Mean Estimation in High-Dimensional Binary Markov Gaussian Mixture Models

Yihan Zhang and Nir Weinberger

Institute of Science and Technology, Austria
and
Technion - Israel Institute of Technology, Israel

Institute of Science and Technology Austria

TECHNION Israel Institute of Technology

NeurIPS 2022

Motivation

- Memory between data samples is ubiquitous

Motivation

- Memory between data samples is ubiquitous
  - Imaging, meteorology, health care, finance, social science ...

Motivation

- Memory between data samples is ubiquitous
  - Imaging, meteorology, health care, finance, social science ...
- Statistical inference algorithms from samples with memory are a developed topic

Motivation

- Memory between data samples is ubiquitous
  - Imaging, meteorology, health care, finance, social science ...
- Statistical inference algorithms from samples with memory are a developed topic
  - Baum-Welch, various message-passing algorithms...

Motivation

- Memory between data samples is ubiquitous
  - Imaging, meteorology, health care, finance, social science ...
- Statistical inference algorithms from samples with memory are a developed topic
  - Baum-Welch, various message-passing algorithms...
- **But:** Memory improves the performance of statistical inference

Motivation

- Memory between data samples is ubiquitous
  - Imaging, meteorology, health care, finance, social science ...
- Statistical inference algorithms from samples with memory are a developed topic
  - Baum-Welch, various message-passing algorithms...
- **But:** Memory improves the performance of statistical inference
  - To what extent?

Motivation

- Memory between data samples is ubiquitous
  - Imaging, meteorology, health care, finance, social science ...
- Statistical inference algorithms from samples with memory are a developed topic
  - Baum-Welch, various message-passing algorithms...
- **But:** Memory improves the performance of statistical inference
  - To what extent?
  - How it interacts with the number of samples? parameters dimension? signal-to-noise?

Motivation

- Memory between data samples is ubiquitous
  - Imaging, meteorology, health care, finance, social science ...
- Statistical inference algorithms from samples with memory are a developed topic
  - Baum-Welch, various message-passing algorithms...
- **But:** Memory improves the performance of statistical inference
  - To what extent?
  - How it interacts with the number of samples? parameters dimension? signal-to-noise?
- **In this talk:** estimation in a basic Gaussian model with memory

Related work – Gaussian mixture model

- Many papers on estimation in Gaussian mixture models **(memoryless** model) via the method-of-moments (MoM) or Expectation-maximization (EM)

Related work – Gaussian mixture model

- Many papers on estimation in Gaussian mixture models **(memoryless** model) via the method-of-moments (MoM) or Expectation-maximization (EM)
- Models with **memory**:

Related work – Gaussian mixture model

- Many papers on estimation in Gaussian mixture models **(memoryless** model) via the method-of-moments (MoM) or Expectation-maximization (EM)
- Models with **memory**:
  - [YBW15] analyzed Baum-Welch for Gaussian HMM

Related work – Gaussian mixture model

- Many papers on estimation in Gaussian mixture models **(memoryless** model) via the method-of-moments (MoM) or Expectation-maximization (EM)
- Models with **memory**:
  - [YBW15] analyzed Baum-Welch for Gaussian HMM
    - Not minimax optimal in the number of samples, dimension, noise level, and the amount of memory

Related work – Gaussian mixture model

- Many papers on estimation in Gaussian mixture models **(memoryless** model) via the method-of-moments (MoM) or Expectation-maximization (EM)
- Models with **memory**:
  - [YBW15] analyzed Baum-Welch for Gaussian HMM
    - Not minimax optimal in the number of samples, dimension, noise level, and the amount of memory
  - Minimax error rates for linear regression with Markovian covariates [Bre+20]

Related work – Gaussian mixture model

- Many papers on estimation in Gaussian mixture models **(memoryless** model) via the method-of-moments (MoM) or Expectation-maximization (EM)
- Models with **memory**:
  - [YBW15] analyzed Baum-Welch for Gaussian HMM
    - Not minimax optimal in the number of samples, dimension, noise level, and the amount of memory
  - Minimax error rates for linear regression with Markovian covariates [Bre+20]
  - Linear and logistic regression with general network dependencies [DDP19; Kan+21]

Related work – Gaussian mixture model

- Many papers on estimation in Gaussian mixture models (**memoryless** model) via the method-of-moments (MoM) or Expectation-maximization (EM)
- Models with **memory**:
  - [YBW15] analyzed Baum-Welch for Gaussian HMM
    - Not minimax optimal in the number of samples, dimension, noise level, and the amount of memory
  - Minimax error rates for linear regression with Markovian covariates [Bre+20]
  - Linear and logistic regression with general network dependencies [DDP19; Kan+21]
  - Learnability and generalization bounds [Dag+19]

