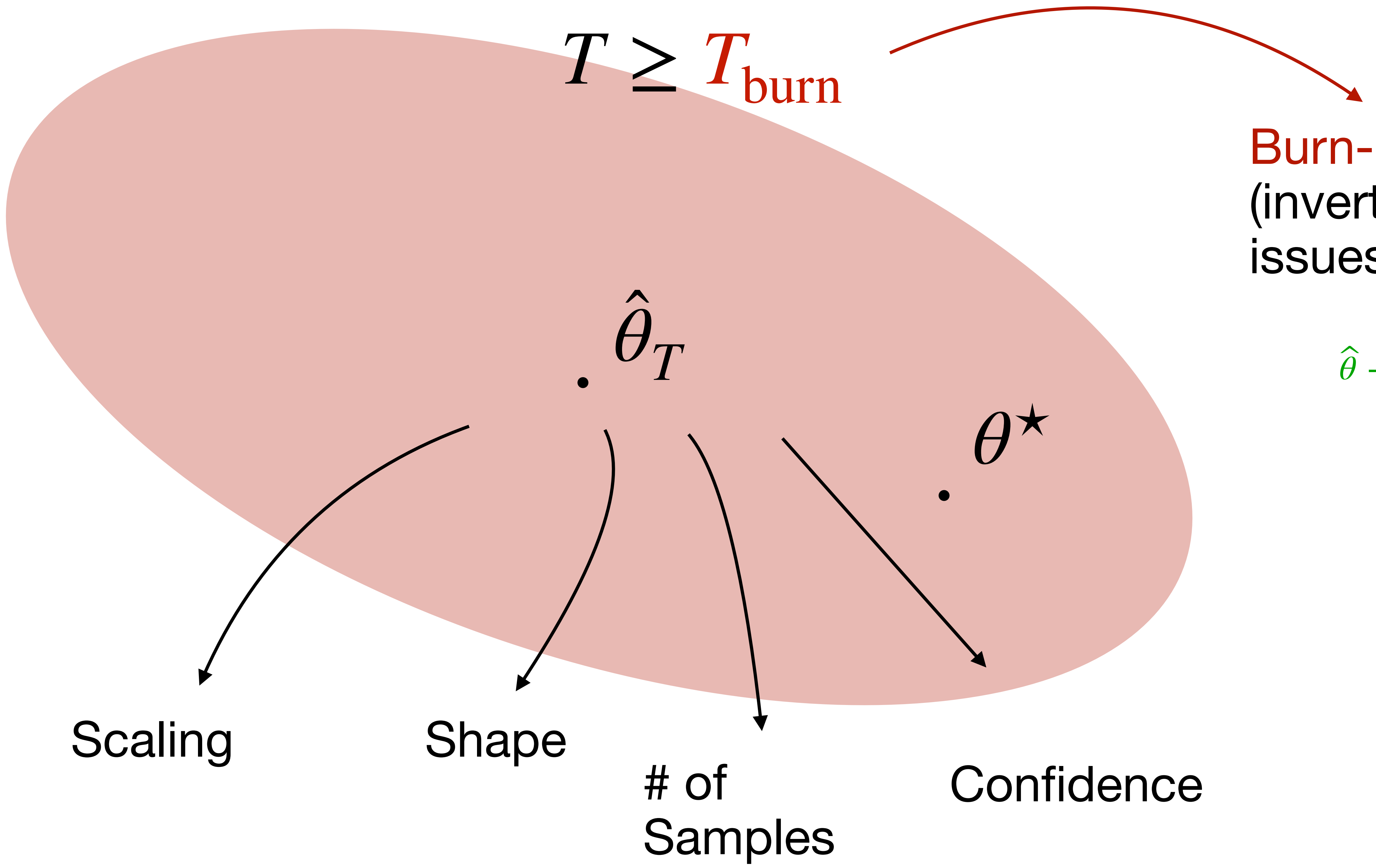
Q3: Confidence?

Q4: Absolute Scaling?



Finite Sample Guarantees

$$\mathrm{Prob}(\theta^\star \in c\mathscr{B}(T,\delta)) \geq 1-\delta$$



Burn-in Time
(invertibility issues)

