Asymptotic Theory for Linear-Chain Conditional Random Fields

Mathieu Sinn, Pascal Poupart
University of Waterloo

14th International Conference on Artificial Intelligence and Statistics, Ft. Lauderlade, Florida

April 11, 2011

Motivation

An **L-CRF** models the distribution of $\mathbf{Y}_n$ **conditional on** $\mathbf{X}_n$ where

- $\mathbf{X}_n = (X_0, \ldots, X_n)$ are the **observations** (ranging in $\mathcal{X}$)
- $\mathbf{Y}_n = (Y_0, \ldots, Y_n)$ are the **labels** (ranging in $\mathcal{Y}$)

Basic assumption: $\mathbf{Y}_n$ conditional on $\mathbf{X}_n$ forms a **Markov Chain**.

Here we assume that

$$P_{\boldsymbol{\lambda}}(\mathbf{Y}_n = \mathbf{y}_n \,|\, \mathbf{X}_n = \mathbf{x}_n) \quad = \quad \frac{1}{Z_{\boldsymbol{\lambda}}(\mathbf{x}_n)} \prod_{t=1}^{n} \exp\left(\boldsymbol{\lambda}^T \mathbf{f}(x_t, y_{t-1}, y_t)\right)$$

for $\mathbf{x}_n = (x_0, \ldots, x_n)$, $\mathbf{y}_n = (y_0, \ldots, y_n)$.

Motivation

Suppose we observe $\mathbf{X}_n$ and $\mathbf{Y}_n$ where

- the distribution of $\mathbf{X}_n$ is **unknown**
- the distribution of $\mathbf{Y}_n$ conditional on $\mathbf{X}_n$ follows an L-CRF with **known** features $\mathbf{f}$ and **unknown** weights $\boldsymbol{\lambda}^*$.

How can we **estimate** $\boldsymbol{\lambda}^*$ from the data?

Common approach is to maximize **conditional log-likelihood**, i.e.,

$$\text{compute} \quad \hat{\boldsymbol{\lambda}}_n = \arg\max_{\boldsymbol{\lambda} \in \boldsymbol{\Theta}} \mathcal{L}_n(\boldsymbol{\lambda})$$

$$\text{where} \quad \mathcal{L}_n(\boldsymbol{\lambda}) = \frac{1}{n}\Big(\sum_{t=1}^{n} \boldsymbol{\lambda}^T \mathbf{f}(X_t, Y_{t-1}, Y_t) - \log Z_{\boldsymbol{\lambda}}(\mathbf{X}_n) \Big).$$

Motivation

Intuitively: the **larger** n, the **closer** $\hat{\boldsymbol{\lambda}}_n$ will be to $\boldsymbol{\lambda}^*$.

We would like to establish conditions for **consistency** of the MLEs, i.e., under which $\hat{\boldsymbol{\lambda}}_n$ **converges to** $\boldsymbol{\lambda}^*$ as $n \longrightarrow \infty$.

Problem: We don't have a model for infinite sequences!

Solution:

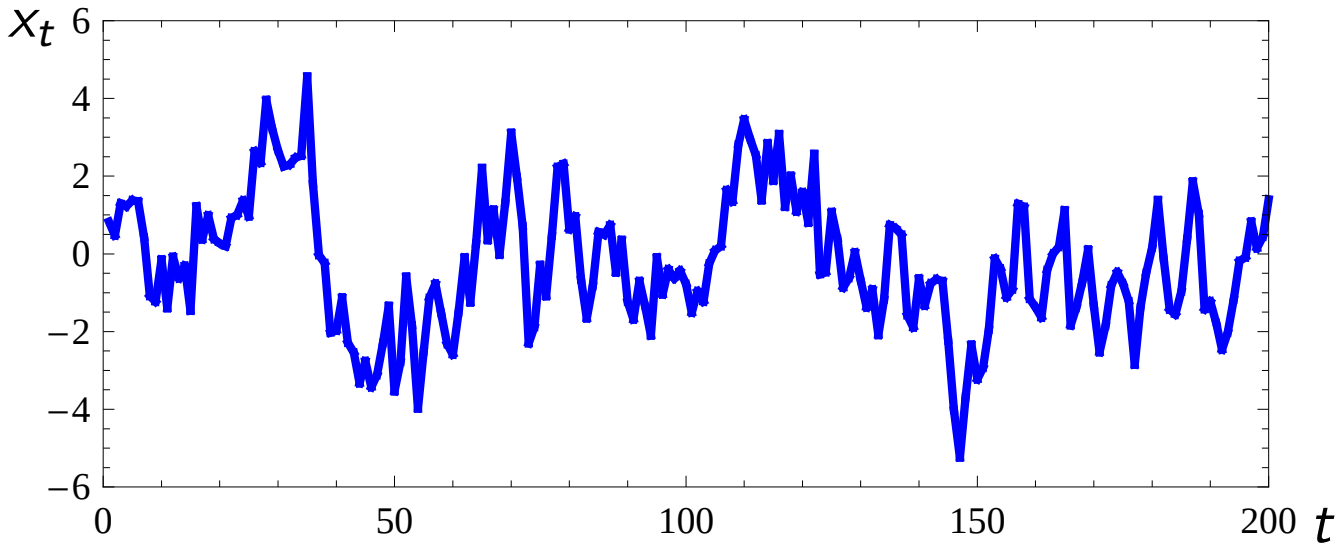
1. Define L-CRFs for double-sided **infinite sequences** $\mathbf{X} = (X_t)_{t \in \mathbb{Z}}$ and $\mathbf{Y} = (Y_t)_{t \in \mathbb{Z}}$.
2. Consider $\mathbf{X}_n$ and $\mathbf{Y}_n$ to be **finite subsequences** of $\mathbf{X}$ and $\mathbf{Y}$.

Motivation

Example: Suppose that $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{1, 2\}$. Let

$$\mathbf{f}(x_t, y_{t-1}, y_t) = \begin{bmatrix} \mathbf{1}(y_t = 1)\, x_t \\ \mathbf{1}(y_t = 2)\, x_t \\ \mathbf{1}(y_{t-1} = 1, y_t = 1) \\ \mathbf{1}(y_{t-1} = 2, y_t = 2) \end{bmatrix} \quad \text{and} \quad \boldsymbol{\lambda}^* = \begin{bmatrix} 1 \\ -1 \\ 2 \\ 2 \end{bmatrix}.$$

Here $\mathbf{X}$ is an AR(1) process: $X_t = 0.75\, X_{t-1} + \epsilon_t$ with $\epsilon_t \sim \mathrm{N}(0, 1)$.

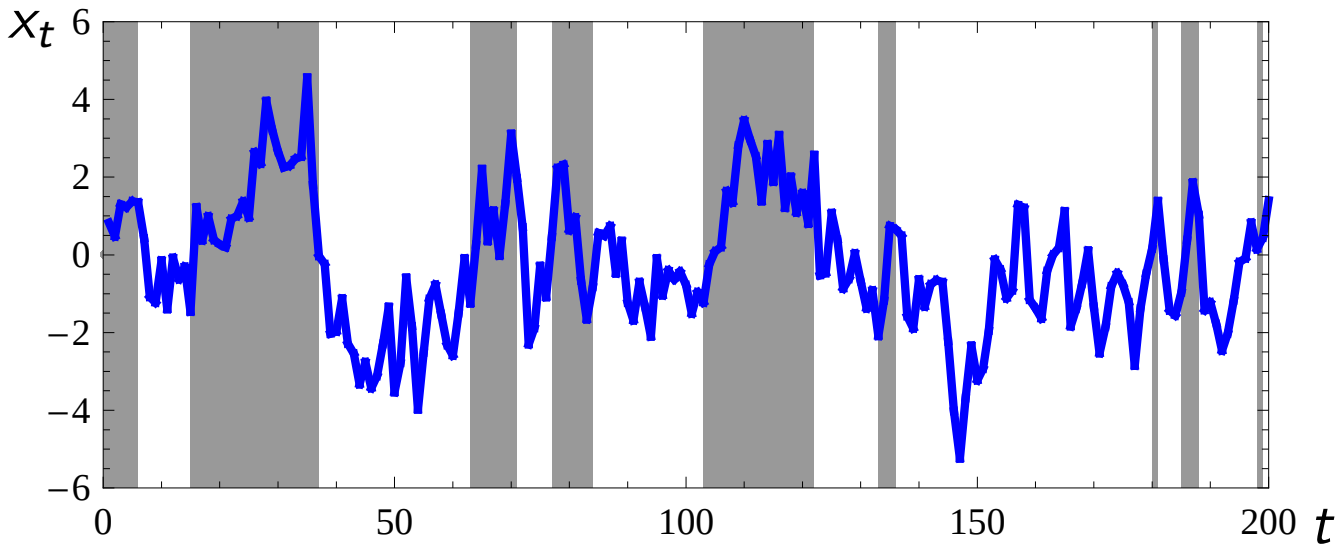


Motivation

Example: Suppose that $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{1, 2\}$. Let

$$\mathbf{f}(x_t, y_{t-1}, y_t) = \begin{bmatrix} \mathbf{1}(y_t = 1)\, x_t \\ \mathbf{1}(y_t = 2)\, x_t \\ \mathbf{1}(y_{t-1} = 1, y_t = 1) \\ \mathbf{1}(y_{t-1} = 2, y_t = 2) \end{bmatrix} \quad \text{and} \quad \boldsymbol{\lambda}^* = \begin{bmatrix} 1 \\ -1 \\ 2 \\ 2 \end{bmatrix}.$$

Y conditional on **X** follows an L-CRF for infinite sequences.

