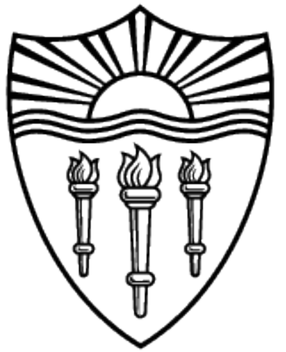
Multi-Trajectory Parameter Recovery via Hellinger Localization

Yichen Zhou
University of Southern California

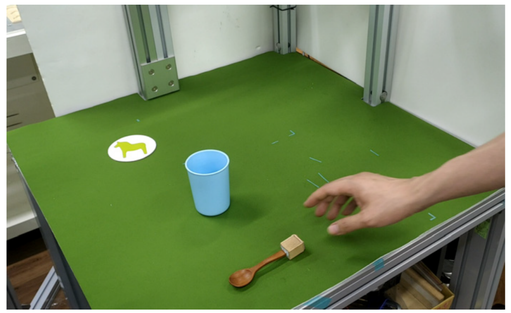
Joint work with:
Eliot Shekhtman
Ingvar Ziemann
Nikolai Matni
Stephen Tu

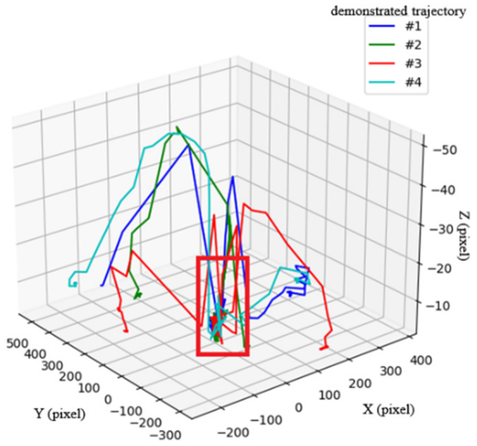


Real Data is often Multiple Trajectories

Scenario: robot learning from visual demonstration

- **Task**: learn to grasp an object from the table
- **Data**: multiple human demos
- **Dependency structure**:
  - Within each demo = **temporarily correlated**
  - Across demos = **independent**





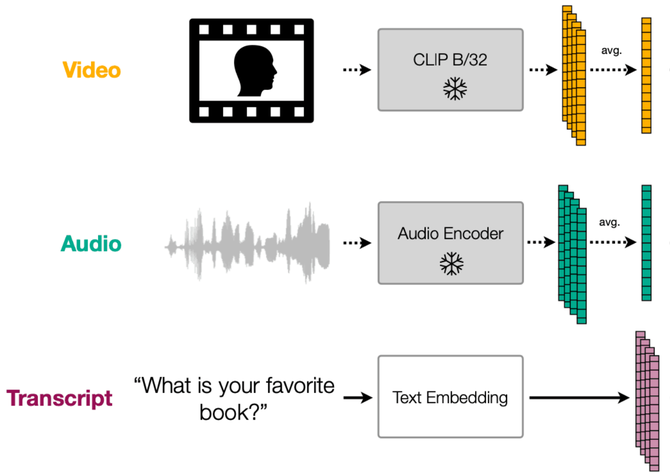
(a) (b)

Credit: Hwang et al., Sensors 2022

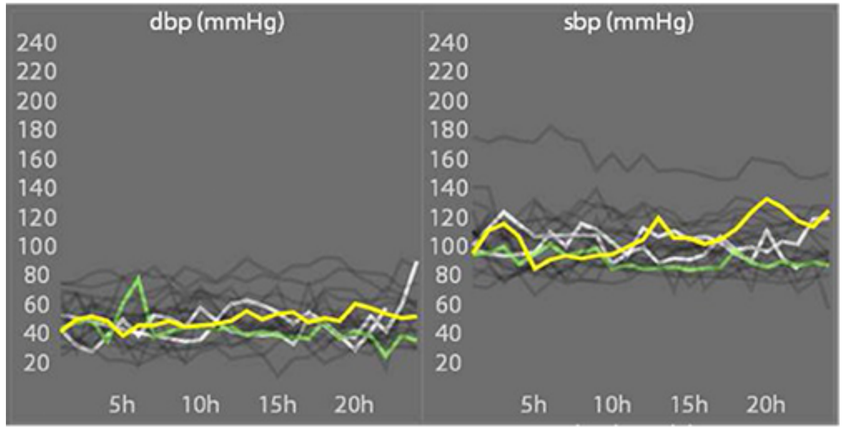
Real Data is often Multiple Trajectories

