Learning Dynamical Systems

Symposium on Machine Learning and Dynamical Systems

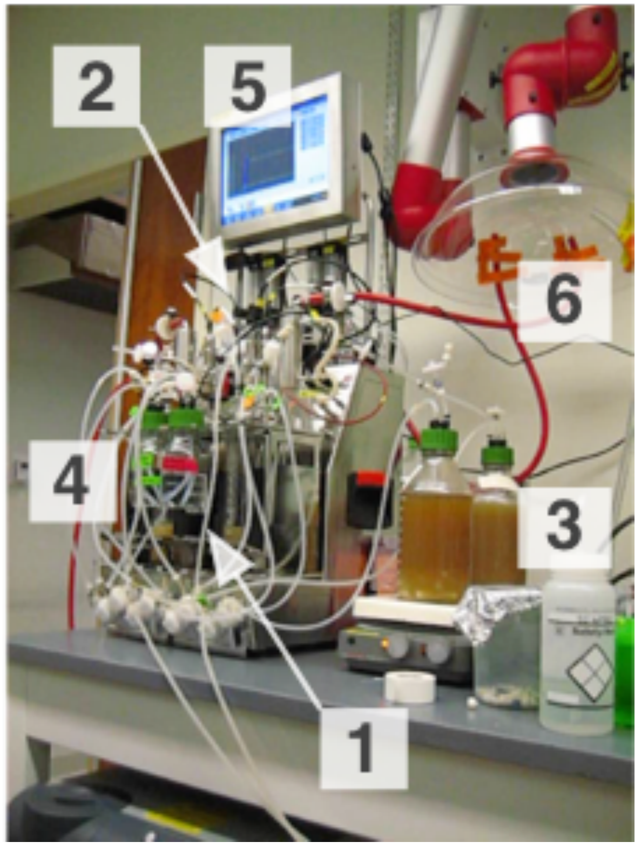
Sayan Mukherjee

Center for Scalable Data Science and AI at Universität Leipzig and MPI-MiS
Duke University
https://sayanmuk.github.io/

Joint work with:

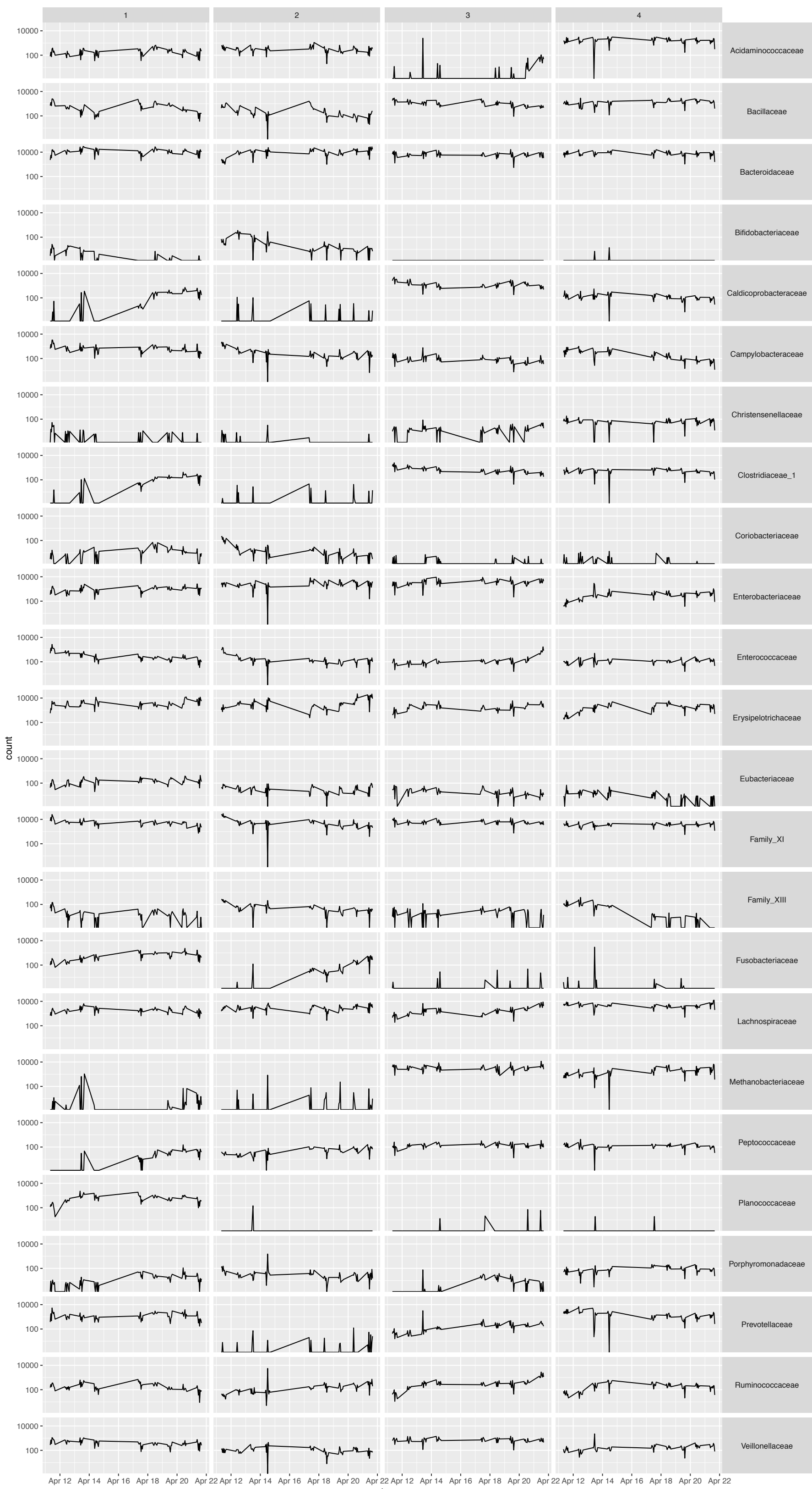
Dynamical systems — K. McGoff (UNC Ch) | A. Nobel (UNC CH) | L. Su (Duke)

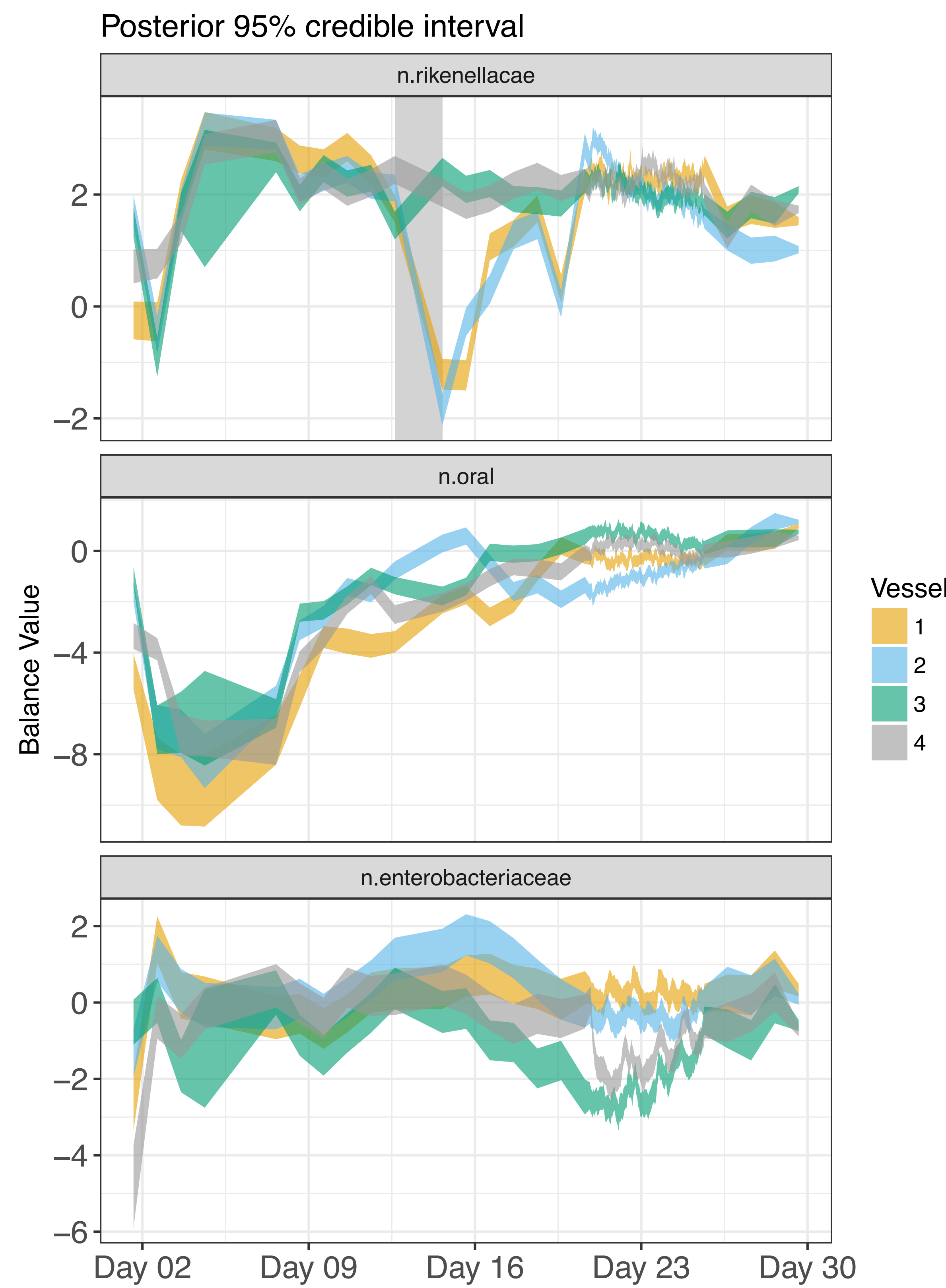
EXAMPLE QUESTIONS OF INTEREST



- How fast does the community change?
- Did a new food change the community? If so, in what way?

Microbial ecology





MODELING TIME-EVOLUTION

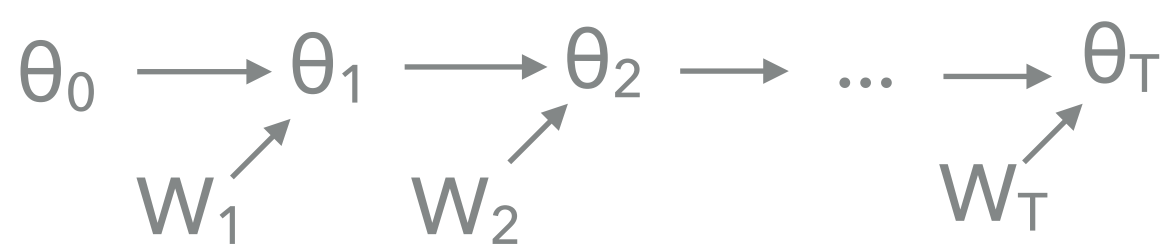
True State with Biological Variation

$$\theta_0 \longrightarrow \theta_1 \longrightarrow \theta_2 \longrightarrow \ldots \longrightarrow \theta_T$$

