Converging Fast and Slow: Different Avatars of EM

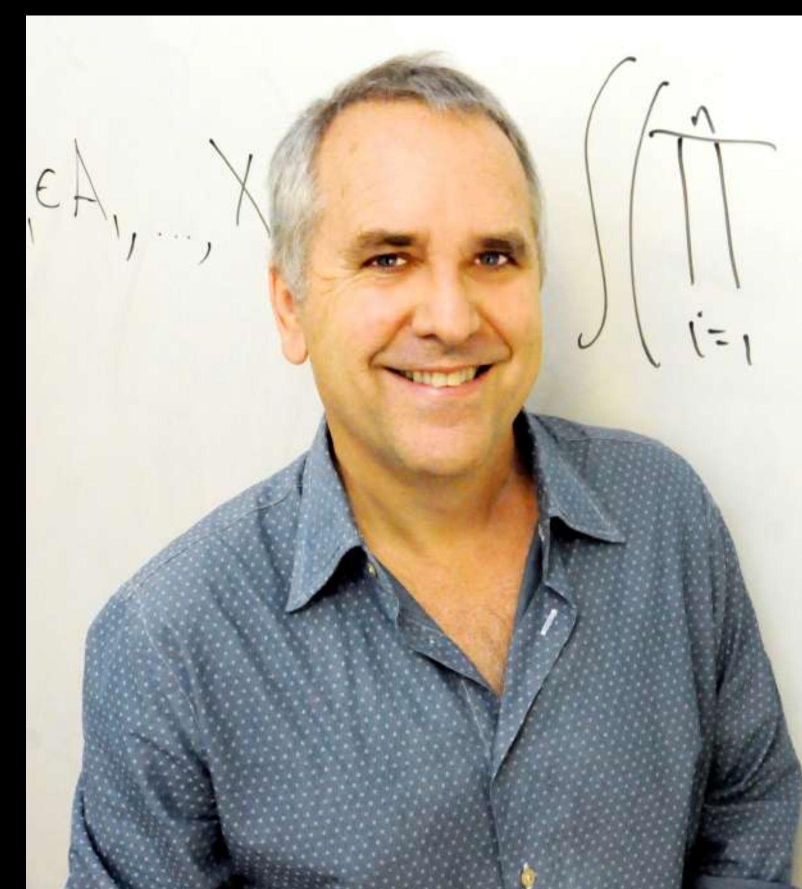
Raaz Dwivedi
Department of EECS, UC Berkeley

Nhat Ho*

Koulik Khamaru*

Martin Wainwright

Bin Yu

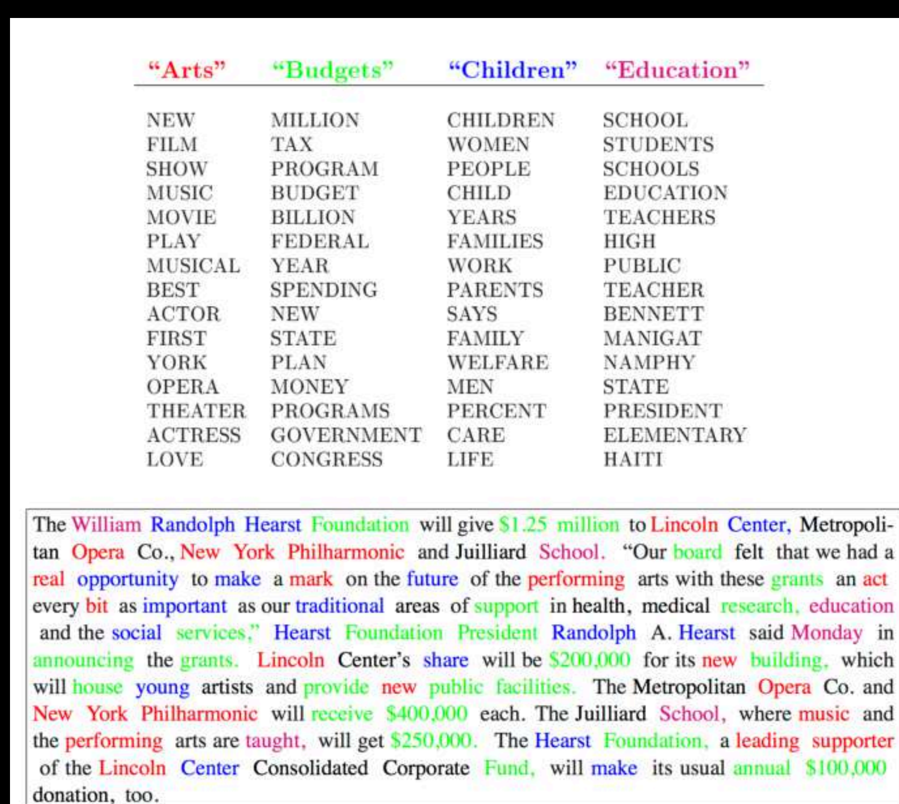


Michael Jordan

Mixture models: Usefulness

- Heterogenous sub-populations in various datasets
  - Topic modeling, Financial returns
  - Image annotation, classification, segmentation

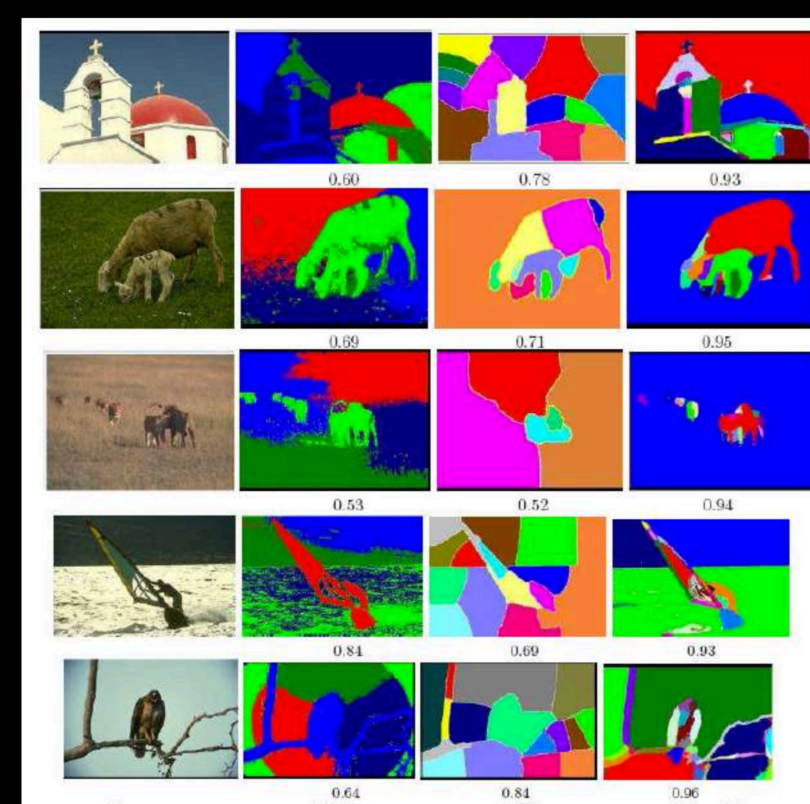
Source: Blei et al. 2003



"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Source: Rotem et al. 2007



Mixture models: Formulation

- Distribution of observed variable X in a latent variable model with labels Z

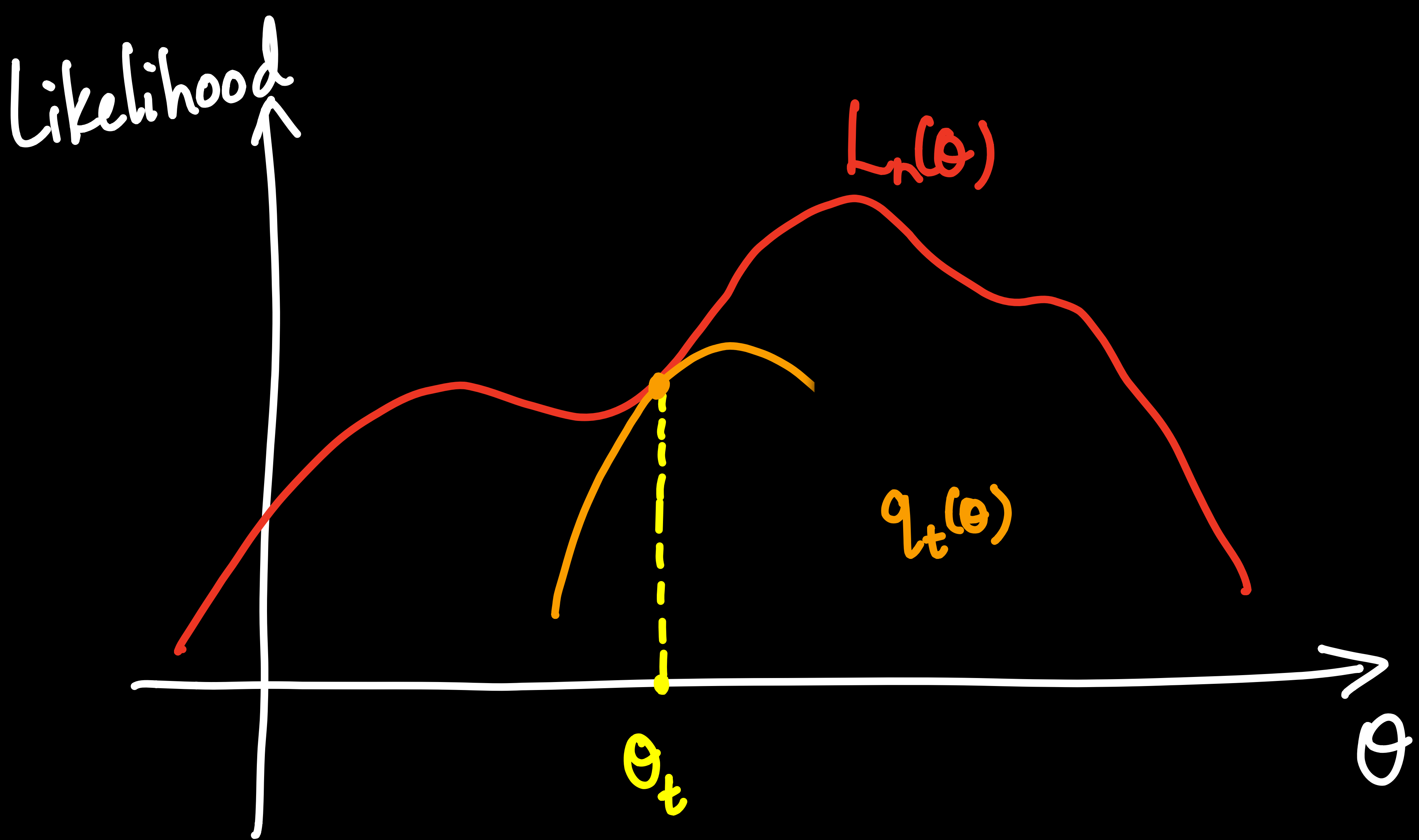
$$Z \sim \text{multinomial}(w_1, \ldots, w_K)$$
$$[X \mid Z = k] \sim \mathscr{P}_k$$
$$X \sim \sum_{k=1}^{K} w_k \mathscr{P}_k$$

