Central Limit Theorems for Conditional Markov Chains

Mathieu Sinn
IBM Research - Ireland

Bei Chen
IBM Research - Ireland

Abstract

This paper studies Central Limit Theorems for real-valued functionals of Conditional Markov Chains. Using a classical result by Dobrushin (1956) for non-stationary Markov chains, a conditional Central Limit Theorem for fixed sequences of observations is established. The asymptotic variance can be estimated by resampling the latent states conditional on the observations. If the conditional means themselves are asymptotically normally distributed, an unconditional Central Limit Theorem can be obtained. The methodology is used to construct a statistical hypothesis test which is applied to synthetically generated environmental data.

1 INTRODUCTION

Conditional Random Fields, introduced by Lafferty et al. (2001), are a widely popular class of undirected graphical models for the distribution of a collection of latent states conditional on observable variables. In the special case of a linear-chain graph structure, the latent states form a Conditional Markov Chain.

Asymptotic statistical properties of Conditional Random Fields and, more specifically, Conditional Markov Chains, have been first investigated by Xiang and Neville (2011), and Sinn and Poupart (2011a). The main focus of this research was on asymptotic properties of Maximum Likelihood (ML) estimates, such as convergence in probability (Xiang and Neville, 2011), or almost sure convergence (Sinn and Poupart, 2011a). While originally the analysis was restricted to models with bounded feature functions, Sinn and Chen (2012) have recently generalized the results to the unbounded case.

This paper studies Central Limit Theorems for real-valued functionals of Conditional Markov Chains. Previous work in this direction is a Central Limit Theorem by Xiang and Neville (2011), however, as this paper points out, the proof of their result is flawed. Central Limit Theorems are of great practical importance because they allow for the construction of confidence intervals and hypothesis tests at predefined coverage and signficance levels. For example, in the experimental part of this paper, it is demonstrated how to construct tests for dependencies between latent states and observable variables.

On the theoretical side, studying Central Limit Theorems gives great insight into the dependency structure of time series and helps to understand long-memory effects. In accordance with the results in (Sinn and Chen, 2012), the main finding in this regard is the importance of the tail distribution of the feature functions, and of concentration inequalities for bounded functionals of the observable variables.

The outline of this paper is as follows: Section 2 reviews the definition and fundamental properties of Conditional Markov Chains, which will serve as the mathematical framework of the analysis. The main mathematical results are presented in Section 3. Using a result by Dobrushin (1956) for non-stationary Markov chains, a conditional Central Limit Theorem (where the sequence of observations is considered fixed) is dervied. If the conditional means themselves are asymptotically normally distributed, an unconditional version of the Central Limit Theorem is obtained. Section 4 presents an algorithm for estimating the asymptotic variance by resampling the latent states conditional on a given sequence of observations. Empirical experiments with synthetically generated environmental data are shown in Section 5. In particular, it is demonstrated how the previous methodology can be used for the construction of hypothesis tests. Section 6 concludes the paper with an outlook on open problems for future research.

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

2 CONDITIONAL MARKOV CHAINS

Preliminaries. Throughout this paper $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{R}$ denote the sets of natural numbers, integers and real numbers, respectively. Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with the following two stochastic processes defined on it:

- $\boldsymbol{X} = (X_t)_{t\in\mathbb{Z}}$ are the *observable variables*, ranging in the metric space $\mathcal{X}$ equipped with the Borel sigma-field $\mathcal{A}$.
- $\boldsymbol{Y} = (Y_t)_{t\in\mathbb{Z}}$ are the *latent states*, ranging in the finite set $\mathcal{Y}$. Let $|\mathcal{Y}|$ denote the cardinality of $\mathcal{Y}$.

It is assumed that $\boldsymbol{Y}$ conditional on $\boldsymbol{X}$ forms a Conditional Markov Chain, the definition of which is given in the following paragraph. The distribution of $\boldsymbol{X}$ is arbitrary for now. Later on, in order to establish statistical properties of the Conditional Markov Chain, certain assumptions on the mixing behavior of $\boldsymbol{X}$ will be made.

Definition. The distribution of $\boldsymbol{Y}$ conditional on $\boldsymbol{X}$ is parameterized by a vector $\boldsymbol{f}$ of *feature functions* $f : \mathcal{X}\times\mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$, and real-valued *model weights* $\boldsymbol{\lambda}$ (of the same dimension as $\boldsymbol{f}$). Throughout this paper it is assumed that the following regularity condition is satisfied:

(A1) The model weights and the feature functions are finite: $|\boldsymbol{\lambda}| < \infty$, and $|\boldsymbol{f}(x, i, j)| < \infty$ for all $x \in \mathcal{X}$ and $i, j \in \mathcal{Y}$.