Problem formulation – Statistical model

- A binary Markov chain

$$\mathbb{P}[S_0 = 1] = 1/2, \quad S_i = \begin{cases} S_{i-1}, & \text{w.p. } 1-\delta \\ -S_{i-1}, & \text{w.p. } \delta \end{cases}, \quad i = 1, \ldots, n$$

Problem formulation – Statistical model

- A binary Markov chain

$$\mathbb{P}[S_0 = 1] = 1/2, \quad S_i = \begin{cases} S_{i-1}, & \text{w.p. } 1-\delta \\ -S_{i-1}, & \text{w.p. } \delta \end{cases}, \quad i = 1, \ldots, n$$

- An unknown mean parameter $\theta_* \in \mathbb{R}^d$ with $\|\theta_*\| = t$

Problem formulation – Statistical model

- A binary Markov chain

$$\mathbb{P}[S_0 = 1] = 1/2, \quad S_i = \begin{cases} S_{i-1}, & \text{w.p. } 1-\delta \\ -S_{i-1}, & \text{w.p. } \delta \end{cases}, \quad i = 1, \ldots, n$$

- An unknown mean parameter $\theta_* \in \mathbb{R}^d$ with $\|\theta_*\| = t$
- Observations

$$X_i = S_i \cdot \theta_* + Z_i, \quad Z_i \stackrel{\text{IID}}{\sim} N(0, I_d), \quad i = 1, \ldots, n$$

Problem formulation – Statistical model

- A binary Markov chain

$$\mathbb{P}[S_0 = 1] = 1/2, \quad S_i = \begin{cases} S_{i-1}, & \text{w.p. } 1-\delta \\ -S_{i-1}, & \text{w.p. } \delta \end{cases}, \quad i = 1, \ldots, n$$

- An unknown mean parameter $\theta_* \in \mathbb{R}^d$ with $\|\theta_*\| = t$
- Observations

$$X_i = S_i \cdot \theta_* + Z_i, \quad Z_i \overset{\text{IID}}{\sim} N(0, I_d), \quad i = 1, \ldots, n$$

- Local minimax rate: For $d \geq 2$

$$\mathsf{M}(n,d,\delta,t) := \inf_{\hat{\theta}(X_1^n)} \sup_{\|\theta_*\|=t} \mathbb{E}\left[\min\{\|\theta_* - \hat{\theta}(X_1^n)\|, \|\theta_* + \hat{\theta}(X_1^n)\|\}\right]$$

Problem formulation – Statistical model

- A binary Markov chain

$$\mathbb{P}[S_0 = 1] = 1/2, \quad S_i = \begin{cases} S_{i-1}, & \text{w.p. } 1-\delta \\ -S_{i-1}, & \text{w.p. } \delta \end{cases}, \quad i = 1, \dots, n$$

- An unknown mean parameter $\theta_* \in \mathbb{R}^d$ with $\|\theta_*\| = t$
- Observations

$$X_i = S_i \cdot \theta_* + Z_i, \quad Z_i \overset{\text{IID}}{\sim} N(0, I_d), \quad i = 1, \dots, n$$

- Local minimax rate: For $d \geq 2$

$$\mathsf{M}(n,d,\delta,t) := \inf_{\hat{\theta}(X_1^n)} \sup_{\|\theta_*\|=t} \mathbb{E}\left[\min\{\|\theta_* - \hat{\theta}(X_1^n)\|, \|\theta_* + \hat{\theta}(X_1^n)\|\}\right]$$

- Extremes are solved:

Problem formulation – Statistical model

- A binary Markov chain

$$\mathbb{P}[S_0 = 1] = 1/2, \quad S_i = \begin{cases} S_{i-1}, & \text{w.p. } 1-\delta \\ -S_{i-1}, & \text{w.p. } \delta \end{cases}, \quad i = 1, \dots, n$$

- An unknown mean parameter $\theta_* \in \mathbb{R}^d$ with $\|\theta_*\| = t$
- Observations

$$X_i = S_i \cdot \theta_* + Z_i, \quad Z_i \overset{\text{IID}}{\sim} N(0, I_d), \quad i = 1, \dots, n$$