Silverman et al., bioRxiv 2018

MODELING TIME-EVOLUTION

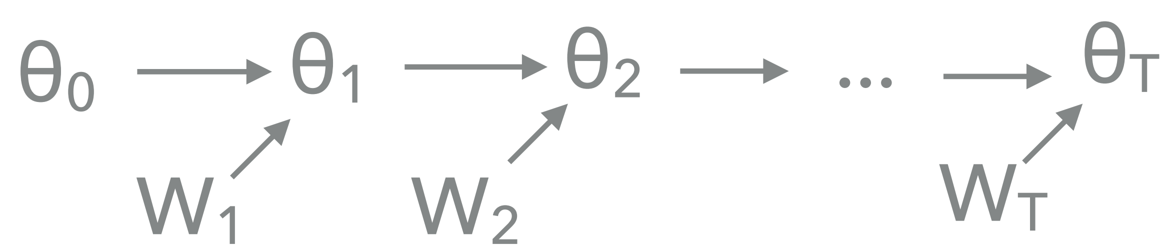
True State with Biological Variation



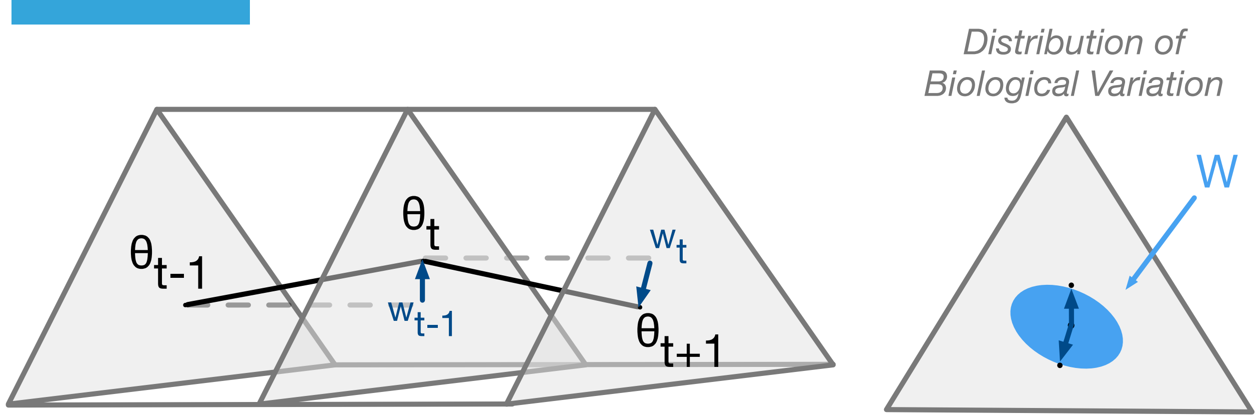
Silverman et al., bioRxiv 2018

MODELING TIME-EVOLUTION

True State with Biological Variation



EXAMPLE



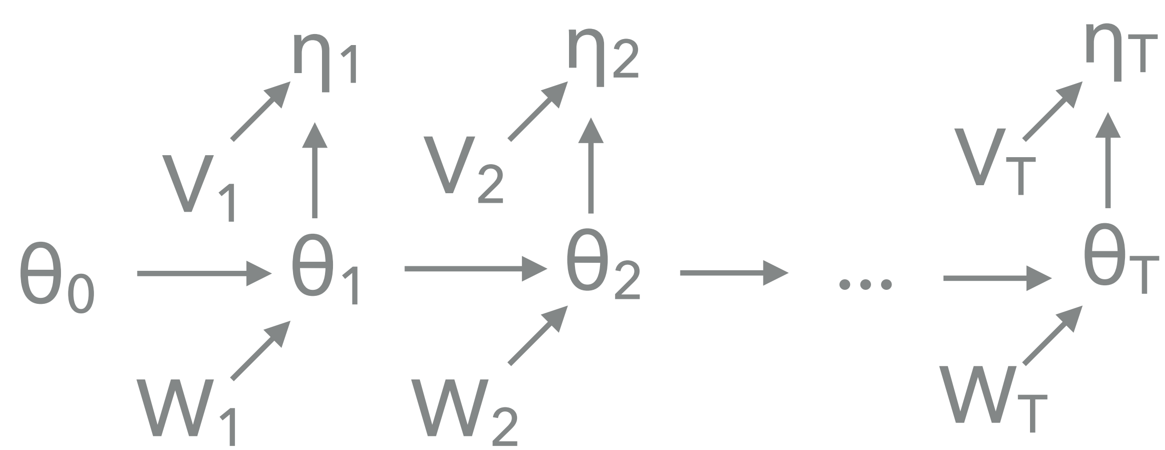
Silverman et al., bioRxiv 2018

MODELING TIME-EVOLUTION

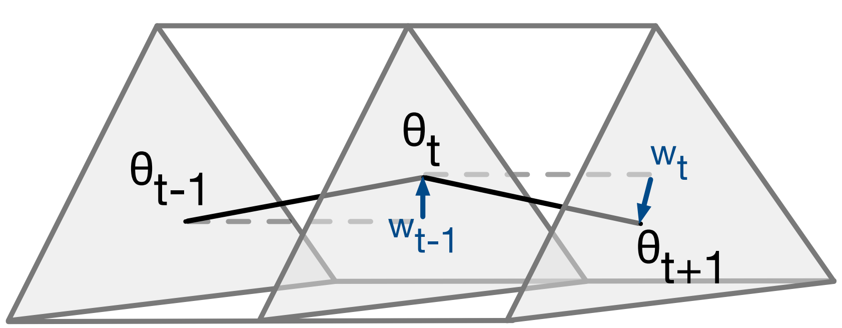
Addition of Technical Noise

↑

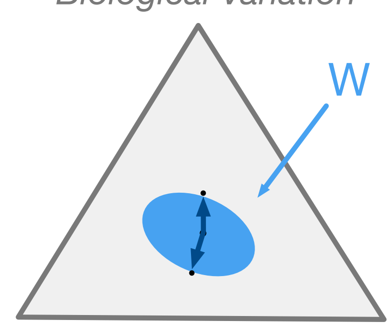
True State with Biological Variation



EXAMPLE



Distribution of Biological Variation



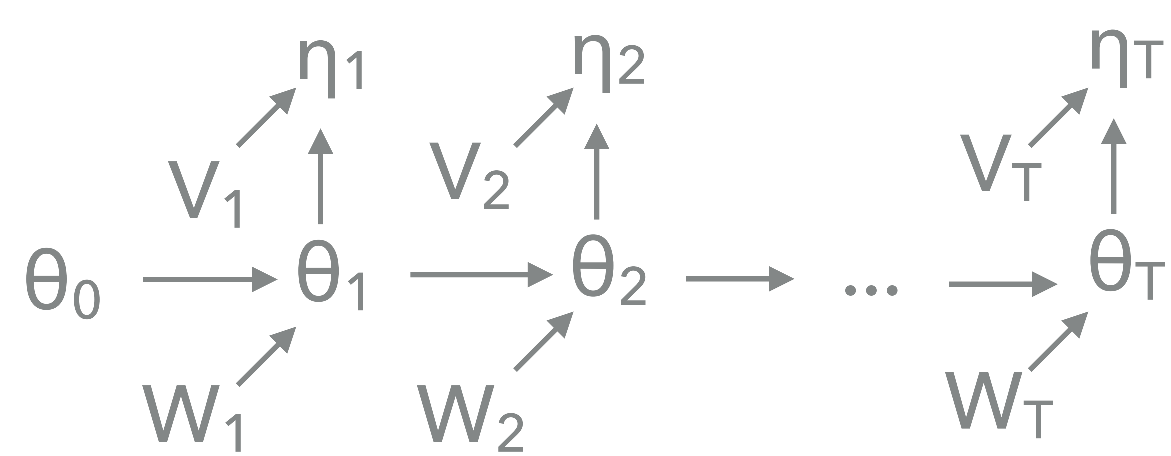
Silverman et al., bioRxiv 2018

MODELING TIME-EVOLUTION

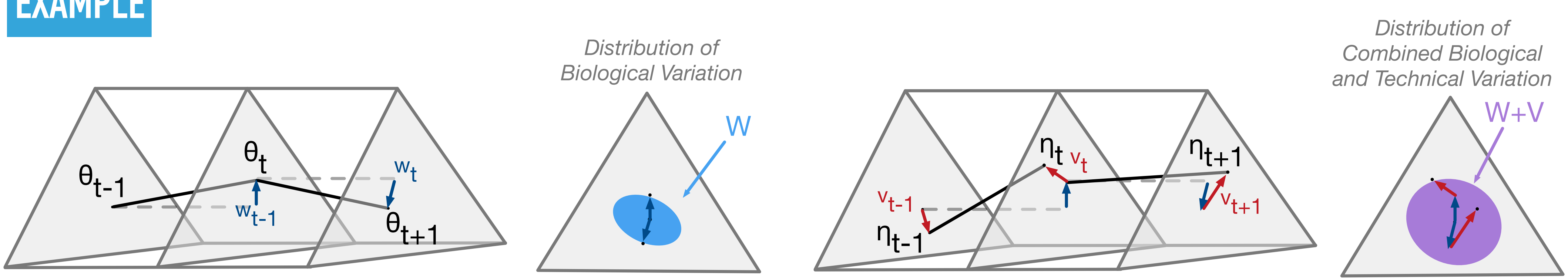
Addition of Technical Noise

↑

True State with Biological Variation



EXAMPLE



Silverman et al., bioRxiv 2018

MODELING TIME-EVOLUTION

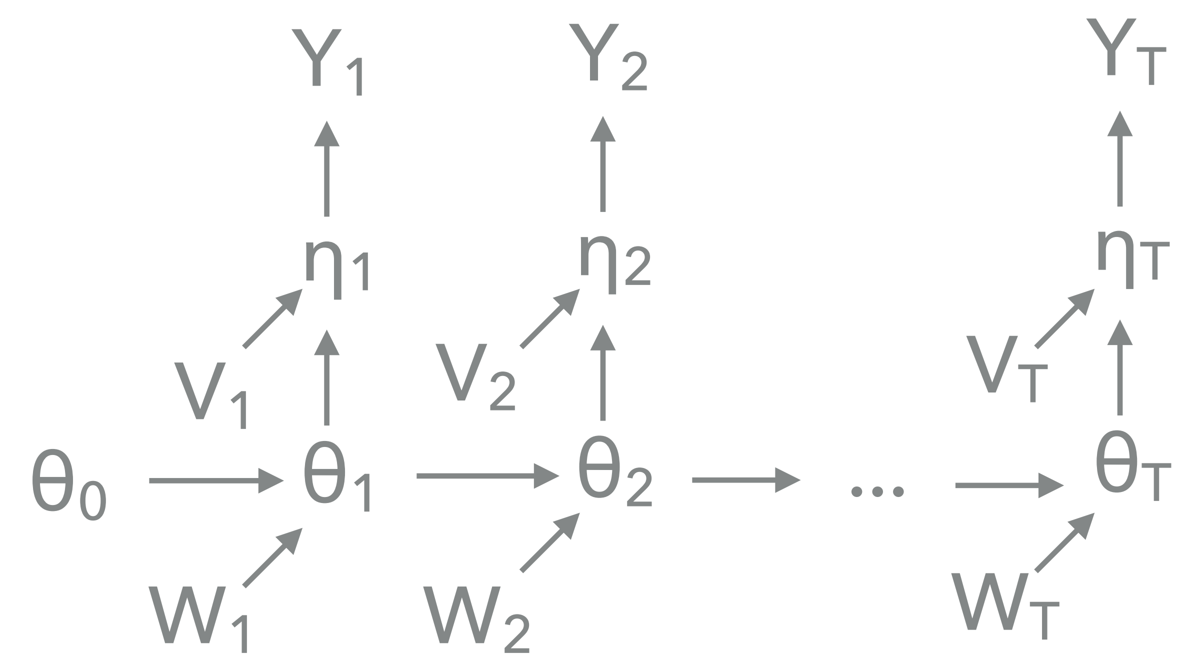
Observed Counts

↑

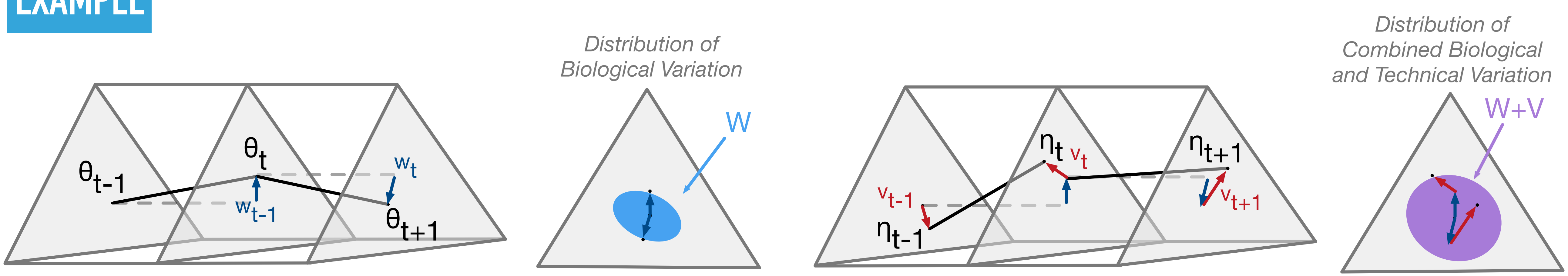
Addition of Technical Noise

↑

True State with Biological Variation

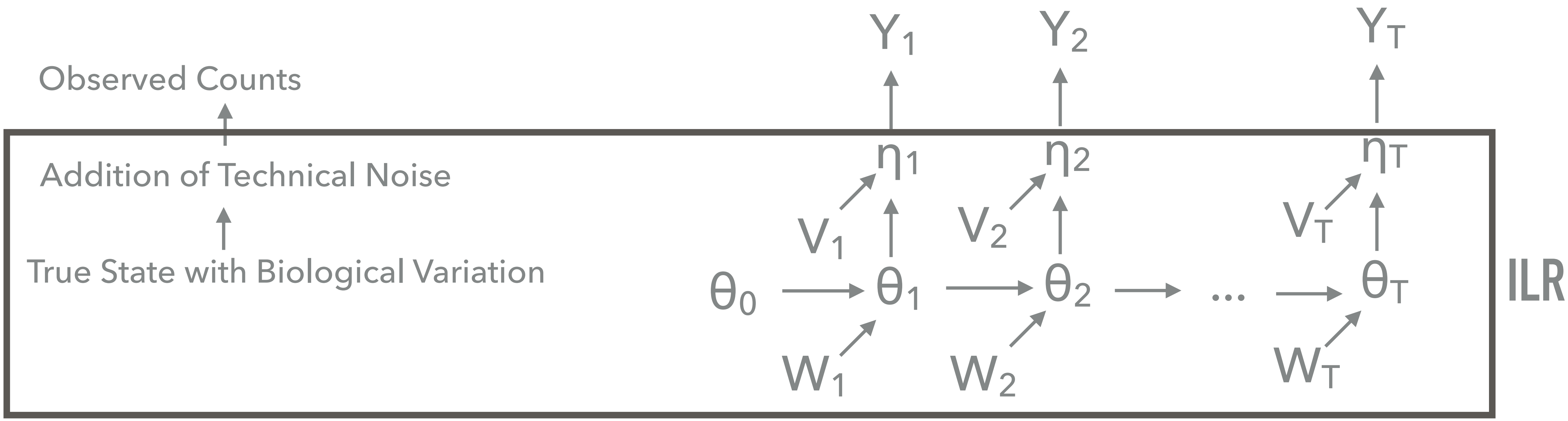


EXAMPLE

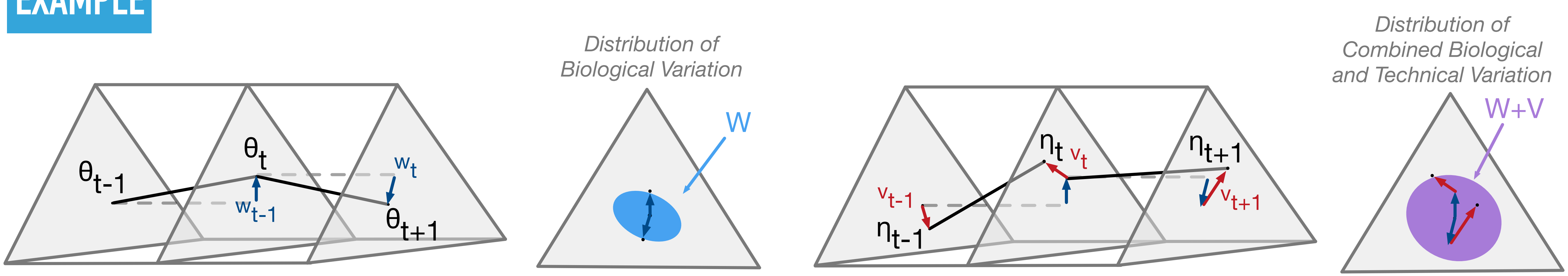


Silverman et al., bioRxiv 2018

MODELING TIME-EVOLUTION

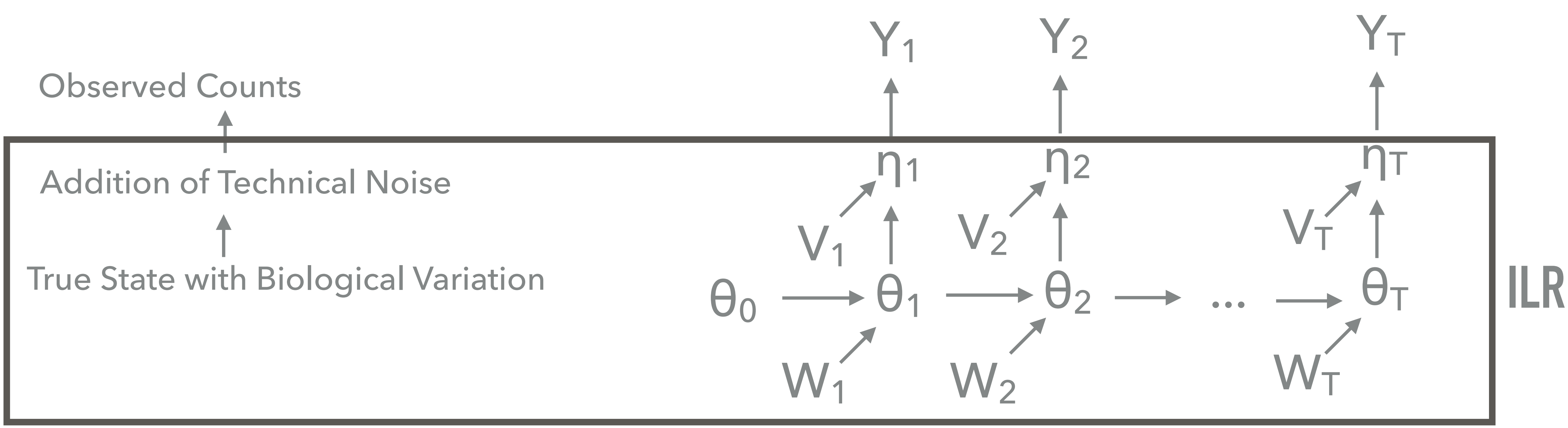


EXAMPLE



Silverman et al., bioRxiv 2018

MODELING TIME-EVOLUTION (LIKELIHOOD MODEL)



Observed Counts

Addition of Technical Noise

True State with Biological Noise

$$\boldsymbol{Y}_t \sim \text{Multinomial}(\boldsymbol{\pi}_t)$$

$$\boldsymbol{\pi}_t = \text{ILR}^{-1}(\boldsymbol{\eta}_t)$$

$$\boldsymbol{\eta}_t = \boldsymbol{F}_t'\boldsymbol{\theta}_t + \boldsymbol{\nu}_t \qquad \boldsymbol{\nu}_t \sim N(0, \boldsymbol{V}_t)$$

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t \qquad \boldsymbol{\omega}_t \sim N(0, \boldsymbol{W}_t)$$

Statistical learning theory

What does learning mean ?