Many more scenarios outside control

- LLMs and VLMs
  - **Data**: independent documents, images, and videos
- Medical time series
  - **Data**: usage from large amount of independent patients



Credit: Palaskar et al., Interspeech 2024

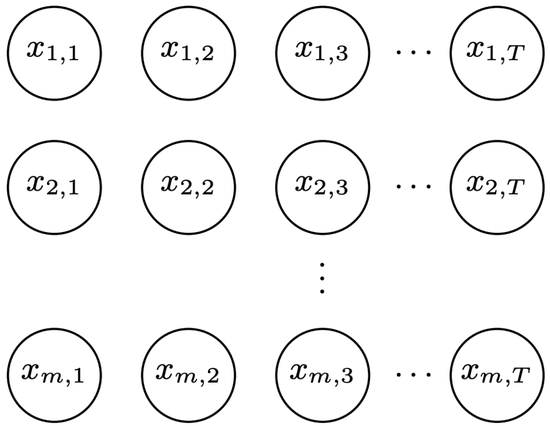


Credit: Scheer et al., JMIR 2022

What is Multi-trajectory

Credit: Tu et al., JMLR 2024, and ChatGPT

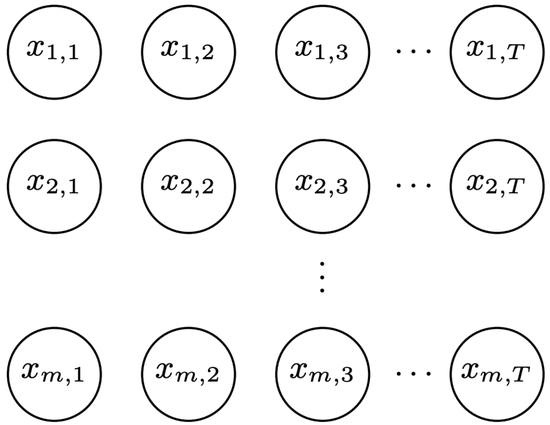
What is Multi-trajectory



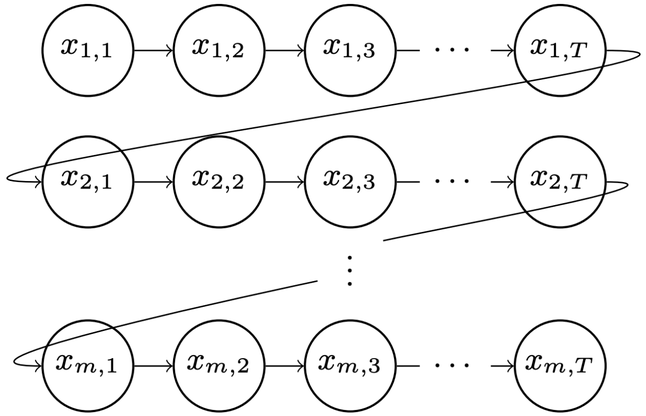
i.i.d.

Credit: Tu et al., JMLR 2024, and ChatGPT

What is Multi-trajectory



i.i.d.