- Local minimax rate: For $d \geq 2$

$$\mathsf{M}(n,d,\delta,t) := \inf_{\hat{\theta}(X_1^n)} \sup_{\|\theta_*\|=t} \mathbb{E}\left[\min\{\|\theta_* - \hat{\theta}(X_1^n)\|, \|\theta_* + \hat{\theta}(X_1^n)\|\}\right]$$

- Extremes are solved:
  - Gaussian location model (GLM, $\delta = 0$); Folklore

Problem formulation – Statistical model

- A binary Markov chain

$$\mathbb{P}[S_0 = 1] = 1/2, \quad S_i = \begin{cases} S_{i-1}, & \text{w.p. } 1-\delta \\ -S_{i-1}, & \text{w.p. } \delta \end{cases}, \quad i = 1, \ldots, n$$

- An unknown mean parameter $\theta_* \in \mathbb{R}^d$ with $\|\theta_*\| = t$
- Observations

$$X_i = S_i \cdot \theta_* + Z_i, \quad Z_i \overset{\text{IID}}{\sim} N(0, I_d), \quad i = 1, \ldots, n$$

- Local minimax rate: For $d \geq 2$

$$\mathsf{M}(n,d,\delta,t) := \inf_{\hat{\theta}(X_1^n)} \sup_{\|\theta_*\|=t} \mathbb{E}\left[\min\{\|\theta_* - \hat{\theta}(X_1^n)\|, \|\theta_* + \hat{\theta}(X_1^n)\|\}\right]$$

- Extremes are solved:
  - Gaussian location model (GLM, $\delta = 0$); Folklore
  - Gaussian mixture model (GMM, $\delta = \frac{1}{2}$); [WZ19]

Minimax rates – Main result

[This work] Up to log-factors:

- For $2 \leq d \leq \delta n$

$$\mathsf{M}(n, d, \delta, t) \asymp \begin{cases} t, & t \leq \left(\frac{\delta d}{n}\right)^{1/4} \\ \frac{1}{t}\sqrt{\frac{\delta d}{n}}, & \left(\frac{\delta d}{n}\right)^{1/4} \leq t \leq \sqrt{\delta} \\ \sqrt{\frac{d}{n}}, & t \geq \sqrt{\delta} \end{cases}$$

Minimax rates – Main result

[This work] Up to log-factors:

- For $2 \leq d \leq \delta n$

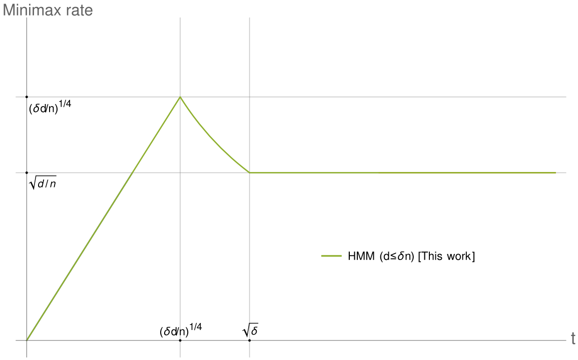
$$\mathsf{M}(n,d,\delta,t) \asymp \begin{cases} t, & t \leq \left(\frac{\delta d}{n}\right)^{1/4} \\ \frac{1}{t}\sqrt{\frac{\delta d}{n}}, & \left(\frac{\delta d}{n}\right)^{1/4} \leq t \leq \sqrt{\delta} \\ \sqrt{\frac{d}{n}}, & t \geq \sqrt{\delta} \end{cases}$$

- For $d \geq \delta n$, $\mathsf{M}(n,d,\delta,t) \asymp \mathsf{M}_{\mathrm{GLM}}(n,d,t)$

Minimax rates – Main result

[This work] Up to log-factors:

- For $2 \le d \le \delta n$
$$\mathsf{M}(n,d,\delta,t) \asymp \begin{cases} t, & t \le \left(\frac{\delta d}{n}\right)^{1/4} \\ \frac{1}{t}\sqrt{\frac{\delta d}{n}}, & \left(\frac{\delta d}{n}\right)^{1/4} \le t \le \sqrt{\delta} \\ \sqrt{\frac{d}{n}}, & t \ge \sqrt{\delta} \end{cases}$$
- For $d \ge \delta n$, $\mathsf{M}(n,d,\delta,t) \asymp \mathsf{M}_{\mathrm{GLM}}(n,d,t)$



