Information-theoretic Bounds on Learning Network Dynamics

Andrea Montanari
[with Jose Bento and Morteza Ibrahimi]

Stanford University

December 11, 2015

What is this talk about?

How long should I observe a system before I can learn its dynamics?

- Information-theoretic tools
- Tutorial (no Information Theory background required)

What is this talk about?

How long should I observe a system before I can learn its dynamics?

- Information-theoretic tools
- Tutorial (no Information Theory background required)

Outline

General approach

General estimation in Hamming metric

- Parameter space Θ, $|\Theta| < \infty$
- Family of probability measures $(\mathrm{P}_\theta)_{\theta \in \Theta}$ on space $\mathcal{X}$

Estimator:

$$\begin{aligned} \widehat{\theta} : \mathcal{X} &\to \Theta \\ X &\mapsto \widehat{\theta}(X) \end{aligned}$$

Minimax risk

$$\mathrm{R}_M(\Theta) = \inf_{\widehat{\theta}} \max_{\theta \in \Theta} \mathrm{P}_\theta\Big(\widehat{\theta}(X) \neq \theta\Big)$$

General estimation in Hamming metric

- Parameter space Θ, $|\Theta| < \infty$
- Family of probability measures $(\mathrm{P}_\theta)_{\theta\in\Theta}$ on space $\mathcal{X}$

Estimator:

$$\begin{aligned}\widehat{\theta} : \mathcal{X} &\to \Theta \\ \boldsymbol{X} &\mapsto \widehat{\theta}(\boldsymbol{X})\end{aligned}$$

Minimax risk

$$\mathrm{R}_M(\Theta) = \inf_{\widehat{\theta}} \max_{\theta\in\Theta} \mathrm{P}_\theta\left(\widehat{\theta}(\boldsymbol{X}) \neq \theta\right)$$

General estimation in Hamming metric

- Parameter space Θ, $|\Theta| < \infty$
- Family of probability measures $(\mathrm{P}_\theta)_{\theta\in\Theta}$ on space $\mathcal{X}$

Estimator:

$$\begin{aligned}\widehat{\theta} : \mathcal{X} &\to \Theta \\ \boldsymbol{X} &\mapsto \widehat{\theta}(\boldsymbol{X})\end{aligned}$$

Minimax risk

$$\mathrm{R}_M(\Theta) = \inf_{\widehat{\theta}} \max_{\theta\in\Theta} \mathrm{P}_\theta\Big(\widehat{\theta}(\boldsymbol{X}) \neq \theta\Big)$$

Entropy and conditional entropy

$X \sim p_X(\,\cdot\,)$, probability measure on $\mathcal{X}$, finite

$$H(X) = -\sum_{x\in\mathcal{X}} p_X(x)\log p_X(x)\,.$$

$(X,Y) \sim p_{X,Y}(\,\cdot\,,\,\cdot\,)$, probability measure on $\mathcal{X}\times\mathcal{Y}$, finite

$$H(X|Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p_{X,Y}(x,y)\log p_{X|Y}(x|y)\,.$$

Entropy and conditional entropy

$X \sim p_X(\,\cdot\,)$, probability measure on $\mathcal{X}$, finite

$$H(X) = -\sum_{x\in\mathcal{X}} p_X(x)\log p_X(x)\,.$$

$(X, Y) \sim p_{X,Y}(\,\cdot\,,\,\cdot\,)$, probability measure on $\mathcal{X}\times\mathcal{Y}$, finite

$$H(X|Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p_{X,Y}(x,y)\log p_{X|Y}(x|y)\,.$$

Two properties

Chain rule

$$H(X, Y) = H(X|Y) + H(Y)$$

Sub-addittivity

$$H(X, Y) \leq H(X) + H(Y)$$

[with = iff independent]

Two properties

Chain rule

$$H(X, Y) = H(X|Y) + H(Y)$$

Sub-addittivity

$$H(X, Y) \leq H(X) + H(Y)$$

[with = iff independent]

