Asymptotic Theory for Linear-Chain Conditional Random Fields

Mathieu Sinn
University of Waterloo

Pascal Poupart
University of Waterloo

Abstract

In this theoretical paper we develop an asymptotic theory for Linear-Chain Conditional Random Fields (L-CRFs) and apply it to derive conditions under which the Maximum Likelihood Estimates (MLEs) of the model weights are strongly consistent. We first define L-CRFs for infinite sequences and analyze some of their basic properties. Then we establish conditions under which ergodicity of the observations implies ergodicity of the joint sequence of observations and labels. This result is the key ingredient to derive conditions for strong consistency of the MLEs. Interesting findings are that the consistency crucially depends on the limit behavior of the Hessian of the likelihood function and that, asymptotically, the state feature functions do not matter.

1 INTRODUCTION

Conditional Random Fields (CRFs) are a widely popular model to describe the statistical dependence between sequences of "observations" and "labels" (Lafferty et al., 2001). Applications include natural language processing (Sutton and McCallum, 2006), the analysis of genome data (Culotta et al., 2005), and human activity recognition (Omar et al., 2010). Extensions of CRFs are hierarchical CRFs (Liao et al. 2007b), relational CRFs (Taskar et al. 2002), and semi-Markov CRFs (Sarawagi and Cohen 2004). The key idea of CRFs is to represent the distribution of the labels conditional on the observations by a Markov random field. The simplest non-trivial class of models is that of Linear-Chain CRFs (L-CRFs) where the random field forms a plain Markov chain.

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

In this paper, we study asymptotical properties of the Maximum Likelihood Estimates (MLEs) for the model weights. More specifically, we assume that we are given a sequence of observations and labels where the distribution of the labels conditional on the observations follows an L-CRF with known feature functions but unknown weights. In this setting, we investigate conditions under which the MLEs converge to the true weights as the length of the sequences goes to infinity. Note that, to state and to analyze this problem, a definition of L-CRFs for infinite sequences is required.

Our research is motivated by the following questions: How can the weights and feature functions be jointly estimated in the case where both are unknown? How robust is the training and inference towards a sampling bias (that is, when training and test data come from different distributions)? How well is the model identifiable in the presence of noisy data? To tackle these problems of great practical importance, the present paper aims to achieve a better understanding of the simplest case, namely, when the feature functions are known and a sampling bias or noisy data is absent. Furthermore, it provides a theoretical framework and useful techniques to study the more complicated cases.

This paper is structured as follows: In Sec. 2 we introduce some notation and review the definition of L-CRFs for finite sequences. In Sec. 3 we define L-CRFs for infinite sequences and derive some of their basic properties. Sec. 4 establishes conditions under which ergodicity of the sequence of observations implies ergodicity of the joint sequence of observations and labels. In Sec. 5 we apply the previous results to derive conditions under which the MLEs are strongly consistent. Sec. 6 concludes the paper.

2 PRELIMINARIES

Throughout this paper, let $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$ denote the sets of natural numbers, integers and real numbers, respectively. Let $\mathcal{X}$ be a metric space with the Borel sigma-field $\mathcal{A}$. To fix ideas, think of $\mathcal{X}$ as $\mathbb{R}^d$ equipped with the Euclidean norm. Furthermore, consider a finite set $\mathcal{Y}$ and let $\mathcal{B}$ denote the power set of $\mathcal{Y}$. In the following

we suppose that $\mathcal{Y} = \{1, \dots, \ell\}$ for some $\ell \in \mathbb{N}$. Now let $(\Omega, \mathcal{F})$ be a measurable space, and let $\boldsymbol{X} = (X_t)_{t\in\mathbb{Z}}$ and $\boldsymbol{Y} = (Y_t)_{t\in\mathbb{Z}}$ be sequences of measurable mappings from $(\Omega, \mathcal{F})$ into $\mathcal{X}$ and $\mathcal{Y}$, respectively. We refer to $\boldsymbol{X}$ and $\boldsymbol{Y}$ as the sequences of *observations* and *labels*, respectively.

A conventional Linear-Chain Conditional Random Field (L-CRF) specifies the conditional distribution of $\boldsymbol{Y}_n = (Y_0, \dots, Y_n)$ given $\boldsymbol{X}_n = (X_0, \dots, X_n)$. The distributions are parameterized as follows: Let $\boldsymbol{f}^{\text{state}}$ and $\boldsymbol{f}^{\text{trans}}$ be vectors of $\mathbb{R}$-valued functions defined on $\mathcal{X}\times\mathcal{Y}$ and $\mathcal{X}\times\mathcal{Y}^2$, respectively, and $\boldsymbol{\mu}$, $\boldsymbol{\nu}$ be $\mathbb{R}$-valued vectors of corresponding lengths. We call $\boldsymbol{f}^{\text{state}}$ and $\boldsymbol{f}^{\text{trans}}$ the *state* and *transition* feature functions, and $\boldsymbol{\mu}$, $\boldsymbol{\nu}$ the *model weights*. Write $\boldsymbol{\lambda}$ for the stacked vector $(\boldsymbol{\mu}, \boldsymbol{\nu})$, and let

$$\begin{aligned}\boldsymbol{\lambda}^T \boldsymbol{f}(\boldsymbol{x}_n, \boldsymbol{y}_n) &= \sum_{t=0}^{n} \boldsymbol{\mu}^T \boldsymbol{f}^{\text{state}}(x_t, y_t) \\ &+ \sum_{t=1}^{n} \boldsymbol{\nu}^T \boldsymbol{f}^{\text{trans}}(x_t, y_{t-1}, y_t) \quad (1)\end{aligned}$$

for $\boldsymbol{x}_n = (x_0, \dots, x_n) \in \mathcal{X}^{n+1}$ and $\boldsymbol{y}_n = (y_0, \dots, y_n) \in \mathcal{Y}^{n+1}$. We make three remarks at this point: First, the feature functions could also depend on t; here we restrict ourselves to models where the dependencies between observations and labels do not vary with time. Second, the assumption that the feature functions only depend on the current observation x_t is without loss of generality; if they depended on, say, x_{t-l}, …, x_{t+l}, then simply consider the modified observations $\tilde{x}_t = (x_{t-l}, \dots, x_{t+l})$ instead of x_t. Third, equation (1) suggests that there might be some overlap between the state and transition feature functions; this is indeed the case and will play an important role later.

Now let $\boldsymbol{x} = (x_t)_{t\in\mathbb{Z}}$ be a sequence in $\mathcal{X}$ and write $\boldsymbol{x}_n$ to denote the projection of $\boldsymbol{x}$ onto the components $(x_0, \dots, x_n)$. For any $\boldsymbol{y}_n = (y_0, \dots, y_n) \in \mathcal{Y}^{n+1}$, define the conditional probability

$$\begin{aligned}&P_{\boldsymbol{\lambda}}^{(0:n)}(\boldsymbol{Y}_n = \boldsymbol{y}_n \,|\, \boldsymbol{X} = \boldsymbol{x}) \\ &\qquad = \frac{1}{Z_{\boldsymbol{\lambda}}(\boldsymbol{x}_n)} \exp\left(\boldsymbol{\lambda}^T \boldsymbol{f}(\boldsymbol{x}_n, \boldsymbol{y}_n)\right) \quad (2)\end{aligned}$$

where $Z_{\boldsymbol{\lambda}}(\boldsymbol{x}_n)$ is the normalizing partition function

$$Z_{\boldsymbol{\lambda}}(\boldsymbol{x}_n) = \sum_{\boldsymbol{y}'_n \in \mathcal{Y}^{n+1}} \exp\left(\boldsymbol{\lambda}^T \boldsymbol{f}(\boldsymbol{x}_n, \boldsymbol{y}'_n)\right). \quad (3)$$

Note that the conditional probability in (2) only depends on the components $x_0, \dots, x_n$ of $\boldsymbol{x}$. Thus, $P_{\boldsymbol{\lambda}}^{(0:n)}$ specifies a conditional distribution for $\boldsymbol{Y}_n$ given $\boldsymbol{X}_n$. In the next section we will define L-CRFs for infinite sequences by considering conditional distributions which depend on the entire sequence $\boldsymbol{x}$.

The key step in applying L-CRFs is the specification of the feature functions and the model weights. A common approach is to select the feature functions manually according to some expert domain knowledge; alternative ways are to iteratively select those functions from a set of candidates which give the largest increase of the conditional log-likelihood (McCallum, 2003), or to use virtual evidence boosting (Liao et al. 2007a). Given a fixed set of feature functions, the model weights are usually learned from labeled training data by maximizing the conditional log-likelihood. In Sec. 5 we suppose that the "true" feature functions are known, and consider asymptotic properties of the Maximum Likelihood estimates.