Our dataset consists of two sets of random variables $X \subseteq \mathbb{R}^d$ and $Y \subseteq R^k$, $k = 1$.

What does learning mean ?

Our dataset consists of two sets of random variables $X \subseteq \mathbb{R}^d$ and $Y \subseteq R^k$, $k = 1$.

Data D – $\{(x_i, y_i)\}_{i=1}^n \overset{iid}{\sim} \rho(x, y)$.

What does learning mean ?

Our dataset consists of two sets of random variables $X \subseteq \mathbb{R}^d$ and $Y \subseteq R^k$, $k = 1$.

Data $D - \{(x_i, y_i)\}_{i=1}^{n} \overset{iid}{\sim} \rho(x, y)$.

Learning means given data find a "good" function: $\widehat{f} : X \to Y$.

What does learning mean ?

Our dataset consists of two sets of random variables $X \subseteq \mathbb{R}^d$ and $Y \subseteq R^k$, $k = 1$.

Data $D - \{(x_i, y_i)\}_{i=1}^n \overset{iid}{\sim} \rho(x, y)$.

Learning means given data find a "good" function: $\widehat{f} : X \to Y$.

A "good" function has the property: $y \approx \widehat{f}(x)$, for most $(x, y) \sim \rho$.

What does learning mean ?

Our dataset consists of two sets of random variables $X \subseteq \mathbb{R}^d$ and $Y \subseteq R^k$, $k = 1$.

Data $D - \{(x_i, y_i)\}_{i=1}^n \overset{iid}{\sim} \rho(x, y)$.

Learning means given data find a "good" function: $\widehat{f} : X \to Y$.

A "good" function has the property: $y \approx \widehat{f}(x)$, for most $(x, y) \sim \rho$.

We really care about performance on unobserved data.

What do we need to learn ?

Under what conditions can we "learn" ?

What do we need to learn ?

Under what conditions can we "learn" ?

Depends on the algorithm to infer the function from data:

$$\mathcal{A} : D \to \widehat{f}.$$

What do we need to learn ?

Under what conditions can we "learn" ?

Depends on the algorithm to infer the function from data:

$$\mathcal{A} : D \to \widehat{f}.$$

What constraints on this algorithm need to be imposed ?

A not so good algorithm

Consider the following algorithm

$$\widehat{f}(x) = \sum_i y_i \delta_{x_i}.$$

A not so good algorithm

Consider the following algorithm

$$\widehat{f}(x) = \sum_i y_i \delta_{x_i}.$$

Well for any $x \notin D$

$$\widehat{f}(x) = 0.$$

A simple learning algorithm

A hypothesis space is a class of functions $f \in \mathcal{H}$, for example the space of square integrable functions.

A simple learning algorithm

A hypothesis space is a class of functions $f \in \mathcal{H}$, for example the space of square integrable functions.

Consider the following learning algorithm $\mathcal{A}$:

$$\widehat{f} = \arg\min_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_i (f(x_i) - y_i)^2 \right].$$

A simple learning algorithm

A hypothesis space is a class of functions $f \in \mathcal{H}$, for example the space of square integrable functions.

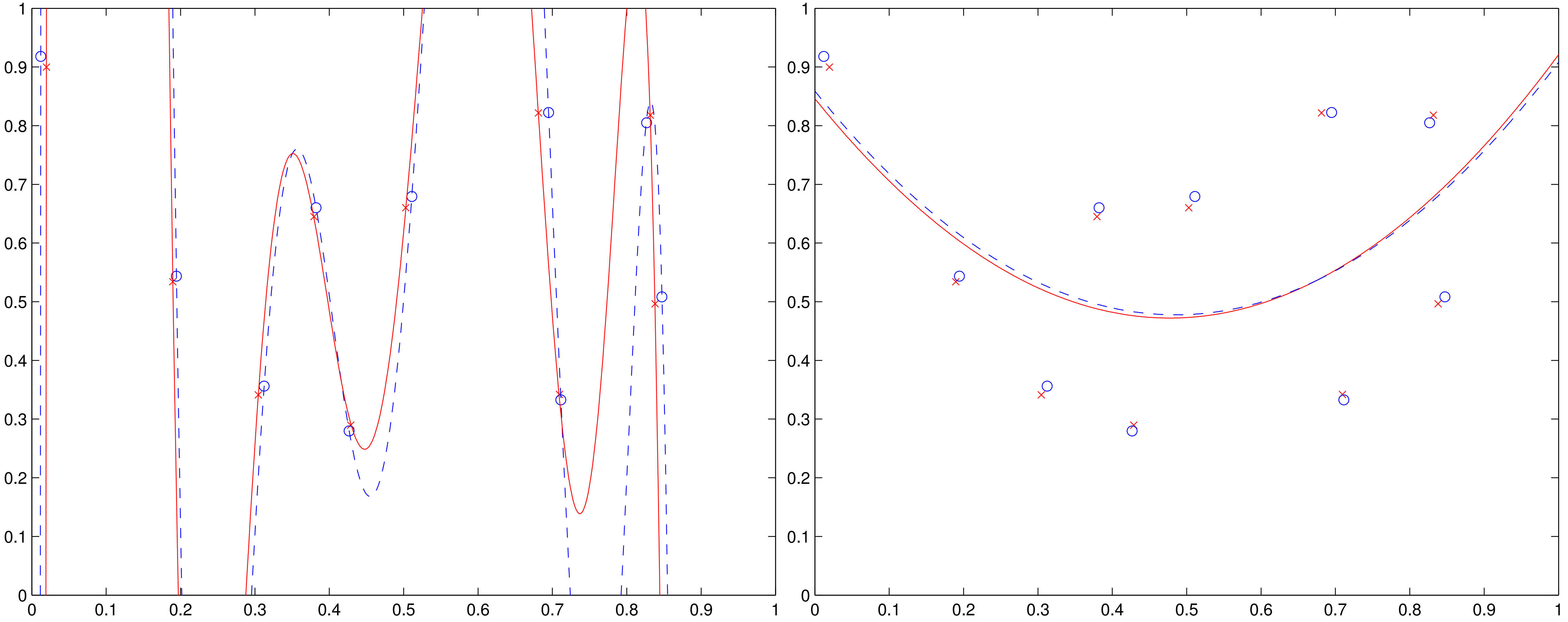
Consider the following learning algorithm $\mathcal{A}$:

$$\widehat{f} = \arg\min_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_i (f(x_i) - y_i)^2 \right].$$

We will want to compare f_s to the best possible function

$$f^* = \arg\min_{f \in \mathcal{H}} \left[\int_{X,Y} (f(x_i) - y_i)^2 \, \mathrm{d}\rho(x, y) \right].$$

Example



Statistical complexity and learnability

For what hypothesis spaces $\mathcal{H}$ can we use our simple algorithm to learn a good $\widehat{f}$.

Statistical complexity and learnability

For what hypothesis spaces $\mathcal{H}$ can we use our simple algorithm to learn a good $\widehat{f}$.

Covering number: Given a hypothesis space $\mathcal{H}$ and the supnorm, the covering number $\mathcal{N}(\mathcal{H}, \epsilon)$ is the minimal number $\ell \in \mathbb{N}$ such that for every $f \in \mathcal{H}$ there exists functions $\{g_i\}_{i=1}^{\ell}$ such that

$$\sup_{x \in \mathcal{X}} |f(x) - g_i(x)| \leq \epsilon \text{ for some } i.$$

Statistical complexity and learnability

For what hypothesis spaces $\mathcal{H}$ can we use our simple algorithm to learn a good $\widehat{f}$.

Covering number: Given a hypothesis space $\mathcal{H}$ and the supnorm, the covering number $\mathcal{N}(\mathcal{H}, \epsilon)$ is the minimal number $\ell \in \mathbb{N}$ such that for every $f \in \mathcal{H}$ there exists functions $\{g_i\}_{i=1}^{\ell}$ such that

$$\sup_{x \in \mathcal{X}} |f(x) - g_i(x)| \leq \epsilon \text{ for some } i.$$

The metric entropy is $\log \mathcal{N}(\mathcal{H}, \epsilon)$ and

Statistical complexity and learnability

For what hypothesis spaces $\mathcal{H}$ can we use our simple algorithm to learn a good $\widehat{f}$.

Covering number: Given a hypothesis space $\mathcal{H}$ and the supnorm, the covering number $\mathcal{N}(\mathcal{H}, \epsilon)$ is the minimal number $\ell \in \mathbb{N}$ such that for every $f \in \mathcal{H}$ there exists functions $\{g_i\}_{i=1}^{\ell}$ such that

$$\sup_{x \in \mathcal{X}} |f(x) - g_i(x)| \leq \epsilon \text{ for some } i.$$

The metric entropy is $\log \mathcal{N}(\mathcal{H}, \epsilon)$ and a formal definition of learnability is

$$\forall \epsilon > 0, \quad \lim_{n \to \infty} \frac{\log \mathcal{N}(\mathcal{H}, \epsilon)}{n} = 0.$$

A learning theory result

Proposition

Under mild conditions, with probability at least $1 - e^{-t}$ *(*$t > 0$*)*

$$\|f_s - f^*\|_{\rho_x} \leq \sqrt{\frac{(\log \mathcal{N}(\mathcal{H}, \varepsilon/8) + t)}{n}}.$$

Learning dynamical systems

Hidden Markov Models

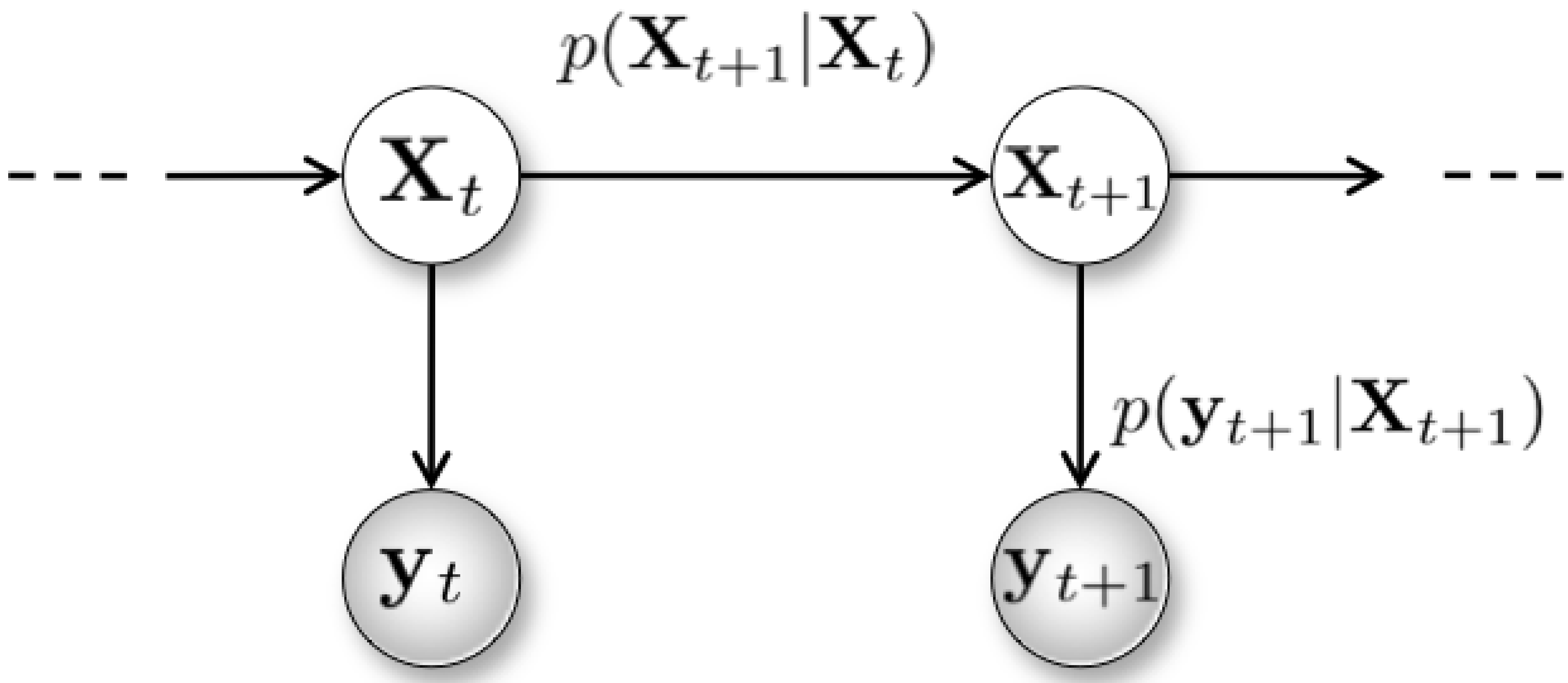
Markov model:

$$x_{t+1} = f(x_t; \theta), \quad \text{state process}$$

Hidden Markov model:

$$\begin{aligned} x_{t+1} &= f(x_t; \theta_1) && \text{hidden state process} \\ y_{t+1} &= g(x_{t+1}; \theta_2) && \text{observation process.} \end{aligned}$$

Hidden Markov Models



Stochastic versus deterministic systems

Should the process $(X_t)_t$ be stochastic or deterministic?

Stochastic versus deterministic systems

Should the process $(X_t)_t$ be stochastic or deterministic?

- If the conditional distribution of X_{t+1} given X_t has positive variance, then we'll say the process $(X_t)_t$ is stochastic.

Stochastic versus deterministic systems

Should the process $(X_t)_t$ be stochastic or deterministic?

- If the conditional distribution of X_{t+1} given X_t has positive variance, then we'll say the process $(X_t)_t$ is stochastic.
- Otherwise, we'll say the process $(X_t)_t$ is deterministic.

In ecology both types of systems are commonly used.

