Sharp analysis of EM for weakly indentifiable models

Koulik Khamaru*

Department of Statistics, UC Berkeley

Based on joint works with
Nhat Ho*, Raaz dwivedi*, Michael I. Jordan, Martin J. Wainwright, and Bin Yu

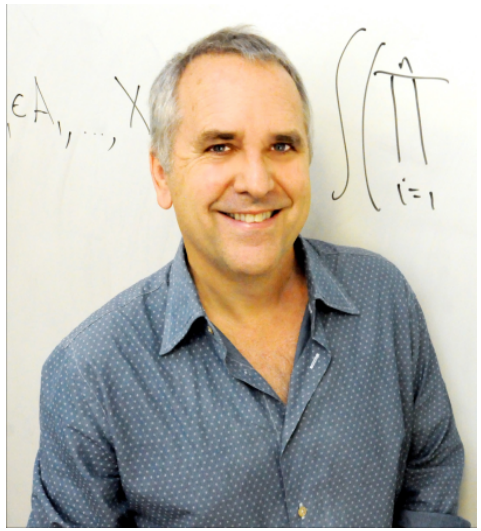
June 17, 2020

Raaz Dwivedi*

Nhat Ho*

Martin Wainwright

Bin Yu



Michael Jordan

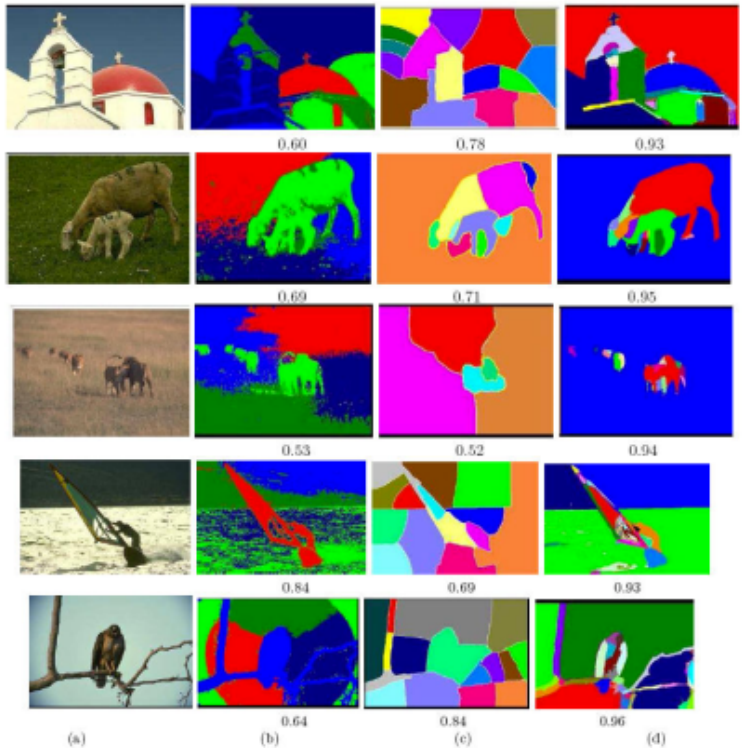
Mixture models: usefulness

- Topic modeling, Financial returns
- Image annotation, classification, segmentation