In the remaining part of this section we introduce an alternative representation for the right hand side of (2), which is fundamental to all that follows. Consider the function $\boldsymbol{\alpha}_{\boldsymbol{\lambda}} : \mathcal{X} \to \mathbb{R}^\ell$ with the ith component

$$\alpha_{\boldsymbol{\lambda}}(x, i) = \exp\left(\boldsymbol{\mu}^T \boldsymbol{f}^{\text{state}}(x, i)\right).$$

Furthermore, let $\boldsymbol{M}_{\boldsymbol{\lambda}}(x)$ be the $\ell \times \ell$-matrix with the (i, j)-th component

$$m_{\boldsymbol{\lambda}}(x, i, j) = \exp\left(\boldsymbol{\mu}^T \boldsymbol{f}^{\text{state}}(x, j) + \boldsymbol{\nu}^T \boldsymbol{f}^{\text{trans}}(x, i, j)\right).$$

In the terminology of Markov Random Fields, $\alpha_{\boldsymbol{\lambda}}(x_t, y_t)$ is the *potential* of the event $X_t = x_t$ and $Y_t = y_t$ (with respect to the model weights $\boldsymbol{\lambda}$), while $m_{\boldsymbol{\lambda}}(x_t, y_{t-1}, y_t)$ is the potential of the event $X_t = x_t$, $Y_{t-1} = y_{t-1}$ and $Y_t = y_t$. For a sequence $\boldsymbol{x} = (x_t)_{t\in\mathbb{Z}}$ in $\mathcal{X}$ and $s, t \in \mathbb{Z}$ with $s \le t$, define

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

Write $\alpha_s^t(\boldsymbol{\lambda}, \boldsymbol{x}, i)$ and $\beta_s^t(\boldsymbol{\lambda}, \boldsymbol{x}, j)$ to denote the ith and jth component of $\boldsymbol{\alpha}_s^t(\boldsymbol{\lambda}, \boldsymbol{x})$ and $\boldsymbol{\beta}_s^t(\boldsymbol{\lambda}, \boldsymbol{x})$. Note that $\alpha_s^t(\boldsymbol{\lambda}, \boldsymbol{x}, y_t)$ is the potential of the event $X_s = x_s, \dots, X_t = x_t$ and $Y_t = y_t$, while $\beta_s^t(\boldsymbol{\lambda}, \boldsymbol{x}, y_s)$ is the potential of the event $X_s = x_s, \dots, X_t = x_t$ and $Y_s = y_s$. The following is well-known (Wallach, 2004):

Proposition 1. *Let $t, k \in \mathbb{N}$ such that $t + k \le n$. Then, for all $y_t, \dots, y_{t+k} \in \mathcal{Y}$,*

$$\begin{aligned}&P_{\boldsymbol{\lambda}}^{(0:n)}(Y_t = y_t, \dots, Y_{t+k} = y_{t+k} \,|\, \boldsymbol{X} = \boldsymbol{x}) \\ &\qquad = \frac{\alpha_1^t(\boldsymbol{\lambda}, \boldsymbol{x}, y_t)\, \beta_{t+k}^n(\boldsymbol{\lambda}, \boldsymbol{x}, y_{t+k})}{\boldsymbol{\alpha}_1^t(\boldsymbol{\lambda}, \boldsymbol{x})^T \boldsymbol{\beta}_t^n(\boldsymbol{\lambda}, \boldsymbol{x})} \\ &\qquad\qquad \times \prod_{i=1}^{k} m_{\boldsymbol{\lambda}}(x_{t+i}, y_{t+i-1}, y_{t+i})\end{aligned}$$

where, as usual, products over empty index sets are equal to 1.

3 L-CRFs FOR INFINITE SEQUENCES

In this section we define L-CRFs for infinite sequences and derive some of their basic properties. We will assume that the following condition is satisfied:

(A1) The feature functions $\boldsymbol{f}^{\text{state}}$ and $\boldsymbol{f}^{\text{trans}}$ are bounded, and the model weights $\boldsymbol{\lambda}$ are finite.

The next lemma states some obvious consequence of assumption (A1) which will be needed below. Let m_{inf} and m_{sup} denote the infimum and the supremum over all $m_{\boldsymbol{\lambda}}(x,i,j)$ with $x \in \mathcal{X}$ and $i,j \in \mathcal{Y}$.

Lemma 1. *Suppose that* (A1) *holds. Then* $m_{\text{inf}} > 0$ *and* $m_{\text{sup}} < \infty$. *In particular, the quantities* φ^2 *and* ψ^2 *defined by*

$$\varphi^2 \;:=\; \inf\left\{ \min_{i,j,k,l\in\mathcal{Y}} \frac{m(x,k,i)\,m(x,l,j)}{m(x,k,j)\,m(x,l,i)} : \quad x \in \mathcal{X} \right\},$$

$$\psi^2 \;:=\; \inf\left\{ \min_{i,j,k,l\in\mathcal{Y}} \frac{m(x,i,k)\,m(x,j,l)}{m(x,j,k)\,m(x,i,l)} : \quad x \in \mathcal{X} \right\},$$

are both strictly greater than zero.

3.1 Definition

Let $\boldsymbol{x} = (x_t)_{t\in\mathbb{Z}}$ be fixed. Our goal is to define the distribution of the infinite sequence $\boldsymbol{Y}$ conditional on $\boldsymbol{X} = \boldsymbol{x}$. Let $t \in \mathbb{Z}$ and $k \in \mathbb{N}$. We first specify the conditional marginal distribution of $(Y_t, \ldots, Y_{t+k})$. Then, by applying Kolmogorov's extension theorem, we obtain the conditional distribution for the entire sequence $\boldsymbol{Y}$. Let $y_t, \ldots, y_{t+k} \in \mathcal{Y}$. For $n \in \mathbb{N}$ such that $-n \le t$ and $n \le t+k$, define

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

Comparing this to Proposition 1 we see that $P_{\boldsymbol{\lambda}}^{(-n,n)}$ specifies the conditional distribution of $(Y_t, \ldots, Y_{t+k})$ in the fashion of a conventional L-CRF considering the finite observational context $(X_{-n}, \ldots, X_n)$. Theorem 2 below shows that this distribution converges as n tends to infinity. For the proof, we will need the following result.

Theorem 1. *Suppose that* (A1) *holds. For a fixed sequence* $(x_t)_{t\in\mathbb{N}}$ *in* $\mathcal{X}$, *let* $\boldsymbol{M}_n = (m_n(i,j))_{i,j\in\mathcal{Y}}$ *be given by* $\boldsymbol{M}_n = \boldsymbol{M}_{\boldsymbol{\lambda}}(x_1)\ldots\boldsymbol{M}_{\boldsymbol{\lambda}}(x_n)$. *Then there exist positive numbers* r_{ij} *such that, for all* $i,j,k \in \mathcal{Y}$,

$$\lim_{n\to\infty} \frac{m_n(i,k)}{m_n(j,k)} \;=\; r_{ij},$$

i.e., the rows of $\boldsymbol{M}_n$ *tend to proportionality as* n *tends to infinity. Moreover,*

$$\max_{k\in\mathcal{Y}}\left(\frac{m_n(i,k)}{m_n(j,k)}\right) \;\ge\; r_{ij} \;\ge\; \min_{k\in\mathcal{Y}}\left(\frac{m_n(i,k)}{m_n(j,k)}\right)$$

for all $n \in \mathbb{N}$, *and*

$$\min_{k\in\mathcal{Y}}\left(\frac{m_n(i,k)}{m_n(j,k)}\right) \;\ge\; \left[1 - 4\left(\frac{1-\psi}{1+\psi}\right)^n\right] \max_{k\in\mathcal{Y}}\left(\frac{m_n(i,k)}{m_n(j,k)}\right)$$

with ψ *given in Lemma 1. The same results apply, with* φ *instead of* ψ, *for* $\boldsymbol{M}_n = \boldsymbol{M}_{\boldsymbol{\lambda}}(x_1)^T\ldots\boldsymbol{M}_{\boldsymbol{\lambda}}(x_n)^T$.

Proof. The proof, which uses well-known results from the theory of weak ergodicity (Seneta, 2006), is included in the supplementary material. □

Now we are prepared to establish the following result.