A key ingredient in the definition of Conditional Markov Chains is the $|\mathcal{Y}| \times |\mathcal{Y}|$-matrix $\boldsymbol{M}(x)$ with the (i, j)-th component given by

$$m(x, i, j) \quad := \quad \exp(\boldsymbol{\lambda}^T \boldsymbol{f}(x, i, j))$$

for $i, j \in \mathcal{Y}$. In terms of statistical physics, $m(x, i, j)$ is the potential of the joint variable assignment $X_t = x$ and $Y_{t-1} = i$, $Y_t = j$. For sequences $\boldsymbol{x} = (x_t)_{t\in\mathbb{Z}}$ in $\mathcal{X}$ and indices $s, t \in \mathbb{Z}$ with $s \le t$ define the vectors

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

Let $\alpha_s^t(\boldsymbol{x}, i)$ and $\beta_s^t(\boldsymbol{x}, j)$ denote the ith and jth components of $\boldsymbol{\alpha}_s^t(\boldsymbol{x})$ and $\boldsymbol{\beta}_s^t(\boldsymbol{x})$, respectively.

The marginal distributions of $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$ are obtained by conditioning on a finite observational context using a conventional Conditional Random Field (see Sutton and McCallum, 2006), and then letting the size of this context going to infinity. Formally, for any $t \in \mathbb{Z}$, $k \ge 0$ and $y_t, \dots, y_{t+k} \in \mathcal{Y}$, let

$$\begin{aligned}&\mathbb{P}(Y_t = y_t, \dots, Y_{t+k} = y_{t+k} \,|\, \boldsymbol{X} = \boldsymbol{x})\\ &\qquad := \prod_{i=1}^{k} m(x_{t+i}, y_{t+i-1}, y_{t+i})\\ &\qquad \times \lim_{n\to\infty} \frac{\alpha_{t-n}^t(\boldsymbol{x}, y_t)\, \beta_{t+k}^{t+k+n}(\boldsymbol{x}, y_{t+k})}{\boldsymbol{\alpha}_{t-n}^t(\boldsymbol{x})^T \boldsymbol{\beta}_t^{t+k+n}(\boldsymbol{x})}.\end{aligned}$$

It can be shown that, under Assumption (A1), the limit on the right hand side is well-defined (Sinn and Poupart, 2011a). Moreover, Kolmogorov's Extension Theorem asserts that the conditional distribution of $\boldsymbol{Y}$ defined by the collection of all such marginal distributions is unique.

Properties. For any matrix $\boldsymbol{\pi} = (\pi_{ij})$ with strictly positive and finite entries, define the *mixing coefficient*

$$\phi(\boldsymbol{\pi}) \quad := \quad \min_{i,j,k,l} \frac{\pi_{ik}\pi_{jl}}{\pi_{jk}\pi_{il}}. \tag{1}$$

This coefficient plays a key role in the theory of nonnegative matrices (Seneta, 2006), which is of great importance for studying mixing properties of Conditional Markov Chains. Note that $0 < \phi(\boldsymbol{\pi}) \le 1$, and

$$\phi(\boldsymbol{\pi}) \quad \ge \quad \prod_{i,j} \pi_{ij}^{-1}. \tag{2}$$

The following proposition reviews fundamental properties of Conditional Markov Chains. For more background, see Proposition 1 in (Sinn and Chen, 2012).

Proposition 1. *Suppose that Assumption* (A1)*holds true. Then the following statements hold true:*

(i) $\boldsymbol{Y}$ *conditional on* $\boldsymbol{X}$ *forms a Markov chain, i.e.*

$$\begin{aligned}&\mathbb{P}(Y_t = y_t \,|\, Y_{t-1} = y_{t-1}, \dots, Y_{t-k} = y_{t-k}, \boldsymbol{X} = \boldsymbol{x})\\ &\qquad = \mathbb{P}(Y_t = y_t \,|\, Y_{t-1} = y_{t-1}, \boldsymbol{X} = \boldsymbol{x})\end{aligned}$$

for every sequence $\boldsymbol{x} = (x_t)_{t\in\mathbb{Z}}$ *in* $\mathcal{X}$.