Information-theoretic approach

- Assume $\theta \sim \mathbb{P}$:

$$\max_{\theta \in \Theta} \mathrm{P}_\theta\Big(\widehat{\theta}(\boldsymbol{X}) \neq \theta\Big) \geq \mathbb{P}\Big(\widehat{\theta}(\boldsymbol{X}) \neq \theta\Big)$$

$$\equiv \sum_{\theta_0 \in \Theta} \mathbb{P}(\theta = \theta_0)\, \mathrm{P}_{\theta_0}\Big(\widehat{\theta}(\boldsymbol{X}) \neq \theta_0\Big)$$

- Fano's inequality

$$\mathbb{P}\Big(\widehat{\theta}(\boldsymbol{X}) \neq \theta\Big) \geq \frac{H(\theta|\boldsymbol{X}) - 1}{\log|\Theta|}$$

For any distribution $\mathbb{P}$,

$$\mathrm{R}_M(\Theta) \geq \frac{H(\theta|\boldsymbol{X}) - 1}{\log|\Theta|}.$$

Information-theoretic approach

- Assume $\theta \sim \mathbb{P}$:

$$\max_{\theta \in \Theta} \mathrm{P}_\theta\Big(\widehat{\theta}(\boldsymbol{X}) \neq \theta\Big) \geq \mathbb{P}\Big(\widehat{\theta}(\boldsymbol{X}) \neq \theta\Big)$$

$$\equiv \sum_{\theta_0 \in \Theta} \mathbb{P}(\theta = \theta_0)\, \mathrm{P}_{\theta_0}\Big(\widehat{\theta}(\boldsymbol{X}) \neq \theta_0\Big)$$

- Fano's inequality

$$\mathbb{P}\Big(\widehat{\theta}(\boldsymbol{X}) \neq \theta\Big) \geq \frac{H(\theta|\boldsymbol{X}) - 1}{\log|\Theta|}$$

For any distribution $\mathbb{P}$,

$$\mathrm{R}_M(\Theta) \geq \frac{H(\theta|\boldsymbol{X}) - 1}{\log|\Theta|}.$$

Information-theoretic approach

- Assume $\theta \sim \mathbb{P}$:

$$\max_{\theta \in \Theta} \mathrm{P}_\theta\Big(\widehat{\theta}(\boldsymbol{X}) \neq \theta\Big) \geq \mathbb{P}\Big(\widehat{\theta}(\boldsymbol{X}) \neq \theta\Big)$$

$$\equiv \sum_{\theta_0 \in \Theta} \mathbb{P}(\theta = \theta_0)\, \mathrm{P}_{\theta_0}\Big(\widehat{\theta}(\boldsymbol{X}) \neq \theta_0\Big)$$

- Fano's inequality

$$\mathbb{P}\Big(\widehat{\theta}(\boldsymbol{X}) \neq \theta\Big) \geq \frac{H(\theta|\boldsymbol{X}) - 1}{\log |\Theta|}$$

For any distribution $\mathbb{P}$,

$$\mathrm{R}_M(\Theta) \geq \frac{H(\theta|\boldsymbol{X}) - 1}{\log |\Theta|}\,.$$

Some simple examples

Toy example

- $\Theta = [2^m] \equiv \{1, 2, \dots, 2^m\}$
- $\boldsymbol{X} = (X_1, \dots, X_T)$, $X_t \in [2^m]$ independent given θ.
-

$$\mathrm{P}_\theta(X_t = \ell) = \begin{cases} 1 - p(1 - 2^{-m}) & \text{if } \ell = \theta, \\ p2^{-m} & \text{otherwise.} \end{cases}$$

How to choose $\mathbb{P}$?

Uniform

Toy example

- $\Theta = [2^m] \equiv \{1, 2, \dots, 2^m\}$
- $\boldsymbol{X} = (X_1, \dots, X_T)$, $X_t \in [2^m]$ independent given θ.
- $$\mathrm{P}_\theta(X_t = \ell) = \begin{cases} 1 - p(1 - 2^{-m}) & \text{if } \ell = \theta, \\ p2^{-m} & \text{otherwise.} \end{cases}$$

How to choose $\mathbb{P}$?