Setting for deterministic dynamics

Suppose that for each θ in Θ (parameter space), we have (X, $\mathcal{X}$, T_θ, μ_θ), where

- X is a complete separable metric space with Borel σ-algebra $\mathcal{X}$

Setting for deterministic dynamics

Suppose that for each θ in Θ (parameter space), we have $(\mathsf{X}, \mathcal{X}, T_\theta, \mu_\theta)$, where

- X is a complete separable metric space with Borel σ-algebra $\mathcal{X}$
- $T_\theta : \mathsf{X} \to \mathsf{X}$ is a measurable map,

Setting for deterministic dynamics

Suppose that for each θ in Θ (parameter space), we have $(\mathsf{X}, \mathcal{X}, T_\theta, \mu_\theta)$, where

- X is a complete separable metric space with Borel σ-algebra $\mathcal{X}$
- $T_\theta : \mathsf{X} \to \mathsf{X}$ is a measurable map,
- μ_θ is a probability measure on $(\mathsf{X}, \mathcal{X})$ is T_θ-invariant if $\mu_\theta(T_\theta^{-1} A) = \mu_\theta(A), \quad \forall A \in \mathcal{X}$

Setting for deterministic dynamics

Suppose that for each θ in Θ (parameter space), we have $(\mathsf{X}, \mathcal{X}, T_\theta, \mu_\theta)$, where

- X is a complete separable metric space with Borel σ-algebra $\mathcal{X}$
- $T_\theta : \mathsf{X} \to \mathsf{X}$ is a measurable map,
- μ_θ is a probability measure on $(\mathsf{X}, \mathcal{X})$ is T_θ-invariant if $\mu_\theta(T_\theta^{-1}A) = \mu_\theta(A), \quad \forall A \in \mathcal{X}$
- the measure preserving system $(\mathsf{X}, \mathcal{X}, T_\theta, \mu_\theta)$ is ergodic if $T_\theta^{-1}A = A$ implies $\mu(A) = \{0, 1\}$.

Setting for deterministic dynamics

Suppose that for each θ in Θ (parameter space), we have $(\mathsf{X}, \mathcal{X}, T_\theta, \mu_\theta)$, where

- X is a complete separable metric space with Borel σ-algebra $\mathcal{X}$
- $T_\theta : \mathsf{X} \to \mathsf{X}$ is a measurable map,
- μ_θ is a probability measure on $(\mathsf{X}, \mathcal{X})$ is T_θ-invariant if $\mu_\theta(T_\theta^{-1}A) = \mu_\theta(A), \quad \forall A \in \mathcal{X}$
- the measure preserving system $(\mathsf{X}, \mathcal{X}, T_\theta, \mu_\theta)$ is ergodic if $T_\theta^{-1}A = A$ implies $\mu(A) = \{0, 1\}$.

Family of systems $(\mathsf{X}, \mathcal{X}, T_\theta, \mu_\theta)_{\theta\in\Theta} \equiv (T_\theta, \mu_\theta)_{\theta\in\Theta}$.

Observational noise

Conditional likelihood: $g_\theta(y \mid x) = f(Y_t = y \mid x_t = x, \theta)$, with

$$\int g_\theta(y \mid x) d\nu(y) = 1.$$

Also $g : \Theta \times \mathsf{X} \times \mathsf{Y} \to \mathbb{R}_+$.

Observational noise

Conditional likelihood: $g_\theta(y \mid x) = f(Y_t = y \mid x_t = x, \theta)$, with

$$\int g_\theta(y \mid x) d\nu(y) = 1.$$

Also $g : \Theta \times \mathsf{X} \times \mathsf{Y} \to \mathbb{R}_+$.

Likelihood for y_0^n in Y^{n+1} conditioned on θ and $X_0 = x$ is

$$p_\theta(y_0^n \mid x) = \prod_{k=0}^{n} g_\theta(y_k \mid T_\theta^k(x)),$$

Observational noise

Conditional likelihood: $g_\theta(y \mid x) = f(Y_t = y \mid x_t = x, \theta)$, with

$$\int g_\theta(y \mid x) d\nu(y) = 1.$$

Also $g : \Theta \times \mathsf{X} \times \mathsf{Y} \to \mathbb{R}_+$.

Likelihood for y_0^n in Y^{n+1} conditioned on θ and $X_0 = x$ is

$$p_\theta(y_0^n \mid x) = \prod_{k=0}^{n} g_\theta(y_k \mid T_\theta^k(x)),$$

and the (marginal) likelihood of observing y_0^n given θ is

$$p_\theta(y_0^n) = \int p_\theta(y_0^n \mid x)\, d\mu_\theta(x).$$

Logistic map

$$
\begin{aligned}
X_0 &\sim U[0, 1] \\
X_{t+1} &= \theta X_t(1 - X_t) \\
Y_{t+1} &\sim N(X_{t+1}, \sigma^2)
\end{aligned}
$$

Dynamic linear models

$$
\begin{aligned}
x_{t+1} &= A_{t+1} x_t \\
y_t &= B_t x_t + v_t,
\end{aligned}
$$

Dynamic linear models

$$\begin{aligned} x_{t+1} &= A_{t+1} x_t \\ y_t &= B_t x_t + v_t, \end{aligned}$$

Here:

- y_t is an observation in $\mathbb{R}^p$;
- x_t is a hidden state in $\mathbb{R}^q$;
- A_t is a $p \times p$ state transition matrix;
- B_t is a $q \times p$ observation matrix;
- v_t is a zero-mean vector in $\mathbb{R}^q$.

Approaches to estimation

There are many approaches to estimation:

- maximum likelihood estimation,
- Bayesian estimation,
- optimization (minimization of a cost function),
- etc.

We'll focus on two approaches:

(1) Bayesian inference;
(2) Empirical risk minimization.

Preliminaries

Observation system $(\mathcal{Y}, T, \nu)$ with $T : \mathcal{Y} \to \mathcal{Y}$

Preliminaries

Observation system $(\mathcal{Y}, T, \nu)$ with $T : \mathcal{Y} \to \mathcal{Y}$

Tracking systems:
Compact metrizable space $\mathcal{X} := \mathsf{X} \times \Theta$ with map $S : \mathcal{X} \to \mathcal{X}$.

$$S : \Theta \times \mathsf{X} \to \mathsf{X}, \quad S_\theta : \mathsf{X} \to \mathsf{X}.$$

Preliminaries

Observation system $(\mathcal{Y}, T, \nu)$ with $T : \mathcal{Y} \to \mathcal{Y}$

Tracking systems:
Compact metrizable space $\mathcal{X} := \mathsf{X} \times \Theta$ with map $S : \mathcal{X} \to \mathcal{X}$.

$$S : \Theta \times \mathsf{X} \to \mathsf{X}, \quad S_\theta : \mathsf{X} \to \mathsf{X}.$$

Loss or regret: $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$. Cost of

$$\ell_n(x, y; \theta) := \ell_n(x_0^{n-1}, y_0^{n-1}) = \sum_{k=0}^{n-1} \ell(x_k, y_k),$$

$x_0^{n-1} = (x, S_\theta x, \ldots, S_\theta^{n-1} x)$ and $y_0^{n-1} = (y, Ty, \ldots, T^{n-1} y)$.

Posterior Consistency

Bayesian inference

Data generating process $y_1^n = (y_1, ..., y_n) \overset{iid}{\sim} f_{\theta^*}$

Bayesian inference

Data generating process $y_1^n = (y_1, ..., y_n) \overset{iid}{\sim} f_{\theta^*}$

Likelihood: $f(y_1^n \mid \theta)$
Prior: $\pi(\theta)$
Marginal likelihood: $f(y_1^n) = \int_\theta f(y_1^n \mid \theta)\pi(\theta)d\theta$

Bayesian inference

Data generating process $y_1^n = (y_1, ..., y_n) \overset{iid}{\sim} f_{\theta^*}$

Likelihood: $f(y_1^n \mid \theta)$
Prior: $\pi(\theta)$
Marginal likelihood: $f(y_1^n) = \int_\theta f(y_1^n \mid \theta)\pi(\theta)d\theta$

Posterior

$$\Pi_n(\theta \mid y_1^n) = \frac{f(y_1^n \mid \theta)\,\pi(\theta)}{f(y_1^n)}.$$

Bayesian inference

Likelihood: $y_1, ..., y_n \overset{iid}{\sim} \text{Ber}(p)$

Prior: $p \sim \text{Beta}(\alpha, \beta)$

Posterior

$$\Pi_n(p \mid y_1^n) = \text{Beta}\left(\alpha + \sum_i y_i, \beta + n - \sum_i y_i\right)$$

Posterior consistency

Does $\lim_{n\to\infty} \Pi_n(\theta \mid y_1^n)$ concentrate around an open neighborhood of θ^* ?

Posterior consistency

Does $\lim_{n\to\infty} \Pi_n(\theta \mid y_1^n)$ concentrate around an open neighborhood of θ^* ?

Neighborhood: $S_\epsilon(\theta^*) = \{\theta \in \Theta : \|\theta - \theta^*\|_1 < \epsilon\}$

Strong posterior consistency

$$\Pi_n(S_\epsilon(\theta^*) \mid y_1^n) \to 1 \text{ a.s. } \forall \epsilon > 0.$$

Classical setting

Consider

$\mathcal{Y}$: a complete metric space endowed with its Borel σ-algebra;

Classical setting

Consider

- $\mathcal{Y}$: a complete metric space endowed with its Borel σ-algebra;
- $\{Y_n\}_{n\geq 0}$: observations as a $\mathcal{Y}$-valued process;

Classical setting

Consider

- $\mathcal{Y}$: a complete metric space endowed with its Borel σ-algebra;
- $\{Y_n\}_{n\geq 0}$: observations as a $\mathcal{Y}$-valued process;
- $(\Theta, \{p_\theta : \theta \in \Theta\})$: a parameter space and a collection of Borel probability densities on $\mathcal{Y}$ (with respect to a common measure);
- $\pi(\theta)$: the prior, a Borel probability distribution on Θ.

Posterior distribution

Let $y_i^j = (y_i, \ldots, y_j)$, and $p_\theta(y_i^j) = \prod_{k=i}^{j} p_\theta(y_k)$.

Posterior distribution

Let $y_i^j = (y_i, \ldots, y_j)$, and $p_\theta(y_i^j) = \prod_{k=i}^{j} p_\theta(y_k)$.

Bayes' rule defines a posterior distribution $\Pi_n(\cdot \mid Y_0^{n-1})$

$$\Pi_n(E \mid Y_0^{n-1}) = \frac{\int_E p_\theta(Y_0^{n-1})\, \pi(d\theta)}{\int_\Theta p_\theta(Y_0^{n-1})\, \pi(d\theta)}, \quad E \subset \Theta.$$

Posterior distribution

Let $y_i^j = (y_i, \ldots, y_j)$, and $p_\theta(y_i^j) = \prod_{k=i}^j p_\theta(y_k)$.

Bayes' rule defines a posterior distribution $\Pi_n(\cdot \mid Y_0^{n-1})$

$$\Pi_n(E \mid Y_0^{n-1}) = \frac{\int_E p_\theta(Y_0^{n-1})\, \pi(d\theta)}{\int_\Theta p_\theta(Y_0^{n-1})\, \pi(d\theta)}, \quad E \subset \Theta.$$

Question: if $\{Y_n\}_{n\geq 0}$ is i.i.d. with density p_{θ_0}, what happens to $\Pi_n(\cdot \mid Y_0^{n-1})$ as n tends to infinity?

Posterior consistency

We say that (θ_0, π) is consistent if for all open neighborhoods U of θ_0,

$$\Pi_n\big(\Theta \setminus U \mid Y_0^{n-1}\big) \to 0, \quad P_{\theta_0}^{\infty} - a.s.$$

Posterior consistency

We say that (θ_0, π) is consistent if for all open neighborhoods U of θ_0,

$$\Pi_n\big(\Theta \setminus U \mid Y_0^{n-1}\big) \to 0, \quad P_{\theta_0}^{\infty} - a.s.$$

Theorem (Doob, 1949)

For π-almost every θ in Θ, the pair (θ, π) is consistent.

Posterior consistency

We say that (θ_0, π) is consistent if for all open neighborhoods U of θ_0,

$$\Pi_n(\Theta \setminus U \mid Y_0^{n-1}) \to 0, \quad P_{\theta_0}^{\infty} - a.s.$$

Theorem (Doob, 1949)

For π-almost every θ in Θ, the pair (θ, π) is consistent.

What about for *every* θ in Θ?

Schwartz conditions

Theorem (Schwartz, 1965)

Let $\theta_0 \in \Theta$. Suppose that

1. for each neighborhood U of θ_0, there exist constants $\beta > 0$ and $C > 0$ and measurable functions $\varphi_n : \mathcal{Y}^n \to [0, 1]$ such that
   a) $\mathbb{E}_{\theta_0}[\varphi_n(Y_0^{n-1})] \leq Ce^{-\beta n}$, and
   b) $\sup_{\theta \notin U} \mathbb{E}_\theta[1 - \varphi_n(Y_0^{n-1})] \leq Ce^{-\beta n}$.
2. for each $\epsilon > 0$,

$$\pi\left(\theta : \mathbb{E}_{\theta_0}[-\log(p_\theta / p_{\theta_0})] < \epsilon\right) > 0.$$

Then (θ_0, π) is consistent.

More recent work

1990's: Inconsistency results for nonparametric models (Θ is infinite dimensional) by Diaconis and Freedman.

2000-2010: Extensive results for nonparametric models, Ghosal and van der Vaart [2017]

2000-2019: Rates of convergence

2000-2019: Convergence with respect to different metrics on Θ (e.g. Hellinger).

Dependence

We would like to consider posterior consistency for stationary processes.

Suppose that $\{Y_n\}_{n\geq 0}$ is stationary (not necessarily i.i.d.).

Dependence

We would like to consider posterior consistency for stationary processes.

Suppose that $\{Y_n\}_{n\geq 0}$ is stationary (not necessarily i.i.d.).

Θ parametrizes a collection of stationary stochastic processes, serving as models of $\{Y_n\}_n$.

Dependence

We would like to consider posterior consistency for stationary processes.

- Suppose that $\{Y_n\}_{n\geq 0}$ is stationary (not necessarily i.i.d.).
- Θ parametrizes a collection of stationary stochastic processes, serving as models of $\{Y_n\}_n$.
- Given a prior distribution π, we'll define a posterior distribution $\Pi_n(\cdot \mid Y_0^{n-1})$.

Dependence

We would like to consider posterior consistency for stationary processes.

Suppose that $\{Y_n\}_{n\geq 0}$ is stationary (not necessarily i.i.d.).

Θ parametrizes a collection of stationary stochastic processes, serving as models of $\{Y_n\}_n$.

Given a prior distribution π, we'll define a posterior distribution $\Pi_n(\cdot \mid Y_0^{n-1})$.

Question: What happens to $\Pi_n(\cdot \mid Y_0^{n-1})$ as n tends to infinity?