"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: Blei et al. 2003



Source: Rotem et al. 2007

Gaussian mixture models: formulation

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

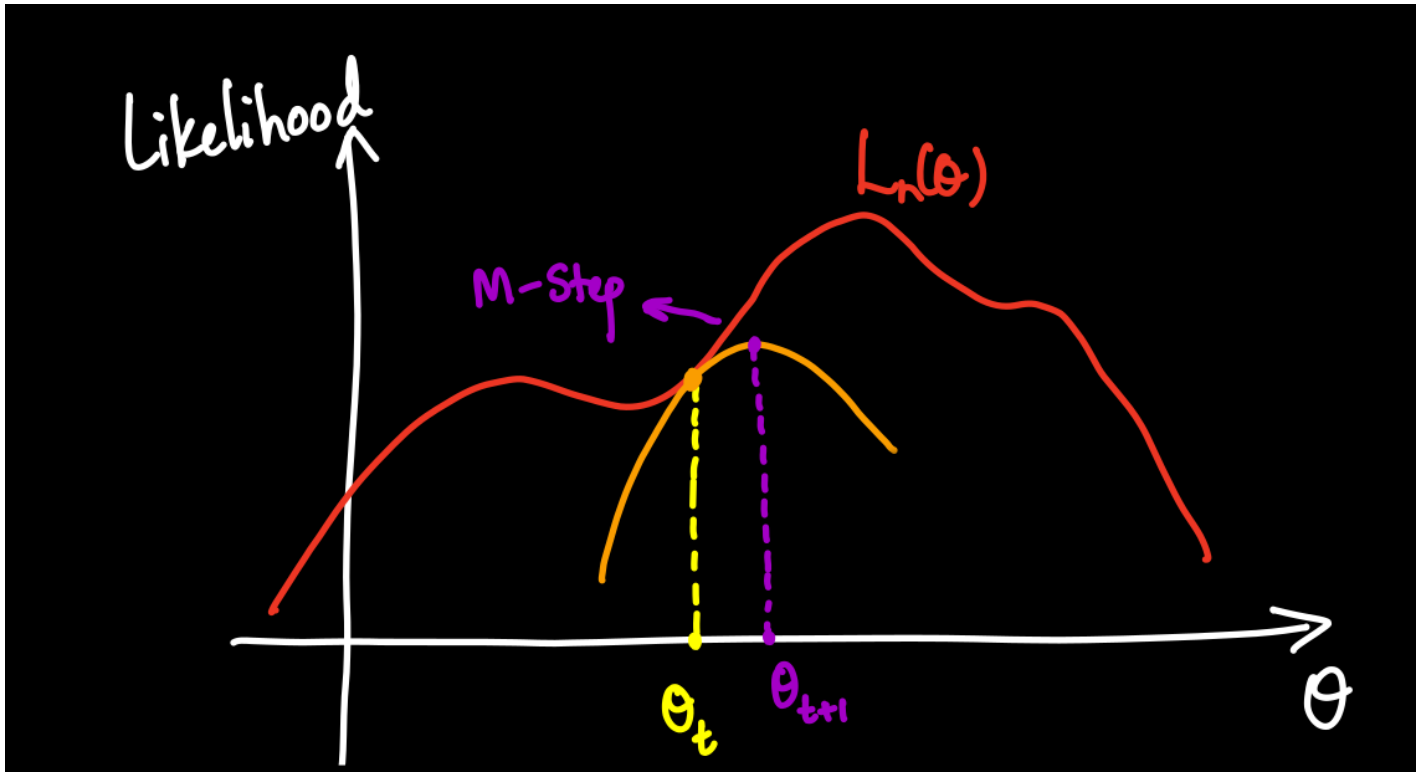
$$Z \sim \text{multinomial}(w_1, \ldots, w_K) \quad \text{and} \quad [X \Big| Z = i] \sim \mathcal{N}(\mu_i, \Sigma_i)$$

- Marginal distribution of X:

$$X \sim \sum_{i=1}^{K} w_i \mathcal{N}(\mu_i, \Sigma_i)$$

- Goal: Given samples $X_1, ..., X_n$, our aim is to estimate the parameters $(\mu_i, \Sigma_i)_{i=i}^{K}$

Estimation by EM algorithm:

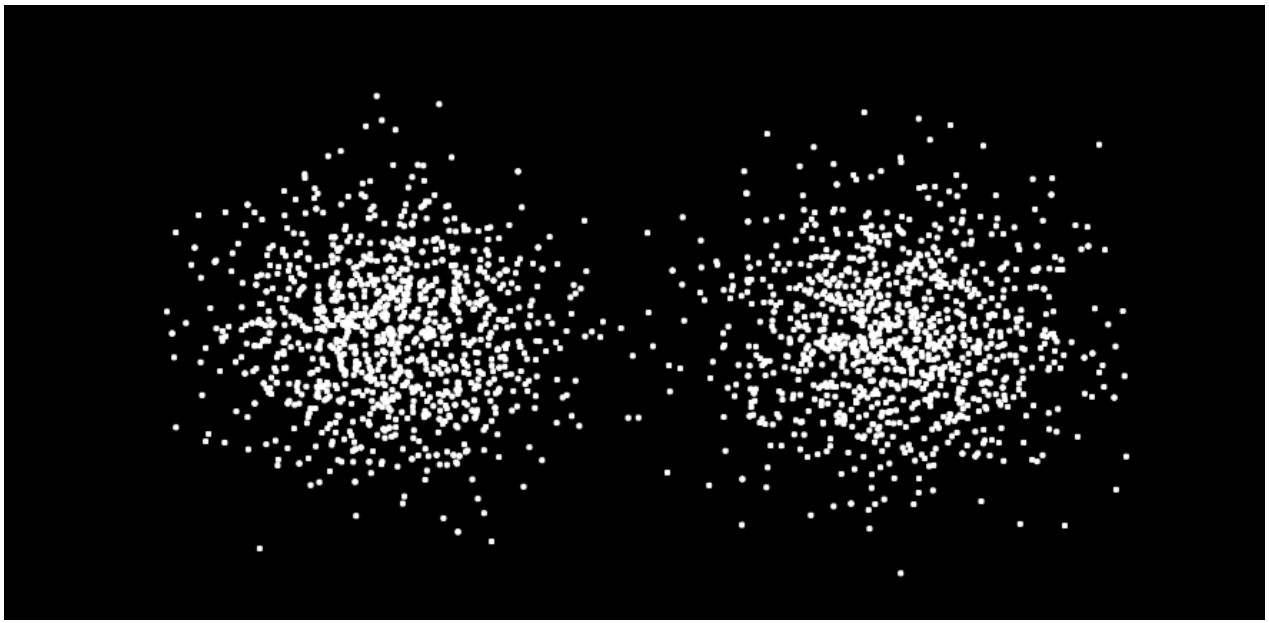


Dempster-Laird-Rubin, Sundberg, Martin-Löf, Jeff Wu 1970-80

Well specified 2-Gaussian mixtures

True model: $\frac{1}{2}\mathcal{N}(-\theta^*, \sigma^{*2} I_d) + \frac{1}{2}\mathcal{N}(\theta^*, \sigma^{*2} I_d)$

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



Earlier work:

- Asymptotic results: Boyles 1983, Neal and Hinton 1995, Ma, Xu and Jordan 1996, 2000, Ruan, Yuan and Zhou 2011, ...

- Non-asymptotic results: Balakrishnan et al. 2017, Yin et al. 2017, Daskalakis et al. 2017, Cai et al. 2019, ...

Strong vs Weak signal

$$\textbf{True model:} \quad \frac{1}{2}\mathcal{N}(-\theta^*, \sigma^{*2} I_d) + \frac{1}{2}\mathcal{N}(\theta^*, \sigma^{*2} I_d)$$

$$\textbf{Fitted model:} \quad \frac{1}{2}\mathcal{N}(-\theta, \sigma^{2} I_d) + \frac{1}{2}\mathcal{N}(\theta, \sigma^{2} I_d)$$

- In strong signal, $\frac{\|\theta^*\|}{\sigma^*}$ is large.
- In weak signal, $\frac{\|\theta^*\|}{\sigma^*} \approx 0$.
- Strong signal case was analyzed by Cai et al. 2019.

Weak signal:

True model: $\mathcal{N}(0, I_d)$

$$\equiv \frac{1}{2}\mathcal{N}(-\theta^*, I_d) + \frac{1}{2}\mathcal{N}(\theta^*, I_d) \text{ with } \theta^* = 0$$

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

Two observations:

- Here $\theta^* = 0$ and $\sigma^* = 1$, the weak signal case.
- Over-specified model.