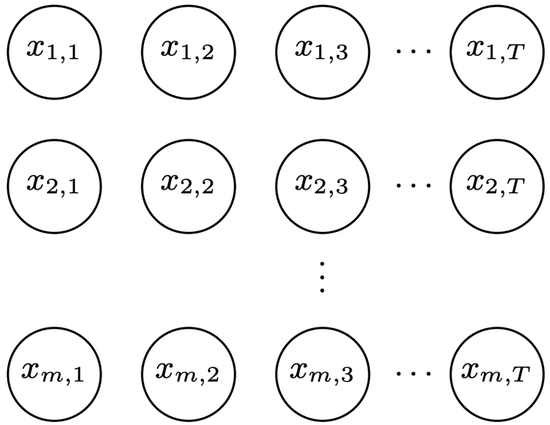


a single trajectory

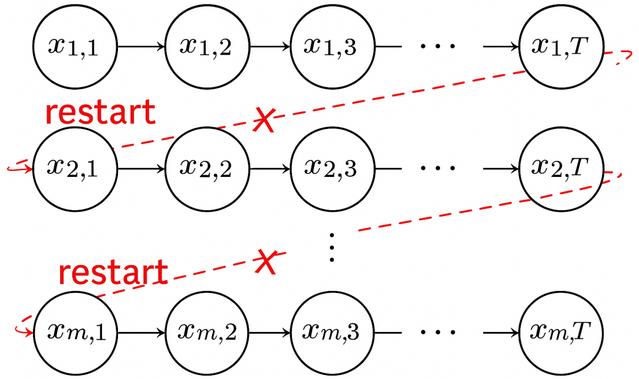
(assuming Markovian for illustration purposes)

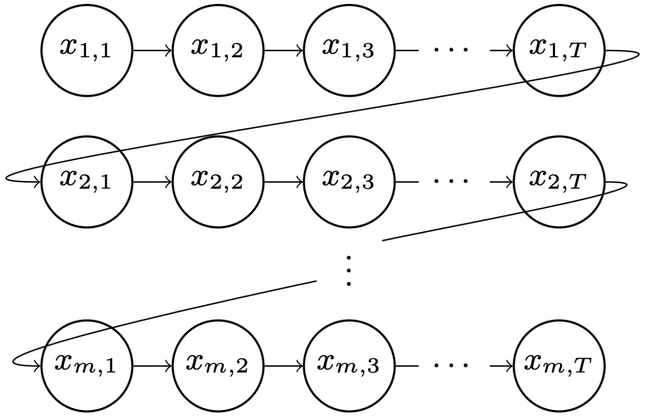
Credit: Tu et al., JMLR 2024, and ChatGPT

What is Multi-trajectory



i.i.d.



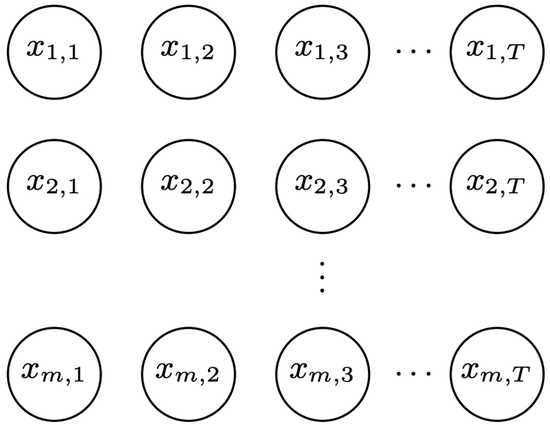


a single trajectory

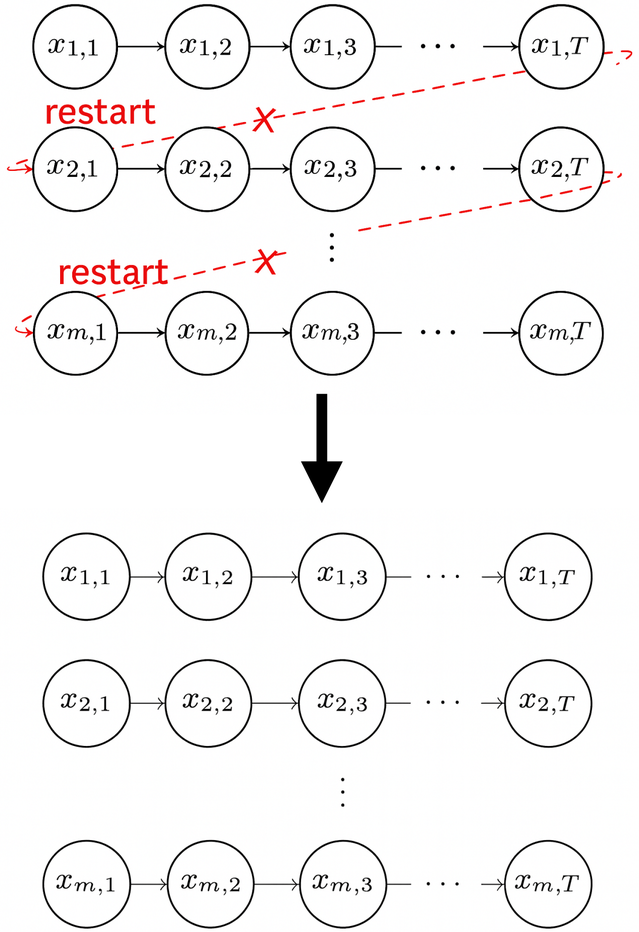
(assuming Markovian for illustration purposes)