Classical Bayesian inference

Likelihood: $f(y_1^n \mid \theta)$
Prior: $\pi(\theta)$
Marginal likelihood: $f(y_1^n) = \int_\theta f(y_1^n \mid \theta)\pi(\theta)d\theta$

Classical Bayesian inference

Likelihood: $f(y_1^n \mid \theta)$
Prior: $\pi(\theta)$
Marginal likelihood: $f(y_1^n) = \int_\theta f(y_1^n \mid \theta)\pi(\theta)d\theta$

Posterior

$$\Pi_n(\theta \mid y_1^n) = \frac{f(y_1^n \mid \theta)\,\pi(\theta)}{f(y_1^n)}.$$

Preliminaries

Observation system $(\mathcal{Y}, T, \nu)$ with $T : \mathcal{Y} \to \mathcal{Y}$

Tracking systems:
Compact metrizable space $\mathcal{X} := \mathsf{X} \times \Theta$ with map $S : \mathcal{X} \to \mathcal{X}$.

$$S : \Theta \times \mathsf{X} \to \mathsf{X}, \quad S_\theta : \mathsf{X} \to \mathsf{X}.$$

Loss or regret: $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$. Cost of

$$\ell_n(x, y; \theta) := \ell_n(x_0^{n-1}, y_0^{n-1}) = \sum_{k=0}^{n-1} \ell(x_k, y_k),$$

$x_0^{n-1} = (x, S_\theta x, \ldots, S_\theta^{n-1} x)$ and $y_0^{n-1} = (y, Ty, \ldots, T^{n-1} y)$.

Gibbs posterior

Given observations y and a prior π on $\mathcal{X}$.

Gibbs posterior

Given observations y and a prior π on $\mathcal{X}$.

The Gibbs posterior is

$$\begin{aligned}
\Pi_n(A \mid y) &= \frac{\int_A \exp\big(-\ell_n(x, y; \theta)\big)\, d\pi(x)}{Z_n(y)}, \quad A \subset \Theta \times \mathrm{X} \\
Z_n(y) &= \int_{\mathcal{X}} \exp\big(-\ell_n(x, y; \theta)\big)\, d\pi(x).
\end{aligned}$$

Gibbs posterior

Given observations y and a prior π on $\mathcal{X}$.

The Gibbs posterior is

$$\begin{aligned} \Pi_n(A \mid y) &= \frac{\int_A \exp\big(-\ell_n(x, y; \theta)\big)\, d\pi(x)}{Z_n(y)}, \quad A \subset \Theta \times \mathrm{X} \\ Z_n(y) &= \int_{\mathcal{X}} \exp\big(-\ell_n(x, y; \theta)\big)\, d\pi(x). \end{aligned}$$

Two questions

(1) Is $\lim_{n\to\infty} \Pi_n(\cdot \mid y)$ unique.

Gibbs posterior

Given observations y and a prior π on $\mathcal{X}$.

The Gibbs posterior is

$$\Pi_n(A \mid y) = \frac{\int_A \exp\big(-\ell_n(x, y; \theta)\big)\, d\pi(x)}{Z_n(y)}, \quad A \subset \Theta \times \mathrm{X}$$

$$Z_n(y) = \int_{\mathcal{X}} \exp\big(-\ell_n(x, y; \theta)\big)\, d\pi(x).$$

Two questions

(1) Is $\lim_{n\to\infty} \Pi_n(\cdot \mid y)$ unique.

(2) Does $\lim_{n\to\infty} \Pi_n(\cdot \mid y)$ concentrate around T.

Gibbs posterior

(1) Decision theoretic perspective of Bayesian inference, coherent inference with respect to a utility.

Gibbs posterior

(1) Decision theoretic perspective of Bayesian inference, coherent inference with respect to a utility.

(2) If ℓ_n is the negative log likelihood then recover standard posterior.

Gibbs posterior

(1) Decision theoretic perspective of Bayesian inference, coherent inference with respect to a utility.

(2) If ℓ_n is the negative log likelihood then recover standard posterior.

(3) Robust to misspecification, robust statistics.

Gibbs posterior

(1) Decision theoretic perspective of Bayesian inference, coherent inference with respect to a utility.

(2) If ℓ_n is the negative log likelihood then recover standard posterior.

(3) Robust to misspecification, robust statistics.

(4) Calibration/violation of likelihood principle

$$\Pi_n(A \mid y) = \frac{\int_A \exp\Big(-\psi \ell_n(x, y; \theta)\Big)\, d\pi(x)}{Z_n(y)}.$$

Gibbs measures

Given $\mathcal{X}$, the map S, a potential function f, and a measure μ_0

$$G_n(x;\mu_0,f) = \frac{\exp\left(\sum_{k=1}^{n} f(S^k x)\right)}{\int_{\mathcal{X}} \exp\left(\sum_{k=1}^{n} f(S^k x)\right) d\mu_0}.$$

The Gibbs measure is the limit point of the sequence $G_n(x;\mu_0,f)$ and the Gibbs measure is denoted as $\mu_0(f)$.

Gibbs measures

Given $\mathcal{X}$, the map S, a potential function f, and a measure μ_0

$$G_n(x;\mu_0,f) = \frac{\exp\big(\sum_{k=1}^n f(S^k x)\big)}{\int_{\mathcal{X}} \exp\big(\sum_{k=1}^n f(S^k x)\big)\, d\mu_0}.$$

The Gibbs measure is the limit point of the sequence $G_n(x;\mu_0,f)$ and the Gibbs measure is denoted as $\mu_0(f)$.

Recall the Gibbs posterior

$$\Pi_n(x \mid y) = \frac{\exp\big(-\sum_{k=1}^n \ell(S^k x, T^k y)\big)}{\int_{\mathcal{X}} \exp\big(-\sum_{k=1}^n \ell(S^k x, T^k y)\big)\, d\pi(x)}.$$

Sequence space model

Alphabet $\mathcal{A}$ is a finite set ($|A| = N$) and $\Sigma = \mathcal{A}^{\mathbb{Z}}$.

Sequence space model

Alphabet $\mathcal{A}$ is a finite set ($|A| = N$) and $\Sigma = \mathcal{A}^{\mathbb{Z}}$.

Left shift operator $\sigma : \mathcal{A}^{\mathbb{Z}} \to \mathcal{A}^{\mathbb{Z}}$ with $(\sigma x)_i = x_{i+1}$.

Sequence space model

Alphabet $\mathcal{A}$ is a finite set ($|A| = N$) and $\Sigma = \mathcal{A}^{\mathbb{Z}}$.

Left shift operator $\sigma : \mathcal{A}^{\mathbb{Z}} \to \mathcal{A}^{\mathbb{Z}}$ with $(\sigma x)_i = x_{i+1}$.

The set obtained by forbidding a finite number of wods $\mathcal{F}$

$$\Sigma_{\mathcal{F}} = \{x \in \mathcal{A}^{\mathbb{Z}} \mid x_{[i,j]} \neq u \forall i, j \in \mathbb{Z}, u \in \mathcal{F}\}$$

is a shift of finite type (SFT)

Sequence space model

The restriction of the shift maps encoded by matrix A

$$\Sigma_A = \{(a_i)_{i=-\infty}^{\infty} \in \Sigma_{\mathcal{F}}, \quad A_{a_i,a_{i+1}} = 1 \quad \forall i \in \mathbb{Z}\}$$

are called a topological Markov chain or a 1-step SFT.
One can similarly define m-step SFT.

Sequence space model

The restriction of the shift maps encoded by matrix A

$$\Sigma_A = \{(a_i)_{i=-\infty}^{\infty} \in \Sigma_{\mathcal{F}}, \quad A_{a_i, a_{i+1}} = 1 \quad \forall i \in \mathbb{Z}\}$$

are called a topological Markov chain or a 1-step SFT.
One can similarly define m-step SFT.

Σ_A is mixing if and only if there exists $n \geq 1$ such that A^n contains all positive entries.

Sequence space model

The restriction of the shift maps encoded by matrix A

$$\Sigma_A = \{(a_i)_{i=-\infty}^{\infty} \in \Sigma_{\mathcal{F}}, \quad A_{a_i, a_{i+1}} = 1 \quad \forall i \in \mathbb{Z}\}$$

are called a topological Markov chain or a 1-step SFT.
One can similarly define m-step SFT.

Σ_A is mixing if and only if there exists $n \geq 1$ such that A^n contains all positive entries.

For $x \in \Sigma_A$, let $x[i, j] = \{y \in \Sigma_A : x_i^j = y_i^j\}$.

Gibbs measure

Definition

Let $f : \Sigma_{\mathcal{F}} \to \mathbb{R}$ be continuous. A measure μ on $\Sigma_{\mathcal{F}}$ has the Gibbs property for f if there exists $K > 1$ and $\mathcal{P} \in \mathbb{R}$ such that for all $x \in \mathcal{A}^{\mathbb{Z}}$ and $m \geq 1$,

$$K^{-1} \leq \frac{\mu(x[0, m-1])}{\exp(-\mathcal{P}m + \sum_{k=0}^{m-1} f(\sigma^k(x)))} \leq K.$$

Gibbs measure

Definition

Let $f : \Sigma_{\mathcal{F}} \to \mathbb{R}$ be continuous. A measure μ on $\Sigma_{\mathcal{F}}$ has the Gibbs property for f if there exists $K > 1$ and $\mathcal{P} \in \mathbb{R}$ such that for all $x \in \mathcal{A}^{\mathbb{Z}}$ and $m \geq 1$,

$$K^{-1} \leq \frac{\mu(x[0, m-1])}{\exp(-\mathcal{P}m + \sum_{k=0}^{m-1} f(\sigma^k(x)))} \leq K.$$

Theorem (Bowen)

If $\Sigma_{\mathcal{F}}$ is a mixing SFT, and $f : \Sigma_{\mathcal{F}} \to \mathbb{R}$ is Hölder continuous, then there exists a unique Gibbs measure for f on $\Sigma_{\mathcal{F}}$.

Gibbs measure

Definition

Let $f : \Sigma_{\mathcal{F}} \to \mathbb{R}$ be continuous. A measure μ on $\Sigma_{\mathcal{F}}$ has the Gibbs property for f if there exists $K > 1$ and $\mathcal{P} \in \mathbb{R}$ such that for all $x \in \mathcal{A}^{\mathbb{Z}}$ and $m \geq 1$,

$$K^{-1} \leq \frac{\mu(x[0, m-1])}{\exp(-\mathcal{P}m + \sum_{k=0}^{m-1} f(\sigma^k(x)))} \leq K.$$

Theorem (Bowen)

If $\Sigma_{\mathcal{F}}$ is a mixing SFT, and $f : \Sigma_{\mathcal{F}} \to \mathbb{R}$ is Hölder continuous, then there exists a unique Gibbs measure for f on $\Sigma_{\mathcal{F}}$.

$f : \Sigma_{\mathcal{F}} \to \mathbb{R}$ is called a potential, and $\mathcal{P} = \mathcal{P}(f)$ is its pressure.

The model class

We consider families of dependent processes as follows.

Let Θ be a compact metric space.

The model class

We consider families of dependent processes as follows.

Let Θ be a compact metric space.

Let $\{f_\theta : \theta \in \Theta\}$ be a continuously parametrized family of Hölder continuous potential functions.

The model class

We consider families of dependent processes as follows.

Let Θ be a compact metric space.

Let $\{f_\theta : \theta \in \Theta\}$ be a continuously parametrized family of Hölder continuous potential functions.

Let $\{\mu_\theta : \theta \in \Theta\}$ be the corresponding family of Gibbs measures.

The model class

We consider families of dependent processes as follows.

Let Θ be a compact metric space.

Let $\{f_\theta : \theta \in \Theta\}$ be a continuously parametrized family of Hölder continuous potential functions.

Let $\{\mu_\theta : \theta \in \Theta\}$ be the corresponding family of Gibbs measures.

Markov chains of all orders are included in these model classes.

Observation densities

We consider a general observational model as follows.

Let λ be a Borel measure on $\mathcal{Y}$

Let $g : \Theta \times \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ be a measurable function such that for all $\theta \in \Theta$ and $x \in \mathcal{X}$,

$$\int g(\theta, x, y) \, \lambda(dy) = 1.$$

- We write $g_\theta(\cdot \mid x)$ instead of $g(\theta, x, \cdot)$, and we interpret it as a conditional density on $\mathcal{Y}$ given θ and x.

Observation densities

We consider a general observational model as follows.

Let λ be a Borel measure on $\mathcal{Y}$

Let $g : \Theta \times \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ be a measurable function such that for all $\theta \in \Theta$ and $x \in \mathcal{X}$,

$$\int g(\theta, x, y)\, \lambda(dy) = 1.$$

- We write $g_\theta(\cdot \mid x)$ instead of $g(\theta, x, \cdot)$, and we interpret it as a conditional density on $\mathcal{Y}$ given θ and x.
- We require several integrability and regularity conditions on g.

Hidden Gibbs processes

Given $\theta \in \Theta$, the marginal likelihood of y_0^{n-1} is

$$p_\theta(y_0^{n-1}) = \int \prod_{k=0}^{n-1} g_\theta(y_k \mid \sigma^k(x))\, \mu_\theta(dx).$$

Equivalently, we have

$$\begin{aligned} X_0 &\sim \mu_\theta \\ X_{n+1} &= \sigma(X_n) \\ Y_n &\sim g_\theta(y \mid X_n)\lambda(dy). \end{aligned}$$

Let $\mathbb{P}_\theta^Y$ denote the distribution of the process $\{Y_n\}_{n\geq 0}$ under θ.

Posterior consistency

For $\theta \in \Theta$, let $[\theta] = \left\{\theta' \in \Theta : \mathbb{P}_{\theta}^{Y} = \mathbb{P}_{\theta'}^{Y}\right\}$.

Posterior consistency

For $\theta \in \Theta$, let $[\theta] = \left\{\theta' \in \Theta : \mathbb{P}_{\theta}^{Y} = \mathbb{P}_{\theta'}^{Y}\right\}$.

Theorem (McGoff-M-Nobel)

Suppose π is fully supported on Θ, and let $\theta_0 \in \Theta$. Then for any neighborhood U of $[\theta_0]$,

$$\Pi_n\left(\Theta \setminus U \mid Y_0^{n-1}\right) \to 0, \quad \mathbb{P}_{\theta_0}^{Y} - a.s.$$

More general setting

We consider

Θ as before;

$\mathcal{X}$ and $\{\mu_\theta : \theta \in \Theta\}$ as before;

$\ell : \Theta \times \mathcal{X} \times \mathcal{X} \to [0, \infty)$ a continuous loss function;

$\{Y_n\}_{n \geq 0}$ an arbitrary stationary ergodic process.

$y_0^{n-1} := (y_0, \ldots, y_{n-1}) \in \mathcal{Y}^n$.

The loss incurred by parameter θ and initial condition x

$$\ell(\theta, x; y_0^{n-1}) = \sum_{k=0}^{n-1} \ell(\theta, \sigma^k(x), y_k).$$

Gibbs posterior distribution

Prior π and same $\{\mu_\theta : \theta \in \Theta\}$ as before.

Gibbs posterior distribution

Prior π and same $\{\mu_\theta : \theta \in \Theta\}$ as before.