Main result

For the no signal case $\theta^* = 0$, for arbitrary initialization, the sample EM iterate satisfies:

$$\|\theta_n^t - \theta^*\|_2 \lesssim \begin{cases} \left(\frac{1}{n}\right)^{\frac{1}{8}} & \text{for} \quad t \gtrsim n^{\frac{3}{4}} \\ \left(\frac{d}{n}\right)^{\frac{1}{4}} & \text{for} \quad t \gtrsim \left(\frac{n}{d}\right)^{\frac{1}{2}} \qquad d \geq 2. \end{cases}$$

Strong vs Weak signal

For the no signal case $\theta^* = 0$, for arbitrary initialization, the sample EM iterate satisfies:

$$\|\theta_n^t - \theta^*\|_2 \lesssim \begin{cases} \left(\frac{1}{n}\right)^{\frac{1}{8}} & \text{for} \quad t \gtrsim n^{\frac{3}{4}} \\ \left(\frac{d}{n}\right)^{\frac{1}{4}} & \text{for} \quad t \gtrsim \left(\frac{n}{d}\right)^{\frac{1}{2}} \end{cases} \qquad d \geq 2.$$

- For the strong signal case, $\|\theta^*\| > C$, [Cai et al. 2019] showed[1]

$$\|\theta_n^t - \theta^*\|_2 \lesssim \left(\frac{d}{n}\right)^{\frac{1}{2}} \qquad \text{for} \qquad t \gtrsim \log(n).$$

[1] Result for a more general probelm an for a variant of EM

Statistical slowdown

Weak signal:

$$\|\theta_n^t - \theta^*\|_2 \lesssim \begin{cases} \left(\frac{1}{n}\right)^{\frac{1}{8}} & \text{for} \quad t \gtrsim n^{\frac{3}{4}} \\ \left(\frac{d}{n}\right)^{\frac{1}{4}} & \text{for} \quad t \gtrsim \left(\frac{n}{d}\right)^{\frac{1}{2}} \quad d \geq 2. \end{cases}$$

Statistical slowdown

Strong signal:

$$\|\theta_n^t - \theta^*\|_2 \lesssim \left(\frac{d}{n}\right)^{\frac{1}{2}} \quad \text{for} \quad t \gtrsim \log(n).$$

Computational slowdown

Weak signal:

$$\|\theta_n^t - \theta^*\|_2 \lesssim \begin{cases} \left(\frac{1}{n}\right)^{\frac{1}{8}} & \text{for} \quad t \gtrsim n^{\frac{3}{4}} \\ \left(\frac{d}{n}\right)^{\frac{1}{4}} & \text{for} \quad t \gtrsim \left(\frac{n}{d}\right)^{\frac{1}{2}} \end{cases} \qquad d \geq 2.$$