Theorem 2. *Suppose that* (A1) *holds. Then the following limit is well-defined:*

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

Moreover, there exist constants $c > 0$ *and* $0 < \kappa < 1$, *which do not depend on* $\boldsymbol{x}$, *such that*

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

Proof. Define $\boldsymbol{G}_n := \boldsymbol{M}_{\boldsymbol{\lambda}}(x_t)^T\ldots\boldsymbol{M}_{\boldsymbol{\lambda}}(x_{-n})^T$ and $\boldsymbol{H}_n := \boldsymbol{M}_{\boldsymbol{\lambda}}(x_{t+k+1})\ldots\boldsymbol{M}_{\boldsymbol{\lambda}}(x_n)$. Write $g_n(i,j)$ and $h_n(i,j)$ for the (i,j)-th components of $\boldsymbol{G}_n$ and $\boldsymbol{H}_n$. Note that $\boldsymbol{\alpha}_{-n}^t(\boldsymbol{\lambda},\boldsymbol{x}) = \boldsymbol{G}_n\boldsymbol{\alpha}_{\boldsymbol{\lambda}}(x_{-n-1})$ and $\boldsymbol{\beta}_{t+k}^n(\boldsymbol{\lambda},\boldsymbol{x}) = \boldsymbol{H}_n(1,1,\ldots,1)^T$. Furthermore, with $\boldsymbol{F} := \boldsymbol{M}_{\boldsymbol{\lambda}}(x_{t+1})\ldots\boldsymbol{M}_{\boldsymbol{\lambda}}(x_{t+k})$, we have

$$\boldsymbol{\alpha}_{-n}^t(\boldsymbol{\lambda},\boldsymbol{x})^T\boldsymbol{\beta}_t^n(\boldsymbol{\lambda},\boldsymbol{x}) \;=\; \boldsymbol{\alpha}_{-n}^t(\boldsymbol{\lambda},\boldsymbol{x})^T\boldsymbol{F}\boldsymbol{\beta}_{t+k}^n(\boldsymbol{\lambda},\boldsymbol{x}).$$

According to Theorem 1, there exist numbers r_{ij}, s_{ij} such that

$$\lim_{n\to\infty}\frac{g_n(i,k)}{g_n(j,k)} = r_{ij} \quad\text{and}\quad \lim_{n\to\infty}\frac{h_n(i,k)}{h_n(j,k)} = s_{ij}$$

for all $k \in \mathcal{Y}$. Consequently, the ratio of $\alpha_{-n}^t(\boldsymbol{\lambda},\boldsymbol{x},i)$ to $\alpha_{-n}^t(\boldsymbol{\lambda},\boldsymbol{x},j)$ converges to r_{ij}, and the ratio of $\beta_{t+k}^n(\boldsymbol{\lambda},\boldsymbol{x},i)$ to $\beta_{t+k}^n(\boldsymbol{\lambda},\boldsymbol{x},j)$ converges to s_{ij}. Hence,

$$\lim_{n\to\infty}\frac{\alpha_{-n}^t(\boldsymbol{\lambda},\boldsymbol{x},i)\,\beta_{t+k}^n(\boldsymbol{\lambda},\boldsymbol{x},j)}{\boldsymbol{\alpha}_{-n}^t(\boldsymbol{\lambda},\boldsymbol{x})^T\boldsymbol{F}\boldsymbol{\beta}_{t+k}^n(\boldsymbol{\lambda},\boldsymbol{x})} \;=\; \frac{1}{\boldsymbol{r}_i^T\boldsymbol{F}\boldsymbol{s}_j}$$

where $\boldsymbol{r}_i = (r_{1i}, \ldots, r_{\ell i})^T$ and $\boldsymbol{s}_j = (s_{1j}, \ldots, s_{\ell j})^T$, which proves the first part of the theorem. For the second part, note that one can choose

$$\kappa \;=\; \frac{(1-\varphi)(1-\psi)}{(1+\varphi)(1+\psi)}$$

with φ^2 and ψ^2 as defined in Lemma 1, and

$$c \;=\; 16 \left(\frac{m_{\sup}}{m_{\inf}}\right)^2 \ell^{k+1} m_{\sup}^{2k}.$$

See the supplementary material for details. □

It is easy to verify that

$$P_{\boldsymbol{\lambda}}(Y_t = y_t, \ldots, Y_{t+k} = y_{t+k} \,|\, \boldsymbol{X} = \boldsymbol{x}) \;\geq\; 0$$

and

$$\sum_{y_t, \ldots, y_{t+k} \in \mathcal{Y}} P_{\boldsymbol{\lambda}}(Y_t = y_t, \ldots, Y_{t+k} = y_{t+k} \,|\, \boldsymbol{X} = \boldsymbol{x}) \;=\; 1,$$

so the distribution of $(Y_t, \ldots, Y_{t+k})$ conditional on $\boldsymbol{X} = \boldsymbol{x}$ is well-defined. Furthermore, the collection of all such marginal distributions with $t \in \mathbb{Z}$ and $k \in \mathbb{N}$ satisfies the consistency conditions of Kolmogorov's extension theorem and hence specifies a unique probability measure on the space of sequences $\boldsymbol{y} = (y_t)_{t \in \mathbb{Z}}$ with $y_t \in \mathcal{Y}$. In this way, we obtain the distribution of the infinite sequence $\boldsymbol{Y}$ conditional on $\boldsymbol{X} = \boldsymbol{x}$.

Throughout the rest of this paper, let $\boldsymbol{\Theta}$ be a set of model weights such that each $\boldsymbol{\lambda} \in \boldsymbol{\Theta}$ satisfies assumption (A1). For simplicity we assume that the distribution of $\boldsymbol{X}$ does not depend on $\boldsymbol{\lambda}$. More precisely, let $\boldsymbol{\mathcal{X}}$ denote the space of sequences $\boldsymbol{x} = (x_t)_{t \in \mathbb{Z}}$ in $\mathcal{X}$, and $\boldsymbol{\mathcal{A}}$ be the corresponding product sigma-field. We assume there exists a probability measure $P_{\boldsymbol{X}}$ on $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{A}})$ such that $P_{\boldsymbol{X}}(A) = P_{\boldsymbol{\lambda}}(\boldsymbol{X} \in A)$ for all $A \in \boldsymbol{\mathcal{A}}$ and $\boldsymbol{\lambda} \in \boldsymbol{\Theta}$.

3.2 Basic Properties

The following corollary summarizes consequences of the previous definition. Statement (*ii*) says that the probability for the transition between any two labels is always bounded away from zero, regardless of the observational context. This result will be of particular importance in Sec. 4 where we investigate under which conditions ergodicity of $\boldsymbol{X}$ implies ergodicity of the joint sequence $(\boldsymbol{X}, \boldsymbol{Y})$.

Corollary 1. *Suppose that condition* (A1) *holds and hence the distribution of $\boldsymbol{Y}$ conditional on $\boldsymbol{X} = \boldsymbol{x}$ is well-defined for every $\boldsymbol{x} = (x_t)_{t \in \mathbb{Z}}$.*

(i) In the definition of the distribution of $\boldsymbol{Y}$ conditional on $\boldsymbol{X} = \boldsymbol{x}$ we may assume, without loss of generality, that the weighted state feature functions are equal to zero, $\boldsymbol{\mu}^T \boldsymbol{f}^{\text{state}} = 0$.

(ii) $\boldsymbol{Y}$ conditional on $\boldsymbol{X} = \boldsymbol{x}$ is a Markov process with

$$P_{\boldsymbol{\lambda}}(Y_{t+1} = y_{t+1} \,|\, Y_t = y_t,\, \boldsymbol{X} = \boldsymbol{x}) \;=\; m_{\boldsymbol{\lambda}}(x_t, y_t, y_{t+1}) \lim_{n \to \infty} \frac{\beta_{t+1}^n(\boldsymbol{\lambda}, \boldsymbol{x}, y_{t+1})}{\beta_t^n(\boldsymbol{\lambda}, \boldsymbol{x}, y_t)}.$$

Furthermore, for all $y_t, y_{t+1} \in \mathcal{Y}$,

$$P_{\boldsymbol{\lambda}}(Y_{t+1} = y_{t+1} \,|\, Y_t = y_t,\, \boldsymbol{X} = \boldsymbol{x}) \;\geq\; \frac{1}{\ell} \left(\frac{m_{\inf}}{m_{\sup}}\right)^2.$$

(iii) If $\boldsymbol{X}$ is stationary, then the joint sequence $(\boldsymbol{X}, \boldsymbol{Y})$ is stationary.