Credit: Tu et al., JMLR 2024, and ChatGPT

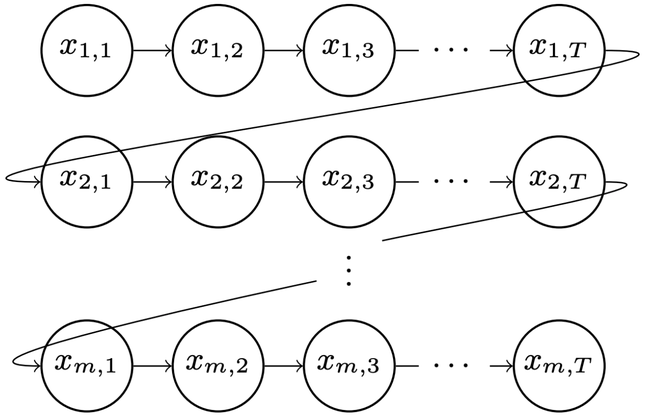
What is Multi-trajectory



i.i.d.



Multi-trajectory



a single trajectory

(assuming Markovian for illustration purposes)

Credit: Tu et al., JMLR 2024, and ChatGPT

Problem Setup and Notation

- **Data**: $z_{1:T}^{(i)}$, $i = 1, \ldots, m$, such that:
  - For each i, $z_{1:T}^{(i)} \sim p_{\theta_\star}(\,\cdot\,)$, where the true parameter $\theta_\star \in \Theta \subset \mathbb{R}^p$, Θ parameterize **temporal dependencies**
    - **Note:** $p_{\theta_\star}(\,\cdot\,)$ is a distribution on length-T trajectories
    - **Example:** Markov chain with transition $M_\star$ and initial distribution π**,**

      $$p_{M_\star}(z_{1:T}) = \pi(z_1) \prod_{i=1}^{T-1} M_\star(z_i, z_{i+1})$$

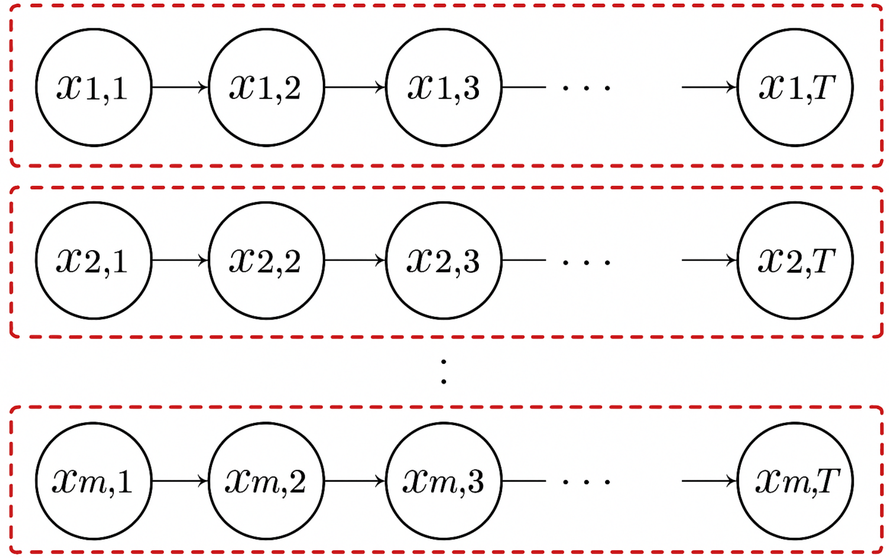
  - For $i \neq j$, $z_{1:T}^{(i)}$ and $z_{1:T}^{(j)}$ are **independent**