- $\mathscr{P}_k = \mathscr{N}(\mu_k, \Sigma_k)$ results in Gaussian mixture model, arguably the most popular in practice

- Given $X_1, \ldots, X_n$, how do we estimate the parameters? Lack of Z makes the problem non-convex

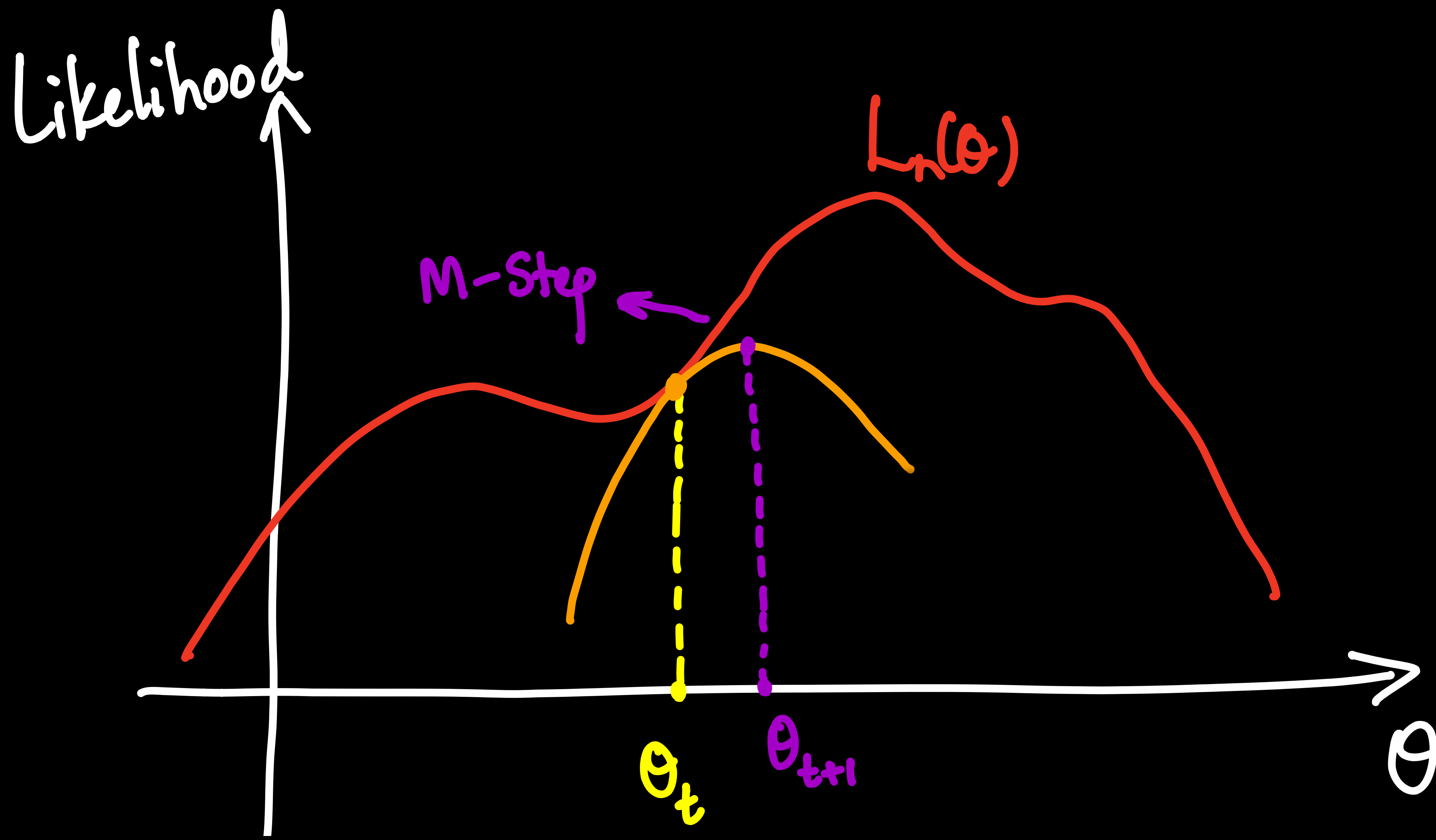
Mixture models: Parameter estimation

- Method of choice: Expectation-Maximization
  (Dempster-Laird-Rubin, Sundberg, Martin-Löf, Jeff Wu 1970-80)



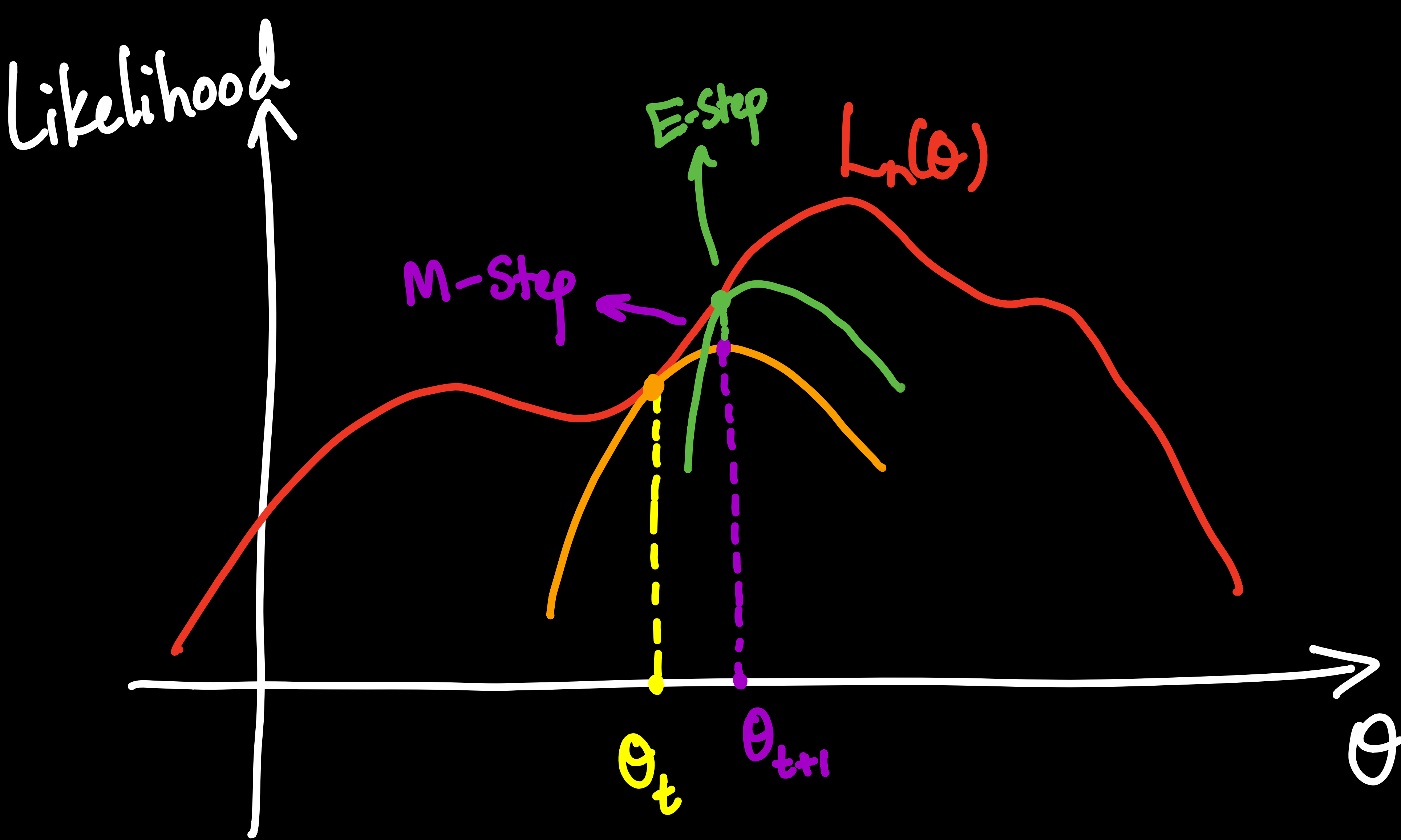
Mixture models: Parameter estimation

- Method of choice: Expectation-Maximization (Dempster-Laird-Rubin, Sundberg, Martin-Löf, Jeff Wu 1970-80)



Mixture models: Parameter estimation

- Method of choice: Expectation-Maximization
  (Dempster-Laird-Rubin, Sundberg, Martin-Löf, Jeff Wu 1970-80)



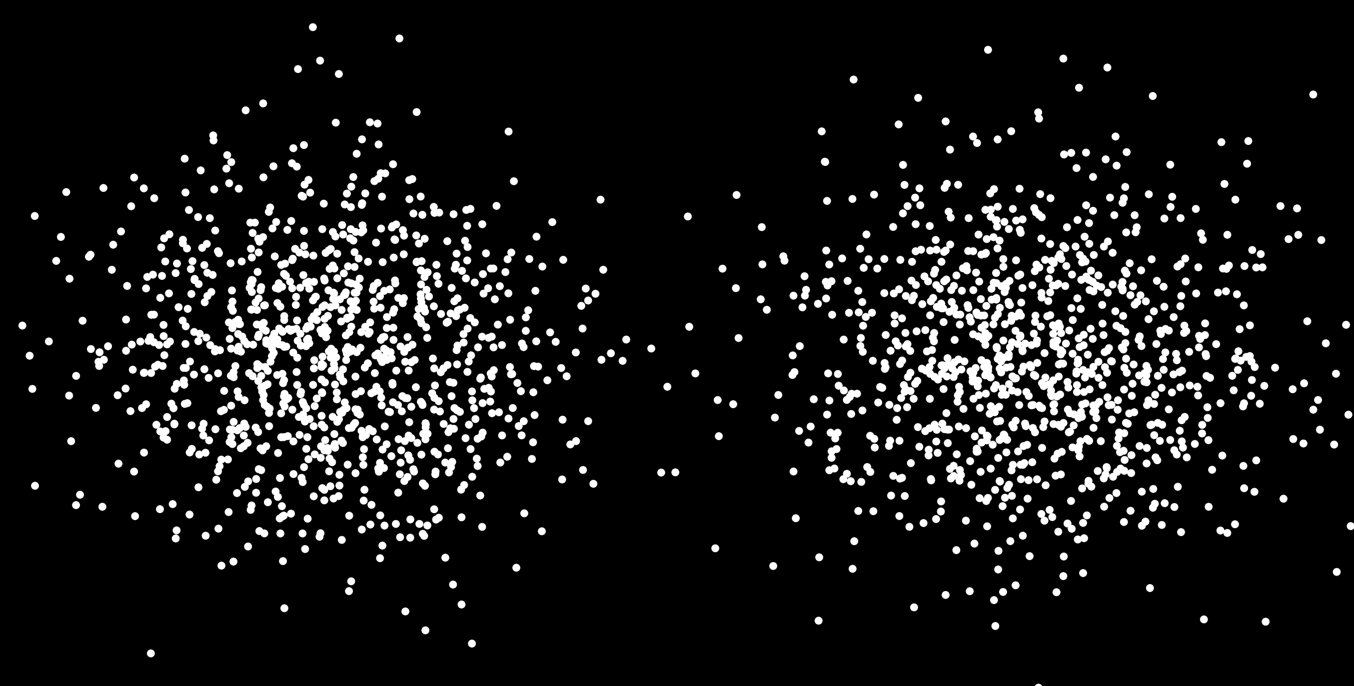
Theoretical Guarantees for EM: Asymptotic and non-asymptotic analysis