P_0 on $\Theta \times \mathcal{X}$ is

$$P_0(A \times B) = \int_A \mu_\theta(B)\, \pi(d\theta).$$

Gibbs posterior distribution

Prior π and same $\{\mu_\theta : \theta \in \Theta\}$ as before.

P_0 on $\Theta \times \mathcal{X}$ is

$$P_0(A \times B) = \int_A \mu_\theta(B)\, \pi(d\theta).$$

The Gibbs posterior is

$$\Pi_n\left(A \mid y_0^{n-1}\right) = \frac{\int_A \exp\left(-\ell\left(\theta, x; y_0^{n-1}\right)\right) P_0(d\theta, dx)}{Z_n\left(y_0^{n-1}\right)}, A \subset \Theta \times \mathcal{X}$$

where $Z_n\left(y_0^{n-1}\right)$ is a normalization constant.

Questions

1. Does the following limit exist with $\mathbb{P}^Y$-probability 1,

$$\lim_n \frac{1}{n} \log Z_n(Y_0^{n-1}),$$

and if so, what is it?

Questions

1. Does the following limit exist with $\mathbb{P}^Y$-probability 1,

$$\lim_n \frac{1}{n} \log Z_n(Y_0^{n-1}),$$

   and if so, what is it?

2. What can be said about the convergence of the posterior distributions $\{\Pi_n\}_n$?

Joinings

Definition (Joining)

Let (X, A, μ, T) and (Y, B, ν, S) be two dynamical systems. A joining of T and S is a probability measure λ on $X \times Y$, with marginals μ and ν respectively, and invariant to the product map $T \times S$.

Joinings

Definition (Joining)

Let (X, A, μ, T) and (Y, B, ν, S) be two dynamical systems. A joining of T and S is a probability measure λ on $X \times Y$, with marginals μ and ν respectively, and invariant to the product map $T \times S$.

Definition (Coupling)

A coupling of two random variable X and X' taking values in $(E, \mathcal{E})$ is any pair of random variables (Y, Y') taking values in $(E \times E, \mathcal{E} \times \mathcal{E})$ whose marginals have the same distribution as X and X', $X \stackrel{D}{=} Y$ and $X' \stackrel{D}{=} Y'$.

Joinings

A stationary $\mathcal{X}$-valued process $\{X_n\}_{n\geq 0}$ is in $\mathcal{P}(\mathcal{X}, \sigma)$ if

$$X_{n+1} = \sigma(X_n), \quad \forall n, \text{ wp } 1.$$

Joinings

A stationary $\mathcal{X}$-valued process $\{X_n\}_{n\geq 0}$ is in $\mathcal{P}(\mathcal{X}, \sigma)$ if

$$X_{n+1} = \sigma(X_n), \quad \forall n, \text{ wp } 1.$$

A joining of $(\mathcal{X}, \sigma)$ with $\{Y_n\}_{n\geq 0}$ is a stationary bi-variate process $(\mathbf{U}, \mathbf{V}) = \{(U_n, V_n)\}_{n\geq 0}$ on $\mathcal{X} \times \mathcal{Y}$ such that

- $\mathbf{U} = \{U_n\}_{n\geq 0}$ is in $\mathcal{P}(\mathcal{X}, \sigma)$, and
- $\mathbf{V} = \{V_n\}_{n\geq 0}$ is equal to $\{Y_n\}_{n\geq 0}$ in distribution.

Joinings

A stationary $\mathcal{X}$-valued process $\{X_n\}_{n\geq 0}$ is in $\mathcal{P}(\mathcal{X}, \sigma)$ if

$$X_{n+1} = \sigma(X_n), \quad \forall n, \text{ wp } 1.$$

A joining of $(\mathcal{X}, \sigma)$ with $\{Y_n\}_{n\geq 0}$ is a stationary bi-variate process $(\mathbf{U}, \mathbf{V}) = \{(U_n, V_n)\}_{n\geq 0}$ on $\mathcal{X} \times \mathcal{Y}$ such that

$\mathbf{U} = \{U_n\}_{n\geq 0}$ is in $\mathcal{P}(\mathcal{X}, \sigma)$, and

$\mathbf{V} = \{V_n\}_{n\geq 0}$ is equal to $\{Y_n\}_{n\geq 0}$ in distribution.

The set of joinings of $(\mathcal{X}, \sigma)$ with $\{Y_n\}_{n\geq 0}$ is denoted by $\mathcal{J}$.

Convergence theorem

Theorem (McGoff-M-Nobel)

Suppose π is fully supported and ℓ satisfies appropriate regularity and integrability conditions. Then there exists a lower semicontinuous function $\phi : \Theta \to \mathbb{R}$ such that with probability 1,

$$\lim_n -\frac{1}{n} \log Z_n(y) = \inf_{\theta \in \Theta} \phi(\theta).$$

Convergence theorem

Theorem (McGoff-M-Nobel)

Suppose π is fully supported and ℓ satisfies appropriate regularity and integrability conditions. Then there exists a lower semicontinuous function $\phi : \Theta \to \mathbb{R}$ such that with probability 1,

$$\lim_n -\frac{1}{n} \log Z_n(y) = \inf_{\theta \in \Theta} \phi(\theta).$$

The above is the rate function in the large deviation sense.

Variational formulation of $Z_n(y)$ – average cost

Limiting average cost

$$\lim_{n\to\infty} \frac{1}{n} \int_{\mathcal{X}} \ell_n(x, y)\, d\lambda_y(x) = \int \ell\, d\lambda.$$

Variational formulation of $Z_n(y)$ – entropy term

Given two Borel probability measures π and μ on $\mathcal{X}$ and a finite measurable partition ξ of $\mathcal{X}$.
Denote $\mu \prec_\xi \pi$ as $\pi(C) = 0 \Rightarrow \mu(C) = 0$ for $C \in \xi$.

Variational formulation of $Z_n(y)$ – entropy term

Given two Borel probability measures π and μ on $\mathcal{X}$ and a finite measurable partition ξ of $\mathcal{X}$.
Denote $\mu \prec_\xi \pi$ as $\pi(C) = 0 \Rightarrow \mu(C) = 0$ for $C \in \xi$.

Define

$$L(\mu \,\|\, \pi, \xi) = \begin{cases} \sum_{C \in \xi} \mu(C) \log \pi(C), & \text{if } \mu \prec_\xi \pi \\ -\infty, & \text{otherwise,} \end{cases}$$

with $0 \cdot \log 0 = 0$.

Variational formulation of $Z_n(y)$ – entropy term

Given two Borel probability measures π and μ on $\mathcal{X}$ and a finite measurable partition ξ of $\mathcal{X}$.
Denote $\mu \prec_\xi \pi$ as $\pi(C) = 0 \Rightarrow \mu(C) = 0$ for $C \in \xi$.

Define

$$L(\mu \,\|\, \pi, \xi) = \begin{cases} \sum_{C \in \xi} \mu(C) \log \pi(C), & \text{if } \mu \prec_\xi \pi \\ -\infty, & \text{otherwise,} \end{cases}$$

with $0 \cdot \log 0 = 0$.

In spirit consider all finite measurable partitions ξ

$$F(\mu, \pi) = \sup_\xi L(\mu \,\|\, \pi, \xi).$$

Convergence

Theorem (McGoff-M.-Nobel)

Suppose a GIbbs prior, then for ν almost every y,

$$\lim_{n\to\infty} -\frac{1}{n}\log Z_n(y) = \inf_{\lambda\in\mathcal{J}}\left\{\int \ell\, d\lambda + F(\lambda,\mu_\theta)\right\},$$

and the infimum in the above expression is attained.

Bayes as a variational problem

Suppose a GIbbs prior, then for ν almost every y,

$$\lim_{n\to\infty} -\frac{1}{n}\log Z_n(y) = \inf_{\lambda\in\mathcal{J}}\left\{\int \ell\, d\lambda + F(\lambda,\mu_\theta)\right\},$$

A way to write Bayes rule

$$\Pi(\theta \mid x) = \arg\min_{\mu}\left\{\int_\theta \ell(\theta,x)d\mu(\theta) + d_{KL}(\mu,\pi)\right\}$$

Convergence

Proposition (McGoff-M.-Nobel)

Suppose a GIbbs prior and consider the pressure

$$\mathcal{P} = \inf_{\lambda \in \mathcal{J}} \left\{ \int \ell \, d\lambda + F(\lambda, \mu_\theta) \right\}$$

$$\theta_* = \arg\min_{\theta \in \Theta} \mathcal{P}.$$

Convergence

Proposition (McGoff-M.-Nobel)

Suppose a GIbbs prior and consider the pressure

$$\begin{aligned} \mathcal{P} &= \inf_{\lambda \in \mathcal{J}} \left\{ \int \ell \, d\lambda + F(\lambda, \mu_\theta) \right\} \\ \theta_* &= \arg\min_{\theta \in \Theta} \mathcal{P}. \end{aligned}$$

For all $\varepsilon > 0$

$$\mathrm{P}(d(S_{\theta_*}, T) < \varepsilon) \to 1 \text{ a.s as } n \to \infty.$$

Toy example: Markov model

- $\{\mu_\theta : \theta \in \Theta\}$ is a collection of Gibbs measures on a common finite state space;
- there exists $\theta^* \in \Theta$ such that $\hat{\lambda} = \mu_{\theta^*}$;
- $\ell(\theta; y_0^{n-1}) = -\log \mu_\theta(y_0^{n-1})$.

The standard Variational Principle for Gibbs measures yields that the posterior distribution converges almost surely to θ^*.

Toy example: Markov model

- $\{\mu_\theta : \theta \in \Theta\}$ is a collection of Gibbs measures on a common finite state space;
- there exists $\theta^* \in \Theta$ such that $\hat{\lambda} = \mu_{\theta^*}$;
- $\ell(\theta; y_0^{n-1}) = -\log \mu_\theta(y_0^{n-1})$.

The standard Variational Principle for Gibbs measures yields that the posterior distribution converges almost surely to θ^*.

More generally: convergence analysis for Gibbs posteriors under dependence.

Ideas used in proofs

The main technical tools include:

(1) The thermodynamic formalism from dynamical systems (as developed by Sinai, Ruelle, Bowen, and others);

(2) The theory of joinings, introduced by Furstenberg;

(3) Aspects of the “random" thermodynamic formalism of Kifer.

Key ideas

1. Posterior consistency as a two-stage process:
   1.1 Find the limiting variational problem.
   1.2 Analyze the variational problem for consistency.

2. A general framework to adapt ideas from the thermodynamic formalism for Bayesian analysis.

A large deviations perspective

Gibb's measures have a large deviation property. Was this exponential scaling driving our convergence results ? If so can we extend the results to other stochastic and deterministic dynamics.

A large deviations perspective

Gibb's measures have a large deviation property. Was this exponential scaling driving our convergence results ? If so can we extend the results to other stochastic and deterministic dynamics.

Yes.

Two conditions

To prove posterior consistency we need to check:

1. Prove a conditional conditional deviation behavior for one empirical process on $\mathcal{X} \times \mathcal{Y}$; that is a conditional large deviation result for a single model process;
2. Prove an exponential continuity condition over the model family; this allows us to prove a large deviation result over the entire model family.

Why

Use large deviations to prove posterior consistency to have a flexible framework that can be applied to a variety of processes without having to study the process in detail.

1. Continuous time hypermixing stochastic processes;
2. Gibbs processes on shifts of finite type.

Large deviations

Given a Polish space $\mathbb{Z}$ and a lower semicontinuous function $\mathcal{I} : \mathbb{Z} \to [0, \infty]$. A family $(\eta_t)_{t \in \mathbb{T}}$ of probability measures satisfies the large deviation principle with rate function $\mathcal{I}$ if for every closed set $E \subset \mathbb{Z}$,

$$\limsup_{t \to \infty} \frac{1}{t} \log \eta_t(E) \leq - \inf_{z \in E} \mathcal{I}(z),$$

Large deviations

Given a Polish space $\mathbb{Z}$ and a lower semicontinuous function $\mathcal{I} : \mathbb{Z} \to [0, \infty]$. A family $(\eta_t)_{t \in \mathbb{T}}$ of probability measures satisfies the large deviation principle with rate function $\mathcal{I}$ if for every closed set $E \subset \mathbb{Z}$,

$$\limsup_{t \to \infty} \frac{1}{t} \log \eta_t(E) \leq - \inf_{z \in E} \mathcal{I}(z),$$

and for every open set $U \subset \mathbb{Z}$,

$$\liminf_{t \to \infty} \frac{1}{t} \log \eta_t(U) \geq - \inf_{z \in U} \mathcal{I}(z).$$

Large deviations perspective

A sequence of measures $\{\mu_t\}$ satisfies the large deviation principle with rate function I if for all $\Gamma \in \mathcal{B}$

$$-\inf_{x \in \Gamma^o} I(x) \leq \liminf_{t \to \infty} \frac{1}{t} \log \mu_t(\Gamma) \leq \limsup_{t \to \infty} \frac{1}{t} \log \mu_t(\Gamma) \leq -\inf_{x \in \overline{\Gamma}} I(x)$$

or

$$\lim_{t \to \infty} \frac{1}{t} \ln p_t(s) = I(s), \quad p_t(s) = e^{-tI(s)+o(t)}.$$

where p_t is the pdf corresponding to μ_t.

Large deviations perspective

A sequence of measures $\{\mu_t\}$ satisfies the large deviation principle with rate function I if for all $\Gamma \in \mathcal{B}$

$$-\inf_{x\in\Gamma^o} I(x) \leq \liminf_{t\to\infty} \frac{1}{t}\log \mu_t(\Gamma) \leq \limsup_{t\to\infty} \frac{1}{t}\log \mu_t(\Gamma) \leq -\inf_{x\in\overline{\Gamma}} I(x)$$

or

$$\lim_{t\to\infty} \frac{1}{t}\ln p_t(s) = I(s), \quad p_t(s) = e^{-tI(s)+o(t)}.$$

where p_t is the pdf corresponding to μ_t.

Laplace principle: X_n is a sequence of r.v.'s on $\mathcal{X}$ that satisfies for all $f \in \mathcal{C}_b(\mathcal{X})$

$$\lim_{t\to\infty} -\frac{1}{t}\mathbb{E}[\exp(-tf(X_n)] = \inf_{x\in\mathcal{X}} f(x) + I(x),$$

I is the rate function.