Proof. (*i*) Consider the state and transition feature functions $\tilde{\boldsymbol{f}}^{\text{state}}$, $\tilde{\boldsymbol{f}}^{\text{trans}}$ and the weights $\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{\nu}}$ given in such a way that $\tilde{\boldsymbol{\mu}}^T \tilde{\boldsymbol{f}}^{\text{state}} = 0$ and

$$\begin{aligned} \tilde{\boldsymbol{\nu}}^T \tilde{\boldsymbol{f}}^{\text{trans}}(x_t, y_{t-1}, y_t) \;&=\; \boldsymbol{\mu}^T \boldsymbol{f}^{\text{state}}(x_t, y_t) \\ &+\; \boldsymbol{\nu}^T \boldsymbol{f}^{\text{trans}}(x_t, y_{t-1}, y_t). \end{aligned}$$

It is easy to see that the resulting distributions for $\boldsymbol{Y}$ conditional on $\boldsymbol{X} = \boldsymbol{x}$ are identical; in particular, the limit in Theorem 2 does not depend on the values of the vectors $\boldsymbol{\alpha}_{\boldsymbol{\lambda}}(x_{t-n})$. (*ii*) The equality is directly obtained by the definition of the conditional distribution of $\boldsymbol{Y}$, and it is easily verified that it holds for any $P_{\boldsymbol{\lambda}}(Y_{t+1} = y_{t+1} \,|\, Y_t = y_t, \ldots, Y_{t-k} = y_{t-k}, \boldsymbol{X} = \boldsymbol{x})$ with $k \in \mathbb{N}$. To establish the lower bound note that, according to Theorem 1,

$$\frac{\beta_{t+1}^n(\boldsymbol{\lambda}, \boldsymbol{x}, y_{t+1})}{\beta_{t+1}^n(\boldsymbol{\lambda}, \boldsymbol{x}, y_t)} \;\geq\; \frac{m_{\inf}}{m_{\sup}}$$

for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ and $n \in \mathbb{N}$. Since $\boldsymbol{\beta}_t^n(\boldsymbol{\lambda}, \boldsymbol{x}) = \boldsymbol{M}_{\boldsymbol{\lambda}}(x_{t+1}) \boldsymbol{\beta}_{t+1}^n(\boldsymbol{\lambda}, \boldsymbol{x})$, the result follows. (*iii*) The stationarity of the joint sequence is obvious because

$$\begin{aligned} &P_{\boldsymbol{\lambda}}(\boldsymbol{X} \in \boldsymbol{A},\, \boldsymbol{Y} \in \boldsymbol{B}) \\ &\qquad = \int_{\boldsymbol{A}} P_{\boldsymbol{\lambda}}(\boldsymbol{Y} \in \boldsymbol{B} \,|\, \boldsymbol{X} = \boldsymbol{x})\, P_{\boldsymbol{X}}(d\boldsymbol{x}) \end{aligned}$$

for any measurable events $\boldsymbol{A}$ and $\boldsymbol{B}$, and by the stationarity of $\boldsymbol{X}$ the integral on the right hand side is invariant with respect to time shifts of $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the integrand. □

Throughout the rest of this paper, any expected value such as $E_{\boldsymbol{\lambda}}[g(X_t, Y_t)]$ or $E_{\boldsymbol{\lambda}}[g(X_t, Y_{t-1}, Y_t)]$ is with respect to the joint stationary distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$, provided that $\boldsymbol{X}$ is stationary.

4 JOINT ERGODICITY

In this section, we establish conditions under which the joint sequence $(\boldsymbol{X}, \boldsymbol{Y})$ is ergodic. The key step is

to embed this joint sequence in a Markov chain in a random environment. Recall that $\boldsymbol{\mathcal{X}}$ is the space of sequences $\boldsymbol{x} = (x_t)_{t\in\mathbb{Z}}$ in $\mathcal{X}$, and $\boldsymbol{\mathcal{A}}$ is the corresponding product sigma-field. Let τ denote the shift operator on $\boldsymbol{\mathcal{X}}$, that is, $\tau^k\boldsymbol{x} = \boldsymbol{x}'$ with $\boldsymbol{x}' = (x'_t)_{t\in\mathbb{Z}}$ given by $x'_t = x_{t+k}$. As we show below, a sufficient condition for joint ergodicity is that besides (A1) the following condition is satisfied:

(A2) $\boldsymbol{X}$ is ergodic, i.e., the probability measure $P_{\boldsymbol{X}}$ on $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{A}})$ satisfies $P_{\boldsymbol{X}}(A) = P_{\boldsymbol{X}}(\tau^{-1}A)$ for every $A \in \boldsymbol{\mathcal{A}}$, and $P_{\boldsymbol{X}}(A) = 0$ or $P_{\boldsymbol{X}}(A) = 1$ for every set $A \in \boldsymbol{\mathcal{A}}$ such that $A = \tau^{-1}A$.

A particular consequence of condition (A2) is that $\boldsymbol{X}$ is stationary. Furthermore, for every measurable function $g : \boldsymbol{\mathcal{X}} \to \mathbb{R}$ such that $E|g(\boldsymbol{X})| < \infty$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} g(\tau^t \boldsymbol{X}) \quad = \quad E[g(\boldsymbol{X})]$$

P-almost surely (Cornfeld et al., 1982), where we omit the subscript $\boldsymbol{\lambda}$ in $P_{\boldsymbol{\lambda}}$ and $E_{\boldsymbol{\lambda}}$ as the distribution of $\boldsymbol{X}$ does not depend on $\boldsymbol{\lambda}$. Note that condition (A2) is not too restrictive. For example, any stationary ARMA process is ergodic, and a sufficient condition for stationary Gaussian processes to be ergodic is that the autocovariances go to zero as the lag goes to infinity.

4.1 Markov Chains in Random Environments

Let us demonstrate how the joint sequence $(\boldsymbol{X}, \boldsymbol{Y})$ can be embedded in a Markov chain. By $\vec{\boldsymbol{\mathcal{X}}}$ we denote the space of sequences $\vec{\boldsymbol{x}} = (\boldsymbol{x}_t)_{t\in\mathbb{Z}}$ with $\boldsymbol{x}_t \in \boldsymbol{\mathcal{X}}$ for $t \in \mathbb{Z}$, that is, each $\boldsymbol{x}_t$ is a sequence in $\mathcal{X}$. Write $\vec{\boldsymbol{\mathcal{A}}}$ for the corresponding product sigma-field. Let $\vec{\tau}$ denote the shift operator on $\vec{\boldsymbol{\mathcal{X}}}$, that is, $\vec{\tau}^k\vec{\boldsymbol{x}} = \vec{\boldsymbol{x}}'$ with $\vec{\boldsymbol{x}}' = (\boldsymbol{x}'_t)_{t\in\mathbb{Z}}$ given by $\boldsymbol{x}'_t = \boldsymbol{x}_{t+k}$. For a given probability measure π on $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{A}})$, let the probability measure $\vec{\pi}$ on $(\vec{\boldsymbol{\mathcal{X}}}, \vec{\boldsymbol{\mathcal{A}}})$ be defined as follows: For $\vec{A} \in \vec{\boldsymbol{\mathcal{A}}}$ write $\vec{A} = \bigtimes_{t\in\mathbb{Z}} A_t$ with $A_t \in \boldsymbol{\mathcal{A}}$. Then define

$$\vec{\pi}(\vec{A}) \quad := \quad \pi\Big(\bigcap_{t\in\mathbb{Z}} \tau^{-t} A_t\Big).$$

We say that $\vec{\pi}$ is $\vec{\tau}$-*ergodic* if $\vec{\pi}(\vec{\tau}^{-1}\vec{A}) = \vec{\pi}(\vec{A})$ for all $\vec{A} \in \vec{\boldsymbol{\mathcal{A}}}$, and $\vec{\pi}(\vec{A}) = 0$ or $\vec{\pi}(\vec{A}) = 1$ for each $\vec{A} \in \vec{\boldsymbol{\mathcal{A}}}$ satisfying $\vec{\tau}^{-1}\vec{A} = \vec{A}$. The proof of the following technical lemma is included in the supplementary material.

Lemma 2. *If π is τ-ergodic, then $\vec{\pi}$ is $\vec{\tau}$-ergodic.*

Note that according to Lemma 2, if the sequence $\boldsymbol{X}$ is ergodic, then the probability measure $\vec{P}_{\boldsymbol{X}}$ on $(\vec{\boldsymbol{\mathcal{X}}}, \vec{\boldsymbol{\mathcal{A}}})$ is $\vec{\tau}$-ergodic.