- Asymptotic results: Boyles 1983, Neal and Hinton 1995, Ma, Xu and Jordan 1996, 2000, …
- Several recent works on the non-asymptotic behavior of EM in $\mathbb{R}^d$ with n samples

Theoretical Guarantees for EM: Well-specified 2-Gaussian Mixtures

True Model: $$\frac{1}{2}\mathcal{N}(-\theta^\star, \mathbb{I}_d) + \frac{1}{2}\mathcal{N}(\theta^\star, \mathbb{I}_d)$$

Fitted model: $$\frac{1}{2}\mathcal{N}(-\theta, \mathbb{I}_d) + \frac{1}{2}\mathcal{N}(\theta, \mathbb{I}_d)$$



Theoretical Guarantees for EM: Well-specified 2-Gaussian Mixtures

True Model: $\frac{1}{2}\mathcal{N}(-\theta^\star, \mathbb{I}_d) + \frac{1}{2}\mathcal{N}(\theta^\star, \mathbb{I}_d)$

Fitted model: $\frac{1}{2}\mathcal{N}(-\theta, \mathbb{I}_d) + \frac{1}{2}\mathcal{N}(\theta, \mathbb{I}_d)$

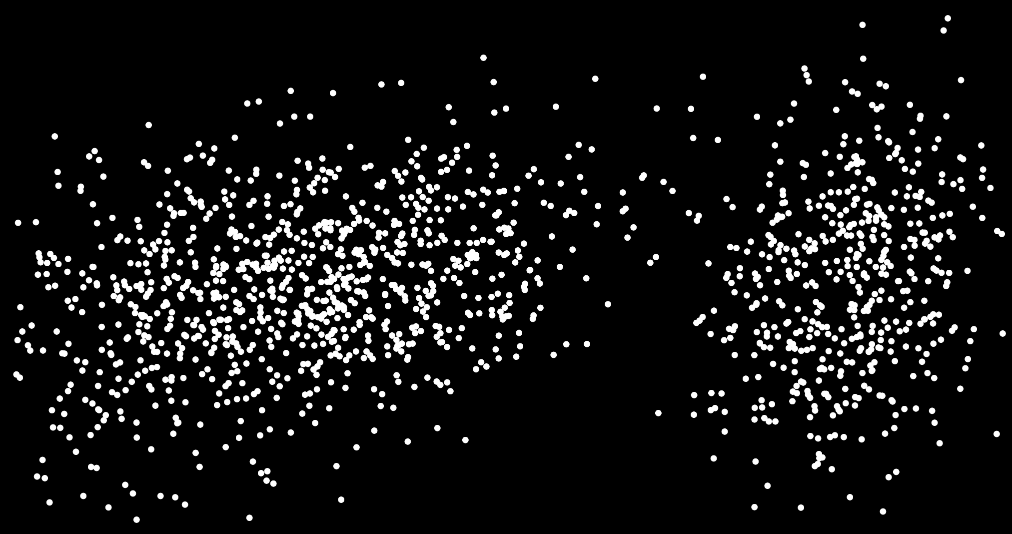
(Balakrishnan, Wainwright, Yu '17)
EM with good initialization + Strong Signal $\|\theta^\star\| > C$:

$$\|\theta_n^t - \theta^\star\|_2 \lesssim \sqrt{\frac{d}{n}} \quad \textit{for} \quad t \gtrsim \log\left(\frac{n}{d}\right) \textit{ and } \; n \gtrsim d$$

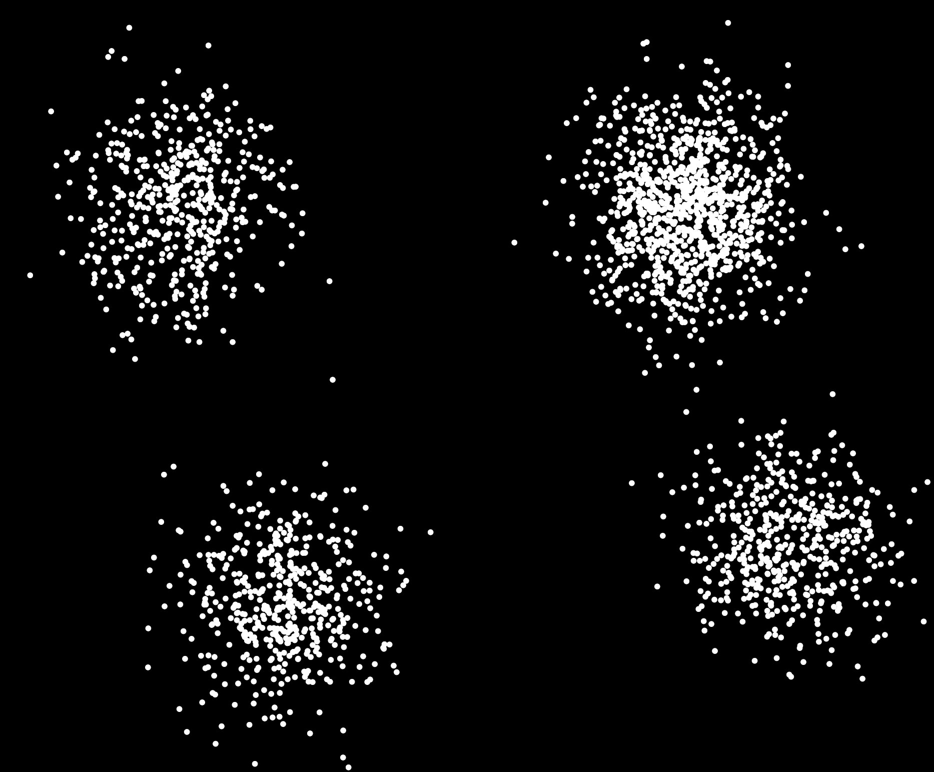
Well-initialized EM on well-specified well-separated mixtures:

$$\sqrt{\frac{d}{n}} \text{ error in } \log \frac{n}{d} \text{ steps}$$

Cai, Ma and Zhang, 2019
General, well-separated 2-mixtures
Fitted with 2-mixtures



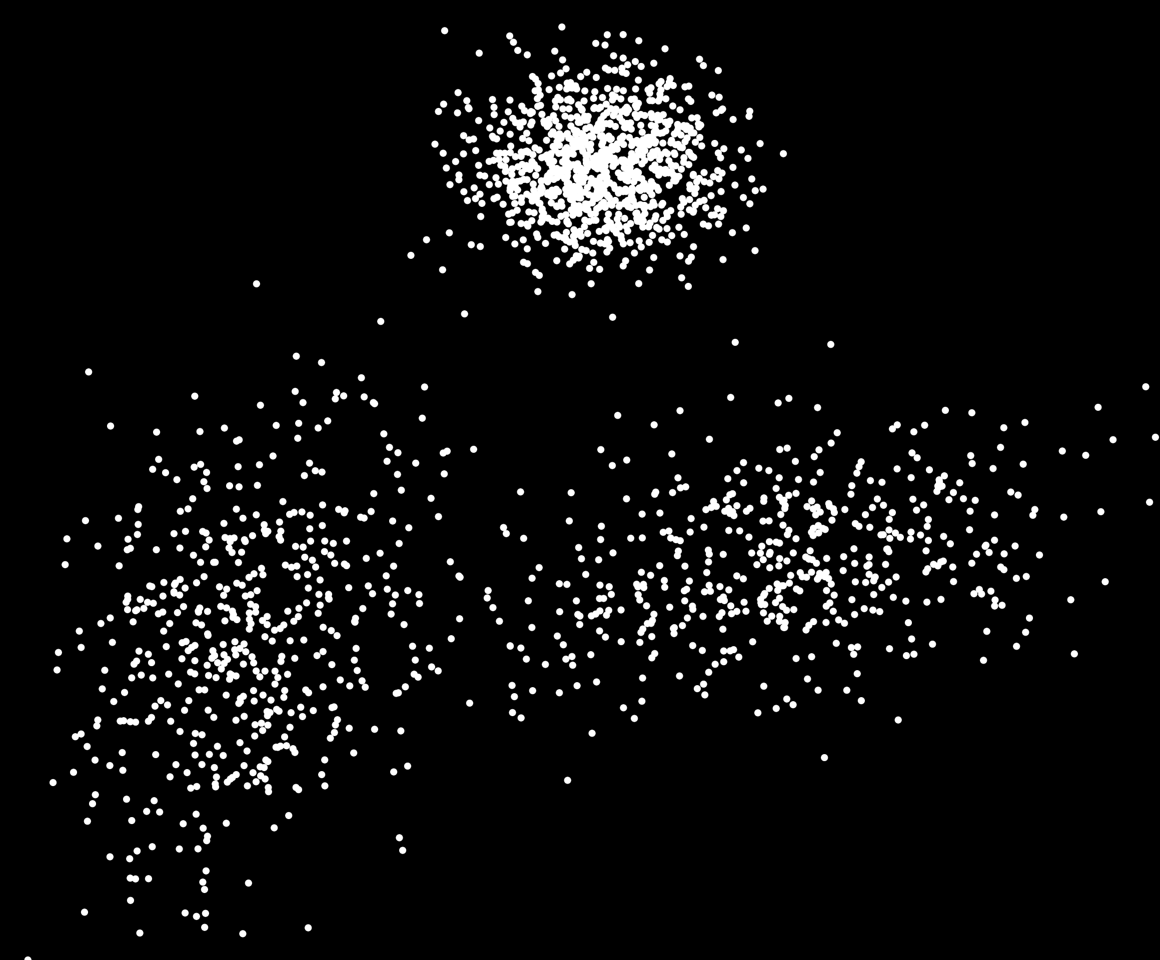
Yan, Yin and Sarkar, 2017
Spherical, well-separated k-mixtures
Fitted with k spherical mixtures



Other works: Wang+ 2015, Daskalakis+ 2017, Hao+ 2018, …

“But what happens when the components are too close to each other?”

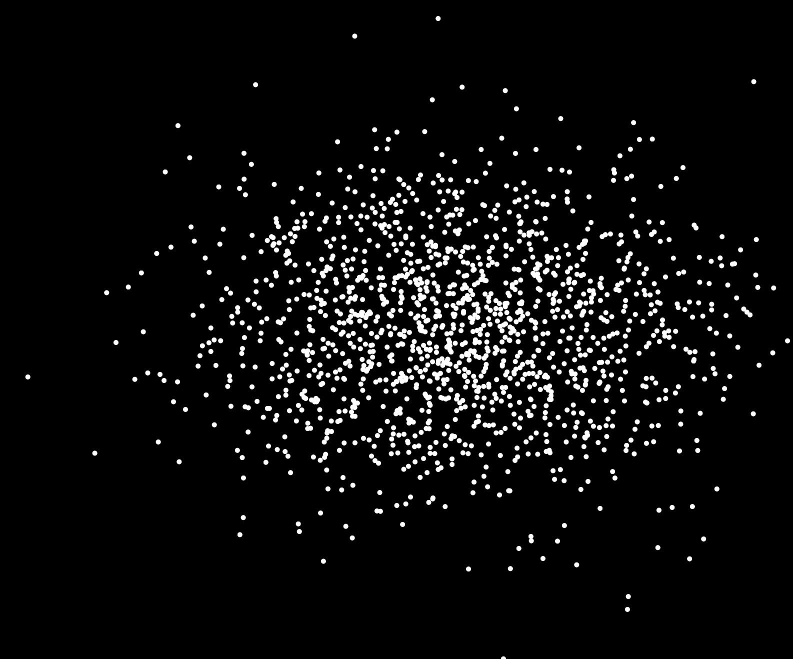
“Or, when the number of components is over-specified in the fitted model?”