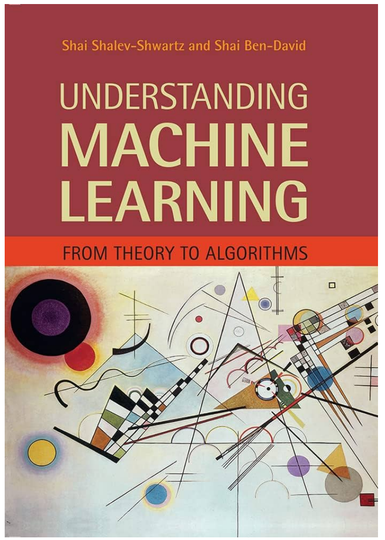
Problem Setup

- **Maximum Likelihood Estimator:** $\theta_{\mathsf{MLE}} \in \arg\max_{\theta \in \Theta} \sum_{i=1}^{m} \log p_\theta \left(z_{1:T}^{(i)} \right)$
  - **Error metric**: the $\mathscr{L}^2$-norm $\left\| \theta_{\mathsf{MLE}} - \theta_\star \right\|$
  - **Examples:**
    - Least squares in LDS with Gaussian noise
    - Kalman filter
- **Goal:** non-asymptotic bound on $\left\| \theta_{\mathsf{MLE}} - \theta_\star \right\|$
  - "With probability at least $1 - \delta$, $\left\| \theta_{\mathsf{MLE}} - \theta_\star \right\| \leq \ldots$"

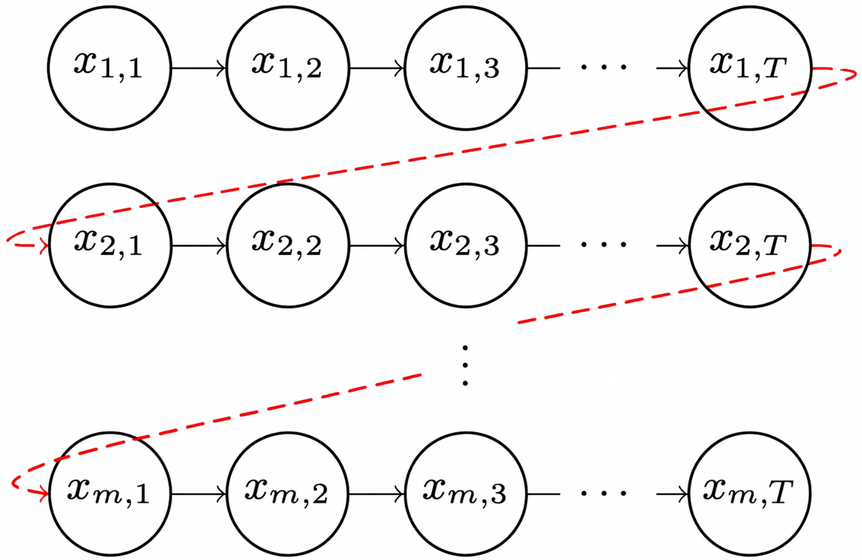
Baselines

- **Approach 1**: reduction to **i.i.d.** estimation
  - **Data**: $z_{1:T}^{(1)}, z_{1:T}^{(2)}, \ldots, z_{1:T}^{(m)}$
    - treat each trajectory as one observation
  - **Rate**: Excess Risk $\lesssim \sqrt{\frac{p}{m}}$ by **naïvely** applying non-asymptotic i.i.d. estimation theory
    - What is **Excess Risk**?
  - **Problem**: rate is not adapted to total number of observations





Baselines



- **Approach 2**: reduction to **single trajectory** estimation
  - **Data**: $z_1^{(1)}, \ldots, z_T^{(1)}, z_1^{(2)}, \ldots, z_1^{(m)}, \ldots, z_T^{(m)}$
    - One sequence of length mT
  - **Rate**: Excess Risk $\lesssim \sqrt{\dfrac{p}{mT/\min(\kappa, T)}}$, where κ = mixing time / “block length”
    - mixing time: how fast processes forget the past
  - **Problem**: mixing assumption + sample size deflation (by κ or T)
    - **Non-mixing dynamics**: unstable LDS, mixture of dynamics, data on the internet