Step 1

For a fixed θ show

$$\lim_{t\to\infty} \frac{1}{t} \log Z_t^{\theta}(y) = \inf_{\lambda\in\mathcal{J}(S:\nu)} \left\{ \int c\, d\lambda + F(\lambda,\pi) \right\} = -V(\theta).$$

Exponential continuity

The set $\{\mu_\theta\}_{\theta\in\Theta}$ is an exponentially continuous family with respect to L if the following holds: for all $\theta \in \Theta$, it holds for ν-a.e. $y \in \mathcal{Y}$ that the following limit exists

$$\lim_{t\to\infty} \frac{1}{t} \log \int_{\mathcal{X}} \exp\left(-L_\theta^t(x,y)\right) d\mu_\theta(x) =: -V(\theta),$$

and if $(\theta_t)_{t\in\mathbb{T}}$ is a family of parameters such that $\theta_t \longrightarrow \theta$ in Θ, then

$$\lim_{t\to\infty} \frac{1}{t} \log \int_{\mathcal{X}} \exp\left(-L_{\theta_t}^t(x,y)\right) d\mu_{\theta_t}(x) = -V(\theta).$$

Step 2

Proposition (M-Su)

Suppose $\{\mu_\theta\}_{\theta \in \Theta}$ is an exponentially continuous family with respect to the loss function L and π is a Borel probability measure on Θ. Then for ν-almost every $y \in \mathcal{Y}$,

$$\lim_{t \to \infty} -\frac{1}{t} \log Z_t^\pi(y) = \inf_{\theta \in supp(\pi)} V(\theta).$$

Examples

1. Mixing shifts of finite type.
2. Hypermixing processes.

Hypermixing Processes

Given a closed interval $I \subset \mathbb{T}$, denote $\mathcal{F}_I = \sigma(X_t \,:\, t \in I)$.

Definition

Given $\ell > 0$, $n \geq 2$, and real-valued functions $f_1, ..., f_n$ on $\mathcal{X}$, we say that $f_1, ..., f_n$ are ℓ-measurably separated if there exist intervals $I_1, .., I_n$ such that $\text{dist}(I_M, I_{m'}) \geq \ell$ for $1 \leq m < m' \leq n$ and f_m is $\mathcal{F}_{I_m}$-measurable for each $1 \leq m \leq n$.

Hypermixing Processes

Definition

A process μ is hypermixing if there exists a number $\ell_0 \geq 0$ and non-increasing $\alpha, \beta : (\ell_0, \infty) \to [1, \infty)$ and $\gamma : (\ell_0, \infty) \to [0, 1]$ for which

$$\lim_{\ell \to \infty} \alpha(\ell) = 1, \quad \lim \sup_{\ell \to \infty} \ell(\beta(\ell)) - 1) < \infty, \quad \lim_{\ell \to 0} \gamma(\ell) = 0$$

$$\|f_1 \cdots f_n\|_{L^1(\mu)} \leq \prod_{k=1}^{n} \|f_k\|_{L^{\alpha(\ell)}(\mu)},$$

whenever $n \geq 2$, $\ell > \ell_0$ and $f_1, ..., f_n$ are ℓ-measurably separated functions and

$$\int_{\mathcal{X}} f\, g \,\mathrm{d}\mu - \left(\int_{\mathcal{X}} f \mathrm{d}\mu\right)\left(\int_{\mathcal{X}} g \mathrm{d}\mu\right) \leq \gamma(\ell) \|f\|_{L^{\beta(\ell)}(\mu)} \|g\|_{L^{\beta(\ell)}(\mu)}$$

when $\ell > \ell_0$ and $f, g \in L^1(\mu)$ are ℓ-measurably separated.

Key ideas

1. A general framework for posterior consistency:
   1.1 Prove a conditional large deviations result for one member in the family.
   1.2 Prove exponential continuity across parameterized family.

2. An approach that can be used for SPDEs to symbolic dynamics,

Large deviations approach by Young

T satisfies $\epsilon > 0$ there exists $p = p(\epsilon) \in \mathbb{Z}^+$ such that given any $x_1, ..x_k \in X$, $n_1, ..., n_k \in \mathbb{Z}^+$, and $p_1, ..., p_{k-1} > p(\epsilon)$ there exists $x \in X$ s.t

$$
\begin{aligned}
d(T^i x, T^i x_1) &< \epsilon, \quad 0 \leq i < n_1 \\
d(T^{n_1+p_1+i} x, T^i x_2) &< \epsilon, \quad 0 \leq i < n_2 \\
&\vdots \\
d(T^{n_1+\cdots+n_{k-1}+p_1+\cdots+p_{k-1}+i} x, T^i x_k) &< \epsilon, \quad 0 \leq i < n_k
\end{aligned}
$$

Large deviations approach by Young

Given the above condition.
Let $h_\mu(T)$ be the Kolmogorov-Sinai entropy and set $f : X \to R$, $S_t f = \sum_{j=0}^{t-1} f \circ T^j$.
Assume $h_\mu(T) < \infty$, for every $\phi \in C(X, \mathbb{R})$ and $c \in \mathbb{R}$

$$\begin{aligned}
\lim \sup_{t\to\infty} \frac{1}{t} \log \mu \left\{ \frac{1}{t} S_t \phi \geq c \right\} &\leq \sup \left\{ h_\nu(T) - \int \xi d\nu : \nu \in M(X,t) \int \phi d\nu \geq c \right\} \\
\lim \sup_{t\to\infty} \frac{1}{t} \log \mu \left\{ \frac{1}{t} S_t \phi \geq c \right\} &\geq \sup \left\{ h_\nu(T) - \int \xi d\nu : \nu \in M(X,t) \int \phi d\nu > c \right\}.
\end{aligned}$$

Empirical Risk Minimization

The empirical minimization framework

We consider the following conditions on our model space (T_θ, g_θ) with the following conditions:

(D1) the index set Θ is a compact metric space;
(D2) the map $(\theta, x) \mapsto S_\theta(x)$ from $\Theta \times \mathcal{X}$ to $\mathcal{X}$ is continuous;
(D3) the map $(\theta, x) \mapsto g_\theta(x)$ from $\Theta \times \mathcal{X}$ to $\mathbb{R}$ is continuous.

Let the loss function ℓ be lower semi-continuous and satisfy (C1)

$$\mathbb{E}\left[\sup_{|u| \leq K_{\mathcal{S}}} \ell(u, Y_0)\right].$$

The error incurred by a $\theta \in \Theta$ and initial $x \in \mathcal{X}$ given $\mathbf{Y}$ is

$$R_n(\theta : x) = \frac{1}{n} \sum_{k=0}^{n-1} \ell\left(g_\theta \circ S_\theta^k(x), Y_k\right).$$

The empirical minimizer

A sequence of measurable functions $\theta_n : \mathbb{R}^n \to \Theta$, $n \geq 1$, will be called empirical minimum risk estimates if

$$\lim_n \inf_x R_n(\hat{\theta}_n : x) = \lim_n \inf_\theta \inf_x R_n(\theta : x) \quad w.p.1,$$

where $\hat{\theta}_n := \theta_n(Y_0, ..., Y_{n-1})$.

Does $\hat{\theta}_n$ converge ?
Does it converge to something meaningful ?

The population minimizer

The ℓ-distortion between two stationary processes **U** and **V** is

$$\gamma_\ell(\mathbf{U}, \mathbf{V}) = \inf_{\mathcal{J}(\mathbf{U},\mathbf{V})} \mathbb{E}[\ell(U_0, V_0)].$$

The population minimizer

The ℓ-distortion between two stationary processes **U** and **V** is

$$\gamma_\ell(\mathbf{U}, \mathbf{V}) = \inf_{\mathcal{J}(\mathbf{U},\mathbf{V})} \mathbb{E}[\ell(U_0, V_0)].$$

A family $\mathcal{S}$ corresponds to a family $\mathcal{Q}_\mathcal{S} = \bigcup_{\theta \in \Theta} \mathcal{Q}_\theta$.

The population minimizer

The ℓ-distortion between two stationary processes **U** and **V** is

$$\gamma_\ell(\mathbf{U}, \mathbf{V}) = \inf_{\mathcal{J}(\mathbf{U},\mathbf{V})} \mathbb{E}[\ell(U_0, V_0)].$$

A family $\mathcal{S}$ corresponds to a family $\mathcal{Q}_{\mathcal{S}} = \bigcup_{\theta \in \Theta} \mathcal{Q}_\theta$.

Given a stationary observation process **Y** the population minimizers are the set

$$\Theta_\ell(\mathbf{Y}) = \operatorname*{argmin}_{\theta \in \Theta} \min_{\mathbf{U} \in \mathcal{Q}_\theta} \gamma_\ell(\mathbf{U}, \mathbf{Y}).$$

The population minimizer

The ℓ-distortion between two stationary processes $\mathbf{U}$ and $\mathbf{V}$ is

$$\gamma_\ell(\mathbf{U}, \mathbf{V}) = \inf_{\mathcal{J}(\mathbf{U},\mathbf{V})} \mathbb{E}[\ell(U_0, V_0)].$$

A family $\mathcal{S}$ corresponds to a family $\mathcal{Q}_\mathcal{S} = \bigcup_{\theta\in\Theta} \mathcal{Q}_\theta$.

Given a stationary observation process $\mathbf{Y}$ the population minimizers are the set

$$\Theta_\ell(\mathbf{Y}) = \operatorname*{argmin}_{\theta\in\Theta} \min_{\mathbf{U}\in\mathcal{Q}_\theta} \gamma_\ell(\mathbf{U}, \mathbf{Y}).$$

Does $\hat{\theta}_n$ converge to Θ_ℓ ?

Convergence

Theorem (McGoff-Nobel)

Let $\mathcal{S}$ satisfy (D1)-(D3), let ℓ be a lower semicontinuous loss function. If $\mathbf{Y}$ is a stationary ergodic process satisfying satisfying (C1) then $\Theta_\ell(\mathbf{Y})$ is non-empty and compact and

Convergence

Theorem (McGoff-Nobel)

Let $\mathcal{S}$ satisfy (D1)-(D3), let ℓ be a lower semicontinuous loss function. If $\mathbf{Y}$ *is a stationary ergodic process satisfying satisfying (C1) then $\Theta_\ell(\mathbf{Y})$ is non-empty and compact and*

1. *any sequence $\{\hat{\theta}_n\}$ of minimum risk estimators converges almost surely to $\Theta_\ell(\mathbf{Y})$*

Convergence

Theorem (McGoff-Nobel)

Let $\mathcal{S}$ satisfy (D1)-(D3), let ℓ be a lower semicontinuous loss function. If $\mathbf{Y}$ is a stationary ergodic process satisfying satisfying (C1) then $\Theta_\ell(\mathbf{Y})$ is non-empty and compact and

1. *any sequence $\{\hat{\theta}_n\}$ of minimum risk estimators converges almost surely to $\Theta_\ell(\mathbf{Y})$*
2. *for each $\theta \in \Theta_\ell(\mathbf{Y})$ there exists a sequence of minimum risk estimators that converges almost surely to θ.*

Convergence

Theorem (McGoff-Nobel)

Let $\mathcal{S}$ satisfy (D1)-(D3), let ℓ be a lower semicontinuous loss function. If $\mathbf{Y}$ is a stationary ergodic process satisfying satisfying (C1) then $\Theta_\ell(\mathbf{Y})$ is non-empty and compact and

1. *any sequence $\{\hat{\theta}_n\}$ of minimum risk estimators converges almost surely to $\Theta_\ell(\mathbf{Y})$*
2. *for each $\theta \in \Theta_\ell(\mathbf{Y})$ there exists a sequence of minimum risk estimators that converges almost surely to θ.*

What drives this result ?

Entropy of a sequence

For two sequences **u** and **v** we denote the pseudo metric

$$d_{n,p}(\mathbf{u},\mathbf{v}) = \left(n^{-1} \sum_{k=0}^{n-1} |u_k - v_k|^p \right)^{1/p}, \quad 1 \leq p < \infty,$$

if $p = \infty$ then $\max_{0 \leq k \leq n-1} |u_k - v_k|$.

Entropy of a sequence

For two sequences **u** and **v** we denote the pseudo metric

$$d_{n,p}(\mathbf{u},\mathbf{v}) = \left(n^{-1} \sum_{k=0}^{n-1} |u_k - v_k|^p \right)^{1/p}, \quad 1 \leq p < \infty,$$

if $p = \infty$ then $\max_{0 \leq k \leq n-1} |u_k - v_k|$.

Let $\mathcal{U} \subseteq \mathbb{R}^N$ be a family of infinite sequences. For each $r > 0$ let $\mathcal{N}(\mathcal{U}, r, d_{n,p})$ be the covering number. We now state two entropy metrics

$$h_p(\mathcal{U}, r) = \limsup_n \frac{1}{n} \log \mathcal{N}(\mathcal{U}, r, d_{n,p}), \quad h_p(\mathcal{U}) = \lim_{r \searrow 0} h_p(\mathcal{U}, r).$$

Entropy of dynamical systems

Definition

The entropy $h(\mathcal{S})$ of a family of dynamical models is the common value of $h_p(\mathcal{S})$, where

$$\mathcal{U}_{\mathcal{S}} = \left\{ (g_\theta \circ S_\theta^k(x))_{k \geq 0} : x \in \mathcal{X}, \theta \in \Theta \right\} \subseteq \mathbb{R}^N.$$

Entropy of dynamical systems

Definition

The entropy $h(\mathcal{S})$ of a family of dynamical models is the common value of $h_p(\mathcal{S})$, where

$$\mathcal{U}_{\mathcal{S}} = \left\{ (g_\theta \circ S_\theta^k(x))_{k \geq 0} : x \in \mathcal{X}, \theta \in \Theta \right\} \subseteq \mathbb{R}^N.$$

Theorem (McGoff-Nobel)

For any family $\mathcal{S}$ of dynamical models satisfying (D1)- (D3), and some regularity conditions on the observation process $\mathbf{Y}$ *if* $h(\mathcal{S}) = 0$ *the any sequence of minimum ℓ-risk estimates converges almost surely to* $\Theta_\ell(\mathbf{Y})$.

Key ideas

1. Entropy condition for learning in dynamical systems:
   1.1 Equivalence between topological entropy and statistical notions of entropy.
   1.2 Relation between topological entropy and n-widths.