EM slows down… some old works? but can we quantify it?

We consider the simplest over-specified case: True model has **one** component and we fit **two** components

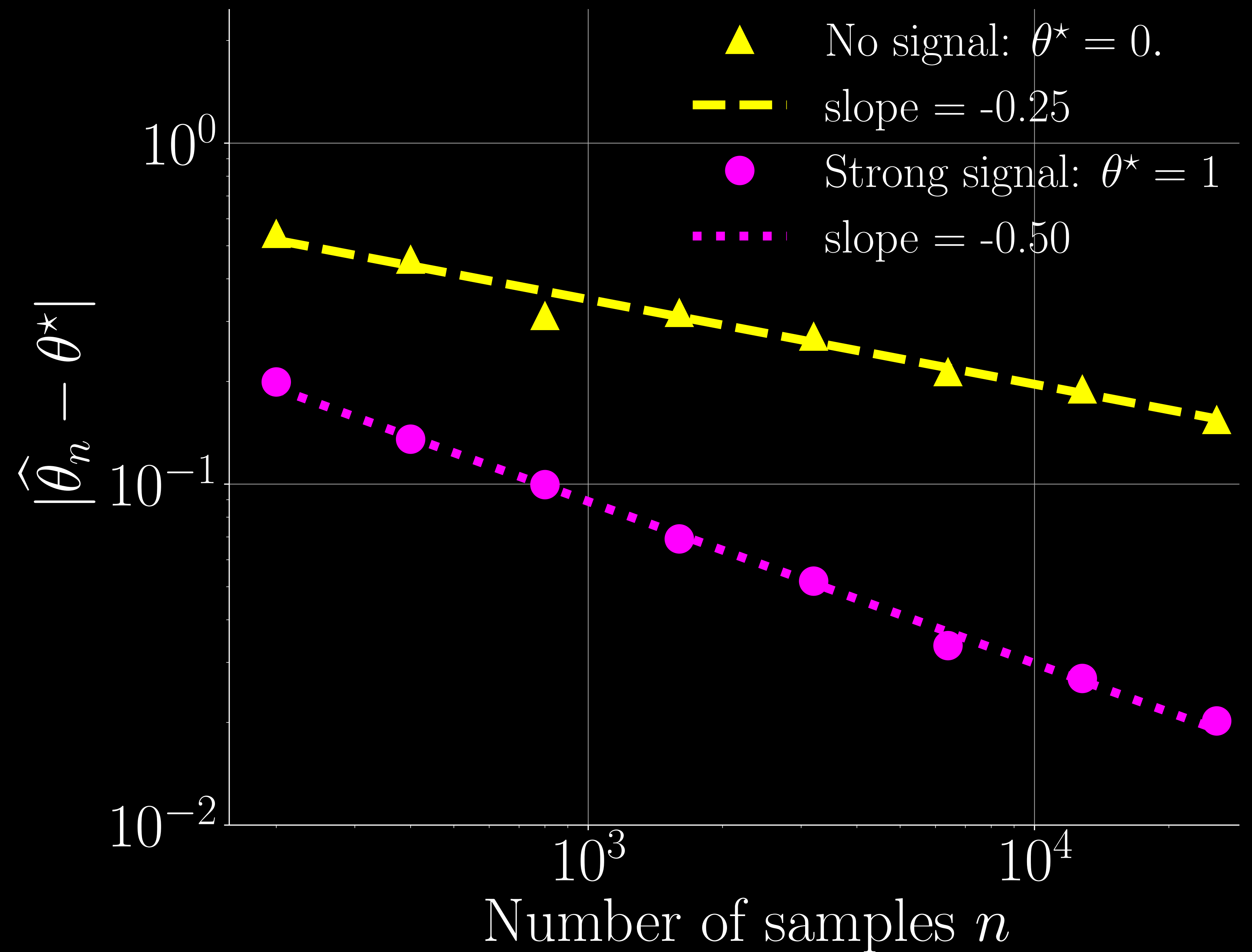
True Model: $\mathcal{N}(0, \mathbb{I}_d)$



$$= \quad \frac{1}{2}\mathcal{N}(\theta^\star, \mathbb{I}_d) + \frac{1}{2}\mathcal{N}(-\theta^\star, \mathbb{I}_d) \quad \textbf{with} \quad \theta^\star = 0$$

Fitted model: $\frac{1}{2}\mathcal{N}(\theta, \mathbb{I}_d) + \frac{1}{2}\mathcal{N}(-\theta, \mathbb{I}_d)$

Converging fast and slow: Statistical rates for EM estimates vs SNR



Our main result: Convergence of sample EM with weak signal

In the case of no signal $\theta^\star = 0$, for arbitrary initialization, the sample EM iterates satisfy

$$\|\theta_n^t - \theta^\star\|_2 \lesssim \left(\frac{d}{n}\right)^{1/4} \quad \textit{for} \quad t \gtrsim \left(\frac{n}{d}\right)^{1/2} \quad \textit{and} \quad n \gtrsim d,$$

Converging fast and slow

In the case of no signal $\theta^\star = 0$, for arbitrary initialization, the sample EM iterates satisfy

$$\|\theta_n^t - \theta^\star\|_2 \lesssim \left(\frac{d}{n}\right)^{1/4} \quad \textit{for} \quad t \gtrsim \left(\frac{n}{d}\right)^{1/2} \quad \textit{and} \quad n \gtrsim d,$$

Balakrishnan+ 2017

For strong signal $\|\theta^\star\| > C$, sample EM iterates satisfy

$$\|\theta_n^t - \theta^\star\|_2 \lesssim \left(\frac{d}{n}\right)^{1/2} \quad \textit{for} \quad t \gtrsim \log\left(\frac{n}{d}\right) \quad \textit{and} \quad n \gtrsim d$$

Converging fast and slow

In the case of no signal $\theta^\star = 0$, for arbitrary initialization, the sample EM iterates satisfy

$$\|\theta_n^t - \theta^\star\|_2 \lesssim \left(\frac{d}{n}\right)^{1/4} \quad \textit{for} \quad t \gtrsim \left(\frac{n}{d}\right)^{1/2} \quad \textit{and} \quad n \gtrsim d,$$