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

$\hat{\theta}_{T_{\text{burn}}}$

Asymptotics

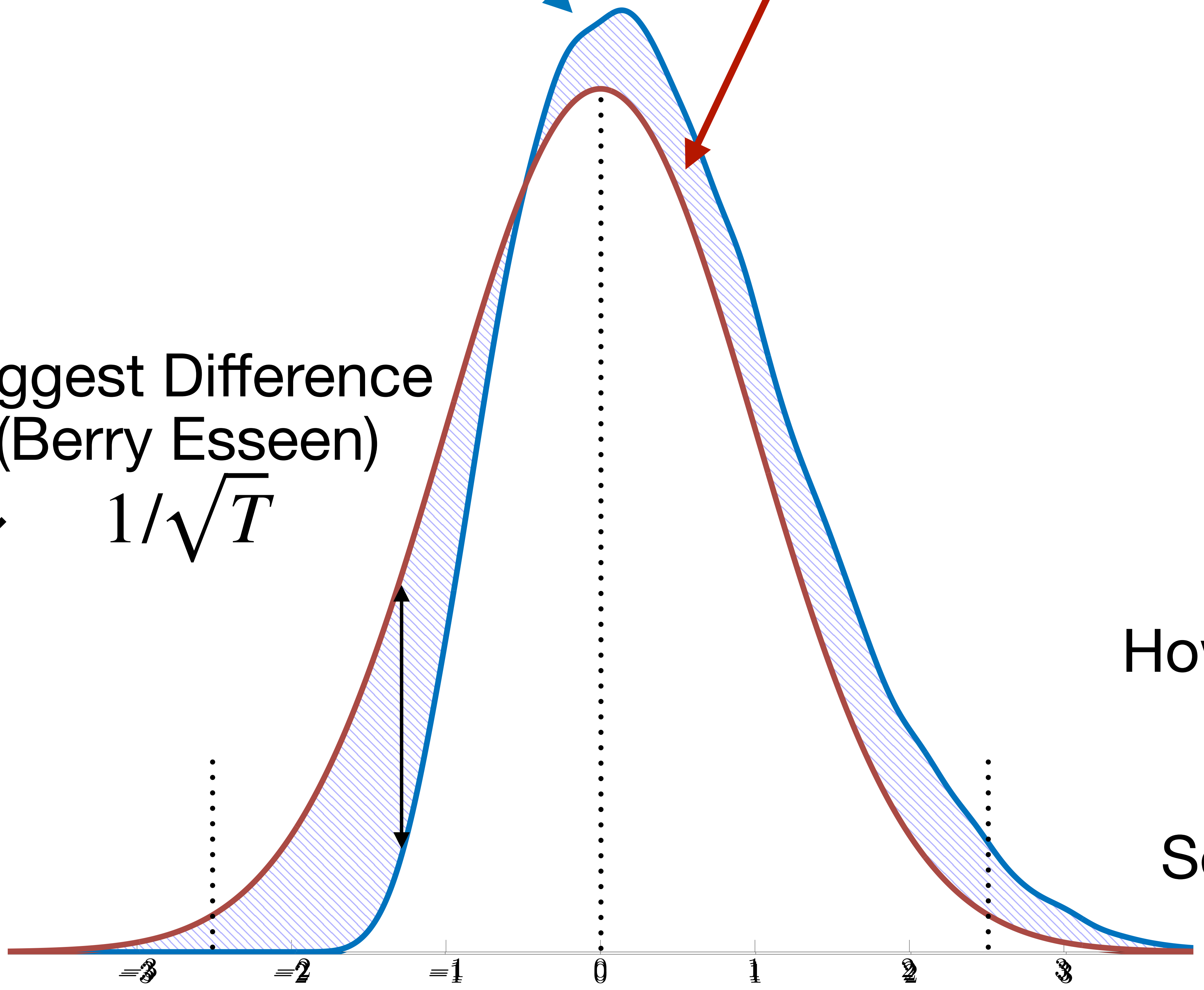
$$\sqrt{T}(\hat{\theta}_T - \theta^\star) \to^d \mathcal{N}(0, \Sigma_\theta)$$

Ljung 1999

Deviation for
$T < \infty$?

Biggest Difference
(Berry Esseen)
$1/\sqrt{T}$

Burn-in
$1/\delta^2$



Using finite samples
$\exp(-T)$

Burn-in
$\log 1/\delta$

However negative events: **tail events**

Seem to decay much faster

Why finite sample?

Complementary tools

	Finite Sample	Asymptotics
Rate	$1/\sqrt{T}$	$1/\sqrt{T}, T \to \infty$
Shape	$\frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top$	$\lim \frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top$
Confidence	$T \geq \log 1/\delta$	$T \geq 1/\delta^2$
Scale	**Conservative universal const.**	**Optimal**
Transient	**Burn-in times**	✗

With Berry Esseen

Renewed Attention

$$\|\hat{\theta}_T - \theta^\star\| = O(1/\sqrt{T})$$

$$\Downarrow$$

$$\|\hat{\theta}_T - \theta^\star\| = \frac{C_{\text{sys}}}{\sqrt{T}}, \quad T \geq T_{\text{burn}}$$

System theoretic properties

What makes learning difficult or hard: $C_{\text{sys}}, T_{\text{burn}}$

Thank you!