Uniform

Evaluating the conditional entropy

$$H(\theta) = \log|\Theta| = m$$

$$\begin{aligned}
H(\theta|\boldsymbol{X}) &= H(\theta, \boldsymbol{X}) - H(\boldsymbol{X}) && \text{(chain rule)}\\
&= H(\boldsymbol{X}|\theta) + H(\theta) - H(\boldsymbol{X}) && \text{(chain rule)}\\
&= \sum_{i=1}^{T} H(X_t|\theta) - H(\boldsymbol{X}) + m && \text{(chain rule)}\\
&\geq \sum_{t=1}^{T} \{H(X_t|\theta) - H(X_t)\} + m && \text{(subadditivity)}\\
&= -T\{H(X_1) - H(X_1|\theta)\} + m
\end{aligned}$$

Evaluating the conditional entropy

$$H(\theta) = \log|\Theta| = m$$

$$\begin{aligned}
H(\theta|\boldsymbol{X}) &= H(\theta, \boldsymbol{X}) - H(\boldsymbol{X}) && \text{(chain rule)}\\
&= H(\boldsymbol{X}|\theta) + H(\theta) - H(\boldsymbol{X}) && \text{(chain rule)}\\
&= \sum_{i=1}^{T} H(X_t|\theta) - H(\boldsymbol{X}) + m && \text{(chain rule)}\\
&\geq \sum_{t=1}^{T} \{H(X_t|\theta) - H(X_t)\} + m && \text{(subadditivity)}\\
&= -T\{H(X_1) - H(X_1|\theta)\} + m
\end{aligned}$$

Evaluating the conditional entropy

$$H(\theta) = \log|\Theta| = m$$

$$\begin{aligned}
H(\theta|\boldsymbol{X}) &= H(\theta, \boldsymbol{X}) - H(\boldsymbol{X}) && \text{(chain rule)}\\
&= H(\boldsymbol{X}|\theta) + H(\theta) - H(\boldsymbol{X}) && \text{(chain rule)}\\
&= \sum_{i=1}^{T} H(X_t|\theta) - H(\boldsymbol{X}) + m && \text{(chain rule)}\\
&\geq \sum_{t=1}^{T} \{H(X_t|\theta) - H(X_t)\} + m && \text{(subadditivity)}\\
&= -T\{H(X_1) - H(X_1|\theta)\} + m
\end{aligned}$$

Evaluating the conditional entropy

$$H(\theta) = \log |\Theta| = m$$

$$\begin{aligned} H(\theta|\boldsymbol{X}) &= H(\theta, \boldsymbol{X}) - H(\boldsymbol{X}) && \text{(chain rule)} \\ &= H(\boldsymbol{X}|\theta) + H(\theta) - H(\boldsymbol{X}) && \text{(chain rule)} \\ &= \sum_{i=1}^{T} H(X_t|\theta) - H(\boldsymbol{X}) + m && \text{(chain rule)} \\ &\geq \sum_{t=1}^{T} \{H(X_t|\theta) - H(X_t)\} + m && \text{(subadditivity)} \\ &= -T\{H(X_1) - H(X_1|\theta)\} + m \end{aligned}$$

Evaluating the conditional entropy

$$H(\theta) = \log|\Theta| = m$$

$$\begin{aligned}
H(\theta|\boldsymbol{X}) &= H(\theta, \boldsymbol{X}) - H(\boldsymbol{X}) && \text{(chain rule)}\\
&= H(\boldsymbol{X}|\theta) + H(\theta) - H(\boldsymbol{X}) && \text{(chain rule)}\\
&= \sum_{i=1}^{T} H(X_t|\theta) - H(\boldsymbol{X}) + m && \text{(chain rule)}\\
&\geq \sum_{t=1}^{T} \{H(X_t|\theta) - H(X_t)\} + m && \text{(subadditivity)}\\
&= -T\{H(X_1) - H(X_1|\theta)\} + m
\end{aligned}$$

Evaluating the conditional entropy

$$H(\theta) = \log|\Theta| = m$$

$$\begin{aligned}
H(\theta|\boldsymbol{X}) &= H(\theta, \boldsymbol{X}) - H(\boldsymbol{X}) && \text{(chain rule)}\\
&= H(\boldsymbol{X}|\theta) + H(\theta) - H(\boldsymbol{X}) && \text{(chain rule)}\\
&= \sum_{i=1}^{T} H(X_t|\theta) - H(\boldsymbol{X}) + m && \text{(chain rule)}\\
&\geq \sum_{t=1}^{T} \{H(X_t|\theta) - H(X_t)\} + m && \text{(subadditivity)}\\
&= -T\{H(X_1) - H(X_1|\theta)\} + m
\end{aligned}$$