statistical slow-down

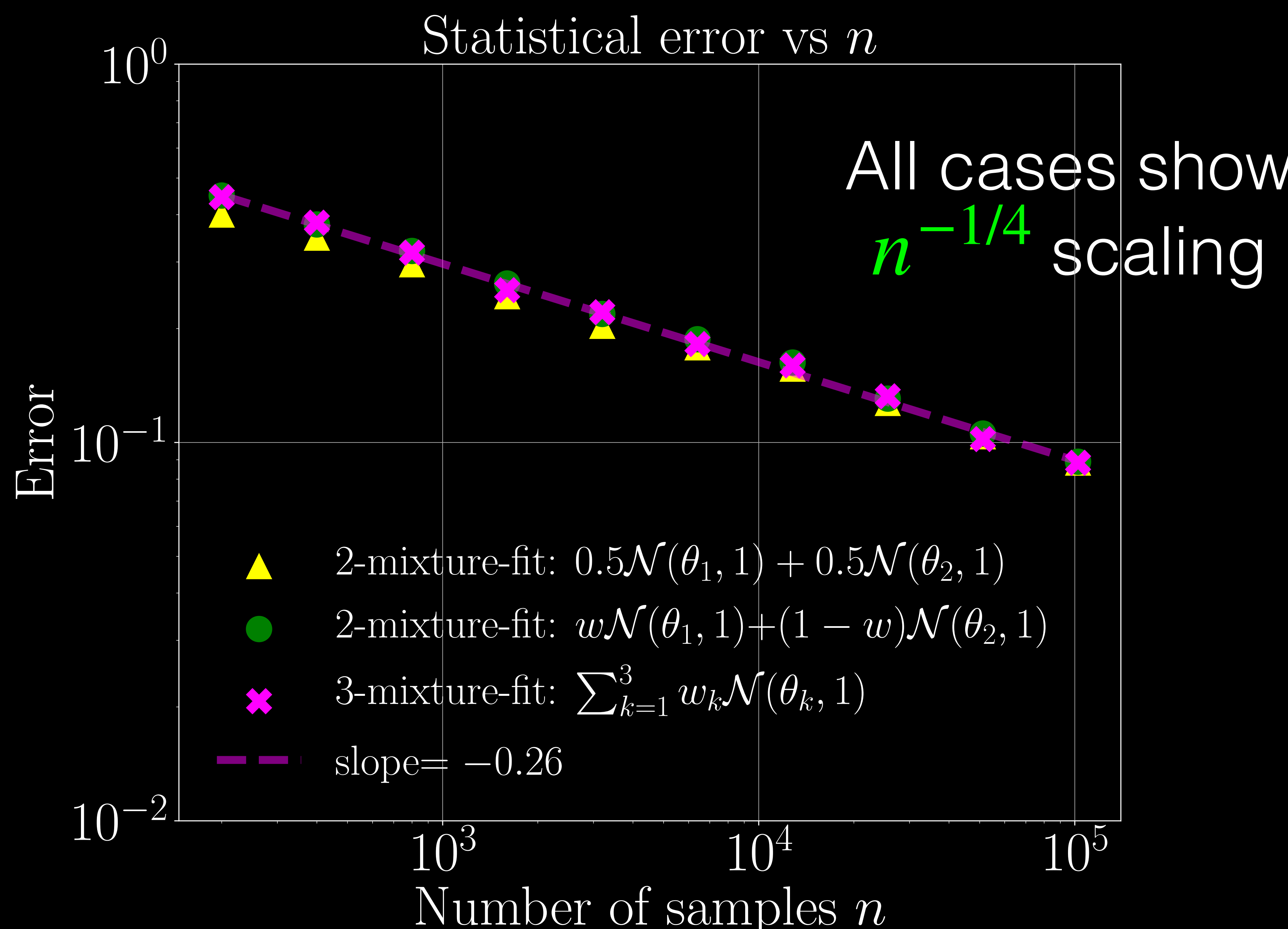
computational slow-down

Balakrishnan+ 2017

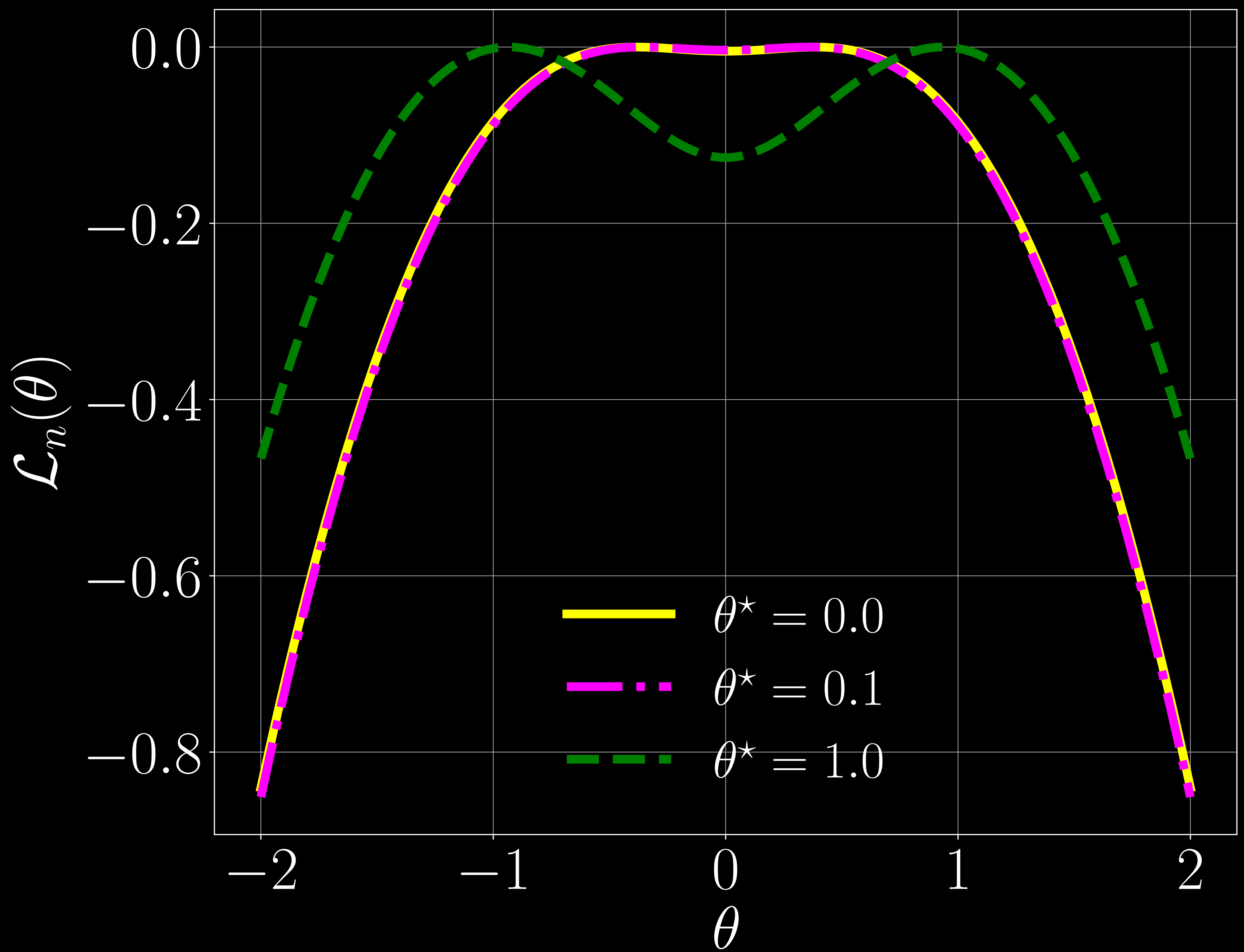
For strong signal $\|\theta^\star\| > C$, sample EM iterates satisfy

$$\|\theta_n^t - \theta^\star\|_2 \lesssim \left(\frac{d}{n}\right)^{1/2} \quad \textit{for} \quad t \gtrsim \log\left(\frac{n}{d}\right) \quad \textit{and} \quad n \gtrsim d$$

Zero SNR:
Statistical rates for non-special fits



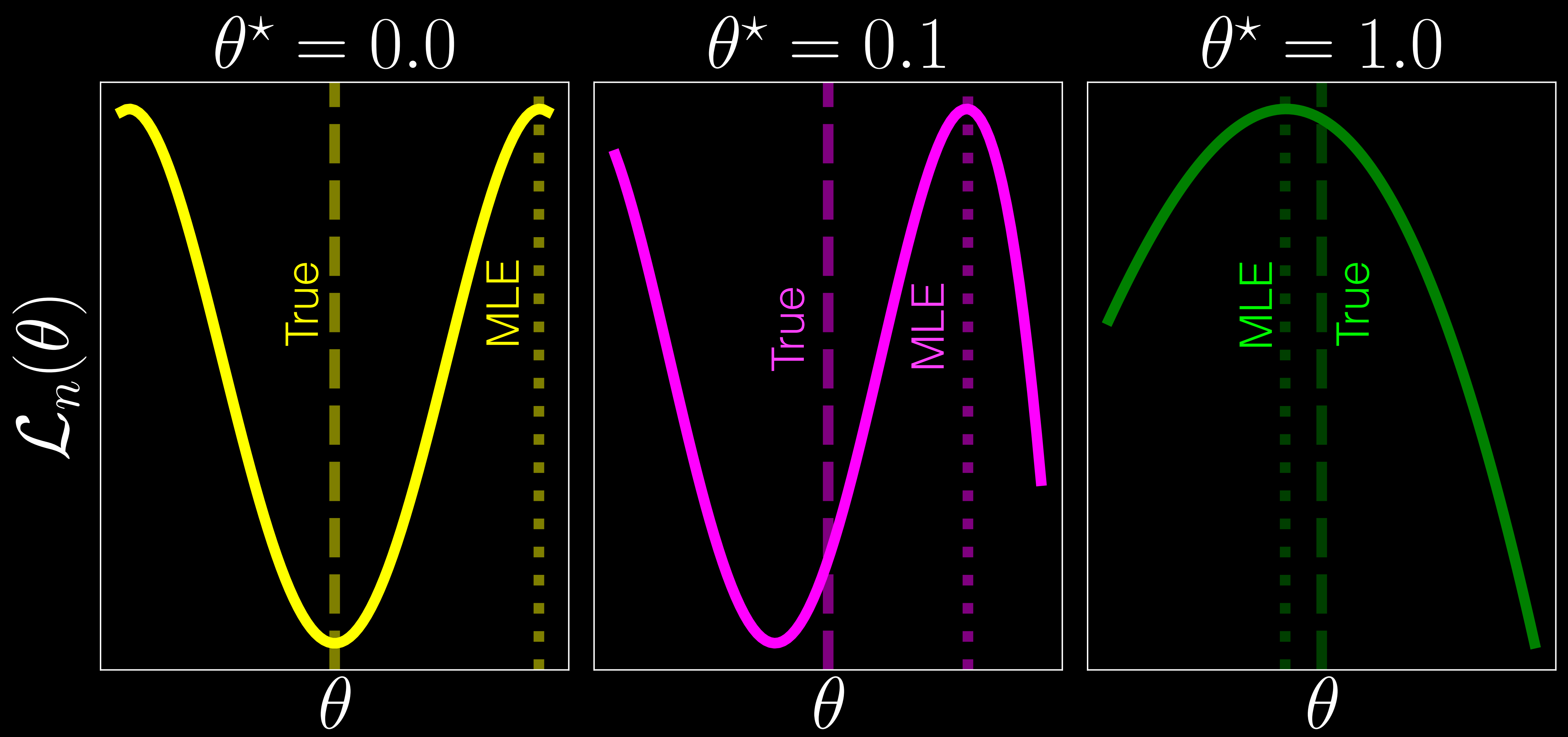
Zero signal
= Degenerate Fisher matrix
= Flatter log-likelihood



Flatness: EM takes more iterations to converge

statistical slow-down

Zero signal
= Degenerate Fisher matrix
= Slow rate for MLE [Chen 1995, Rousseau 2011, Nguyen 2013, Ho+ 2018]



MLE farther from $\theta^\star$: slower statistical rate for EM estimates

Proving the slow rates

Closed form updates for EM

Fitted model: $\frac{1}{2}\mathcal{N}(\theta,1) + \frac{1}{2}\mathcal{N}(-\theta,1)$

Population EM iteration:
$$\begin{aligned}\theta^{t+1} &= \mathbb{E}[X \tanh(X^\top \theta^t)] \\ &=: M(\theta^t)\end{aligned}$$

Sample EM iteration:
$$\begin{aligned}\theta_n^{t+1} &= \frac{1}{n}\sum_{i=1}^{n} X_i \tanh(X_i^\top \theta_n^t) \\ &=: M_n(\theta_n^t)\end{aligned}$$

Can study the updates via the operators M and M_n

Proof strategy: From population to sample analysis

$$\|\theta_n^{t+1} - \theta^\star\| = \|M_n(\theta_n^t) - \theta^\star\|$$
$$\leq \|M(\theta_n^t) - \theta^\star\| + \|M_n(\theta_n^t) - M(\theta_n^t)\|$$

- **Population**-level behavior
- **Deterministic** analysis
- Characterizes the **"algorithmic"** rate of convergence

- **Finite sample** perturbation error
- **Probabilistic** analysis
- Characterizes the **"statistical"** rate of convergence

computational slow-down

Proof strategy: From population to sample analysis

$$\|\theta_n^{t+1} - \theta^\star\| = \|M_n(\theta_n^t) - \theta^\star\|$$
$$\leq \|M(\theta_n^t) - \theta^\star\| + \|M_n(\theta_n^t) - M(\theta_n^t)\|$$