Now consider the measurable space $(\mathcal{Z}, \mathcal{C})$ with $\mathcal{Z} = \vec{\boldsymbol{\mathcal{X}}} \times \mathcal{Y} \times \mathcal{Y}$ and $\mathcal{C} = \vec{\boldsymbol{\mathcal{A}}} \times \mathcal{B} \times \mathcal{B}$. Let $\boldsymbol{\lambda} \in \boldsymbol{\Theta}$ be fixed. We are going to define a Markov sequence $\boldsymbol{Z} = (Z_t)_{t\in\mathbb{N}}$ with values in $\mathcal{Z}$ such that $\boldsymbol{Z}$ has the same distribution as the sequence $((\tau^{k+t}\boldsymbol{X})_{k\in\mathbb{Z}}, Y_{t-1}, Y_t)_{t\in\mathbb{N}}$ measured with respect to $P_{\boldsymbol{\lambda}}$. Using results on invariant measures of Markov processes we are then going to show that the sequence $\boldsymbol{Z}$ is ergodic and, consequently,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} g((\tau^{t+k}\boldsymbol{X})_{k\in\mathbb{Z}}, Y_{t-1}, Y_t) \\ = \; E_{\boldsymbol{\lambda}}\big[g((\tau^{t+k}\boldsymbol{X})_{k\in\mathbb{Z}}, Y_{t-1}, Y_t)\big] \quad (4)$$

$P_{\boldsymbol{\lambda}}$-almost surely for every measurable $g : \mathcal{Z} \to \mathbb{R}$ for which $E_{\boldsymbol{\lambda}}|g((\tau^{t+k}\boldsymbol{X})_{k\in\mathbb{Z}}, Y_{t-1}, Y_t)| < \infty$. As a special case, we obtain the following result which will be of great importance for analyzing the asymptotical properties of Maximum Likelihood estimates.

Theorem 3. *If conditions* (A1) *and* (A2) *hold, and $g : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfies $E_{\boldsymbol{\lambda}}|g(X_t, Y_{t-1}, Y_t)| < \infty$, then*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} g(X_t, Y_{t-1}, Y_t) \quad = \quad E_{\boldsymbol{\lambda}}[g(X_t, Y_{t-1}, Y_t)]$$

$P_{\boldsymbol{\lambda}}$-almost surely.

In order to establish this theorem, let us first consider how to define $\boldsymbol{Z}$. Let $\mu_{\boldsymbol{\lambda}}$ be the distribution on $(\mathcal{Z}, \mathcal{C})$ induced by $((\tau^{k+1}\boldsymbol{X})_{k\in\mathbb{Z}}, Y_0, Y_1)$ measured with respect to $P_{\boldsymbol{\lambda}}$, that is,

$$\mu_{\boldsymbol{\lambda}}(\vec{A} \times \{y_0\} \times \{y_1\}) \\ := \; P_{\boldsymbol{\lambda}}((\tau^{k+1}\boldsymbol{X})_{k\in\mathbb{Z}} \in \vec{A},\, Y_0 = y_0,\, Y_1 = y_1).$$

This will serve us as the initial distribution of $\boldsymbol{Z}$. In order to specify a Markov kernel on $(\mathcal{Z}, \mathcal{C})$, let $\boldsymbol{Q}(\boldsymbol{\lambda}, \boldsymbol{x})$ with $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ denote the $\ell \times \ell$-matrix with the (i,j)-th component

$$Q(\boldsymbol{\lambda}, \boldsymbol{x}, i, j) \quad = \quad m_{\boldsymbol{\lambda}}(x_0, i, j)\, \lim_{n\to\infty} \frac{\beta_1^n(\boldsymbol{\lambda}, \boldsymbol{x}, j)}{\beta_0^n(\boldsymbol{\lambda}, \boldsymbol{x}, i)}.$$

Note that, for this matrix to be well-defined, it is sufficient to assume that condition (A1) holds. Now we define the Markov kernel $Q_{\boldsymbol{\lambda}}$ from $(\mathcal{Z}, \mathcal{C})$ into itself,

$$Q_{\boldsymbol{\lambda}}((\vec{\boldsymbol{x}}, y'_0, y'_1), \vec{A} \times \{y_0\} \times \{y_1\}) \\ := \; \begin{cases} Q(\boldsymbol{\lambda}, \boldsymbol{x}_0, y'_1, y_1) & \text{if } y_0 = y'_1 \text{ and } \vec{\tau}\vec{\boldsymbol{x}} \in \vec{A}, \\ 0 & \text{otherwise.} \end{cases}$$

It is not difficult to see that $\boldsymbol{Z}$ with the initial distribution $\mu_{\boldsymbol{\lambda}}$ and the Markov kernel $Q_{\boldsymbol{\lambda}}$ has the same distribution as $((\tau^{k+t}\boldsymbol{X})_{k\in\mathbb{Z}}, Y_{t-1}, Y_t)_{t\in\mathbb{N}}$ measured with respect to $P_{\boldsymbol{\lambda}}$. Note that $\boldsymbol{Z}$ can be regarded as a Markov chain in a random environment (Cogburn, 1984; Orey, 1991). In particular, the pairs (Y_{t-1}, Y_t) with $t \in \mathbb{N}$ form a Markov chain where the transition probabilities from (Y_{t-1}, Y_t) to (Y_t, Y_{t+1}) are governed by the stationary environment $\tau^t\boldsymbol{X}$.

4.2 Invariant Measures

For the proof that $\boldsymbol{Z}$ is ergodic and hence (4) applies, we use results on invariant measures of Markov chains. In fact, by the stationarity of $\boldsymbol{X}$ and Corollary 1 (*iii*), it is easily verified that the measure $\mu_{\boldsymbol{\lambda}}$ is *invariant for* $Q_{\boldsymbol{\lambda}}$, that is,

$$\mu_{\boldsymbol{\lambda}}(C) \quad = \quad \int_{\mathcal{Z}} Q_{\boldsymbol{\lambda}}(z, C)\, \mu_{\boldsymbol{\lambda}}(dz)$$

for every $C \in \mathcal{C}$. It only remains to show that the invariant measure $\mu_{\boldsymbol{\lambda}}$ for $Q_{\boldsymbol{\lambda}}$ is unique (Corollary 2.5.2, Hernández-Lerma and Lasserre, 2003). This fact is stated in the next lemma, the proof of which is included in the supplementary material.

Lemma 3. *Suppose that conditions* (A1) *and* (A2) *hold. Then the invariant measure $\mu_{\boldsymbol{\lambda}}$ for $Q_{\boldsymbol{\lambda}}$ is unique.*

5 CONSISTENCY OF MAXIMUM LIKELIHOOD ESTIMATES

In this section, we apply the previous results to study the following problem: Suppose that the distribution of the sequences $\boldsymbol{X}$ and $\boldsymbol{Y}$ is governed by $P_{\boldsymbol{\lambda}_0}$ with $\boldsymbol{\lambda}_0 \in \boldsymbol{\Theta}$ unknown, and we observe realizations of the subsequences $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ and $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)$. Under which conditions can we identify $\boldsymbol{\lambda}_0$ as the sample length n tends to infinity?

According to Corollary 1 (*i*), we may assume in our analysis that the weighted state feature functions are equal to zero. Therefore, to simplify notation, we write $\boldsymbol{f}$ instead of $\boldsymbol{f}^{\text{trans}}$ in the following, so the weighted feature functions in (1) are given by

$$\boldsymbol{\lambda}^T \boldsymbol{f}(\boldsymbol{x}_n, \boldsymbol{y}_n) \quad = \quad \sum_{t=1}^{n} \boldsymbol{\lambda}^T \boldsymbol{f}(x_t, y_{t-1}, y_t).$$

In order to estimate $\boldsymbol{\lambda}_0$, consider the objective function

$$\mathcal{L}_n(\boldsymbol{\lambda}) \quad = \quad \frac{1}{n}\Big(\boldsymbol{\lambda}^T \boldsymbol{f}(\boldsymbol{X}_n, \boldsymbol{Y}_n) - \log Z_{\boldsymbol{\lambda}}(\boldsymbol{X}_n)\Big) \quad (5)$$

with the partition function $Z_{\boldsymbol{\lambda}}(\boldsymbol{X}_n)$ as in (3). Note that $\mathcal{L}_n(\boldsymbol{\lambda})$ is the average conditional log-likelihood of $\boldsymbol{Y}_n$ given $\boldsymbol{X}_n$ based on the finite L-CRF model $P_{\boldsymbol{\lambda}}^{(0:n)}$. Now consider the estimate $\hat{\boldsymbol{\lambda}}_n$ of $\boldsymbol{\lambda}_0$ obtained by maximizing the conditional log-likelihood,

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

If $\mathcal{L}_n(\boldsymbol{\lambda})$ is strictly concave, then the arg max is unique and can be found using gradient-based search (Sha and Pereira, 2003). Obviously, a necessary and sufficient condition for $\mathcal{L}_n(\boldsymbol{\lambda})$ to be strictly concave is that the number of labels ℓ is greater than or equal to 2, and there exists a $\boldsymbol{y}_n \in \mathcal{Y}^{n+1}$ such that at least one component of $\boldsymbol{f}(\boldsymbol{X}_n, \boldsymbol{y}_n)$ is non-zero.