Fano's inequality

$$\mathrm{R}_M(\Theta) \geq \frac{H(\theta|\boldsymbol{X}) - 1}{\log|\Theta|}$$
$$\geq 1 - \frac{T}{m}[H(X_1) - H(X_1|\theta)] - m^{-1}.$$

Toy Theorem

Minimax error probability larger than $(1/2) - (1/m)$ unless

$$T \geq \frac{m}{2[H(X_1) - H(X_1|\theta)]} = \frac{1}{2}\,\frac{\log|\Theta|}{I(X_1;\theta)}$$

Fano's inequality

$$\mathrm{R}_M(\Theta) \geq \frac{H(\theta|\boldsymbol{X}) - 1}{\log|\Theta|}$$
$$\geq 1 - \frac{T}{m}[H(X_1) - H(X_1|\theta)] - m^{-1}\,.$$

Toy Theorem

Minimax error probability larger than $(1/2) - (1/m)$ unless

$$T \geq \frac{m}{2[H(X_1) - H(X_1|\theta)]} = \frac{1}{2}\,\frac{\log|\Theta|}{I(X_1;\theta)}$$

Fano's inequality

$$\mathrm{R}_M(\Theta) \geq \frac{H(\theta|\boldsymbol{X}) - 1}{\log|\Theta|}$$
$$\geq 1 - \frac{T}{m}[H(X_1) - H(X_1|\theta)] - m^{-1}\,.$$

Toy Theorem

Minimax error probability larger than $(1/2) - (1/m)$ unless

$$T \geq \frac{m}{2[H(X_1) - H(X_1|\theta)]} = \frac{1}{2}\,\frac{\log|\Theta|}{I(X_1;\theta)}$$

Fano's inequality

$$\mathrm{R}_M(\Theta) \geq \frac{H(\theta|\boldsymbol{X}) - 1}{\log|\Theta|}$$
$$\geq 1 - \frac{T}{m}[H(X_1) - H(X_1|\theta)] - m^{-1}.$$

Toy Theorem

Minimax error probability larger than $(1/2) - (1/m)$ unless

$$T \geq \frac{m}{2[H(X_1) - H(X_1|\theta)]} = \frac{1}{2}\frac{\log|\Theta|}{I(X_1;\theta)}$$

Interpretation

$$T \geq \frac{1}{2} \frac{\log |\Theta|}{I(X_1;\theta)}$$

- Each observation yields $I(X_1;\theta) = H(\theta) - H(\theta|X_1)$ bits
- Need to accumulate $\log|\Theta|$ bits

A more 'dynamical' example

$t \in \{1, \dots, T-1\}$:

$$X_{t+1} = \sqrt{1-\theta^2}\, X_t + \theta\, Z_t$$

- $\theta \in \Theta \subseteq (0,1)$
- $Z_t \sim_{i.i.d.} \mathrm{N}(0,1)$
- $X_t \sim \mathrm{N}(0,1)$ **dependent**

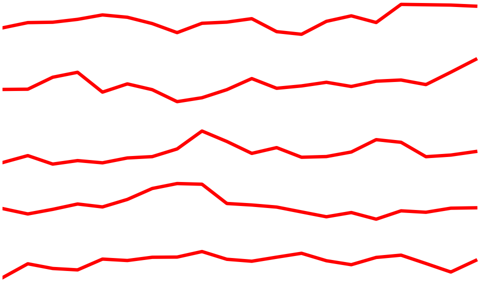
A more 'dynamical' example

$t \in \{1, \ldots, T-1\}$:

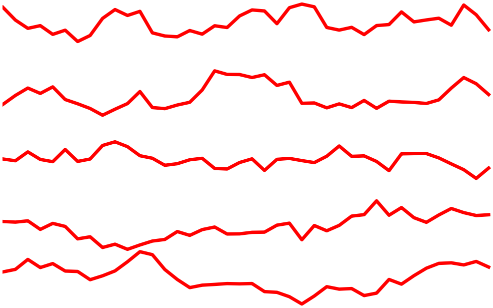
$$X_{t+1} = \sqrt{1-\theta^2}\, X_t + \theta\, Z_t$$

- $\theta \in \Theta \subseteq (0, 1)$
- $Z_t \sim_{i.i.d.} \mathrm{N}(0, 1)$
- $X_t \sim \mathrm{N}(0, 1)$ **dependent**

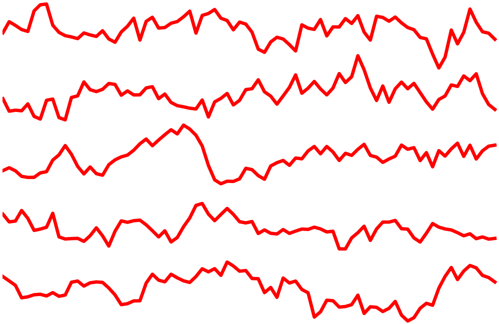
$|\Theta| = 5, \; T = 20$



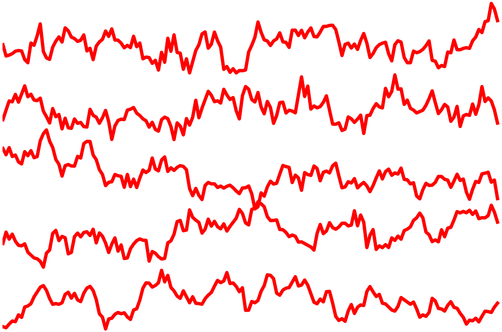
$|\Theta| = 5, \ T = 40$



$|\Theta| = 5, \ T = 80$



$|\Theta| = 5,\ T = 160$



Conditional entropy

$$
\begin{aligned}
H(\theta|\boldsymbol{X}) &= H(\theta) + h(\boldsymbol{X}|\theta) - h(\boldsymbol{X}) \\
&= H(\theta) + \sum_{t=1}^{T} h(X_t|\theta, X_{t-1}) - \sum_{t=1}^{T} h(X_t|X_1, \dots, X_{t-1}) \\
&\geq H(\theta) + \sum_{t=1}^{T} \{h(X_t|\theta, X_{t-1}) - h(X_t|X_{t-1})\} \\
&= H(\theta) - T\{h(X_t|\theta, X_{t-1}) - h(X_t|X_{t-1})\} \qquad \text{(stationarity)} \\
&= H(\theta) - T\, I(\theta; X_t|X_{t-1})
\end{aligned}
$$

Conditional entropy

$$
\begin{aligned}
H(\theta|\boldsymbol{X}) &= H(\theta) + h(\boldsymbol{X}|\theta) - h(\boldsymbol{X}) \\
&= H(\theta) + \sum_{t=1}^{T} h(X_t|\theta, X_{t-1}) - \sum_{t=1}^{T} h(X_t|X_1, \dots, X_{t-1}) \\
&\geq H(\theta) + \sum_{t=1}^{T} \{h(X_t|\theta, X_{t-1}) - h(X_t|X_{t-1})\} \\
&= H(\theta) - T\{h(X_t|\theta, X_{t-1}) - h(X_t|X_{t-1})\} \qquad \text{(stationarity)} \\
&= H(\theta) - T\, I(\theta; X_t|X_{t-1})
\end{aligned}
$$

Conditional entropy

$$
\begin{aligned}
H(\theta|\boldsymbol{X}) &= H(\theta) + h(\boldsymbol{X}|\theta) - h(\boldsymbol{X}) \\
&= H(\theta) + \sum_{t=1}^{T} h(X_t|\theta, X_{t-1}) - \sum_{t=1}^{T} h(X_t|X_1, \dots, X_{t-1}) \\
&\geq H(\theta) + \sum_{t=1}^{T} \{h(X_t|\theta, X_{t-1}) - h(X_t|X_{t-1})\} \\
&= H(\theta) - T\{h(X_t|\theta, X_{t-1}) - h(X_t|X_{t-1})\} \qquad \text{(stationarity)} \\
&= H(\theta) - T\, I(\theta; X_t|X_{t-1})
\end{aligned}
$$