Balakrishnan+ 2017: For strong signal

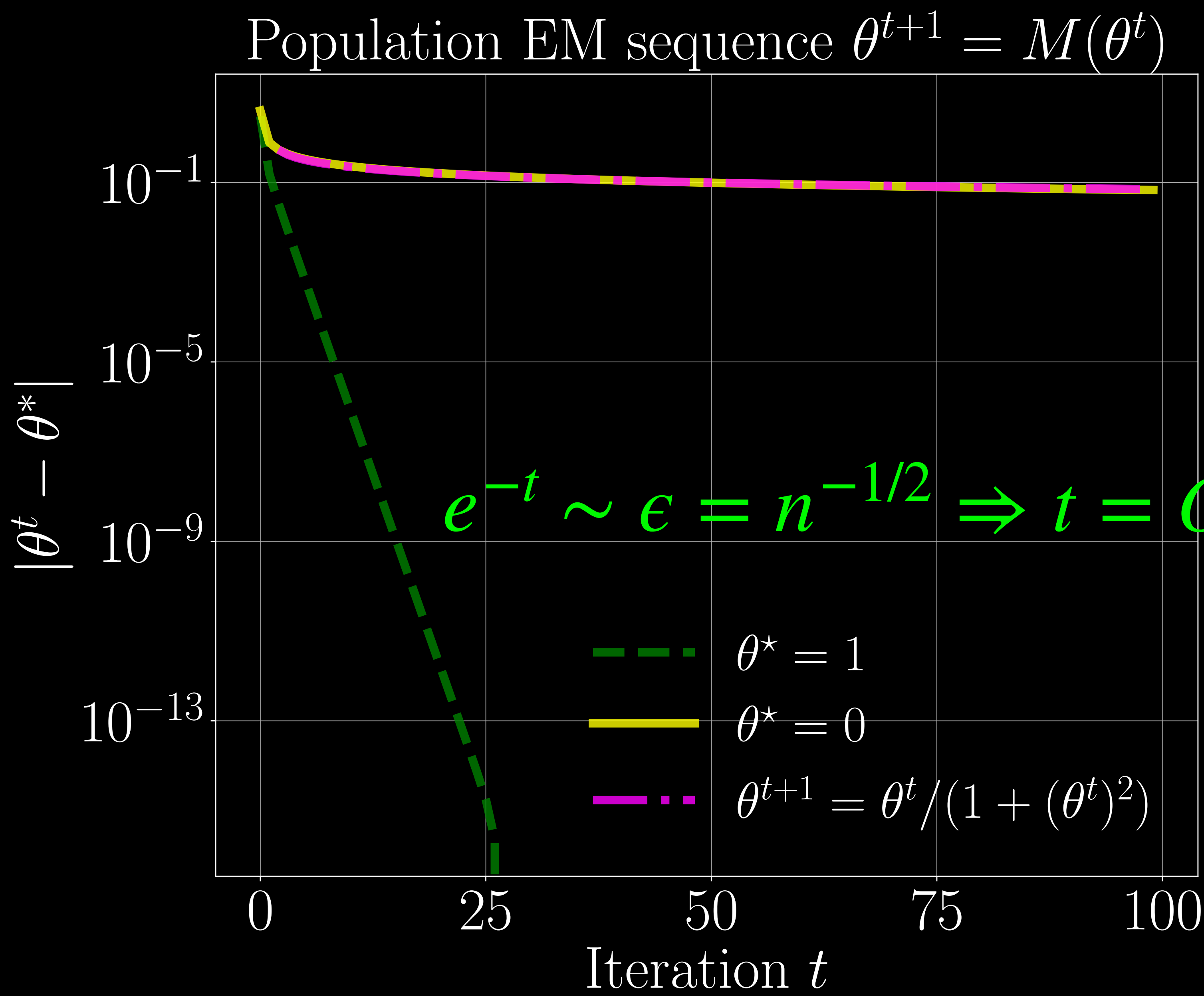
$$\|M(\theta) - \theta^\star\| \leq \kappa\|\theta - \theta^\star\|$$
$$(\kappa < 1 - c)$$

Our work: for no signal

$$\|M(\theta) - \theta^\star\| \asymp (1 - c\|\theta - \theta^\star\|^2) \cdot \|\theta - \theta^\star\|$$
$\kappa(\theta) \to 1$ as $\theta \to \theta^\star$

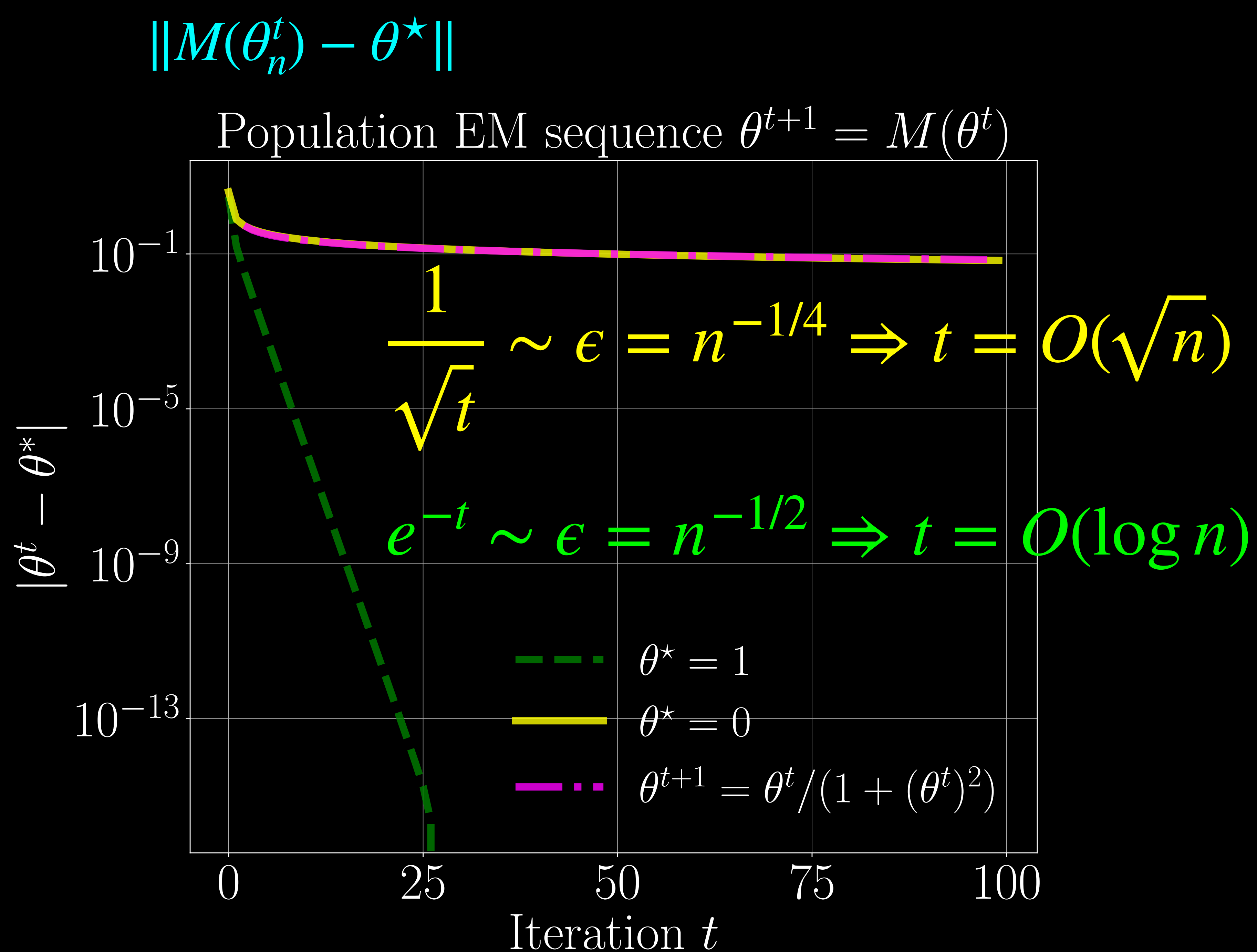
Proof strategy:
From population to sample analysis

$\|M(\theta_n^t) - \theta^\star\|$



$e^{-t} \sim \epsilon = n^{-1/2} \Rightarrow t = O(\log n)$

Proof strategy: From population to sample analysis



statistical slow-down

Proof strategy: From population to sample analysis

$$\|\theta_n^{t+1} - \theta^\star\| = \|M_n(\theta_n^t) - \theta^\star\|$$
$$\leq \|M(\theta_n^t) - \theta^\star\| + \|M_n(\theta_n^t) - M(\theta_n^t)\|$$

Strong Signal

$$\leq \kappa\|\theta_n^t - \theta^\star\| + C\sqrt{\frac{d}{n}}$$

statistical slow-down

Proof strategy: From population to sample analysis

$$\|\theta_n^{t+1} - \theta^\star\| = \|M_n(\theta_n^t) - \theta^\star\|$$

$$\leq \|M(\theta_n^t) - \theta^\star\| + \|M_n(\theta_n^t) - M(\theta_n^t)\|$$

Strong Signal

$$\leq \kappa\|\theta_n^t - \theta^\star\| + C\sqrt{\frac{d}{n}}$$

$$\lesssim \sqrt{\frac{d}{n}} \cdot \frac{1}{1-\kappa}$$

for $t \gtrsim \log_{1/\kappa}\left(\frac{n}{d} \cdot \|\theta^0 - \theta^\star\|\right)$

we are done since $1 - \kappa > c > 0$

statistical slow-down

Proof strategy: From population to sample analysis

$$\|\theta_n^{t+1} - \theta^\star\| = \|M_n(\theta_n^t) - \theta^\star\|$$
$$\leq \|M(\theta_n^t) - \theta^\star\| + \|M_n(\theta_n^t) - M(\theta_n^t)\|$$

Strong Signal

$$\leq \kappa\|\theta_n^t - \theta^\star\| + C\sqrt{\frac{d}{n}}$$

$$\lesssim \sqrt{\frac{d}{n}} \cdot \frac{1}{1-\kappa}$$

for $t \gtrsim \log_{1/\kappa}\left(\frac{n}{d} \cdot \|\theta^0 - \theta^\star\|\right)$

we are done since $1 - \kappa > c > 0$

Weak Signal

$$1 - \kappa(\theta) \approx \|\theta - \theta^\star\|^2$$

$\Downarrow$ (implicit equation)

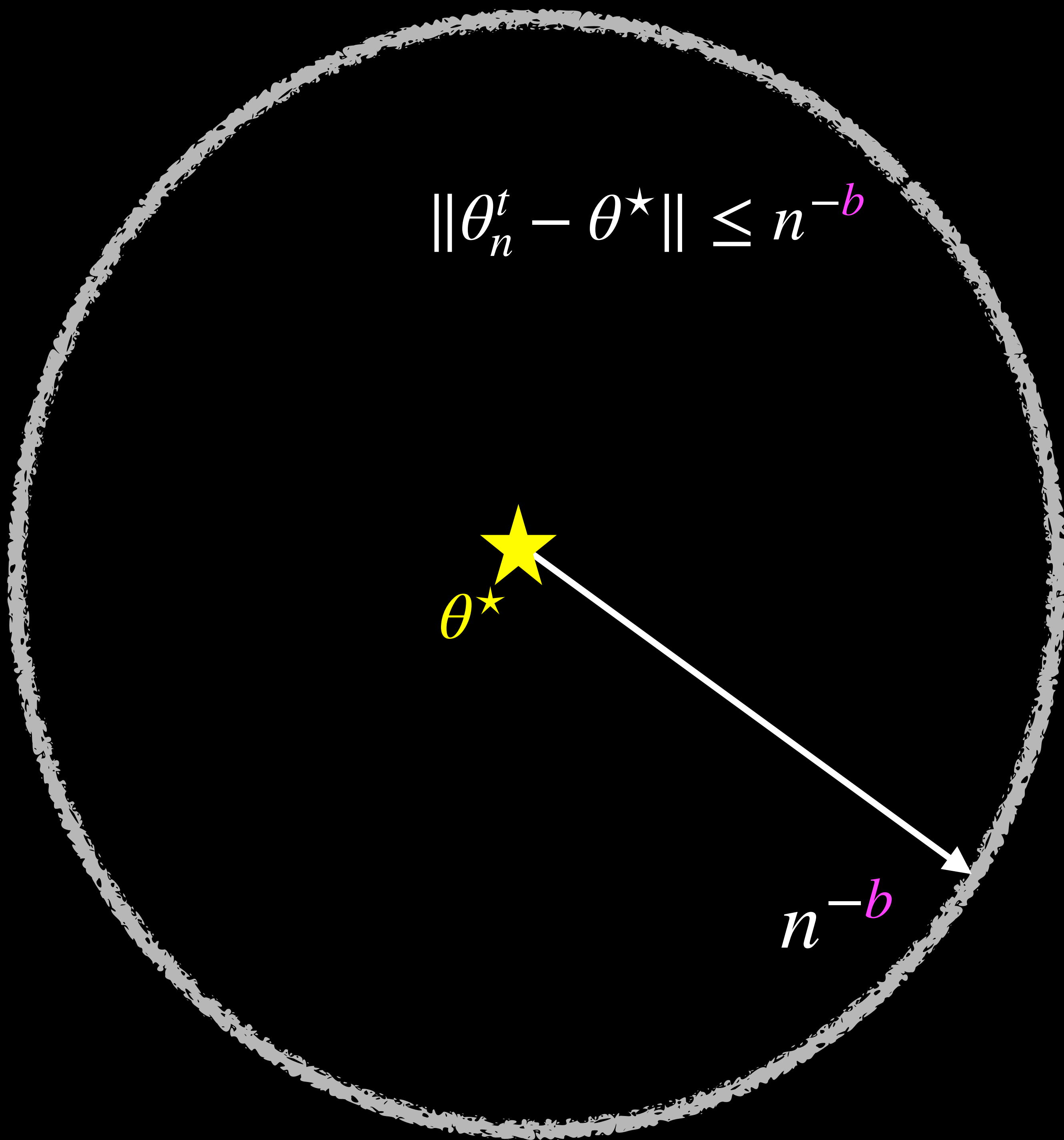
$$\|\widehat{\theta}_n - \theta^\star\| \lesssim \sqrt{\frac{d}{n}} \cdot \frac{1}{\|\widehat{\theta}_n - \theta^\star\|^2}$$