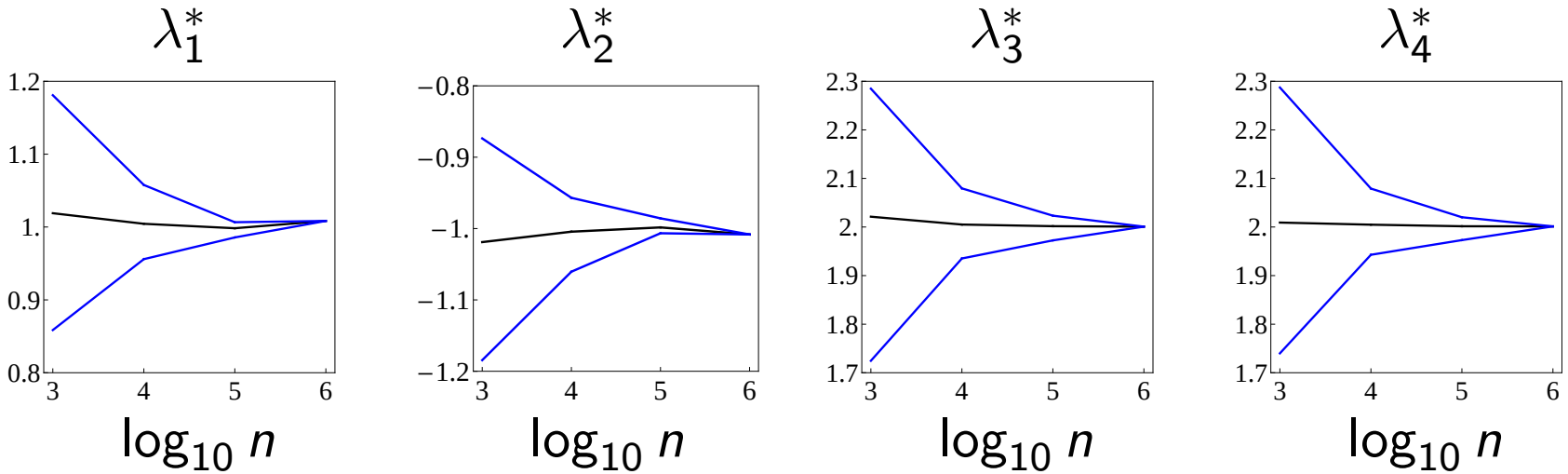


Motivation

Based on this **subsequence**, we obtain the MLE

$$\hat{\boldsymbol{\lambda}}_n = \begin{bmatrix} 0.7 \\ -0.7 \\ 2.1 \\ 2.2 \end{bmatrix} \quad \text{of} \quad \boldsymbol{\lambda}^* = \begin{bmatrix} 1 \\ -1 \\ 2 \\ 2 \end{bmatrix}.$$

As $n \to \infty$, the MLEs converge to $\boldsymbol{\lambda}^*$:



Motivation

What if **f** is **unknown**?

- Use **prior knowledge** to obtain a reasonable **"guess"**
- Use **parametric approximations**, e.g., linear CRFs

What can we say about the limit of $\hat{\boldsymbol{\lambda}}_n$ in those cases?

Our present work

- confirms that, if **f** is known, **"everything works fine"**
- provides a **framework** to study the asymptotics of $\hat{\boldsymbol{\lambda}}_n$ in more difficult cases
- establishes connections between L-CRFs and the theories of **weak ergodicity** and Markov chains in **random environments**

L-CRFs for Infinite Sequences

L-CRFs for Infinite Sequences

Goal Given a sequence $\mathbf{x} = (x_t)_{t\in\mathbb{Z}}$ in $\mathcal{X}$, define the distribution of $\mathbf{Y}$ conditional on $\mathbf{X} = \mathbf{x}$.

Approach Consider the matrix of **potentials**,

$$\mathbf{M}(x) \;=\; \Big[\, \exp\left(\boldsymbol{\lambda}^T \mathbf{f}(x, i, j)\right) \Big]_{i,j\in\mathcal{Y}}$$

For $s, t \in \mathbb{Z}$ with $s \le t$, let

$$\begin{aligned} \boldsymbol{\alpha}_s^t(\mathbf{x}) &= \mathbf{M}(x_t)^T \ldots \mathbf{M}(x_s)^T (1, 1, \ldots, 1)^T \\ \boldsymbol{\beta}_s^t(\mathbf{x}) &= \mathbf{M}(x_{s+1}) \ldots \mathbf{M}(x_t)\, (1, 1, \ldots, 1)^T. \end{aligned}$$

Proposition 1 [Lafferty et al. (2001), Wallach (2004)]

$$\begin{aligned} &P_{\boldsymbol{\lambda}}^{(0,n)}(Y_t = y_t, \ldots, Y_{t+k} = y_{t+k} \,|\, \mathbf{X} = \mathbf{x}) \\ &\quad = \frac{\alpha_1^t(\mathbf{x}, y_t)\, \beta_{t+k}^n(\mathbf{x}, y_{t+k})}{\boldsymbol{\alpha}_1^t(\mathbf{x})^T \boldsymbol{\beta}_t^n(\mathbf{x})} \prod_{i=1}^{k} \exp\left(\boldsymbol{\lambda}^T \mathbf{f}(x_{t+i}, y_{t+i-1}, y_{t+i})\right). \end{aligned}$$

L-CRFs for Infinite Sequences

Take into account the **observational context** $(X_{-n}, \ldots, X_n)$:

$$P_{\boldsymbol{\lambda}}^{(-n,n)}(Y_t = y_t, \ldots, Y_{t+k} = y_{t+k} \,|\, \mathbf{X} = \mathbf{x}) := \frac{\alpha_{-n}^{t}(\mathbf{x}, y_t)\, \beta_{t+k}^{n}(\mathbf{x}, y_{t+k})}{\boldsymbol{\alpha}_{-n}^{t}(\mathbf{x})^T \boldsymbol{\beta}_{t}^{n}(\mathbf{x})} \prod_{i=1}^{k} \exp\left(\boldsymbol{\lambda}^T \mathbf{f}(x_{t+i}, y_{t+i-1}, y_{t+i})\right).$$

Theorem 1 Suppose the features $\mathbf{f}$ are **bounded** and the weights $\boldsymbol{\lambda}$ are **finite**. Then the following limit is well-defined:

$$P_{\boldsymbol{\lambda}}(Y_t = y_t, \ldots, Y_{t+k} = y_{t+k} \,|\, \mathbf{X} = \mathbf{x}) := \lim_{n \to \infty} P_{\boldsymbol{\lambda}}^{(-n,n)}(Y_t = y_t, \ldots, Y_{t+k} = y_{t+k} \,|\, \mathbf{X} = \mathbf{x}).$$

Moreover, there exists a $\kappa \in (0, 1)$ (not depending on $\mathbf{x}$) such that

$$\left| P_{\boldsymbol{\lambda}}^{(-n,n)}(Y_t = y_t, \ldots, Y_{t+k} = y_{t+k} \,|\, \mathbf{X} = \mathbf{x}) - P_{\boldsymbol{\lambda}}(Y_t = y_t, \ldots, Y_{t+k} = y_{t+k} \,|\, \mathbf{X} = \mathbf{x}) \right| = O(\kappa^n).$$

L-CRFs for Infinite Sequences

Proof of Theorem 1 Recall the definition

$$P_{\boldsymbol{\lambda}}^{(-n,n)}(Y_t = y_t, \ldots, Y_{t+k} = y_{t+k} \,|\, \mathbf{X} = \mathbf{x})$$
$$:= \frac{\alpha_{-n}^{t}(\mathbf{x}, y_t)\, \beta_{t+k}^{n}(\mathbf{x}, y_{t+k})}{\boldsymbol{\alpha}_{-n}^{t}(\mathbf{x})^T \boldsymbol{\beta}_{t}^{n}(\mathbf{x})} \prod_{i=1}^{k} \exp\left(\boldsymbol{\lambda}^T \mathbf{f}(x_{t+i}, y_{t+i-1}, y_{t+i})\right).$$

The rows of the matrices

$$\mathbf{M}(x_t)^T \ldots \mathbf{M}(x_{-n})^T \quad \text{and} \quad \mathbf{M}(x_{t+1}) \ldots \mathbf{M}(x_n)$$