(ii) The transition probabilities of the Markov chain, $P_t(\boldsymbol{x}, i, j) := \mathbb{P}(Y_t = j \,|\, Y_{t-1} = i, \boldsymbol{X} = \boldsymbol{x})$ *have the following form:*

$$P_t(\boldsymbol{x}, i, j) \quad = \quad m(x_t, i, j) \lim_{n\to\infty} \frac{\beta_t^n(\boldsymbol{x}, j)}{\beta_{t-1}^n(\boldsymbol{x}, i)}.$$

A lower bound for $P_t(\boldsymbol{x}, i, j)$ *is given by*

$$P_t(\boldsymbol{x}, i, j) \quad \ge \quad \frac{\ell(x_t)\,\ell(x_{t+1})}{|\mathcal{Y}|}$$

where

$$\ell(x) \quad = \quad \frac{\min_{k,l\in\mathcal{Y}} m(x, k, l)}{\max_{k,l\in\mathcal{Y}} m(x, k, l)}.$$

(iii) The transition matrix $\boldsymbol{P}_t(\boldsymbol{x})$ with the (i,j)-th component $P_t(\boldsymbol{x},i,j)$ has the mixing coefficient

$$\phi(\boldsymbol{P}_t(\boldsymbol{x})) \quad = \quad \phi(\boldsymbol{M}(x_t)).$$

Note that, in general, $\boldsymbol{Y}$ conditional on $\boldsymbol{X} = \boldsymbol{x}$ forms a non-stationary Markov chain because the transition matrices $\boldsymbol{P}_t(\boldsymbol{x})$ vary with t.

3 CENTRAL LIMIT THEOREMS

The goal of this paper is to establish Central Limit Theorems for real-valued functionals of Conditional Markov Chains. More precisely, we consider partial sums of the form $\sum_{t=1}^n g(X_t, Y_t)$ where $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and study conditions under which these partial sums (after suitable scaling) converge to a Normal Distribution. A review of the theory of convergence in distribution can be found in (Lehmann, 1999).

Our main result is a Central Limit Theorem for $\sum_{t=1}^n g(X_t, Y_t)$ *conditional* on the sequence of observations $\boldsymbol{X}$, which is stated and proved in Section 3.2. It builds on a result for non-stationary Markov chains due to Dobrushin (1956), which we review in Section 3.1. In Section 3.3 we discuss generalizations of our result, e.g., to vector-valued functions or to functions depending on more than one observation and latent state. Finally, an unconditional version of the Central Limit Theorem is formulated in Section 3.4.

Note that a Central Limit Theorem for Conditional Random Fields has previously been stated by Xiang and Neville (2011), however, the proof of their result is flawed. See the extended version of this paper for more details.

3.1 Dobrushin's Central Limit Theorem

A key tool in our analysis is a Central Limit Theorem for real-valued functionals of non-stationary Markov chains by Dobrushin (1956). We first review his result, closely following the exposition in (Sethuraman and Varadhan, 2004). Let $(\mathcal{Z}, \mathcal{C})$ be a measurable space. For any Markov kernel $\boldsymbol{\pi} = \boldsymbol{\pi}(z, \cdot)$ on $(\mathcal{Z}, \mathcal{C})$, let $\gamma(\boldsymbol{\pi})$ denote the *contraction coefficient*,

$$\gamma(\boldsymbol{\pi}) \quad := \quad \sup_{z_1, z_2 \in \mathcal{Z}, C \in \mathcal{C}} |\boldsymbol{\pi}(z_1, C) - \boldsymbol{\pi}(z_2, C)|. \quad (3)$$

Intuitively, $\gamma(\boldsymbol{\pi})$ measures the maximum distance between conditional distributions $\boldsymbol{\pi}(z_1, \cdot)$, $\boldsymbol{\pi}(z_2, \cdot)$ with $z_1, z_2 \in \mathcal{Z}$. Clearly, $0 \leq \gamma(\boldsymbol{\pi}) \leq 1$, and $\gamma(\boldsymbol{\pi}) = 0$ if and only if the conditional distribution $\boldsymbol{\pi}(z, \cdot)$ does not depend on z. Now let $(\boldsymbol{\pi}_t)$ with $t > 1$ be a sequence of Markov kernels, and $\boldsymbol{\nu}$ a probability measure on $(\mathcal{Z}, \mathcal{C})$. For $n \in \mathbb{N}$ define