In the following we investigate conditions under which the estimates $\hat{\boldsymbol{\lambda}}_n$ are strongly consistent, that is,

$$\lim_{n\to\infty} \hat{\boldsymbol{\lambda}}_n \quad = \quad \boldsymbol{\lambda}_0$$

$P_{\boldsymbol{\lambda}_0}$-almost surely (Lehmann, 1999). Sufficient conditions will be given in Theorem 4 below. The key step is to establish conditions under which $\mathcal{L}_n(\boldsymbol{\lambda})$ converges uniformly to a function $\mathcal{L}(\boldsymbol{\lambda})$, and $\mathcal{L}(\boldsymbol{\lambda})$ has a unique maximum in $\boldsymbol{\lambda}_0$. To establish uniform convergence, we need to make the following assumption on the parameter space:

(A3) The parameter space $\boldsymbol{\Theta}$ is compact.

In our case where the model parameters are $\mathbb{R}$-valued vectors, a sufficient condition for (A3) is that $\boldsymbol{\Theta}$ is the Cartesian product of finite closed intervals.

5.1 Convergence of the Likelihood Function

First, we show that $\mathcal{L}_n(\boldsymbol{\lambda})$ converges for every $\boldsymbol{\lambda} \in \boldsymbol{\Theta}$.

Lemma 4. *Suppose that assumptions* (A1) *and* (A2) *hold. Then there exists a function $\mathcal{L}(\boldsymbol{\lambda})$ such that, for every $\boldsymbol{\lambda} \in \boldsymbol{\Theta}$,*

$$\lim_{n\to\infty} \mathcal{L}_n(\boldsymbol{\lambda}) \quad = \quad \mathcal{L}(\boldsymbol{\lambda})$$

$P_{\boldsymbol{\lambda}_0}$-almost surely.

Proof. Let $\boldsymbol{\lambda} \in \boldsymbol{\Theta}$. We show convergence separately for both terms on the right hand side of (5). For the first term we obtain, according to Theorem 3,

$$\lim_{n\to\infty} \frac{1}{n} \boldsymbol{\lambda}^T \boldsymbol{f}(\boldsymbol{X}_n, \boldsymbol{Y}_n) \quad = \quad \boldsymbol{\lambda}^T E_{\boldsymbol{\lambda}_0}[\boldsymbol{f}(X_t, Y_{t-1}, Y_t)]$$

$P_{\boldsymbol{\lambda}_0}$-almost surely. For the second term, note that

$$\frac{1}{n} \log Z_{\boldsymbol{\lambda}}(\boldsymbol{X}_n) \quad \sim \quad \frac{1}{n} \log \big\| \boldsymbol{M}_{\boldsymbol{\lambda}}(X_1) \ldots \boldsymbol{M}_{\boldsymbol{\lambda}}(X_n) \big\|$$

as $n \to \infty$, where $\sim$ denotes asymptotical equivalence and $\|(m_{ij})\| = \sum_{i,j} |m_{ij}|$. Now, let $L_1(x_1) := \log \|\boldsymbol{M}_{\boldsymbol{\lambda}}(x_1)\|$ and, for $t > 1$,

$$L_t(x_t, \ldots, x_1) \quad := \quad \log \frac{\big\| \boldsymbol{M}_{\boldsymbol{\lambda}}(x_1) \ldots \boldsymbol{M}_{\boldsymbol{\lambda}}(x_t) \big\|}{\big\| \boldsymbol{M}_{\boldsymbol{\lambda}}(x_1) \ldots \boldsymbol{M}_{\boldsymbol{\lambda}}(x_{t-1}) \big\|}.$$

By the same arguments as in the proof of Theorem 2, the rows of the matrices $\boldsymbol{M}_{\boldsymbol{\lambda}}(x_{t-k}) \ldots \boldsymbol{M}_{\boldsymbol{\lambda}}(x_t)$ and $\boldsymbol{M}_{\boldsymbol{\lambda}}(x_{t-k}) \ldots \boldsymbol{M}_{\boldsymbol{\lambda}}(x_{t-1})$ tend to proportionality as k goes to ∞. Therefore,

$$L(x_t, x_{t-1}, \ldots) \quad := \quad \lim_{k\to\infty} L_k(x_t, \ldots, x_{t-k+1})$$

is well-defined. Putting all together, we obtain

$$\frac{1}{n}\log Z_{\boldsymbol{\lambda}}(\boldsymbol{X}_n) \quad \sim \quad \frac{1}{n}\sum_{t=1}^{n} L(X_t, X_{t-1}, \ldots),$$

and by the ergodicity of $\boldsymbol{X}$ the latter expression converges $P_{\boldsymbol{\lambda}_0}$-almost surely to $E_{\boldsymbol{\lambda}_0}[L(X_t, X_{t-1}, \ldots)]$. Hence the proof is complete. □

For the proof that the convergence of $\mathcal{L}_n(\boldsymbol{\lambda})$ to $\mathcal{L}(\boldsymbol{\lambda})$ is uniform on $\boldsymbol{\Theta}$, we need to consider the gradient of $\mathcal{L}_n(\boldsymbol{\lambda})$, which is given by

$$\begin{aligned}\nabla\mathcal{L}_n(\boldsymbol{\lambda}) &= \frac{1}{n}\boldsymbol{f}(\boldsymbol{X}_n, \boldsymbol{Y}_n) \\ &- \frac{1}{n}\sum_{\boldsymbol{y}_n \in \mathcal{Y}^n} \frac{\exp\big(\boldsymbol{\lambda}^T \boldsymbol{f}(\boldsymbol{X}_n, \boldsymbol{y}_n)\big)}{Z_{\boldsymbol{\lambda}}(\boldsymbol{X}_n)}\boldsymbol{f}(\boldsymbol{X}_n, \boldsymbol{y}_n).\end{aligned}$$

Lemma 5. *Suppose that assumptions* (A1) *and* (A2) *hold. Then, for every* $\boldsymbol{\lambda} \in \boldsymbol{\Theta}$*,*

$$\begin{aligned}\lim_{n\to\infty} \nabla\mathcal{L}_n(\boldsymbol{\lambda}) &= E_{\boldsymbol{\lambda}_0}[\boldsymbol{f}(X_t, Y_{t-1}, Y_t)] \\ &\quad - E_{\boldsymbol{\lambda}}[\boldsymbol{f}(X_t, Y_{t-1}, Y_t)]\end{aligned}$$

$P_{\boldsymbol{\lambda}_0}$*-almost surely.*

Proof. The convergence of the first term again follows by Theorem 3. For the second term, note that

$$\begin{aligned}&\frac{1}{n}\sum_{\boldsymbol{y}_n \in \mathcal{Y}^n} \frac{\exp\big(\boldsymbol{\lambda}^T \boldsymbol{f}(\boldsymbol{X}_n, \boldsymbol{y}_n)\big)}{Z_{\boldsymbol{\lambda}}(\boldsymbol{X}_n)}\boldsymbol{f}(\boldsymbol{X}_n, \boldsymbol{y}_n) \\ &\qquad = \frac{1}{n}\sum_{t=1}^{n} E_{\boldsymbol{\lambda}}^{(0:n)}\big[\boldsymbol{f}(X_t, Y_{t-1}, Y_t) \,|\, \boldsymbol{X}\big],\end{aligned}$$

where $E_{\boldsymbol{\lambda}}^{(0:n)}\big[\boldsymbol{f}(X_t, Y_{t-1}, Y_t) \,|\, \boldsymbol{X}\big]$ is the conditional expectation of $\boldsymbol{f}(X_t, Y_{t-1}, Y_t)$ given $\boldsymbol{X}$ under the finite L-CRF model $P_{\boldsymbol{\lambda}}^{(0:n)}$. Using arguments similar to the proof of the uniform bound in Theorem 2, one can show that $E_{\boldsymbol{\lambda}}^{(0:n)}\big[\boldsymbol{f}(X_t, Y_{t-1}, Y_t) \,|\, \boldsymbol{X}\big]$ can be replaced by the conditional expectation of $\boldsymbol{f}(X_t, Y_{t-1}, Y_t)$ given $\boldsymbol{X}$ under the infinite L-CRF model $P_{\boldsymbol{\lambda}}$, that is,

$$\begin{aligned}&\frac{1}{n}\sum_{t=1}^{n} E_{\boldsymbol{\lambda}}^{(0:n)}\big[\boldsymbol{f}(X_t, Y_{t-1}, Y_t) \,|\, \boldsymbol{X}\big] \\ &\qquad \sim \frac{1}{n}\sum_{t=1}^{n} E_{\boldsymbol{\lambda}}\big[\boldsymbol{f}(X_t, Y_{t-1}, Y_t) \,|\, \boldsymbol{X}\big].\end{aligned}$$