$\Downarrow$

$$\|\widehat{\theta}_n - \theta^\star\| \lesssim \left(\frac{d}{n}\right)^{1/6}$$

sub-optimal compared to $n^{-1/4}$

Sharpening the proof: Localize the estimates in a ball

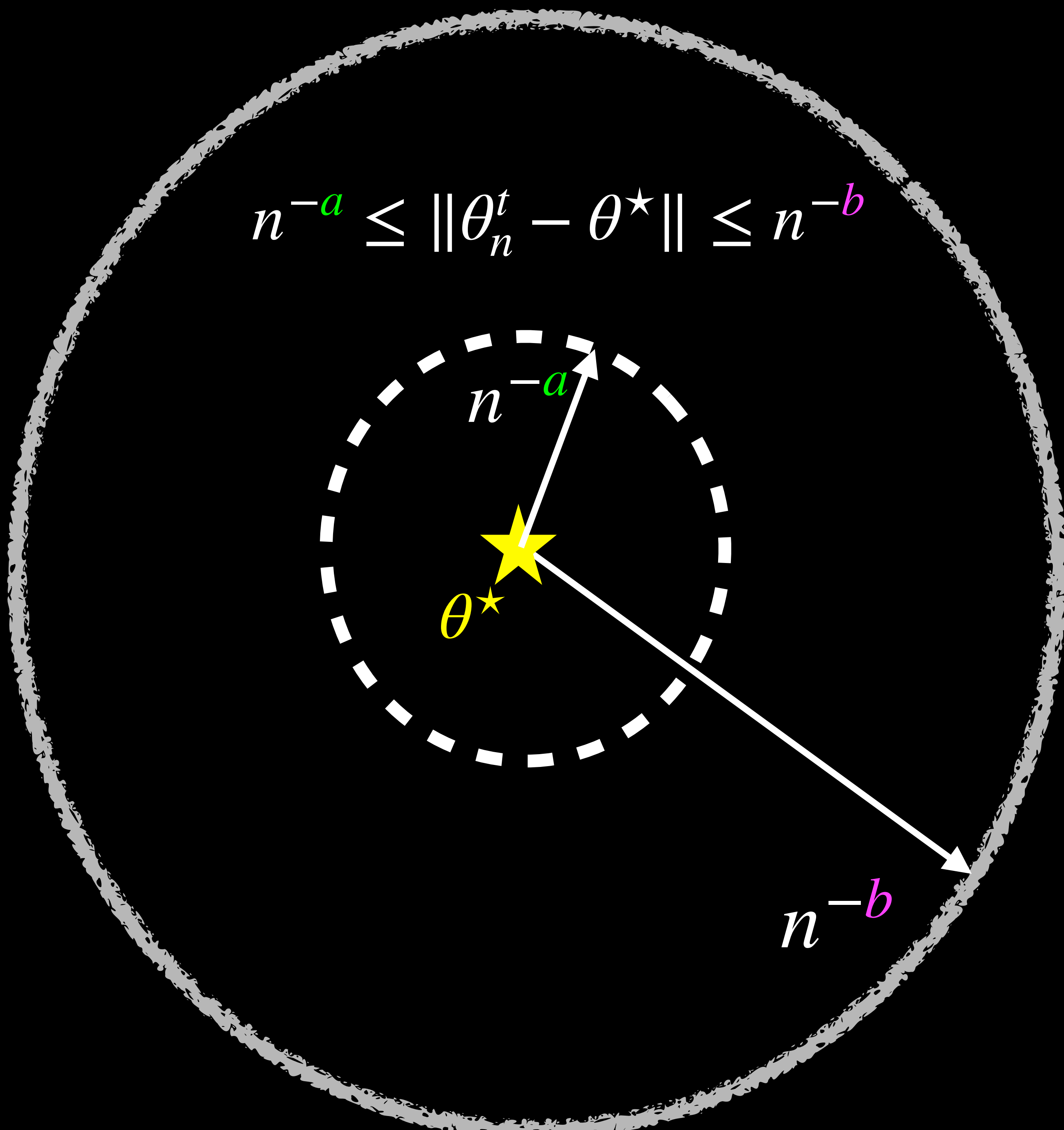


A standard technique in empirical process theory to derive sharp minimax rates

But κ gets too close to 1 if θ_n^t is too close to $\theta^\star$

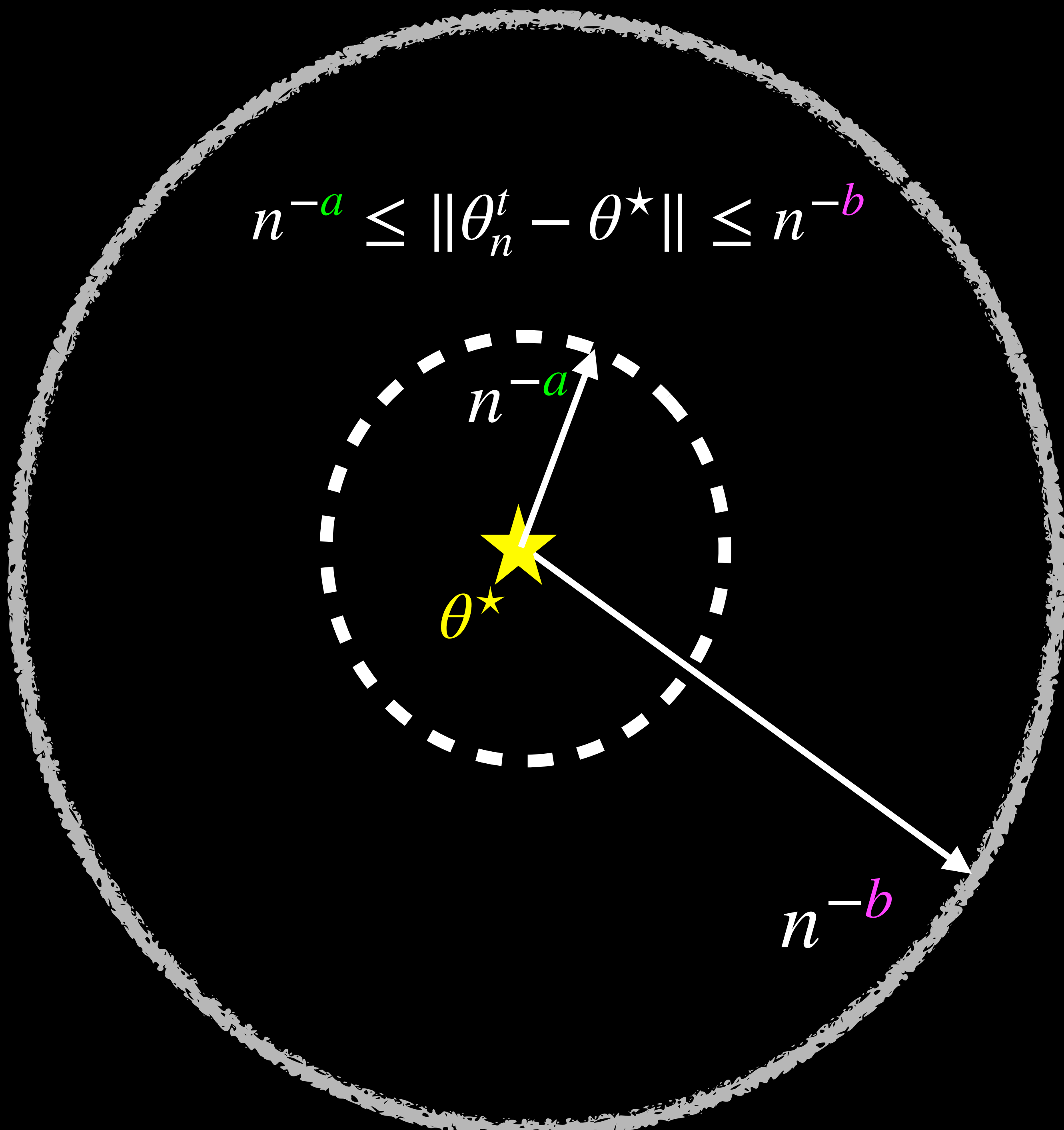
statistical
slow-down

Sharpening the proof:
Localize the estimates in an **annulus**



statistical slow-down

Sharpening the proof: Localize the estimates in an **annulus**



Outer radius provides a control on the perturbation error

$$\|M(\theta_n^t) - M_n(\theta_n^t)\| \leq \frac{n^{-b}}{\sqrt{n}}$$

Inner radius helps to control the contraction

$$1 - \kappa(\theta_n^t) \geq n^{-2a}$$

Leads to a recursion between a and b with a unique fixed point **1/4**

$$a = \frac{1}{3}(b + \frac{1}{2})$$

Summary

Over-specification / weak signal is a double-edged sword

statistical slow-down

$n^{-1/4}$ vs $n^{-1/2}$