$$\gamma_n \quad := \quad \max_{1 < t \leq n} \gamma(\boldsymbol{\pi}_t).$$

Consider a Markov chain $(Z_t)_{t \in \mathbb{N}}$ on $(\mathcal{Z}, \mathcal{C})$ with the initial distribution $\mathbb{P}(Z_1 \in C) = \boldsymbol{\nu}(C)$, and the transition probabilities $\mathbb{P}(Z_t \in C \,|\, Z_{t-1} = z) = \boldsymbol{\pi}_t(z, C)$. Furthermore, let $(g_t)_{t \in \mathbb{N}}$ be a sequence of measurable real-valued functions on $(\mathcal{Z}, \mathcal{C})$, and let S_n denote the partial sum

$$S_n \quad := \quad \sum_{t=1}^n g_t(Z_t). \quad (4)$$

The following theorem due to Dobrushin (1956) establishes conditions under which the standardized partial sums S_n converge to a standard Normal Distribution. Note that the original result applies more generally to triangular sequences of Markov kernels $\boldsymbol{\pi}_t$ and measurable functions g_t. Here we present a simplified version which suffices for our purpose.

Theorem 1. *Suppose there exists a sequence of finite constants $(c_n)_{n \in \mathbb{N}}$ such that*

$$\sup_{1 \leq t \leq n} \sup_{z \in \mathcal{Z}} |g_t(z)| \quad \leq \quad c_n.$$

Furthermore, suppose the following condition holds:

$$\lim_{n \to \infty} c_n^2 (1 - \gamma_n)^{-3} \Big[\sum_{t=1}^n \mathrm{Var}(g_t(Z_t)) \Big]^{-1} \; = \; 0. \quad (5)$$

Then the standardized partial sum S_n in (4) *converges to a standard Normal Distribution:*

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}(S_n)}} \quad \xrightarrow{d} \quad \mathcal{N}(0, 1). \quad (6)$$

The following corollary considers the special case where the functions g_t and variances $\mathrm{Var}(g_t(Z_t))$ are uniformly bounded.

Corollary 1. *Suppose that the functions g_t are bounded so that $c_n \leq c < \infty$ for all $n \in \mathbb{N}$, and the variances* $\mathrm{Var}(g_t(Z_t))$ *are bounded away from zero,* $\mathrm{Var}(g_t(Z_t)) \geq v > 0$ *for all $t \in \mathbb{N}$. Then the convergence statement* (6) *holds provided that*

$$\lim_{n \to \infty} n^{\frac{1}{3}} (1 - \gamma_n) \quad = \quad \infty.$$

3.2 The Conditional Case

Next, we apply Dobrushin's result to establish a Central Limit Theorem for $\sum_{t=1}^n g(X_t, Y_t)$ *conditional* on the observations $\boldsymbol{X}$. For now, let us assume that the function $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ has the following form:

$$g(x, y) \quad = \quad g(x)\, \mathbf{1}(y = i)$$

where $i \in \mathcal{Y}$ and $g : \mathcal{X} \to \mathbb{R}$ are fixed. (In Section 3.3 we will show how to derive the general case.) Let us rewrite the conditional partial sums in order to see how Dobrushin's result comes into play:

$$\sum_{t=1}^{n} g(X_t, Y_t) \,\Big|\, \boldsymbol{X} \quad = \quad \sum_{t=1}^{n} g_t(Y_t) \,\Big|\, \boldsymbol{X}$$

where $g_t(y) := g(X_t)\, \mathbf{1}(y = i)$ for $y \in \mathcal{Y}$. Note that, conditional on $\boldsymbol{X}$, the mappings $g_t : \mathcal{Y} \to \mathbb{R}$ are deterministic. Moreover, according to Proposition 1, $\boldsymbol{Y}$ conditional on $\boldsymbol{X}$ forms a Markov chain. Hence, substituting $(Z_t)_{t\in\mathbb{N}}$ by $(Y_t)_{t\in\mathbb{N}}$ conditional on $\boldsymbol{X}$, we have the same setting as in Dobrushin's theorem.

Assumptions. Let us formulate the assumptions on the observations $\boldsymbol{X}$ that we will use in the proof of our main result. First, let us introduce some measure-theoretic notation: By $\boldsymbol{\mathcal{X}}$ we denote the space of sequences $\boldsymbol{x} = (x_t)_{t\in\mathbb{Z}}$ in $\mathcal{X}$, and by $\boldsymbol{\mathcal{A}}$ the corresponding product σ-field. We write $P_{\boldsymbol{X}}$ for the probability measure on $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{A}})$ defined by $P_{\boldsymbol{X}}(\boldsymbol{A}) := \mathbb{P}(\boldsymbol{X} \in \boldsymbol{A})$ for $\boldsymbol{A} \in \boldsymbol{\mathcal{A}}$, and τ for the shift operator $\tau\boldsymbol{x} := (x_{t+1})_{t\in\mathbb{Z}}$.