What's Achievable

- Central Limit Theorem (CLT) says

$$\left\| \theta_{\text{MLE}} - \theta_\star \right\| = O_p\left(\sqrt{\frac{\text{tr}\left(I_T\left(\theta_\star\right)^{-1}\right)}{m}}\right)$$

- $I_T\left(\theta_\star\right)^{-1}$ is the inverse **Fisher information**; it measures estimation complexity
  - For length-T trajectories, $\text{tr}\left(I_T\left(\theta_\star\right)^{-1}\right)$ scales as **problem complexity**$/T$
  - Thus RHS is $O_p\left(\sqrt{\frac{\text{Complexity}}{mT}}\right)$; Good!
- But CLT only provides an asymptotic rate

Q: Can we get a non-asymptotic

$$\sqrt{\text{Complexity}\left(\theta_\star, \Theta\right) \Big/ mT}$$

decaying rate?

Hellinger Localization:

Yes, up to a log factor

Hellinger Localization 1

Intuition: multi-trajectory allows us to do i.i.d. estimation

Theorem 1 [S**Z**ZMT25]: Given multiple trajectories $\left(z_{1:T}^{(i)}\right)_{i=1}^{m}$, with $z_{1:T}^{(i)} \sim p_{\theta_\star}(\,\cdot\,)$, $\theta_\star \in \Theta \subset \mathbb{R}^p$, then with probability at least $1-\delta$,

$$d_H^2\left(p_{\theta_{\mathsf{MLE}}}, p_{\theta_\star}\right) \lesssim \frac{p \log(C(\Theta) mT/\delta)}{m}$$

Inspired by Foster et al., NeurIPS 2024, exact statement in paper

Recall: $p_{\theta_{\mathsf{MLE}}}$ and $p_{\theta_\star}$ are distributions on length-T trajectories

What's special: squared Hellinger distance is an f**-divergence**

Background: f-divergence

- $D_f(p\|q) := \int f\left(\frac{p(x)}{q(x)}\right) q(x)dx$
  - KL-divergence $D_{KL}(p\|q)$: D_f with $f(x) = \log(x)$
  - squared Hellinger distance $d_H^2(p,q)$: D_f with $f(x) = \frac{1}{2}(\sqrt{x}-1)^2$

Tensorization

- A feature of KL-divergence between **length-T** distributions
  - $$D_{KL}\left(p_{\theta_\star} \,\middle\|\, p_{\theta_{\mathsf{MLE}}}\right) = \sum_{t=1}^{T} \mathbb{E}_{z_{1:t-1}}\left[D_{KL}\left(p_{\theta_\star}(\cdot \mid z_{1:t-1}) \,\middle\|\, p_{\theta_{\mathsf{MLE}}}(\cdot \mid z_{1:t-1})\right)\right]$$
  - Example: for LDS with standard Gaussian noise, the summand is $$\mathbb{E}_{z_t}\left[D_{KL}\left(\mathcal{N}\left(A_\star z_t, 1\right) \,\middle\|\, \mathcal{N}\left(A_{\mathsf{MLE}} z_t, 1\right)\right)\right] \gtrsim \lambda_{\min}\left(\mathbb{E}\left[z_t z_t^\mathsf{T}\right]\right) \left\| A_\star - A_{\mathsf{MLE}} \right\|_F^2$$
  - Can often show $D_{KL}\left(p_{\theta_\star} \,\middle\|\, p_{\theta_{\mathsf{MLE}}}\right)$ scales with $T \left\| \theta_\star - \theta_{\mathsf{MLE}} \right\|^2$

Tensorization is Nice

- **Idea 1**: replace squared Hellinger in Theorem 1 with KL, then we are done?
  - Answer: not true without strong assumptions on tail behavior of $\log p_{\theta_\star}$
- **Idea 2**: can $d_H^2\left(p_{\theta_{\mathsf{MLE}}}, p_{\theta_\star}\right)$ tensorize similarly?
  - Answer: no in general; **but yes locally!**