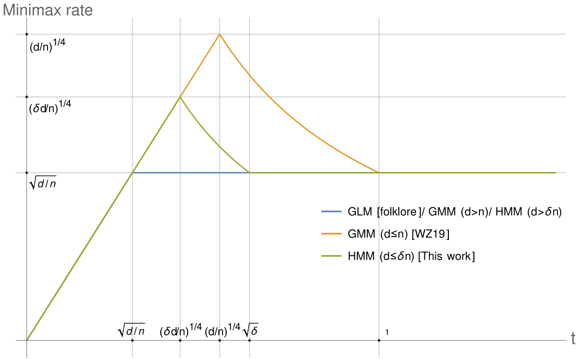
Minimax rates – Main result

[This work] Up to log-factors:

- For $2 \le d \le \delta n$

$$\mathsf{M}(n,d,\delta,t) \asymp \begin{cases} t, & t \le \left(\frac{\delta d}{n}\right)^{1/4} \\ \frac{1}{t}\sqrt{\frac{\delta d}{n}}, & \left(\frac{\delta d}{n}\right)^{1/4} \le t \le \sqrt{\delta} \\ \sqrt{\frac{d}{n}}, & t \ge \sqrt{\delta} \end{cases}$$

- For $d \ge \delta n$, $\mathsf{M}(n,d,\delta,t) \asymp \mathsf{M}_{\mathrm{GLM}}(n,d,t)$



The effect of memory

Global minimax rate $d \lesssim \delta n$	$\Theta\left(\left(\frac{\delta d}{n}\right)^{1/4}\right)$
Minimal SNR for parametric rate $d \lesssim \delta n$	$t \gtrsim \sqrt{\delta}$
Transition to high-dim	$d \asymp \delta n$

The case of unknown δ

- Estimation of δ under an approximation $\theta_\sharp$

The case of unknown δ

- Estimation of δ under an approximation $\theta_\sharp$
  - We propose a MoM estimator, and upper bound its loss

The case of unknown δ

- Estimation of δ under an approximation $\theta_\sharp$
  - We propose a MoM estimator, and upper bound its loss
  - The importance of accurate knowledge of θ_*

The case of unknown δ

- Estimation of δ under an approximation $\theta_\sharp$
  - We propose a MoM estimator, and upper bound its loss
  - The importance of accurate knowledge of θ_*
  - Impossibility result for the matched case $\theta_\sharp = \theta_*$

The case of unknown δ

- Estimation of δ under an approximation $\theta_\sharp$
  - We propose a MoM estimator, and upper bound its loss
  - The importance of accurate knowledge of θ_*
  - Impossibility result for the matched case $\theta_\sharp = \theta_*$
- Estimation of θ_* with an unknown δ

The case of unknown δ

- Estimation of δ under an approximation $\theta_\sharp$
  - We propose a MoM estimator, and upper bound its loss
  - The importance of accurate knowledge of θ_*
  - Impossibility result for the matched case $\theta_\sharp = \theta_*$
- Estimation of θ_* with an unknown δ
  - We propose a three-step algorithm

The case of unknown δ

- Estimation of δ under an approximation $\theta_\sharp$
  - We propose a MoM estimator, and upper bound its loss
  - The importance of accurate knowledge of θ_*
  - Impossibility result for the matched case $\theta_\sharp = \theta_*$
- Estimation of θ_* with an unknown δ
  - We propose a three-step algorithm
  - We prove that it adaptively achieves minimax rates of known δ at some regimes

Yihan Zhang and Nir Weinberger
“Mean Estimation in High-Dimensional Binary Markov Gaussian Mixture Models”
arXiv:2206.02455

References I

Bresler, Guy et al. (2020). "Least Squares Regression with Markovian Data: Fundamental Limits and Algorithms". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc. ISBN: 9781713829546.

Dagan, Yuval et al. (2019). "Learning from weakly dependent data under Dobrushin's condition". In: *CoRR* abs/1906.09247. arXiv: 1906.09247. URL: http://arxiv.org/abs/1906.09247.

Daskalakis, Constantinos, Nishanth Dikkala, and Ioannis Panageas (2019). "Regression from dependent observations". In: *STOC'19—Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, pp. 881–889.

References II

Kandiros, Vardis et al. (18–24 Jul 2021). “Statistical Estimation from Dependent Data”. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 5269–5278. URL: https://proceedings.mlr.press/v139/kandiros21a.html.

Wu, Yihong and Harrison H. Zhou (2019). “Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations”. In: *arXiv preprint arXiv:1908.10935*.

Yang, Fanny, Sivaraman Balakrishnan, and Martin J. Wainwright (2015). “Statistical and computational guarantees for the Baum-Welch algorithm”. In: *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, pp. 658–665.