Conditional entropy

$$
\begin{aligned}
H(\theta|\boldsymbol{X}) &= H(\theta) + h(\boldsymbol{X}|\theta) - h(\boldsymbol{X}) \\
&= H(\theta) + \sum_{t=1}^{T} h(X_t|\theta, X_{t-1}) - \sum_{t=1}^{T} h(X_t|X_1, \dots, X_{t-1}) \\
&\geq H(\theta) + \sum_{t=1}^{T} \{h(X_t|\theta, X_{t-1}) - h(X_t|X_{t-1})\} \\
&= H(\theta) - T\{h(X_t|\theta, X_{t-1}) - h(X_t|X_{t-1})\} \qquad \text{(stationarity)} \\
&= H(\theta) - T\, I(\theta; X_t|X_{t-1})
\end{aligned}
$$

Conditional entropy

$$
\begin{aligned}
H(\theta|\boldsymbol{X}) &= H(\theta) + h(\boldsymbol{X}|\theta) - h(\boldsymbol{X}) \\
&= H(\theta) + \sum_{t=1}^{T} h(X_t|\theta, X_{t-1}) - \sum_{t=1}^{T} h(X_t|X_1, \dots, X_{t-1}) \\
&\geq H(\theta) + \sum_{t=1}^{T} \{h(X_t|\theta, X_{t-1}) - h(X_t|X_{t-1})\} \\
&= H(\theta) - T\{h(X_t|\theta, X_{t-1}) - h(X_t|X_{t-1})\} \qquad \text{(stationarity)} \\
&= H(\theta) - T\, I(\theta; X_t|X_{t-1})
\end{aligned}
$$

A general lemma

Lemma

If P_θ is a stationary Markov process, then error probability $\geq (1/2) - (1/m)$ unless

$$T \geq \frac{1}{2} \frac{\log|\Theta|}{I(\theta; X_t | X_{t-1})} .$$

- $I(\theta; X_t | X_{t-1})$: 'new' information

A general lemma

Lemma

If P_θ *is a stationary Markov process, then error probability* $\geq (1/2) - (1/m)$ *unless*

$$T \geq \frac{1}{2} \frac{\log|\Theta|}{I(\theta; X_t | X_{t-1})}\,.$$

- $I(\theta; X_t|X_{t-1})$: 'new' information

Bounding the mutual information

$$
\begin{aligned}
I(\theta; X_t|X_{t-1}) &= h(X_t|X_{t-1}) - h(X_t|\theta, X_{t-1}) \\
&\le h(X_t) - \mathbb{E}_\theta h(\theta Z_t) \\
&= \mathbb{E}_\theta\{h(\mathrm{N}(0,1)) - h(\mathrm{N}(0,\theta^2))\} \\
&= \mathbb{E}_\theta \log(1/\theta)
\end{aligned}
$$

Bounding the mutual information

$$
\begin{aligned}
I(\theta; X_t | X_{t-1}) &= h(X_t | X_{t-1}) - h(X_t | \theta, X_{t-1}) \\
&\leq h(X_t) - \mathbb{E}_\theta h(\theta Z_t) \\
&= \mathbb{E}_\theta \{ h(\mathrm{N}(0,1)) - h(\mathrm{N}(0,\theta^2)) \\
&= \mathbb{E}_\theta \log(1/\theta)
\end{aligned}
$$

Bounding the mutual information

$$
\begin{aligned}
I(\theta; X_t | X_{t-1}) &= h(X_t | X_{t-1}) - h(X_t | \theta, X_{t-1}) \\
&\le h(X_t) - \mathbb{E}_\theta h(\theta Z_t) \\
&= \mathbb{E}_\theta \{ h(\mathrm{N}(0,1)) - h(\mathrm{N}(0,\theta^2)) \\
&= \mathbb{E}_\theta \log(1/\theta)
\end{aligned}
$$