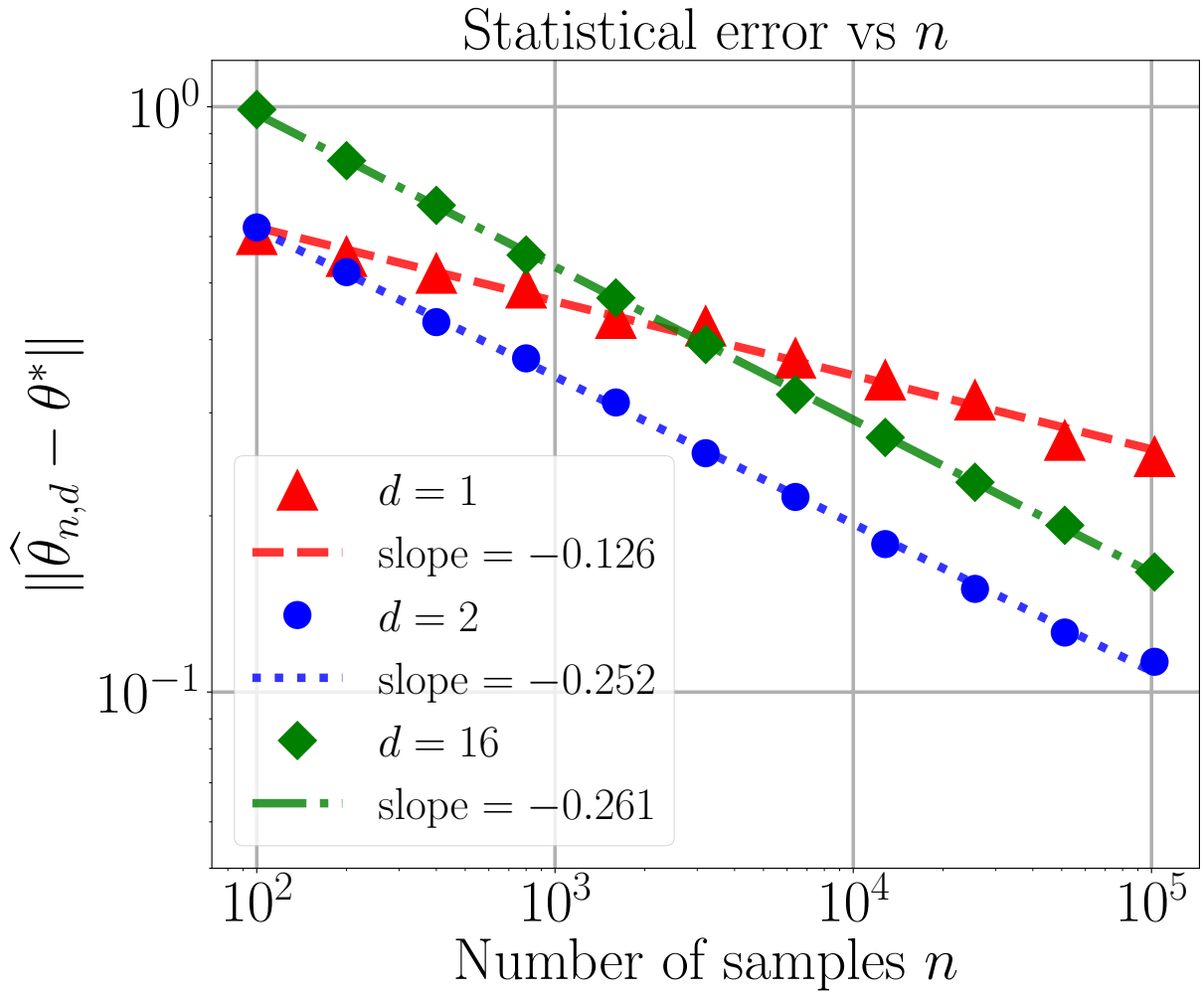
Statistical slowdown

Computational slowdown

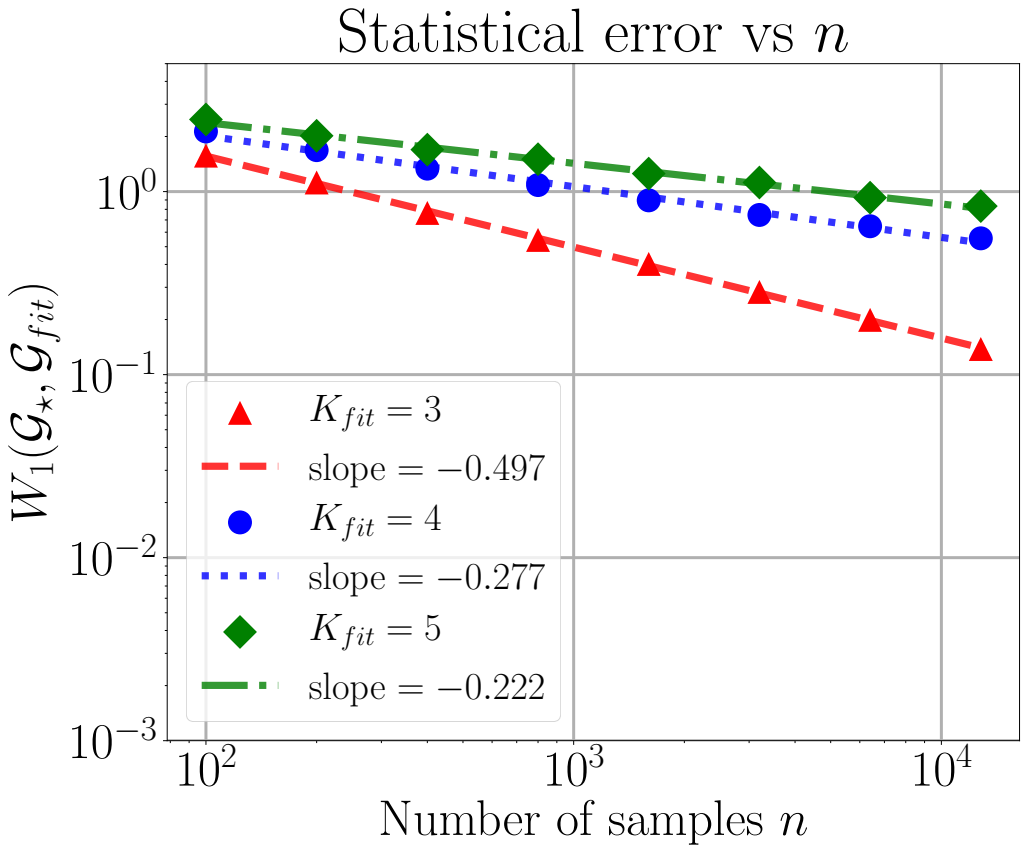
Strong signal:

$$\|\theta_n^t - \theta^*\|_2 \lesssim \left(\frac{d}{n}\right)^{\frac{1}{2}} \qquad \text{for} \qquad t \gtrsim \log(n).$$

Matching simulations:



Do the results generalize?



- True model: 3-Gaussian mixture.
- Fitted model: Mixture of 3 or more Gaussians.

Key ideas

Notation

- EM iterates: $\theta_1^n, \ldots \theta_t^n$
- $M_n(\cdot)$ is the EM operator, i.e., $\theta_{t+1}^n = M_n(\theta_t^n)$
- $M(\cdot)$ is a suitably chosen operator.

Proof strategy $d = 1$

$$\begin{aligned}\|\theta_{t+1}^n - \theta^*\|_2 &:= \|M_n(\theta_t^n) - \theta^*\|_2 \\ &\leq \|M(\theta_t^n) - \theta^*\|_2 + \|M_n(\theta_t^n) - M(\theta_t^n)\|_2 \\ &\leq \kappa(\theta_t^n, \theta^*)\|\theta_t^n - \theta^*\|_2 + \epsilon_n(\theta_t^n, \theta^*)\end{aligned}$$

Key recursion:

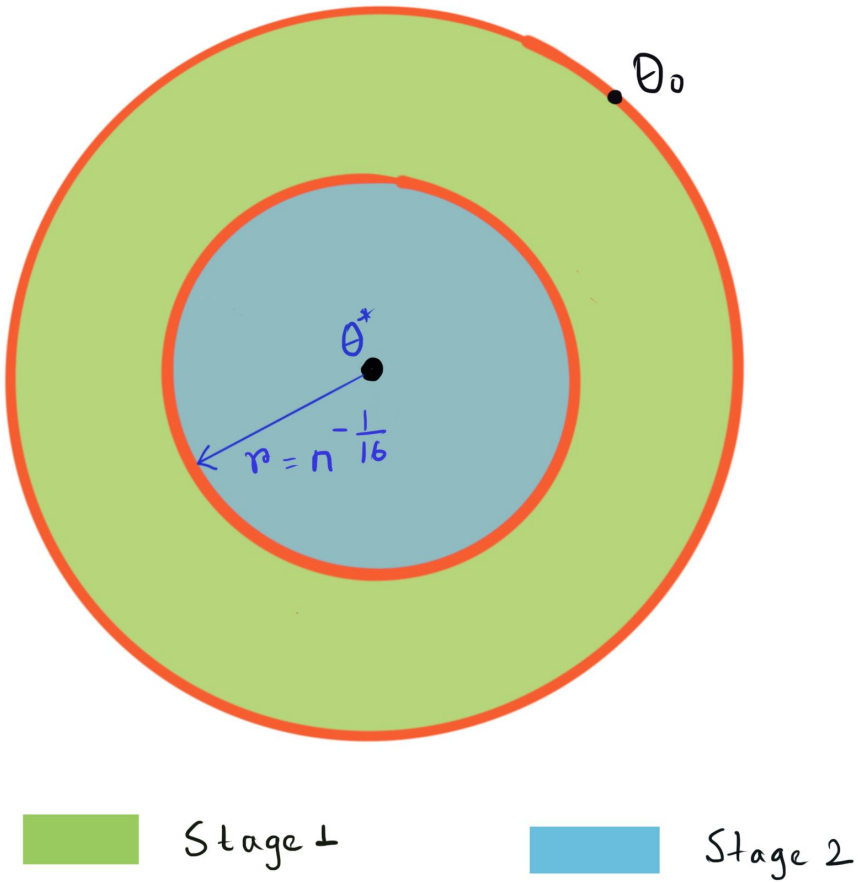
$$\|\theta_{t+1}^n - \theta^*\|_2 \leq \kappa(\theta_t^n, \theta^*)\|\theta_t^n - \theta^*\|_2 \quad + \quad \epsilon_n(\theta_t^n, \theta^*).$$