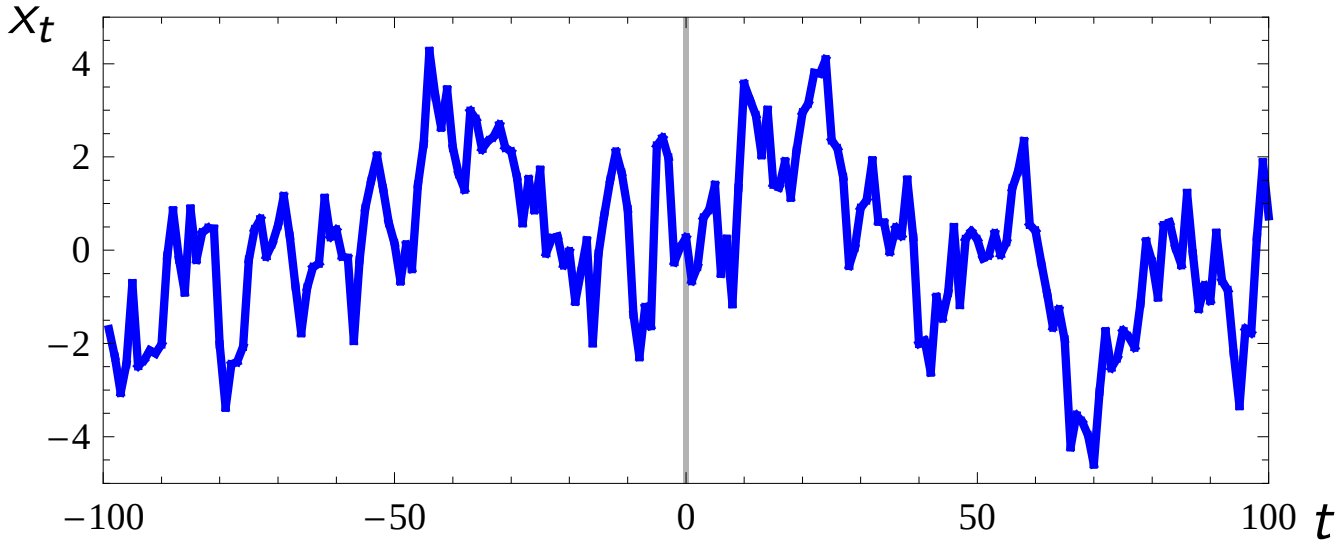
tend to **proportionality** as $n \to \infty$ [Seneta, 2006]. Hence

$$\lim_{n\to\infty} \frac{\boldsymbol{\alpha}_{-n}^{t}(\mathbf{x}, i)}{\boldsymbol{\alpha}_{-n}^{t}(\mathbf{x}, j)} = r_{ij} \quad \text{and} \quad \lim_{n\to\infty} \frac{\boldsymbol{\beta}_{t+k}^{n}(\mathbf{x}, i)}{\boldsymbol{\beta}_{t+k}^{n}(\mathbf{x}, j)} = s_{ij}.$$

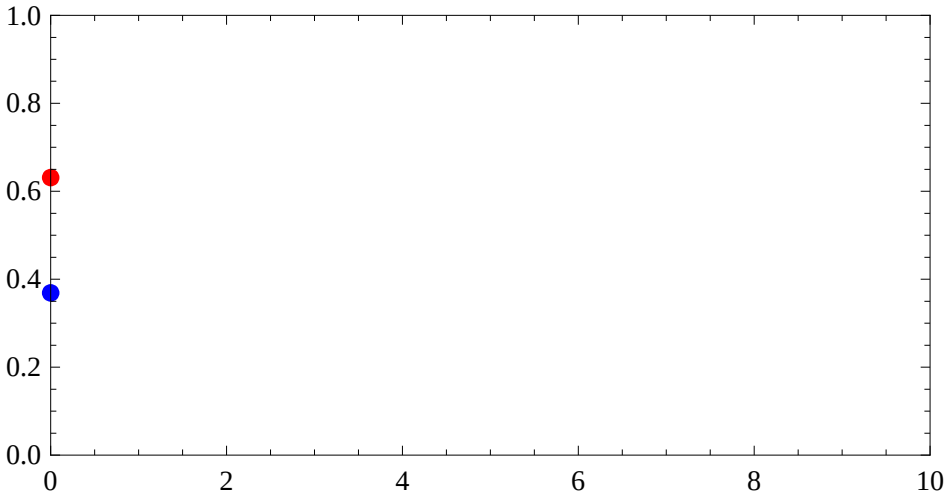
Kolmogorov's extension theorem: The collection of all marginal distributions specifies the distribution of $\mathbf{Y}$ conditional on $\mathbf{X} = \mathbf{x}$.

L-CRFs for Infinite Sequences

Example: Consider the distribution of Y_t at $t = 0$:



As n **increases**, the distribution **converges**:

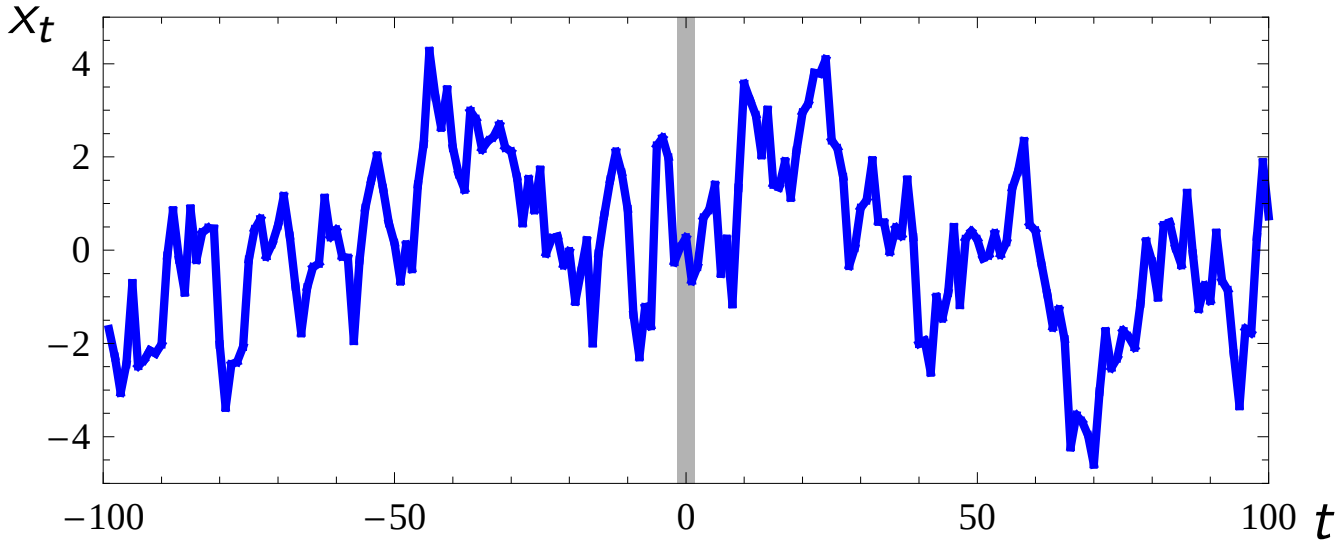


$P_{\lambda}^{(-n,n)}(Y_t = 1 \mid \mathbf{X} = \mathbf{x})$

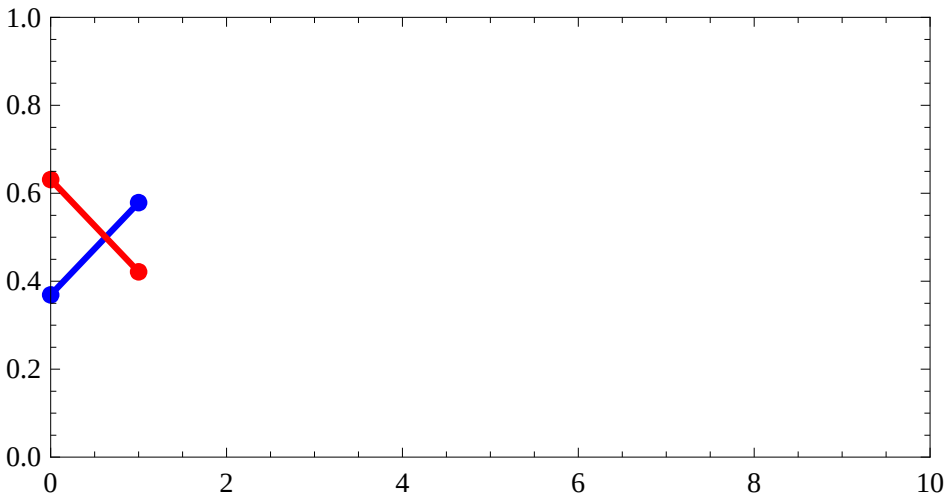
$P_{\lambda}^{(-n,n)}(Y_t = 2 \mid \mathbf{X} = \mathbf{x})$

L-CRFs for Infinite Sequences

Example: Consider the distribution of Y_t at $t = 0$:



As n **increases**, the distribution **converges**:

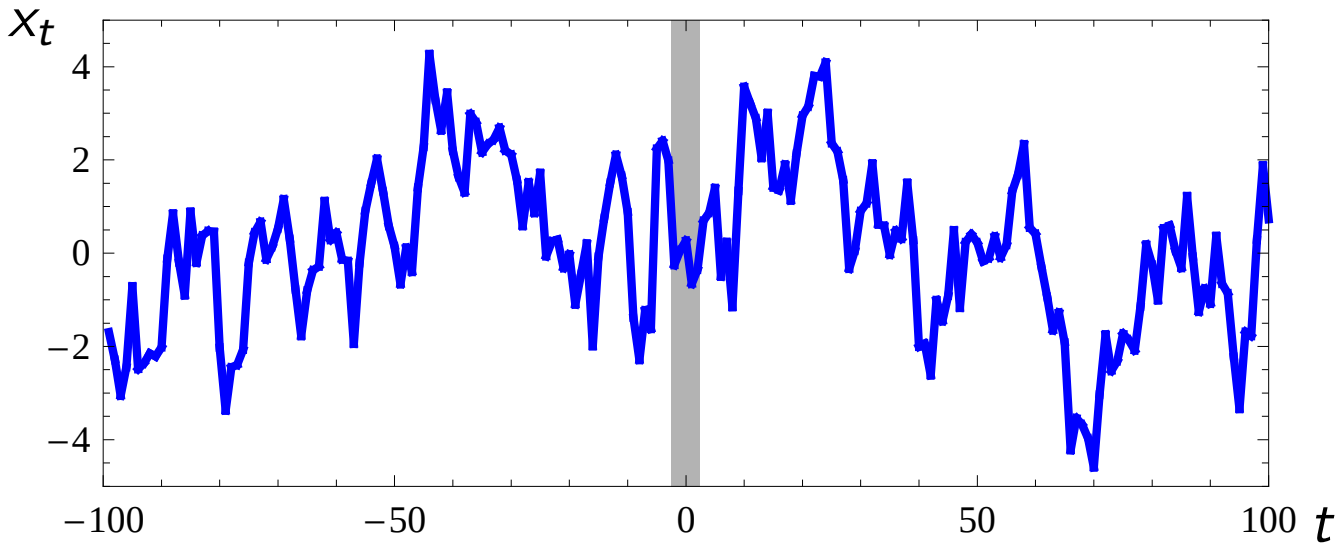


$P_\lambda^{(-n,n)}(Y_t = 1 \mid \mathbf{X} = \mathbf{x})$

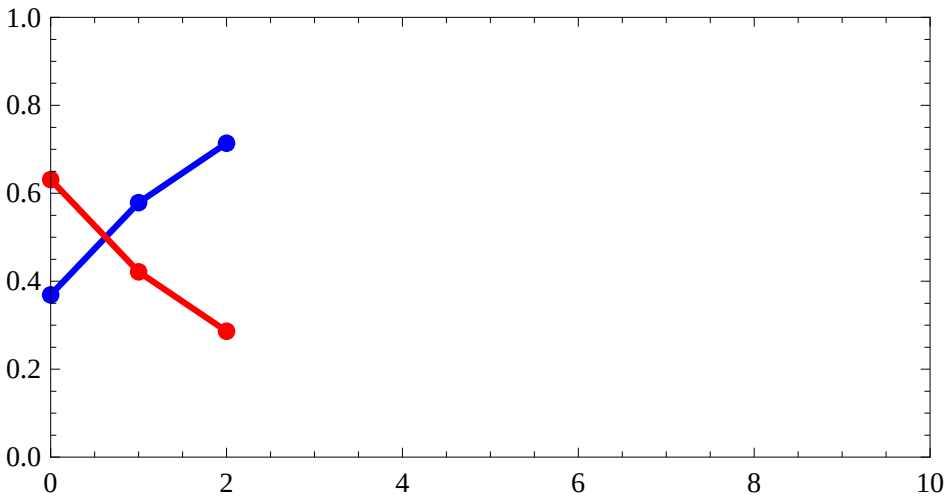
$P_\lambda^{(-n,n)}(Y_t = 2 \mid \mathbf{X} = \mathbf{x})$

L-CRFs for Infinite Sequences

Example: Consider the distribution of Y_t at $t = 0$:



As n **increases**, the distribution **converges**:

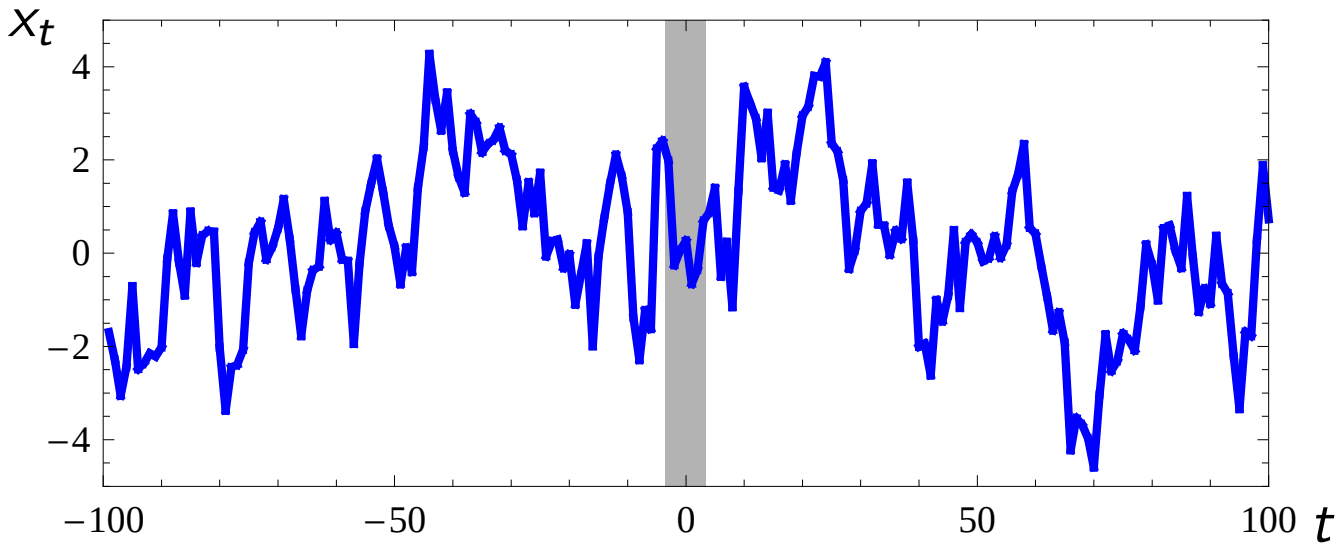


$P_\lambda^{(-n,n)}(Y_t = 1 \mid \mathbf{X} = \mathbf{x})$

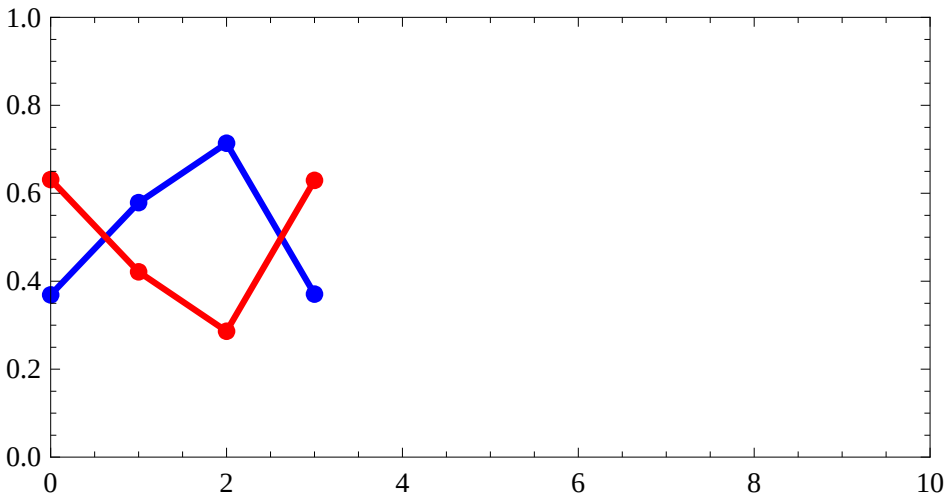
$P_\lambda^{(-n,n)}(Y_t = 2 \mid \mathbf{X} = \mathbf{x})$

L-CRFs for Infinite Sequences

Example: Consider the distribution of Y_t at $t = 0$:



As n **increases**, the distribution **converges**:

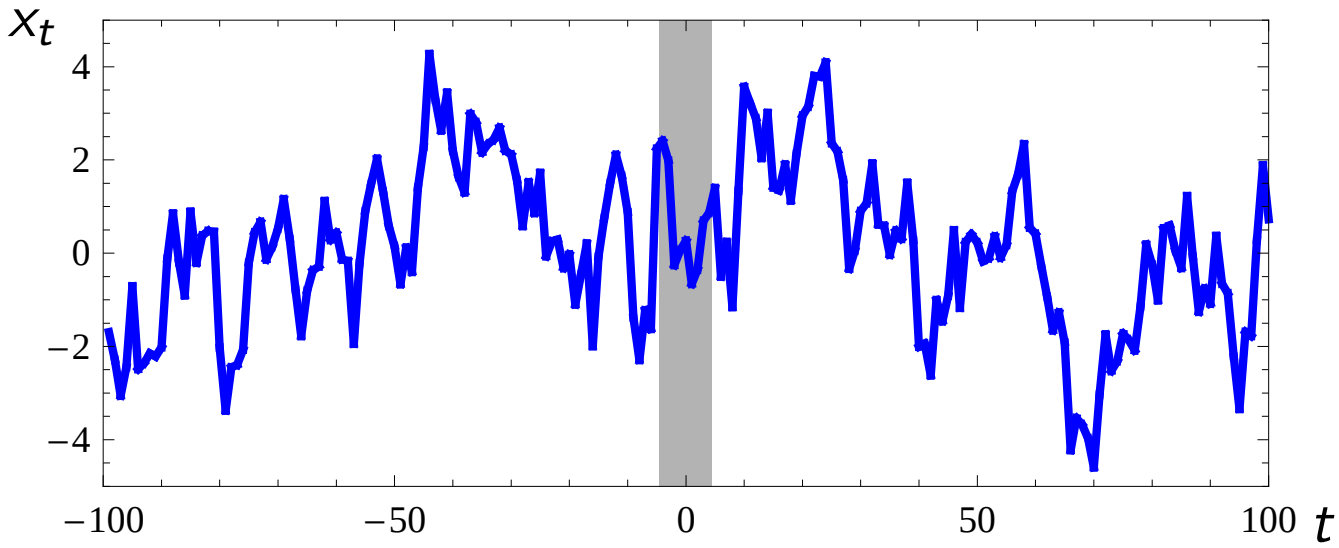


$P_\lambda^{(-n,n)}(Y_t = 1 \mid \mathbf{X} = \mathbf{x})$

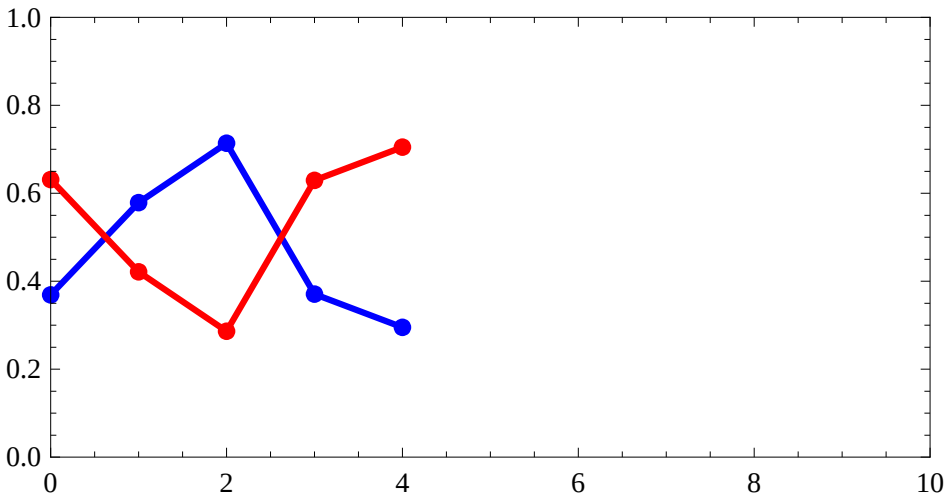
$P_\lambda^{(-n,n)}(Y_t = 2 \mid \mathbf{X} = \mathbf{x})$

L-CRFs for Infinite Sequences

Example: Consider the distribution of Y_t at $t = 0$:



As n **increases**, the distribution **converges**:

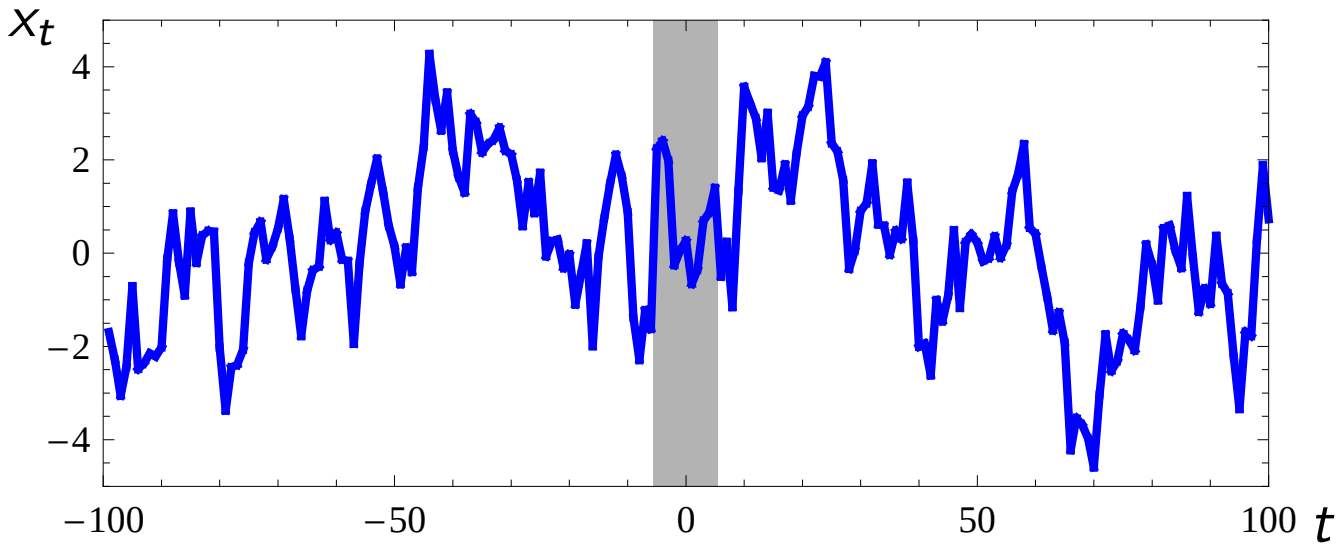


$P_\lambda^{(-n,n)}(Y_t = 1 \mid \mathbf{X} = \mathbf{x})$

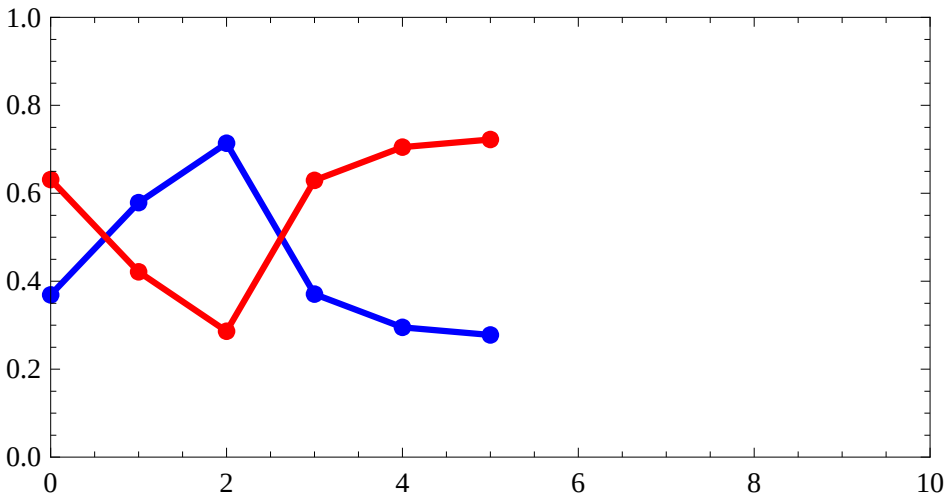
$P_\lambda^{(-n,n)}(Y_t = 2 \mid \mathbf{X} = \mathbf{x})$

L-CRFs for Infinite Sequences

Example: Consider the distribution of Y_t at $t = 0$:



As n **increases**, the distribution **converges**:

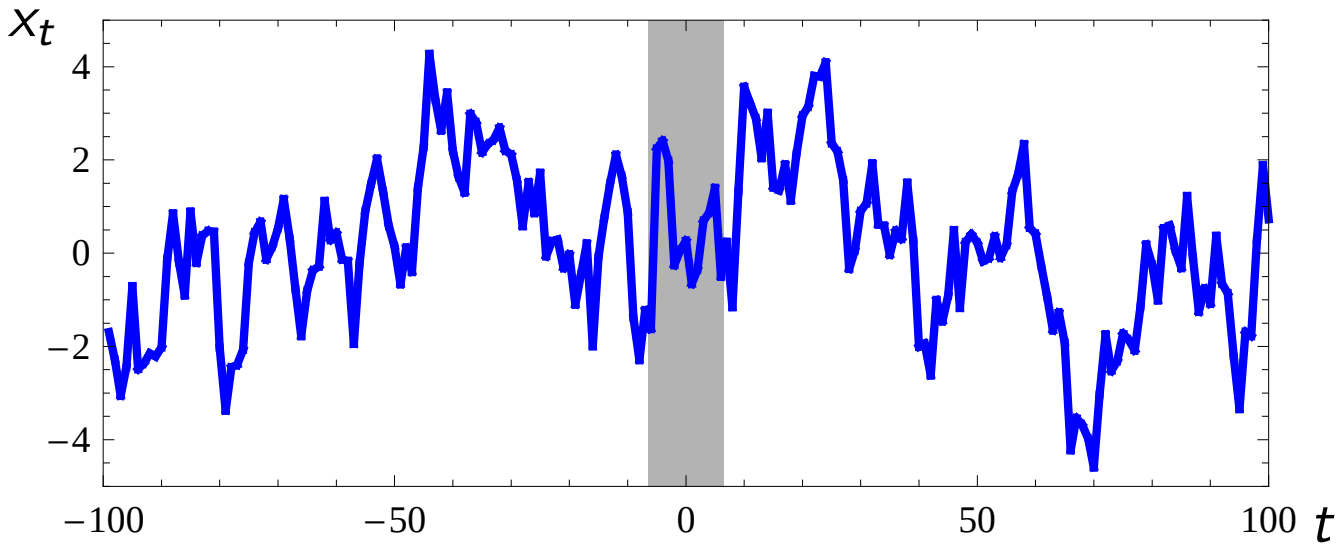


$P_\lambda^{(-n,n)}(Y_t = 1 \mid \mathbf{X} = \mathbf{x})$

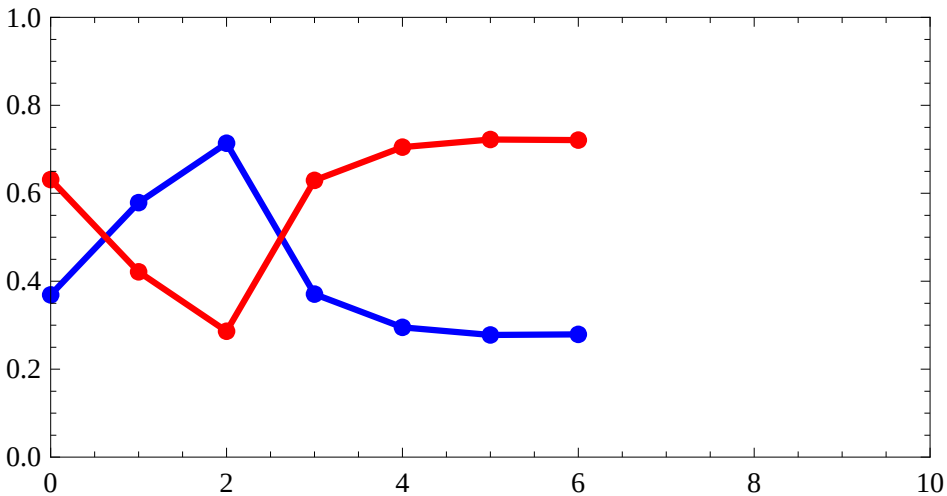
$P_\lambda^{(-n,n)}(Y_t = 2 \mid \mathbf{X} = \mathbf{x})$

L-CRFs for Infinite Sequences

Example: Consider the distribution of Y_t at $t = 0$:



As n **increases**, the distribution **converges**:

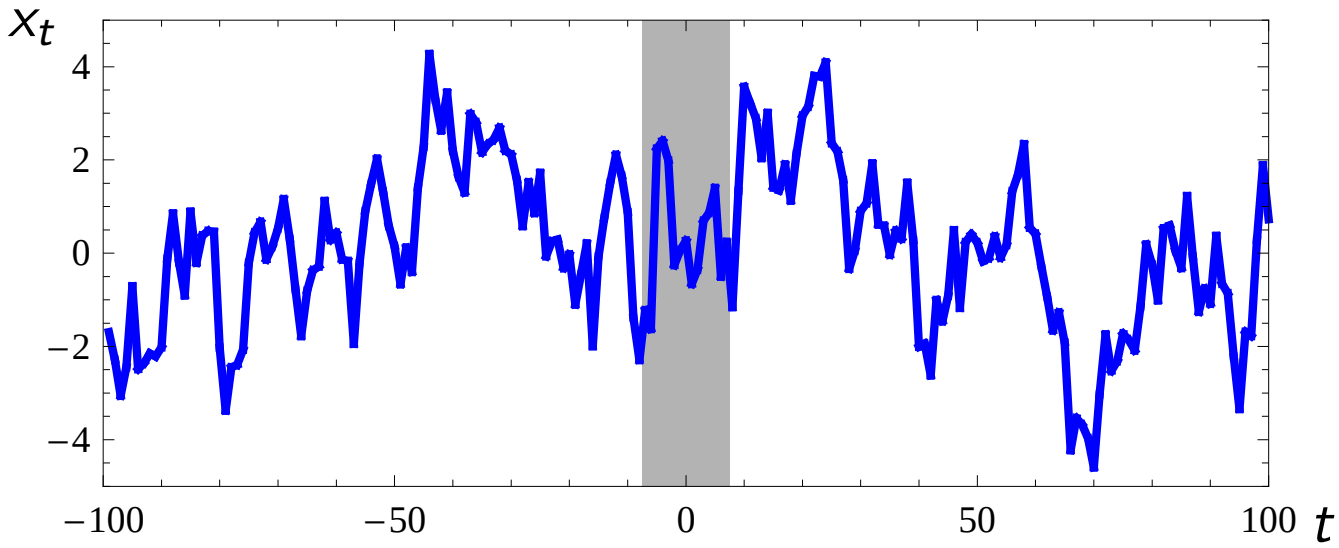


$P_\lambda^{(-n,n)}(Y_t = 1 \mid \mathbf{X} = \mathbf{x})$

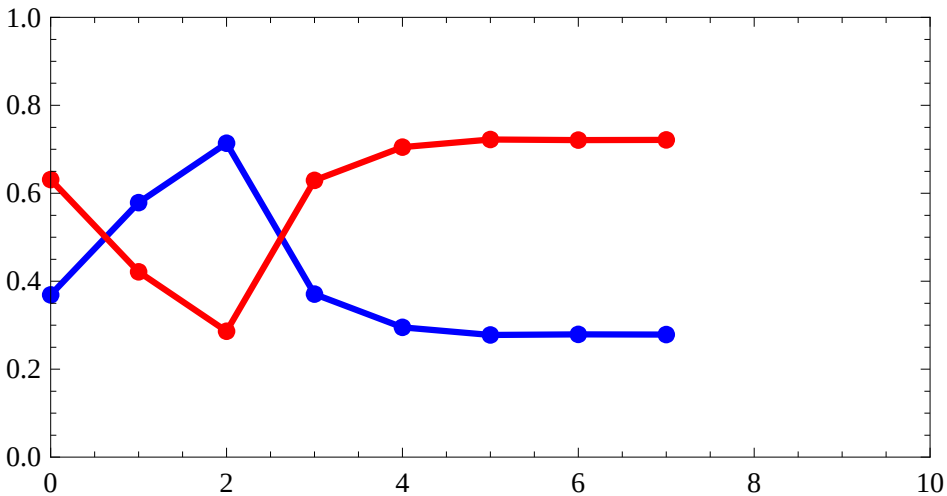
$P_\lambda^{(-n,n)}(Y_t = 2 \mid \mathbf{X} = \mathbf{x})$

L-CRFs for Infinite Sequences

Example: Consider the distribution of Y_t at $t = 0$:



As n **increases**, the distribution **converges**:

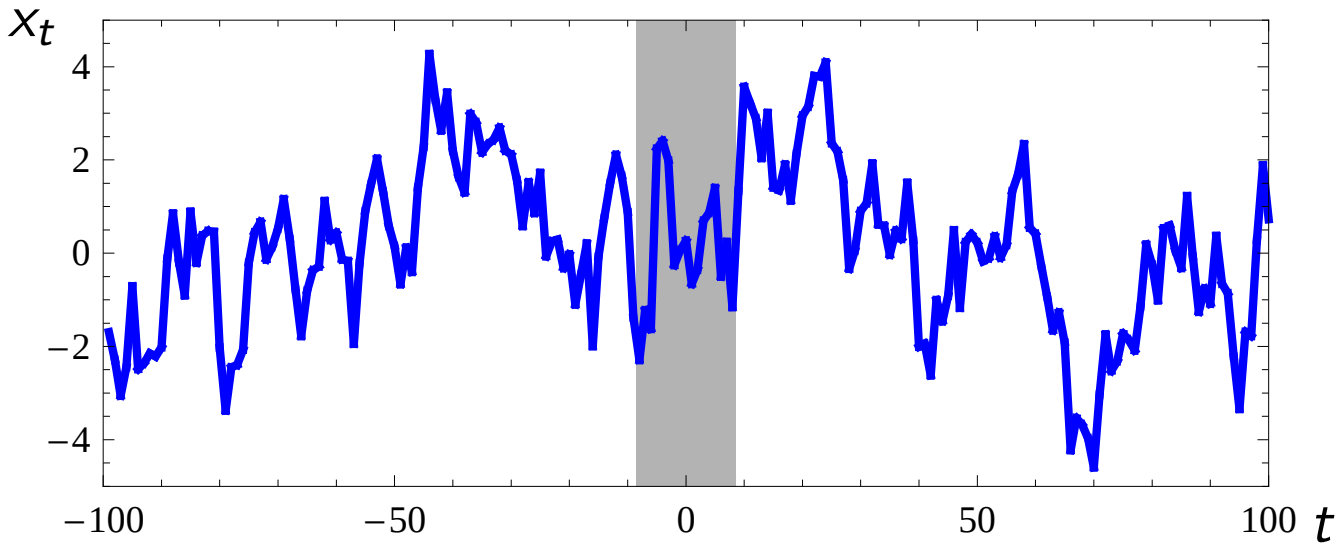


$P_\lambda^{(-n,n)}(Y_t = 1 \mid \mathbf{X} = \mathbf{x})$

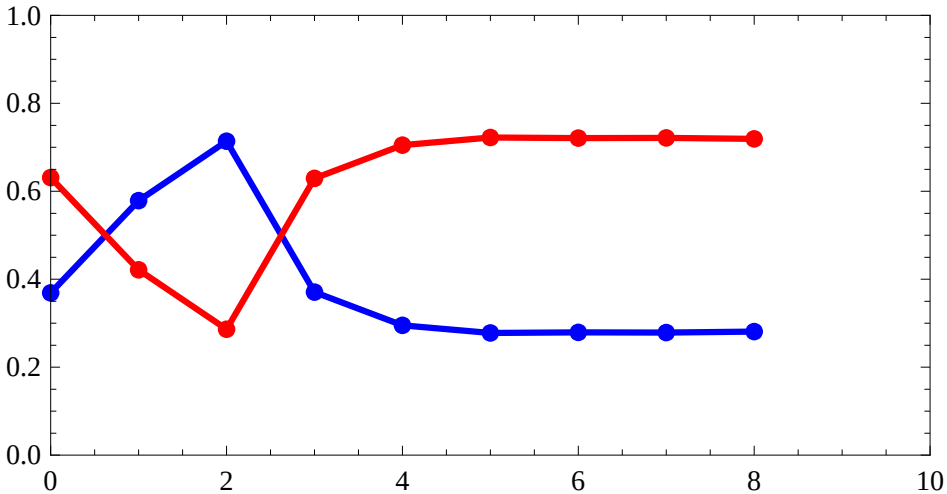
$P_\lambda^{(-n,n)}(Y_t = 2 \mid \mathbf{X} = \mathbf{x})$

L-CRFs for Infinite Sequences

Example: Consider the distribution of Y_t at $t = 0$:



As n **increases**, the distribution **converges**:

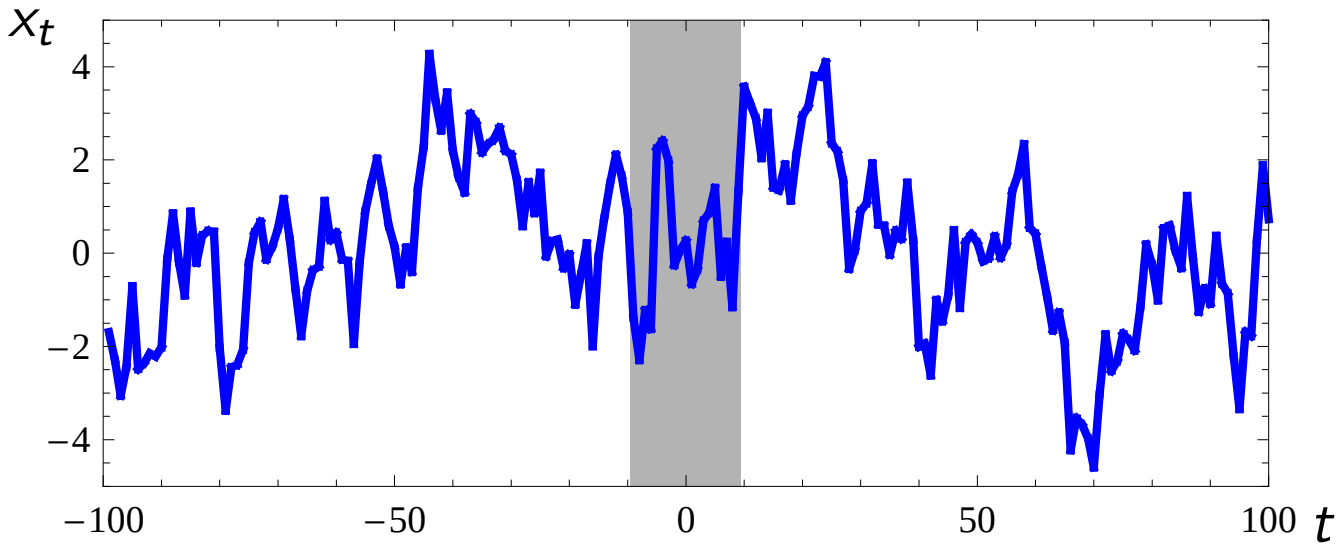


$P_\lambda^{(-n,n)}(Y_t = 1 \mid \mathbf{X} = \mathbf{x})$

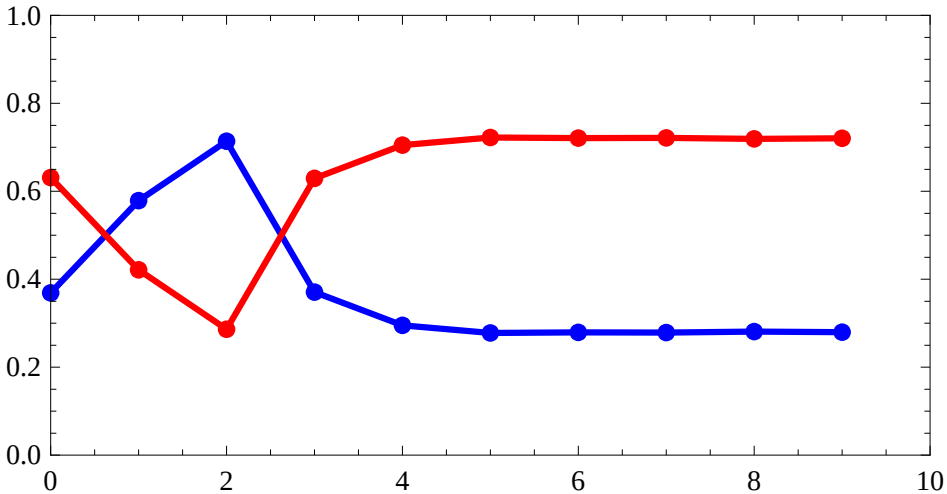
$P_\lambda^{(-n,n)}(Y_t = 2 \mid \mathbf{X} = \mathbf{x})$

L-CRFs for Infinite Sequences

Example: Consider the distribution of Y_t at $t = 0$:



As n **increases**, the distribution **converges**:

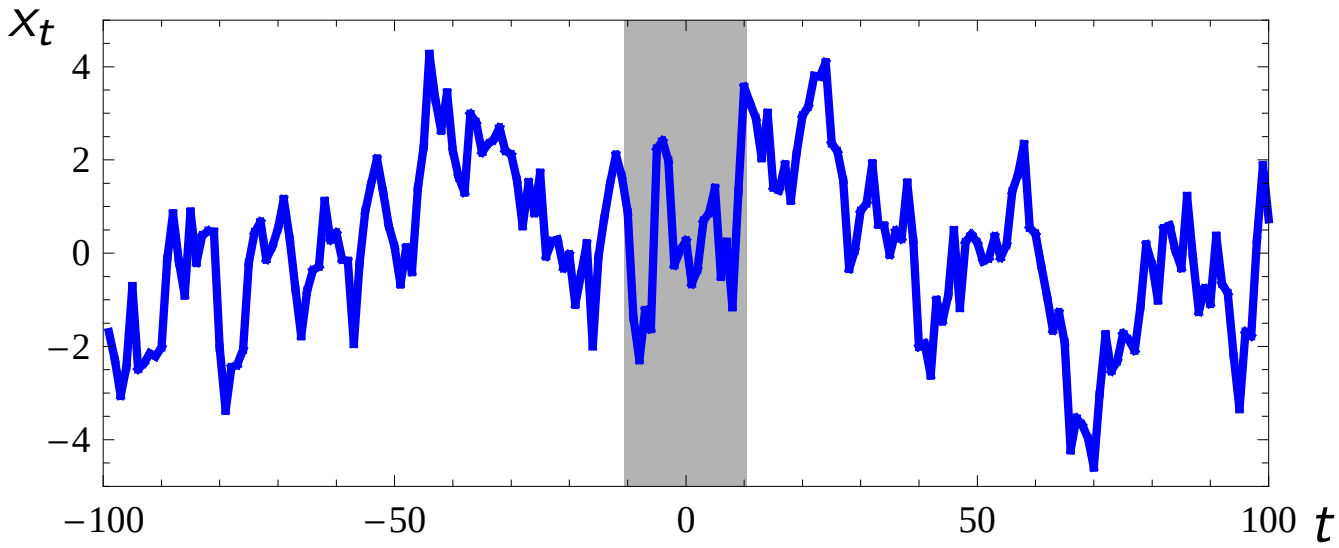


$P_\lambda^{(-n,n)}(Y_t = 1 \mid \mathbf{X} = \mathbf{x})$

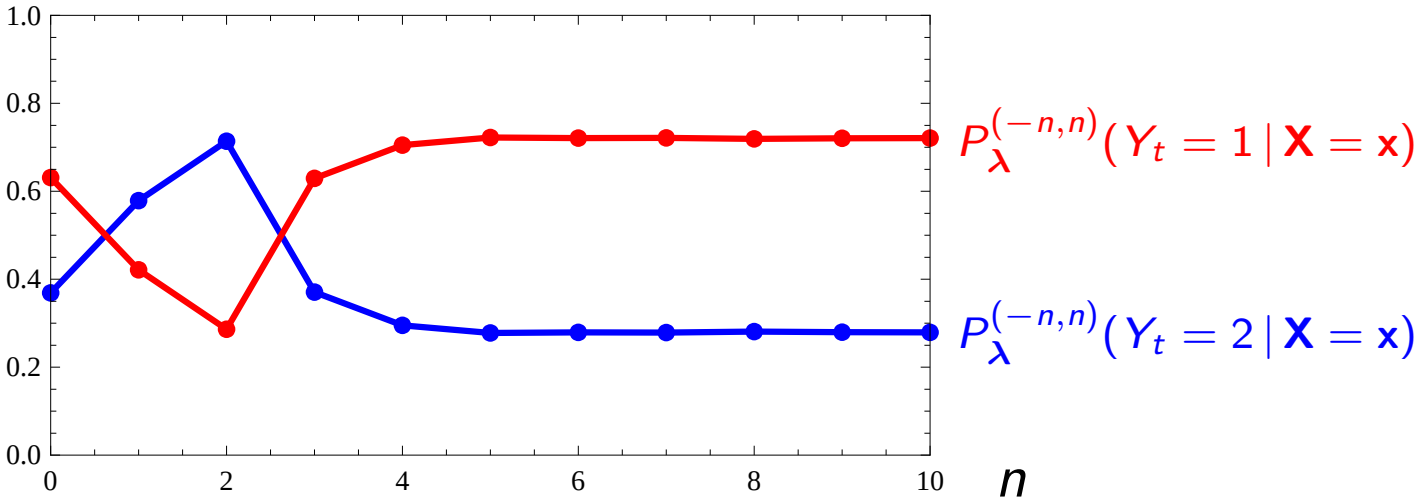
$P_\lambda^{(-n,n)}(Y_t = 2 \mid \mathbf{X} = \mathbf{x})$

L-CRFs for Infinite Sequences

Example: Consider the distribution of Y_t at $t = 0$:



As n **increases**, the distribution **converges**:



L-CRFs for Infinite Sequences

Corollary 1 Suppose the assumptions of Theorem 1 hold.

(i) $\mathbf{Y}$ conditional on $\mathbf{X} = \mathbf{x}$ is a **Markov chain** with

$$P_{\boldsymbol{\lambda}}(Y_{t+1} = y_{t+1} \,|\, Y_t = y_t, \mathbf{X} = \mathbf{x}) = \exp\left(\boldsymbol{\lambda}^T \mathbf{f}(x_{t+1}, y_t, y_{t+1})\right) \lim_{n\to\infty} \frac{\beta^n_{t+1}(\mathbf{x}, y_{t+1})}{\beta^n_t(\mathbf{x}, y_t)}.$$

In particular, $P_{\boldsymbol{\lambda}}(Y_{t+1} = y_{t+1} \,|\, Y_t = y_t, \mathbf{X} = \mathbf{x}) \geq c$ for some $c > 0$ not depending on $\mathbf{x}$.

(ii) If $\mathbf{X}$ is stationary, then the joint sequence $(\mathbf{X}, \mathbf{Y})$ is **stationary**.

(iii) If $\mathbf{X}$ is ergodic, then $(\mathbf{X}, \mathbf{Y})$ is **ergodic**. In particular, for any $g : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfying $E_{\boldsymbol{\lambda}}|g(X_t, Y_{t-1}, Y_t)| < \infty$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} g(X_t, Y_{t-1}, Y_t) = E_{\boldsymbol{\lambda}}[\, g(X_t, Y_{t-1}, Y_t)] \qquad P_{\boldsymbol{\lambda}}\text{-a.s.}$$

Consistency of the MLEs

Consistency of the MLEs

Suppose we observe the **finite subsequences** $\mathbf{X}_n = (X_0, \ldots, X_n)$ and $\mathbf{Y}_n = (Y_0, \ldots, Y_n)$ of $\mathbf{X}$ and $\mathbf{Y}$ where

- the distribution of $\mathbf{X}$ is **unknown**
- the distribution of $\mathbf{Y}$ conditional on $\mathbf{X}$ follows an L-CRF for infinite sequences with **known** features $\mathbf{f}$ and **unknown** weights $\boldsymbol{\lambda}^*$.

In order to **estimate** $\boldsymbol{\lambda}^*$,

$$\text{compute} \quad \hat{\boldsymbol{\lambda}}_n = \arg\max_{\boldsymbol{\lambda} \in \boldsymbol{\Theta}} \mathcal{L}_n(\boldsymbol{\lambda})$$

$$\text{where} \quad \mathcal{L}_n(\boldsymbol{\lambda}) = \frac{1}{n}\Big(\sum_{t=1}^{n} \boldsymbol{\lambda}^T \mathbf{f}(X_t, Y_{t-1}, Y_t) - \log Z_{\boldsymbol{\lambda}}(\mathbf{X}_n) \Big).$$

Under which conditions $\lim_{n\to\infty} \hat{\boldsymbol{\lambda}}_n = \boldsymbol{\lambda}^*$ $P_{\boldsymbol{\lambda}^*}$-a.s.?

Consistency of the MLEs

Lemma 1 Suppose that (A1) the features $\mathbf{f}$ are **bounded** and the weights $\boldsymbol{\lambda}$ are **finite**, (A2) the process $\mathbf{X}$ is **ergodic**, (A3) the parameter space $\boldsymbol{\Theta}$ is **compact**. Then the following holds:

(i) There exists a function $\mathcal{L}(\boldsymbol{\lambda})$ such that, for every $\boldsymbol{\lambda} \in \boldsymbol{\Theta}$,

$$\lim_{n\to\infty} \mathcal{L}_n(\boldsymbol{\lambda}) = \mathcal{L}(\boldsymbol{\lambda}) \qquad P_{\boldsymbol{\lambda}^*}\text{-a.s.}$$

(ii) The convergence of $\mathcal{L}_n(\boldsymbol{\lambda})$ to $\mathcal{L}(\boldsymbol{\lambda})$ is **uniform** on $\boldsymbol{\Theta}$.

(iii) The limit of $\nabla^2 \mathcal{L}_n(\boldsymbol{\lambda})$ is **finite**:

$$\lim_{n\to\infty} \nabla^2 \mathcal{L}_n(\boldsymbol{\lambda}) = -\left(\gamma_{\boldsymbol{\lambda}}(0) + 2\sum_{k=1}^{\infty} \gamma_{\boldsymbol{\lambda}}(k)\right) \qquad P_{\boldsymbol{\lambda}^*}\text{-a.s.}$$

where $\gamma_{\boldsymbol{\lambda}}(k) = \mathrm{Cov}_{\boldsymbol{\lambda}}\big[\mathbf{f}(X_t, Y_{t-1}, Y_t), \mathbf{f}(X_{t+k}, Y_{t+k-1}, Y_{t+k})\big]$.

Consistency of the MLEs

Theorem 2 Suppose that (A1)-(A3) hold. Moreover, suppose that the limit of $\nabla^2 \mathcal{L}_n(\boldsymbol{\lambda})$ is **non-singular** for every $\boldsymbol{\lambda} \in \boldsymbol{\Theta}$. Then $\mathcal{L}(\boldsymbol{\lambda})$ is **strictly concave** and hence

$$\lim_{n \to \infty} \hat{\boldsymbol{\lambda}}_n = \boldsymbol{\lambda}^* \qquad P_{\boldsymbol{\lambda}^*}\text{-a.s.}$$

What if the limit of $\nabla^2 \mathcal{L}_n(\boldsymbol{\lambda})$ is **singular**?

Example: Suppose that $\mathbf{f}$ contains two **identical features** f_i, f_j. Then $\mathcal{L}(\boldsymbol{\lambda}) = \mathcal{L}(\boldsymbol{\lambda}^*)$ for any $\boldsymbol{\lambda} \in \boldsymbol{\Theta}$ with $\lambda_i + \lambda_j = \lambda_i^* + \lambda_j^*$ and $\lambda_k = \lambda_k^*$ for all other k.

Discussion

Discussion

Open questions for **future research**:

- What are **sufficient conditions** for non-singularity of the limit of $\nabla^2 \mathcal{L}(\boldsymbol{\lambda})$ and hence **identifiability** of $\boldsymbol{\lambda}^*$?
- Can we weaken the assumption of **boundedness** of $\mathbf{f}$?
- Straightforward: generalization to L-CRFs of **higher orders**.
- What can we say about the quality of the **approximations**, e.g., by linear CRFs, when the true features are **unknown**?
- In the context of **human activity recognition**: what features should we use to obtain models that are **robust** towards **individual differences**?

References

[1] R. Cogburn (1984). The ergodic theory of Markov chains in random environments. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 66:109-128.

[2] O. Hernández-Lerma and J.B. Lasserre (2003). *Markov chains and invariant probabilities.* Basel, Switzerland: Birkhäuser.

[3] J. Lafferty, A. McCallum and F.C.N. Pereira (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Proceedings of the IEEE International Conference on Machine Learning (ICML)*.

[4] E. Seneta (2006). *Non-negative matrices and Markov chains. Revised Edition.* New York, NY: Springer.

[5] H. Wallach (2004). Conditional random fields: an introduction. *Technical Report MS-CIS-04-21*. University of Pennsylvania, Philadelphia.