Bounding the mutual information

$$
\begin{aligned}
I(\theta; X_t | X_{t-1}) &= h(X_t | X_{t-1}) - h(X_t | \theta, X_{t-1}) \\
&\le h(X_t) - \mathbb{E}_\theta h(\theta Z_t) \\
&= \mathbb{E}_\theta \{ h(\mathrm{N}(0,1)) - h(\mathrm{N}(0,\theta^2)) \\
&= \mathbb{E}_\theta \log(1/\theta)
\end{aligned}
$$

Hence

$$X_{t+1} = \sqrt{1-\theta^2}\, X_t + \theta\, Z_t$$

We need to observe for

$$T \geq \frac{\log|\Theta|}{2\mathbb{E}_\theta \log(1/\theta)}$$

- $T_{\min} \to \infty$ as $\theta \approx 1$ ($\{X_t\}$ independnet
- Bad bound, cannot take $\theta \to 0$

Hence

$$X_{t+1} = \sqrt{1-\theta^2}\, X_t + \theta\, Z_t$$

We need to observe for

$$T \geq \frac{\log|\Theta|}{2\mathbb{E}_\theta \log(1/\theta)}$$

- $T_{\rm min} \to \infty$ as $\theta \approx 1$ ($\{X_t\}$ independnet
- Bad bound, cannot take $\theta \to 0$

A more advanced application

High-dimensional SDE

$$\frac{\mathrm{d}\boldsymbol{X}}{\mathrm{d}t}(t) = \mathbf{F}(\boldsymbol{X}(t);\theta) + \dot{\boldsymbol{B}}(t)$$

- $\boldsymbol{X} = \boldsymbol{X}_0^T = \{\boldsymbol{X}(t)\}_{t\in[0,T]}$, $\boldsymbol{X}(t) \in \mathbb{R}^d$
- $\dot{\boldsymbol{B}} = d-$dimensional white noise: $\mathbb{E}\{\dot{B}_i(t)\dot{B}_j(s)\} = \delta_{ij}\delta(t-s)$
- For each $\theta \in \Theta$, $\mathbf{F}(\,\cdot\,;\theta) : \mathbb{R}^d \to \mathbb{R}^d$

Example

Langevin dynamics

$$\frac{\mathrm{d}\boldsymbol{X}}{\mathrm{d}t}(t) = -\nabla H(\boldsymbol{X}(t);\theta) + \dot{\boldsymbol{B}}(t)$$

Spin model

$$H(x,\theta) = -\sum_{(i,j)} \theta_{ij} x_i x_i + \sum_{i=1}^{n} V(x_i)\,,$$

$$\frac{\mathrm{d}X_i}{\mathrm{d}t}(t) = -V'(x_i(t)) + \sum_{j=1}^{n} \theta_{ij} x_j(t) + \dot{B}_i(t)$$

Example

Langevin dynamics

$$\frac{\mathrm{d}\boldsymbol{X}}{\mathrm{d}t}(t) = -\nabla H(\boldsymbol{X}(t);\theta) + \dot{\boldsymbol{B}}(t)$$

Spin model

$$H(x,\theta) = -\sum_{(i,j)} \theta_{ij} x_i x_i + \sum_{i=1}^{n} V(x_i)\,,$$

$$\frac{\mathrm{d}X_i}{\mathrm{d}t}(t) = -V'(x_i(t)) + \sum_{j=1}^{n} \theta_{ij} x_j(t) + \dot{B}_i(t)$$

Information-theoretic lower bound

Fano

$$T \geq \frac{1}{2} \frac{\log |\Theta|}{I(\boldsymbol{X}_0^T;\theta)/T}$$

Duncan 1970; Kadota, Zakai, Ziv, 1971

$$I(\boldsymbol{X}_0^T;\theta) = \frac{1}{2}\int_0^T \mathbb{E}\Big\{\mathsf{Var}\big(\mathbf{F}(t)|\boldsymbol{X}_0^t\big)\Big\}\mathrm{d}t\,,$$
$$\mathbf{F}(t) \equiv \mathbf{F}(\boldsymbol{X}(t),\theta)$$

- **Interpretation:** New information $= \mathsf{Var}\big(\mathbf{F}(t)|\boldsymbol{X}_0^t\big)$