computational slow-down

$n^{1/2}$ vs $\log n$

Summary

Over-specification / weak signal is a double-edged sword

statistical slow-down

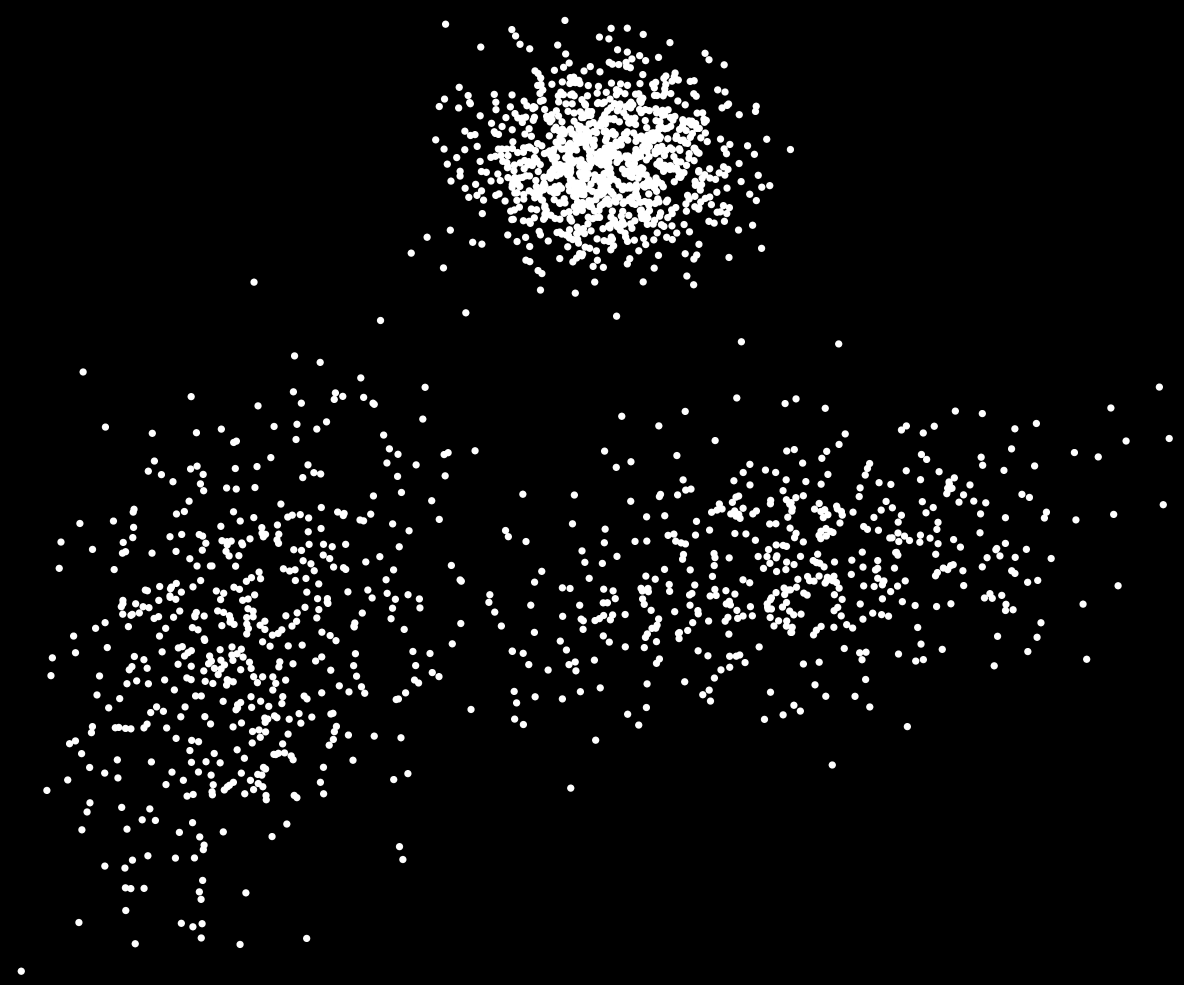
$n^{-1/4}$ vs $n^{-1/2}$

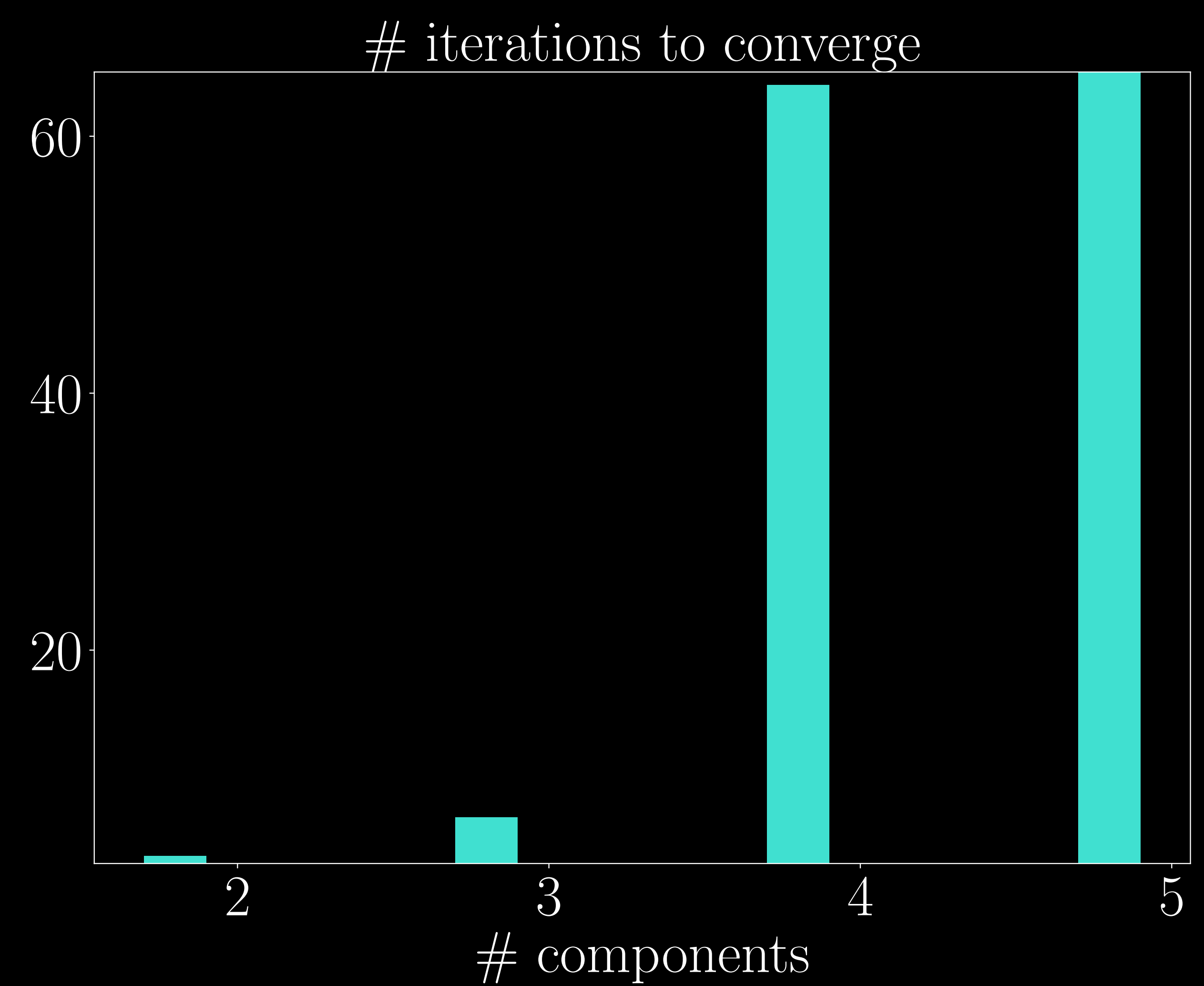
computational slow-down

$n^{1/2}$ vs $\log n$

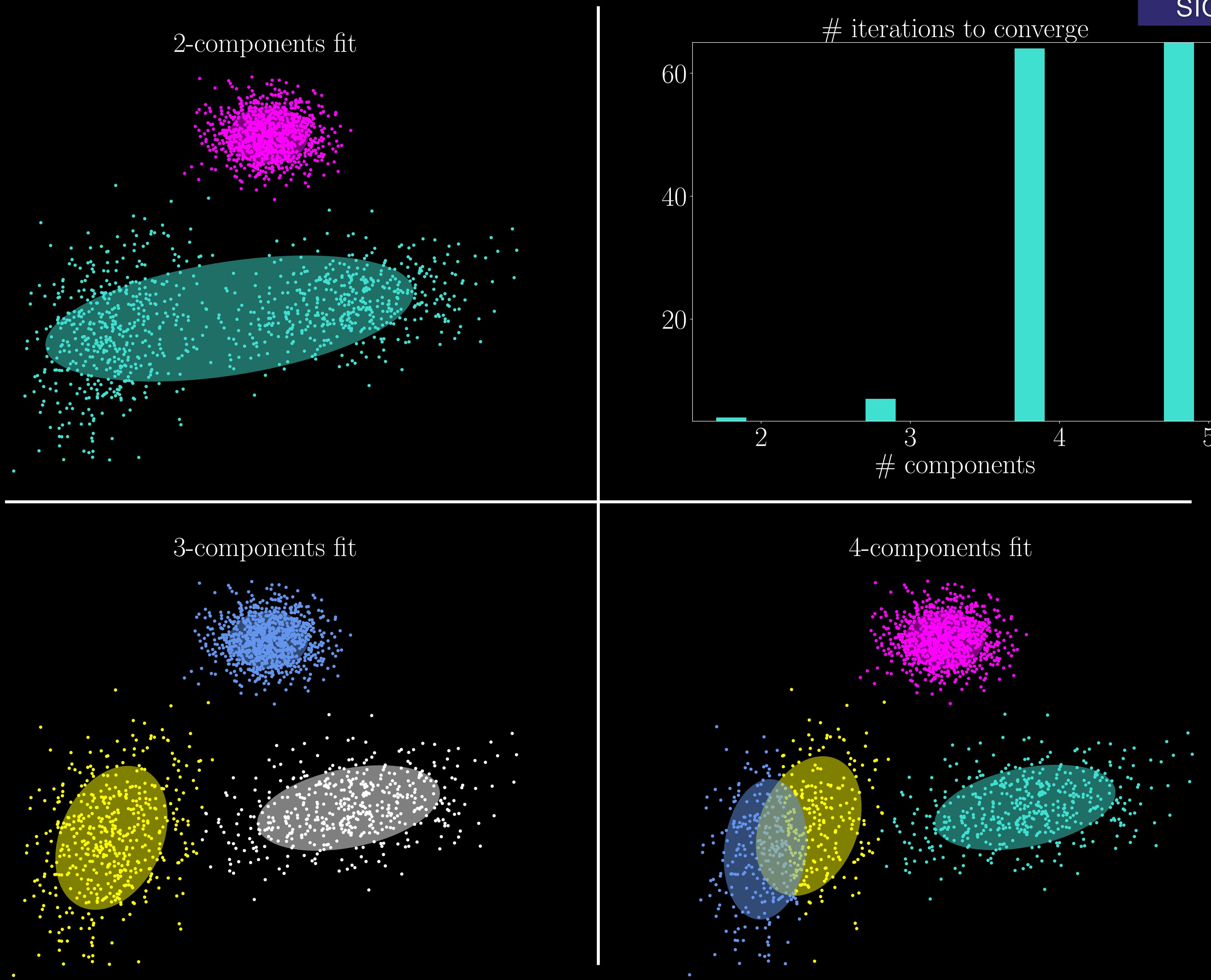
Blessing in disguise?

A future recipe for model selection: Look at EM iterations





leveraging computational slow-down



Follow-up work

We assume zero signal:

Wu and Zhou [2019] generalize it to a minimax weak signal setting
(under restrictive initialization conditions)

We assume known variance:

Our recent work shows that fitting an over-specified model with unknown variance may lead to further slow-down ($n^{-1/8}$)

Localization beyond EM:

We employ localization techniques to derive sharp rates beyond mixture models
(draft in progress)

Thank you!

Over-specification / weak signal is a double-edged sword

statistical slow-down

$n^{-1/4}$ vs $n^{-1/2}$

computational slow-down

$n^{1/2}$ vs $\log n$