Sharp bound on $\kappa(\theta, \theta^*)$

$$\kappa(\theta, \theta^*) \asymp 1 - c\|\theta - \theta^*\|^6 \qquad \text{for} \qquad d = 1.$$

- Observe, $\kappa(\theta, \theta^*) < 1$ (Only locally).
- $\kappa(\theta, \theta^*) \to 1$ as $\theta \to \theta^*$.

Two stages of analysis



Stage-1 :

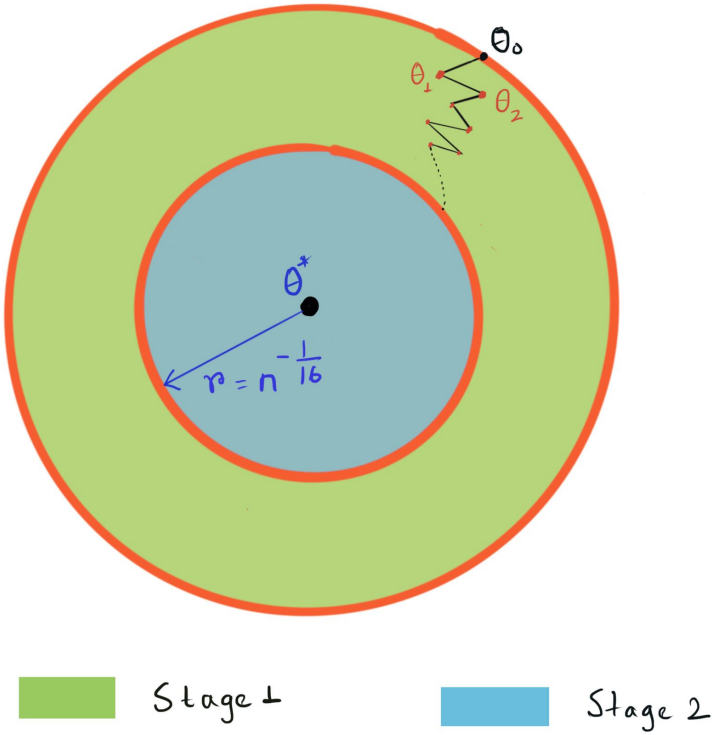
$$\sup_{\|\theta-\theta^*\|\leq r} \epsilon_n(\theta, \theta^*) \leq \frac{r}{\sqrt{n}}$$

Satge-2:

$$\sup_{\|\theta-\theta^*\|\leq r} \epsilon_n(\theta, \theta^*) \leq \frac{r^3}{\sqrt{n}}$$