Information-theoretic lower bound

Fano

$$T \geq \frac{1}{2}\frac{\log|\Theta|}{I(\boldsymbol{X}_0^T;\theta)/T}$$

Duncan 1970; Kadota, Zakai, Ziv, 1971

$$I(\boldsymbol{X}_0^T;\theta) = \frac{1}{2}\int_0^T \mathbb{E}\Big\{\mathsf{Var}\big(\mathbf{F}(t)|\boldsymbol{X}_0^t\big)\Big\}\mathrm{d}t\,,$$
$$\mathbf{F}(t) \equiv \mathbf{F}(\boldsymbol{X}(t),\theta)$$

- **Interpretation:** New information $= \mathsf{Var}\big(\mathbf{F}(t)|\boldsymbol{X}_0^t\big)$

A general lower bound

Theorem (Bento, Ibrahimi, Montanari, 2011; Bento, Ibahimi 2014)

If $\boldsymbol{X}$ is stationary, then

$$T \geq \frac{\log|\Theta|}{\mathbb{E}\Big\{\mathsf{Var}\big(\mathbf{F}(t)\big|\boldsymbol{X}_0^t\big)\Big\}}$$

Application: Sparse, linear model

$$\frac{\mathrm{d}\boldsymbol{X}}{\mathrm{d}t}(t) = -\boldsymbol{X}(t) + \mu\,\boldsymbol{A}_G\boldsymbol{X}(t) + \dot{\boldsymbol{B}}(t)$$

- $\boldsymbol{A}_G \in \{0, +1, -1\}^{d\times d}$ adjacency matrix of a (directed, signed) graph
- $\deg(i) \le k$
- $\lambda_{min}(\mathbf{I} - \mu A_G^{\rm symm}) \equiv 1/\tau > 0$

Application: Sparse, linear model

$$\frac{\mathrm{d}\boldsymbol{X}}{\mathrm{d}t}(t) = -\boldsymbol{X}(t) + \mu\,\boldsymbol{A}_G\boldsymbol{X}(t) + \dot{\boldsymbol{B}}(t)$$

- $\boldsymbol{A}_G \in \{0,+1,-1\}^{d\times d}$ adjacency matrix of a (directed, signed) graph
- $\deg(i) \le k$
- $\lambda_{min}(\mathbf{I} - \mu\boldsymbol{A}_G^{\rm symm}) \equiv 1/\tau > 0$

Application: Sparse, linear model

$$\frac{\mathrm{d}\boldsymbol{X}}{\mathrm{d}t}(t) = -\boldsymbol{X}(t) + \mu\, \boldsymbol{A}\boldsymbol{X}(t) + \dot{\boldsymbol{B}}(t)$$

Theorem (Bento, Ibrahimi, Montanari, 2011)

In order to learn $\boldsymbol{A}_G$, we need time at least

$$T \geq C(k) \max\left\{\frac{1}{\mu}, \frac{\tau}{\mu^2}\right\} \log p\,.$$

- Regularized maximum likelihood gets the right scaling [cf. Jose Bento's talk]

Application: Sparse, linear model

$$\frac{\mathrm{d}\boldsymbol{X}}{\mathrm{d}t}(t) = -\boldsymbol{X}(t) + \mu\,\boldsymbol{A}\boldsymbol{X}(t) + \dot{\boldsymbol{B}}(t)$$

Theorem (Bento, Ibrahimi, Montanari, 2011)

In order to learn $\boldsymbol{A}_G$, we need time at least

$$T \geq C(k) \max\left\{\frac{1}{\mu}, \frac{\tau}{\mu^2}\right\} \log p\,.$$

- Regularized maximum likelihood gets the right scaling [cf. Jose Bento's talk]

Conclusion

Conclusion

- Learning dynamical systems from data
- **Largely open**
- Information theory gives useful lower bounds

Thanks!

Conclusion

- Learning dynamical systems from data
- **Largely open**
- Information theory gives useful lower bounds

Thanks!

Conclusion

- Learning dynamical systems from data
- **Largely open**
- Information theory gives useful lower bounds

Thanks!

Conclusion

- Learning dynamical systems from data
- **Largely open**
- Information theory gives useful lower bounds

Thanks!