f-divergence Tensorizes Locally

- **Taylor expansion:**
  - $d_H^2\left(p_{\theta_{\mathsf{MLE}}}, p_{\theta_\star}\right) = \frac{1}{4} \left\| \theta_{\mathsf{MLE}} - \theta_\star \right\|_{I_T(\theta_\star)}^2 + o\left(\left\| \theta_{\mathsf{MLE}} - \theta_\star \right\|^2 \right)$
- **Fisher information tensorizes:**
  - $I_T(\theta_\star) = \sum_{t=1}^{T} \mathbb{E}_{z_{1:t}} \left[-\nabla_\theta^2 \log p_\theta(z_t \,|\, z_{1:t-1}) \Big|_{\theta=\theta_\star} \right] =: \sum_{t=1}^{T} \mathbb{E}_{z_{1:t-1}} \left[I\left(\theta_\star \,|\, z_{1:t-1}\right) \right]$
  - $\lambda_{\min}\left(I_T(\theta_\star)\right) \geq \sum_{t=1}^{T} \lambda_{\min}\left(\mathbb{E}_{z_{1:t-1}} \left[I\left(\theta_\star \,|\, z_{1:t-1}\right) \right] \right) \approx \frac{T}{\textbf{complexity}}$
  - Unweighted Euclidean: $d_H^2\left(p_{\theta_{\mathsf{MLE}}}, p_{\theta_\star}\right) \gtrsim \lambda_{\min}\left(I_T\left(\theta_\star\right)\right) \left\| \theta_{\mathsf{MLE}} - \theta_\star \right\|^2 \approx \frac{T \left\| \theta_{\mathsf{MLE}} - \theta_\star \right\|^2}{\text{complexity}}$
- **Takeaway:** we just need a non-asymptotic Taylor expansion!

Hellinger Localization 2

Intuition: square Hellinger distance tensorizes locally

Theorem 2 [S**Z**ZMT25]: Assuming for all θ in a neighborhood around $\theta_\star$, $\nabla_\theta \log p_\theta$ and $\nabla^2_\theta \log p_\theta$ satisfy some moment regularity conditions, and $I_T(\theta)$ satisfies a Lipschitz-like condition, and further assuming $d_H^2\left(p_{\theta_\star}, p_{\theta_{\mathsf{MLE}}}\right)$ is small enough, we have

$$d_H^2\left(p_{\theta_\star}, p_{\theta_{\mathsf{MLE}}}\right) \asymp \left\| \theta_\star - \theta_{\mathsf{MLE}} \right\|^2_{I_T(\theta_\star)}$$

See paper for exact conditions

Putting the Pieces Together

- **Theorem 1:** we can make $d_H^2\left(p_{\theta_\star}, p_{\theta_{\text{MLE}}}\right)$ small by requiring large m, **for free!**
- **Theorem 2:** when $d_H^2\left(p_{\theta_\star}, p_{\theta_{\text{MLE}}}\right)$ is small, it scales like $T \left\| \theta_{\text{MLE}} - \theta_\star \right\|^2$
- **Together:** when $m \gtrsim \text{poly}(\theta_\star, \Theta, T)$, with probability at least $1 - \delta$

$$\left\| \theta_\star - \theta_{\text{MLE}} \right\| \lesssim \sqrt{\frac{\text{Complexity}(\theta_\star, \Theta)}{mT} \log(C(\Theta) mT/\delta)}$$

- In many cases, $\text{Complexity}(\theta_\star, \Theta) = p^2$; generally, can avoid dependence on Θ
- Often need $T \gtrsim \text{poly}(\theta_\star, \Theta)$ for regularity conditions; commonly $T \gtrsim \text{poly}(p)$ suffices

Example: Sinusoidal GLM Dynamics

- **System:** $z_{t+1} = \sin\left(A_\star z_t\right) + w_t,\ w_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
  - $A_\star \in \mathbb{A} \subset \mathbb{R}^{d\times d}$ is the parameter of interest ($p = d^2$)
- **Hellinger localization:** assuming $\mathbb{A}$ is compact with radius R under $\|\cdot\|_F$, where $R \geq 1$, $A_\star$ is fully coupled, and $m \gtrsim \text{poly}(d)T$, $T \gtrsim \text{poly}(d)$, then

$$\left\| A_{\mathsf{MLE}} - A_\star \right\|_F \lesssim \frac{d^2}{\sqrt{mT}}\sqrt{\log(RdmT/\delta)}$$

Example: Sinusoidal GLM Dynamics cont.