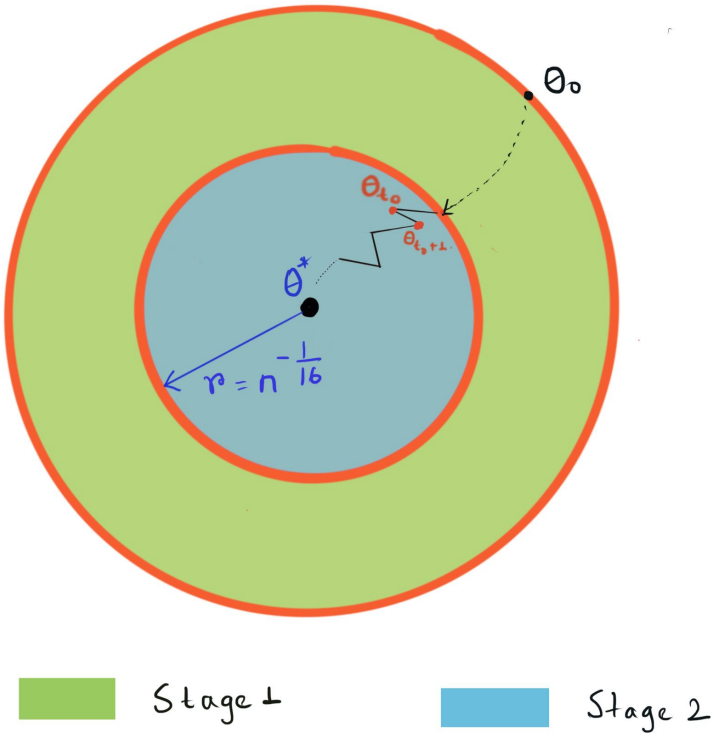
Analysis in stage 1

$$\|\theta_{t+1}^n - \theta^*\|_2 \lesssim \left(1 - c\|\theta - \theta^*\|^6\right)\|\theta_t^n - \theta^*\|_2 \quad + \quad \frac{r}{\sqrt{n}}.$$

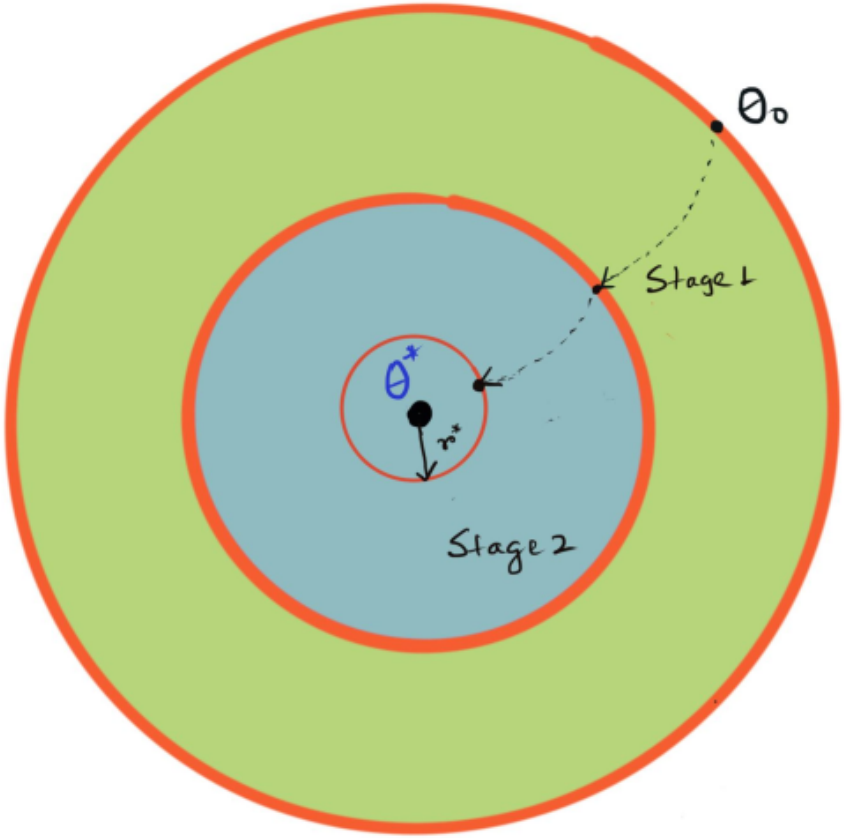


Analysis in stage 2

$$\|\theta_{t+1}^n - \theta^*\|_2 \lesssim \left(1 - c\|\theta - \theta^*\|^6\right)\|\theta_t^n - \theta^*\|_2 + \frac{r^3}{\sqrt{n}}.$$



Satge 1 + Stage 2



Final radius $= r^* = n^{-\frac{1}{8}}$.