2. Empirical learning for dynamical systems,

Conclusion

Open problems and extensions

1. Nonstationary processes

Open problems and extensions

1. Nonstationary processes
2. Rates of convergence for a family of dynamical systems

Open problems and extensions

1. Nonstationary processes
2. Rates of convergence for a family of dynamical systems
3. Computational challenges and motivations

Open problems and extensions

1. Nonstationary processes
2. Rates of convergence for a family of dynamical systems
3. Computational challenges and motivations
4. Mean field limits for RNNs based on SFTs

Open problems and extensions

1. Nonstationary processes
2. Rates of convergence for a family of dynamical systems
3. Computational challenges and motivations
4. Mean field limits for RNNs based on SFTs
5. Necessary and sufficient conditions for ERM

Open problems and extensions

1. Nonstationary processes
2. Rates of convergence for a family of dynamical systems
3. Computational challenges and motivations
4. Mean field limits for RNNs based on SFTs
5. Necessary and sufficient conditions for ERM
6. Thermodynamic formalism for Bayesian analysis

Open problems and extensions

1. Nonstationary processes
2. Rates of convergence for a family of dynamical systems
3. Computational challenges and motivations
4. Mean field limits for RNNs based on SFTs
5. Necessary and sufficient conditions for ERM
6. Thermodynamic formalism for Bayesian analysis
7. When is there a limiting variational form

Open problems and extensions

1. Nonstationary processes
2. Rates of convergence for a family of dynamical systems
3. Computational challenges and motivations
4. Mean field limits for RNNs based on SFTs
5. Necessary and sufficient conditions for ERM
6. Thermodynamic formalism for Bayesian analysis
7. When is there a limiting variational form
8. When is there an equilibrium joining

Open problems and extensions

1. Nonstationary processes
2. Rates of convergence for a family of dynamical systems
3. Computational challenges and motivations
4. Mean field limits for RNNs based on SFTs
5. Necessary and sufficient conditions for ERM
6. Thermodynamic formalism for Bayesian analysis
7. When is there a limiting variational form
8. When is there an equilibrium joining
9. Large deviations for Bayesian analysis

Open problems and extensions

1. Nonstationary processes
2. Rates of convergence for a family of dynamical systems
3. Computational challenges and motivations
4. Mean field limits for RNNs based on SFTs
5. Necessary and sufficient conditions for ERM
6. Thermodynamic formalism for Bayesian analysis
7. When is there a limiting variational form
8. When is there an equilibrium joining
9. Large deviations for Bayesian analysis
10. Medium deviations type results

Acknowledgements

Thanks:
Konstantin Mischaikow, Ramon van Handel, Steve Lalley, Jonathan Mattingly, Karl Petersen, Ioanna Manolopoulou, Jim Berger, Beth Archie, Justing Silverman, Lawrence David, Andrea Agazzi

Acknowledgements

Thanks:
Konstantin Mischaikow, Ramon van Handel, Steve Lalley, Jonathan Mattingly, Karl Petersen, Ioanna Manolopoulou, Jim Berger, Beth Archie, Justing Silverman, Lawrence David, Andrea Agazzi

Funding:

- NSF DMS, CCF, CISE
- AFOSR
- DARPA
- NIH
- Alexander von Humboldt Stifung
- BMBF
- Sächsische Staatsministerum für Wissenschaft

Examples of consistent dynamical systems

Classes of systems with good deterministic mixing are good candidates:

- Axiom A systems;
- symbolic dynamics with Gibbs measures;

Axiom A systems

Given a Riemannian manifold $\mathcal{M}$ with a diffeomorphism $f : \mathcal{M} \to \mathcal{M}$. Then f is an an axiom A system if the following hold:

Axiom A systems

Given a Riemannian manifold $\mathcal{M}$ with a diffeomorphism $f : \mathcal{M} \to \mathcal{M}$. Then f is an an axiom A system if the following hold:

(1) The nonwandering set $\Omega(f)$ is a hyperbolic set and compact.

Axiom A systems

Given a Riemannian manifold $\mathcal{M}$ with a diffeomorphism $f : \mathcal{M} \to \mathcal{M}$. Then f is an an axiom A system if the following hold:

(1) The nonwandering set $\Omega(f)$ is a hyperbolic set and compact.

(2) The set of periodic points of f is dense in $\Omega(f)$.

Axiom A systems

A point $x \in \mathcal{M}$ is non-wandering if for each neighborhood $\mathcal{V}$ of x

$$\mathcal{V} \cap \bigcup_{t>0} f^t(\mathcal{V}) \neq \emptyset.$$

Axiom A systems

A point $x \in \mathcal{M}$ is non-wandering if for each neighborhood $\mathcal{V}$ of x

$$\mathcal{V} \cap \bigcup_{t>0} f^t(\mathcal{V}) \neq \emptyset.$$

A closed subset $\Lambda \subset \mathcal{M}$ is hyperbolic if $f(\Lambda) = \Lambda$ and for each $x \in \Lambda$ there exists E^s_X and E^u_x of $T_x\mathcal{M}$ such that

Axiom A systems

A point $x \in \mathcal{M}$ is non-wandering if for each neighborhood $\mathcal{V}$ of x

$$\mathcal{V} \cap \bigcup_{t>0} f^t(\mathcal{V}) \neq \emptyset.$$

A closed subset $\Lambda \subset \mathcal{M}$ is hyperbolic if $f(\Lambda) = \Lambda$ and for each $x \in \Lambda$ there exists E_X^s and E_x^u of $T_x\mathcal{M}$ such that

i) $T_x\mathcal{M} = E_x^s \bigoplus E_x^u$

ii) $Df(E_x^s) = E_{f(x)}^s$ and $Df(E_x^u) = E_{f(x)}^u$

iii) there exists $c > 0$ and $\lambda \in (0, 1)$ s.t.

$$\|Df^n v\| \leq c\lambda^n \|v\| \qquad \text{for all } n \geq 0 \text{ and } v \in E_x^s$$

$$\|Df^{-n} v\| \leq c\lambda^n \|v\| \qquad \text{for all } n \geq 0 \text{ and } v \in E_x^u$$

Axiom A families

$f : \Theta \times X \to X$ is a parameterized family of diffeomorphisms with

i) $\theta \mapsto f_\theta$ is Hölder continuous;

Axiom A families

$f : \Theta \times X \to X$ is a parameterized family of diffeomorphisms with

i) $\theta \mapsto f_\theta$ is Hölder continuous;

ii) $\alpha > 0$ such that $\forall \theta$ f_θ is $C^{1+\alpha}$;

Axiom A families

$f : \Theta \times X \to X$ is a parameterized family of diffeomorphisms with

i) $\theta \mapsto f_\theta$ is Hölder continuous;

ii) $\alpha > 0$ such that $\forall\theta$ f_θ is $C^{1+\alpha}$;

iii) for each θ, $\Omega(f_\theta)$ is an Axiom A attractor and $f_\theta|_{\Omega(f_\theta)}$ is topologically mixing;

Axiom A families

$f : \Theta \times X \to X$ is a parameterized family of diffeomorphisms with

i) $\theta \mapsto f_\theta$ is Hölder continuous;

ii) $\alpha > 0$ such that $\forall \theta$ f_θ is $C^{1+\alpha}$;

iii) for each θ, $\Omega(f_\theta)$ is an Axiom A attractor and $f_\theta|_{\Omega(f_\theta)}$ is topologically mixing;

iv) for each θ, the measure μ_θ is the unique SRB measure.

Axiom A families

$f : \Theta \times X \to X$ is a parameterized family of diffeomorphisms with

i) $\theta \mapsto f_\theta$ is Hölder continuous;
ii) $\alpha > 0$ such that $\forall \theta$ f_θ is $C^{1+\alpha}$;
iii) for each θ, $\Omega(f_\theta)$ is an Axiom A attractor and $f_\theta|_{\Omega(f_\theta)}$ is topologically mixing;
iv) for each θ, the measure μ_θ is the unique SRB measure.

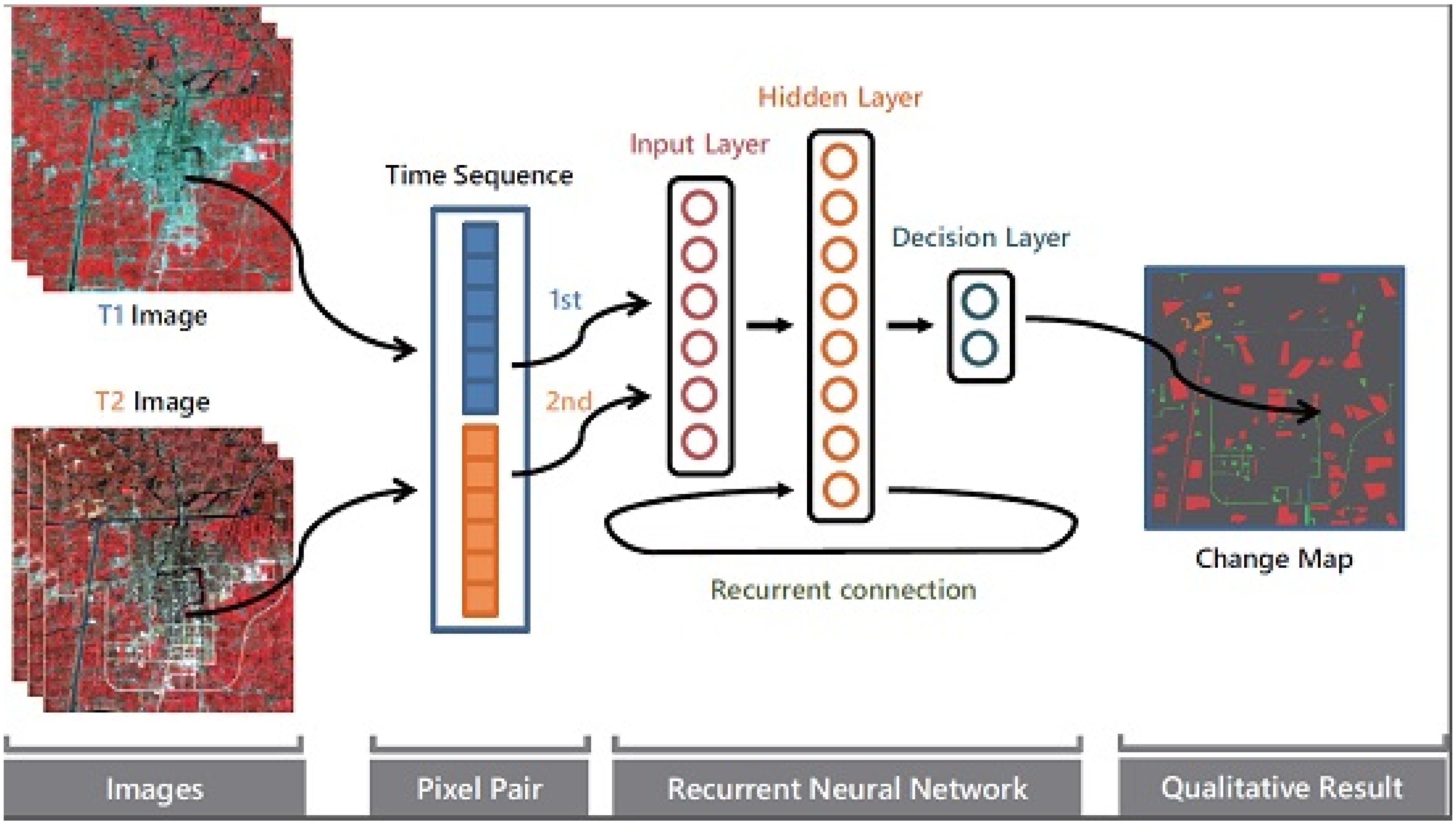
If the above conditions are met $(f_\theta, \mu_\theta)_{\theta \in \Theta}$ is a parameterized family of Axiom A systems.

Consistency of axiom A

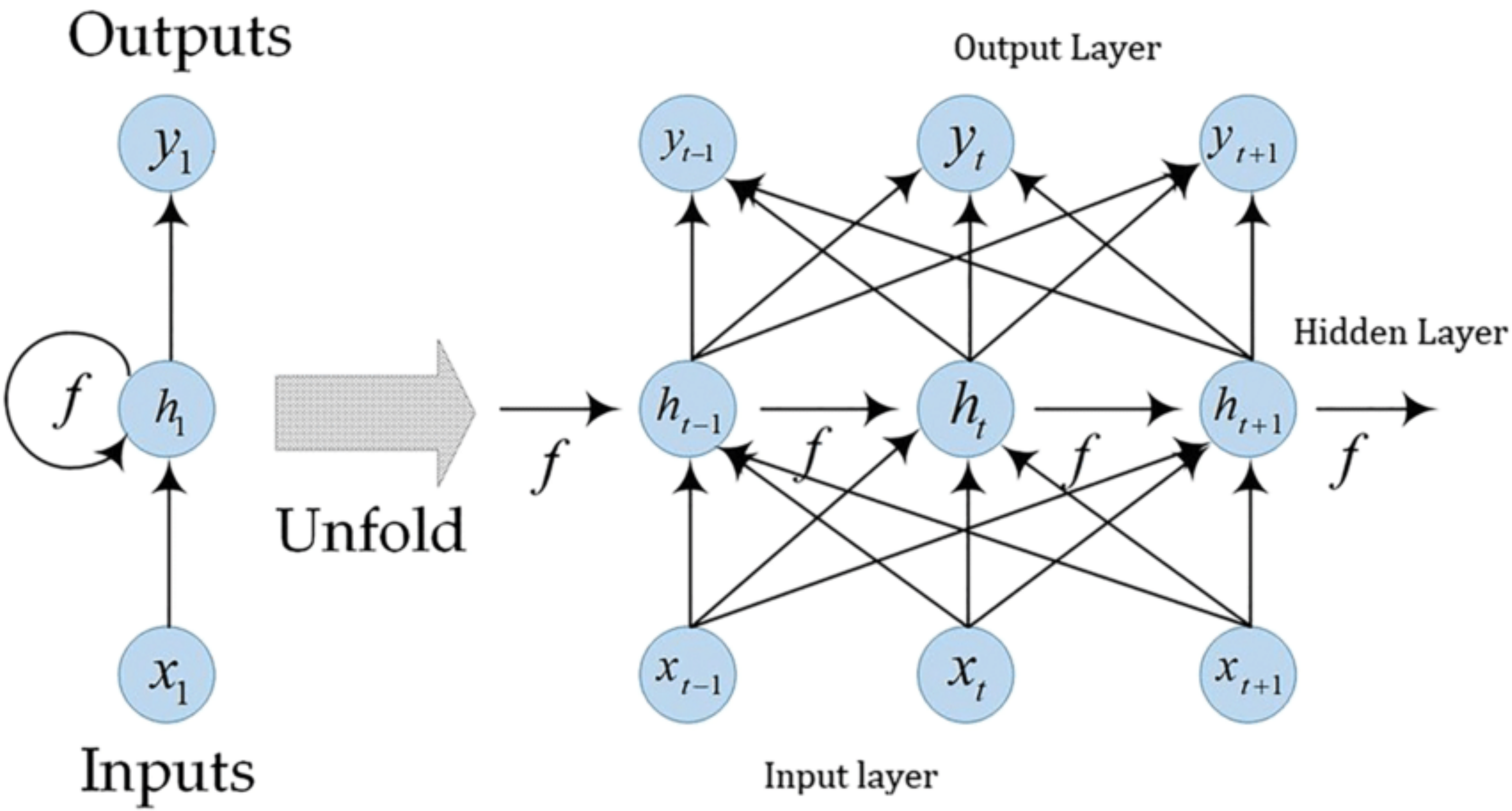
Theorem (McGoff-M.-Nobel-Pillai)

Assume $(T_\theta, \mu_\theta)_{\theta \in \Theta}$ is parameterized family of Axiom A systems with observation densities $(g_\theta)_{\theta \in \Theta}$. If observation integrability (C2) and (C3) hold and observation regularity (M3) and (L2) hold then MLE is consistent.

What about recurrent neural networks



What about recurrent neural networks



Our results

Agazzi-M-Lu (2022) – We state conditions under which we can prove

1. The convergence of the dynamics of the finite-width RNN to its infinite-width limit (the mean-field limit) using a coupling argument.
2. Gradient descent trains these networks to optimal fixed points given infinite training time. This optimality result holds in the feature-learning regime, as opposed to previous results that hold in the NTK regime.
3. There is a limiting stochastic ordinary differential equation that characterizes the dynamics of the network, in particular the weights. There is existence, uniqueness, and stability for the solution of the underlying ordinary differential equation