- **Prior works:** only considered **single trajectory** setting of GLM $z_{t+1} = \phi(A_\star z_t) + w_t$, and assumed:
  - ϕ is expansive ($\sin$ **is not**)
  - A satisfies certain stability condition stronger than $\rho(A) < 1$
  - More complex trajectory regularity assumptions

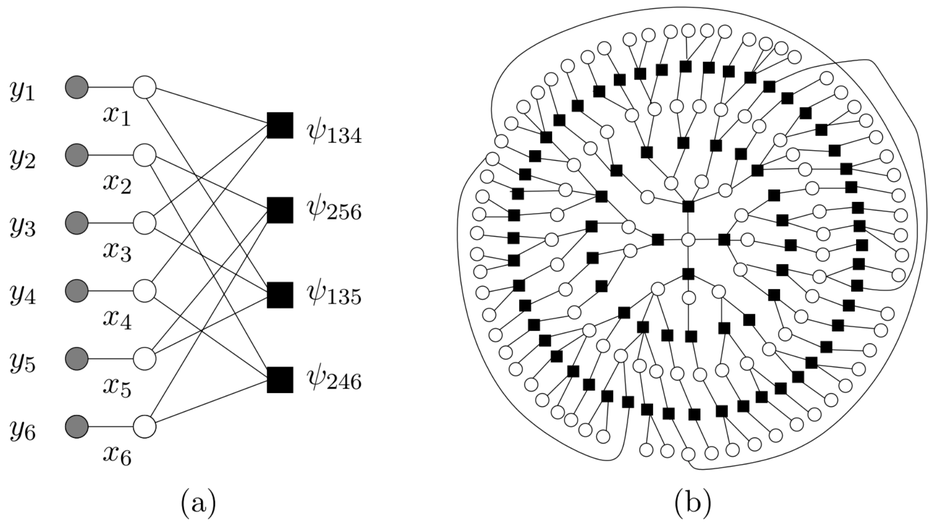
- **Message:** multi-trajectories waives the need for mixing / stability
  - Hellinger localization captures that

More Applications

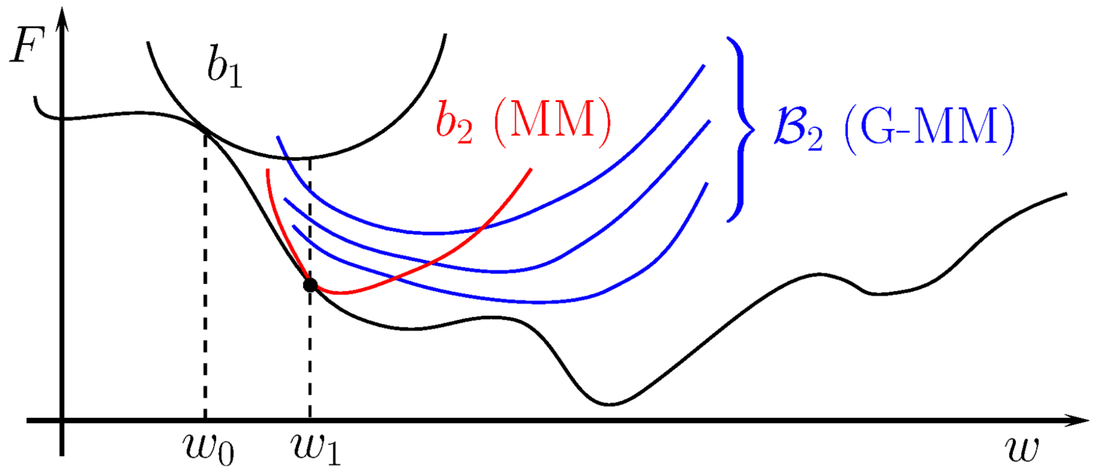
- **Mixture of Markov chains**
- **Dependent regression with heavy-tailed noise**
- **Next token prediction with linear attention**
- **More details in the paper**

Future Directions

- Instantiation in control such as state estimation
- Approximate MLE with EM or variational inference
- Multiple trajectory with non-uniform length
- Data with non-temporal dependence



Credit: Wainwright and Jordan 2008



Credit: Naderi et al., ICML 2019

Thank you!