See the supplementary material for details. Now, as $E_{\boldsymbol{\lambda}}\big[\boldsymbol{f}(X_t, Y_{t-1}, Y_t) \,|\, \boldsymbol{X}\big] = E_{\boldsymbol{\lambda}}\big[\boldsymbol{f}(X_0, Y_{-1}, Y_0) \,|\, \tau^t \boldsymbol{X}\big]$ for every t, we obtain

$$\frac{1}{n}\sum_{t=1}^{n} E_{\boldsymbol{\lambda}}\big[\boldsymbol{f}(X_t, Y_{t-1}, Y_t) \,|\, \boldsymbol{X}\big] = E_{\boldsymbol{\lambda}}\big[\boldsymbol{f}(X_t, Y_{t-1}, Y_t)\big]$$

$P_{\boldsymbol{\lambda}}$-almost surely by the ergodicity of $\boldsymbol{X}$. □

Now we are ready to prove that $\mathcal{L}_n(\boldsymbol{\lambda})$ converges to $\mathcal{L}(\boldsymbol{\lambda})$ uniformly on $\boldsymbol{\Theta}$.

Lemma 6. *Suppose that conditions* (A1)-(A3) *hold. Then the convergence of* $\mathcal{L}_n(\boldsymbol{\lambda})$ *to* $\mathcal{L}(\boldsymbol{\lambda})$ *is uniform on* $\boldsymbol{\Theta}$*, that is,*

$$\lim_{n\to\infty} \sup_{\boldsymbol{\lambda}\in\boldsymbol{\Theta}} \big|\mathcal{L}_n(\boldsymbol{\lambda}) - \mathcal{L}(\boldsymbol{\lambda})\big| \quad = \quad 0$$

$P_{\boldsymbol{\lambda}_0}$*-almost surely.*

Proof. Since $\boldsymbol{\Theta}$ is compact, it is sufficient to show that $\mathcal{L}_n(\boldsymbol{\lambda})$ is stochastically equicontinuous, i.e., for $P_{\boldsymbol{\lambda}}$-almost every $\omega \in \Omega$ and every $\epsilon > 0$, there exists a $\delta > 0$ and an $n_0(\omega)$ such that

$$\sup_{\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\| \le \delta} \big|\mathcal{L}_n(\boldsymbol{\lambda}_1)(\omega) - \mathcal{L}_n(\boldsymbol{\lambda}_2)(\omega)\big| \quad \le \quad \epsilon$$

for all $n \ge n_0(\omega)$. By the Mean value theorem, there exists a (random) $h \in [0, 1]$ such that

$$\begin{aligned}\big|\mathcal{L}_n(\boldsymbol{\lambda}_1) - \mathcal{L}_n(\boldsymbol{\lambda}_2)\big| &\le \big\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\big\| \\ &\quad \times \big\|\nabla\mathcal{L}_n((1-h)\boldsymbol{\lambda}_1 + h\boldsymbol{\lambda}_2)\big\|.\end{aligned}$$

To bound the second factor on the right hand side note that for any $\boldsymbol{\lambda}$ (not necessarily lying in $\boldsymbol{\Theta}$),

$$\begin{aligned}\big\|\nabla\mathcal{L}_n(\boldsymbol{\lambda})\big\| &\le \frac{1}{n}\sum_{t=1}^{n}\big\|\boldsymbol{f}(X_t, Y_{t-1}, Y_t)\big\| \\ &\quad + \frac{1}{n}\sum_{t=1}^{n}\sum_{i,j\in\mathcal{Y}}\big\|\boldsymbol{f}(X_t, i, j)\big\|.\end{aligned}$$

Let U_n denote this upper bound. By the ergodicity of $\boldsymbol{X}$ we obtain that U_n converges $P_{\boldsymbol{\lambda}_0}$-almost surely to a finite limit, which we denote by U. Now, for $P_{\boldsymbol{\lambda}_0}$-almost every $\omega \in \Omega$ there exists an $n_0(\omega)$ such that $\|U_n(\omega) - U\| \le \epsilon$ for all $n \ge n_0(\omega)$. Substituting this into the above inequality, we obtain

$$\big|\mathcal{L}_n(\boldsymbol{\lambda}_1)(\omega) - \mathcal{L}_n(\boldsymbol{\lambda}_2)(\omega)\big| \quad \le \quad \big(U + \epsilon\big)\big\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\big\|$$

for all $n \ge n_0(\omega)$. Setting $\delta := \epsilon/(U + \epsilon)$, we see that the sequence $\mathcal{L}_n(\boldsymbol{\lambda})$ is stochastically equicontinuous. □

5.2 Convergence of the Hessian

Based on the previous results, we are now able to state the following sufficient conditions for strong consistency of $\hat{\boldsymbol{\lambda}}_n$.

Theorem 4. *Suppose that conditions* (A1)-(A3) *hold, and the limit of* $\nabla^2\mathcal{L}_n(\boldsymbol{\lambda})$ *is finite and strictly negative definite. Then* $\mathcal{L}(\boldsymbol{\lambda})$ *is strictly concave on* $\boldsymbol{\Theta}$*, and*

$$\lim_{n\to\infty} \hat{\boldsymbol{\lambda}}_n \quad = \quad \boldsymbol{\lambda}_0$$

$P_{\boldsymbol{\lambda}_0}$*-almost surely.*

Proof. According to Lemma 6, we have uniform convergence of $\mathcal{L}_n(\boldsymbol{\lambda})$ to $\mathcal{L}(\boldsymbol{\lambda})$ on $\boldsymbol{\Theta}$. Thus, if the limit of $\nabla^2\mathcal{L}_n(\boldsymbol{\lambda})$ is strictly negative definite, $\mathcal{L}(\boldsymbol{\lambda})$ is strictly concave and hence has a unique maximum. It only remains to show that this maximum is $\mathcal{L}(\boldsymbol{\lambda}_0)$. Under the assumption that the limit of $\nabla^2\mathcal{L}_n(\boldsymbol{\lambda})$ is finite, the gradient of $\mathcal{L}(\boldsymbol{\lambda})$ is given by the limit of $\nabla\mathcal{L}_n(\boldsymbol{\lambda})$. According to Lemma 5, this limit is zero if $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$, hence $\mathcal{L}(\boldsymbol{\lambda}_0)$ is the unique maximum of $\mathcal{L}(\boldsymbol{\lambda})$ on $\boldsymbol{\Theta}$. □

Let us analyze the asymptotic behavior of $\nabla^2\mathcal{L}_n(\boldsymbol{\lambda})$. In order to compute the second partial derivatives, let $n \in \mathbb{N}$ and write $\boldsymbol{\lambda}$ as a stacked vector $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n)$ with $\boldsymbol{\lambda}_1 = \ldots = \boldsymbol{\lambda}_n$. Correspondingly, consider the stacked feature functions $\boldsymbol{f}(\boldsymbol{X}_n, \boldsymbol{Y}_n) = (\boldsymbol{f}(X_1, Y_0, Y_1), \ldots, \boldsymbol{f}(X_n, Y_{n-1}, Y_n))$, so that

$$\boldsymbol{\lambda}^T \boldsymbol{f}(\boldsymbol{X}_n, \boldsymbol{Y}_n) \;=\; \sum_{t=1}^{n} \boldsymbol{\lambda}_t^T \boldsymbol{f}(X_t, Y_{t-1}, Y_t).$$

Note that the first partial derivatives of $\mathcal{L}_n(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}_t$ are given by

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\lambda}_t}\mathcal{L}_n(\boldsymbol{\lambda}) \;&=\; \frac{1}{n}\,\boldsymbol{f}(X_t, Y_{t-1}, Y_t) \\ &\quad - \frac{1}{n}\,E_{\boldsymbol{\lambda}}^{(0:n)}\big[\boldsymbol{f}(X_t, Y_{t-1}, Y_t)\,|\,\boldsymbol{X}\big].\end{aligned}$$

By further differentiating this expression with respect to $\boldsymbol{\lambda}_{t+k}$ (for k such that $1 \le t+k \le n$), we obtain

$$\begin{aligned}\frac{\partial^2}{\partial\boldsymbol{\lambda}_t\partial\boldsymbol{\lambda}_{t+k}}\mathcal{L}_n(\boldsymbol{\lambda}) \;&=\; -\frac{1}{n}\,E_{\boldsymbol{\lambda}}^{(0:n)}\big[\boldsymbol{f}(X_t, Y_{t-1}, Y_t)\,|\,\boldsymbol{X}\big] \\ &\quad \times E_{\boldsymbol{\lambda}}^{(0:n)}\big[\boldsymbol{f}(X_{t+k}, Y_{t+k-1}, Y_{t+k})\,|\,\boldsymbol{X}\big]^T \\ + \frac{1}{n}\,E_{\boldsymbol{\lambda}}^{(0:n)}&\big[\boldsymbol{f}(X_t, Y_{t-1}, Y_t)\,\boldsymbol{f}(X_{t+k}, Y_{t+k-1}, Y_{t+k})^T\,|\,\boldsymbol{X}\big].\end{aligned}$$

For $k = 0, 1, \ldots, n-1$, consider the sum of all second partial derivatives with the lag k,

$$\hat{\gamma}_{\boldsymbol{\lambda}}^{(n)}(k) \;=\; \sum_{t=1}^{n-k} \frac{\partial^2}{\partial\boldsymbol{\lambda}_t\partial\boldsymbol{\lambda}_{t+k}}\mathcal{L}_n(\boldsymbol{\lambda}),$$

and note that the Hessian of $\mathcal{L}_n(\boldsymbol{\lambda})$ can be written as

$$\nabla^2\mathcal{L}_n(\boldsymbol{\lambda}) \;=\; -\left(\hat{\gamma}_{\boldsymbol{\lambda}}^{(n)}(0) \,+\, 2\sum_{k=1}^{n-1}\hat{\gamma}_{\boldsymbol{\lambda}}^{(n)}(k)\right).$$

The following lemma shows that, if conditions (A1) and (A2) are satisfied, the limit of $\nabla^2\mathcal{L}_n(\boldsymbol{\lambda})$ is finite. The proof is included in the supplementary material. According to the proof of Theorem 4, we obtain that $\boldsymbol{\lambda}_0$ is *one* solution of $\arg\max_{\boldsymbol{\lambda}\in\boldsymbol{\Theta}}\mathcal{L}(\boldsymbol{\lambda})$, however, this solution is not unique unless the limit of $\nabla^2\mathcal{L}_n(\boldsymbol{\lambda})$ is non-singular.

Lemma 7. *Suppose that conditions* (A1) *and* (A2) *hold. Then*

$$\lim_{n\to\infty}\nabla^2\mathcal{L}_n(\boldsymbol{\lambda}) \;=\; -\left(\gamma_{\boldsymbol{\lambda}}(0) \,+\, 2\sum_{k=1}^{\infty}\gamma_{\boldsymbol{\lambda}}(k)\right)$$

$P_{\boldsymbol{\lambda}_0}$-almost surely, where

$$\begin{aligned}\gamma_{\boldsymbol{\lambda}}(k) \;&=\; \\ &\mathrm{Cov}_{\boldsymbol{\lambda}}\big[\boldsymbol{f}(X_t, Y_{t-1}, Y_t),\, \boldsymbol{f}(X_{t+k}, Y_{t+k-1}, Y_{t+k})\big]\end{aligned}$$

is the matrix of covariances between $\boldsymbol{f}(X_t, Y_{t-1}, Y_t)$ and $\boldsymbol{f}(X_{t+k}, Y_{t+k-1}, Y_{t+k})$ measured with respect to $P_{\boldsymbol{\lambda}}$. In particular, the limit of $\nabla^2\mathcal{L}_n(\boldsymbol{\lambda})$ is finite.

The following corollary states a simple necessary condition for the limit of the Hessian $\nabla^2\mathcal{L}_n(\boldsymbol{\lambda})$ to be non-singular and hence for the solution of $\arg\max_{\boldsymbol{\lambda}\in\boldsymbol{\Theta}}\mathcal{L}(\boldsymbol{\lambda})$ to be unique.

Corollary 2. *Suppose that conditions* (A1) *and* (A2) *hold and the vector of feature functions $\boldsymbol{f}$ has dimensionality d. Then a necessary condition for the limit of $\nabla^2\mathcal{L}_n(\boldsymbol{\lambda})$ to be non-singular is that for every pair $a \in \mathbb{R}$, $\boldsymbol{b} \in \mathbb{R}^d$ with $\boldsymbol{b} \neq \boldsymbol{0}$ there exist $i, j \in \mathcal{Y}$ and a subset $A \subset \mathcal{X}$ with $P_{\boldsymbol{X}}(A) > 0$ such that $\boldsymbol{b}^T\boldsymbol{f}(x, i, j) \neq a$ for all $x \in A$.*

In particular, the solution of $\arg\max_{\boldsymbol{\lambda}\in\boldsymbol{\Theta}}\mathcal{L}(\boldsymbol{\lambda})$ fails to be unique if any of the components of $\boldsymbol{f}$ can be expressed as linear combinations of each other. We leave it as an open problem whether the conditions in Corollary 2 are also sufficient for non-singularity. Note that the answer is affirmative when the feature functions $\boldsymbol{f}(x_t, y_{t-1}, y_t)$ do not depend on y_{t-1}. In this case the matrices $\gamma_{\boldsymbol{\lambda}}(k)$ are zero for all $k > 0$, and hence the limit of $\nabla^2\mathcal{L}_n(\boldsymbol{\lambda})$ is equal to $\gamma_{\boldsymbol{\lambda}}(0)$.

6 CONCLUSIONS

We have taken a first step to a rigorous study of asymptotic properties of Maximum Likelihood Estimates (MLEs) in Linear-Chain Conditional Random Fields (L-CRFs). Our analysis is based on L-CRFs for infinite sequences, which are defined by the limit distributions of conventional L-CRFs as the length of the observational context tends to infinity. We have derived basic properties of these L-CRFs and shown that ergodicity of the observation sequence implies ergodicity of the joint sequence of observations and labels. Based on these results, we have established uniform convergence of the average conditional log-likelihood and of the gradient, and pointwise convergence of the Hessian. Under the assumption that the limit of the Hessian is non-singular, our results show that the MLEs are strongly consistent. The question under which conditions non-singularity holds is an open problem for future research.

References

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

I.P. Cornfeld, S.V. Fomin and Y.G. Sinai (1982). *Ergodic theory.* Berlin, Germany: Springer.

A. Culotta, D. Kulp and A. McCallum (2005). Gene prediction with conditional random fields. *Technical Report UM-CS-2005-028.* University of Massachusetts, Amherst.

S.R. Foguel (1969). *The ergodic theory of Markov processes.* Princeton, NJ: Van Nostrand.

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

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).*

E.L. Lehmann (1999). *Elements of large-sample theory.* New York, NY: Springer.

L. Liao, D. Choudhury, D. Fox, H. Kautz and B. Limketkai (2007a). Training conditional random fields using virtual evidence boosting. *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI).*

L. Liao, D. Fox and H. Kautz (2007b). Extracting places and activities from GPS traces using hierarchical conditional random fields. *International Journal of Robotics Research* **26**(1):119-138

A. McCallum (2003). Efficiently inducing features of conditional random fields. *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI).*

F. Omar, M. Sinn, J. Truszkowski, P. Poupart, J. Tung and A. Caine (2010). Comparative analysis of probabilistic models for activity recognition with an instrumented walker. *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI).*

S. Orey (1991). Markov chains with stochastically stationary transition probabilities. *The Annals of Probability* **19**(3):907-928

S. Sarawagi and W.W. Cohen (2004). Semi-markov conditional random fields for information extraction. *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS).*

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

F. Sha and F. Pereira (2003). Shallow parsing with conditional random fields. *Proceedings of HLT-NAACL.*

C. Sutton and A. McCallum (2006). An introduction to Conditional random fields for relational learning. In L. Getoor and B. Taskar (eds.): *Introduction to statistical relational learning.* Cambridge, MA: MIT Press.

B. Taskar, P. Abbeel and D. Koller (2002). Discriminative probabilistic models for relational data. *Proceedings of the Conference on Uncertainty in Artifical Intelligence (UAI).*

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