(A2) $\boldsymbol{X}$ is ergodic, i.e. $P_{\boldsymbol{X}}(\boldsymbol{A}) = P_{\boldsymbol{X}}(\tau^{-1}\boldsymbol{A})$ for each $\boldsymbol{A} \in \boldsymbol{\mathcal{A}}$, and $P_{\boldsymbol{X}}(\boldsymbol{A}) \in \{0, 1\}$ for every $\boldsymbol{A} \in \boldsymbol{\mathcal{A}}$ which satisfyies $\boldsymbol{A} = \tau^{-1}\boldsymbol{A}$.

(A3) The set $\mathcal{Y}$ has cardinality $|\mathcal{Y}| > 1$, and there exists a measurable set $A \in \mathcal{A}$ with $\mathbb{P}(X_t \in A) > 0$ such that $|g(x)| > 0$ for all $x \in A$.

(A4) Let $F : \mathcal{X} \to \mathbb{R}$ denote the mapping

$$F(x) \quad := \quad \sum_{i,j\in\mathcal{Y}} |\boldsymbol{\lambda}^T \boldsymbol{f}(x, i, j)|$$

where $\boldsymbol{f}$ and $\boldsymbol{\lambda}$ are the feature functions and model weights of the Conditional Markov Chain. Then there exist $p, q > 0$ with $\frac{1}{p} + \frac{1}{q} = 1$ such that

$$\mathbb{E}[|g(X_t)|^{2p}] < \infty \quad \text{and} \quad \mathbb{E}[\exp(F(X_t))^{3q}] < \infty.$$

(A5) Let $A \in \mathcal{A}$ be a fixed measurable set. Define $p := \mathbb{P}(X_t \in A)$, and $S_n(\boldsymbol{x}) := \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}(x_t \in A)$ for $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$. Then there exists a constant $\kappa > 0$ such that, for all $n \in \mathbb{N}$ and $\epsilon > 0$,

$$\mathbb{P}(|S_n(\boldsymbol{X}) - p| \geq \epsilon) \quad \leq \quad \exp(-\kappa\epsilon^2 n).$$

Let us explain the rationale behind (A2)-(A5). The ergodicity assumption (A2) implies laws of large numbers for integrable functionals of $\boldsymbol{X}$ (Cornfeld et al., 1982). As a consequence of the invariance property $P_{\boldsymbol{X}}(\boldsymbol{A}) = P_{\boldsymbol{X}}(\tau^{-1}\boldsymbol{A})$, the sequence $\boldsymbol{X}$ is strictly stationary, and hence the moments and probabilities in (A3)-(A5) do not depend on t. (A3) ensures that the conditional variance of $\sum_{t=1}^{n} g(X_t, Y_t)$ is strictly positive. (A4) relates the tail behavior of g and of the feature functions $\boldsymbol{f}$ with the distribution of X_t. Finally, (A5) ensures that indicator functions of X_t satisfy Hoeffding-type concentration inequalities. Sufficient conditions for Φ-mixing processes and martingales can be found in (Samson, 2000; Kontorovich and Ramanan, 2008).

Main result. Next we state our main result. Possible generalizations, e.g., to vector-valued functions or to functions depending on more than one observation and latent state are discussed in Section 3.3.

Theorem 2. (*i*) *Suppose that Assumption* (A1) - (A4) *are satisfied. Then the following statement holds true* $\mathbb{P}$*-almost surely:*

$$\frac{1}{\sigma_n(\boldsymbol{X})} \sum_{t=1}^{n} \big(g(X_t, Y_t) - \mathbb{E}[g(X_t, Y_t) \,|\, \boldsymbol{X}]\big) \,\Big|\, \boldsymbol{X} \quad \xrightarrow{d} \mathcal{N}(0, 1)$$

with the asymptotic variance $\sigma_n^2(\boldsymbol{X})$ *given by*

$$\sigma_n^2(\boldsymbol{X}) \quad = \quad \mathrm{Var}\Big(\sum_{t=1}^{n} g(X_t, Y_t) \,\Big|\, \boldsymbol{X}\Big).$$

(*ii*) *If additionally* (A5) *holds, then* $\frac{1}{n}\sigma_n^2(\boldsymbol{X})$ *converges* $\mathbb{P}$*-almost surely to a constant* $\sigma^2 < \infty$ *given by*

$$\sigma^2 \quad = \quad \sum_{k=-\infty}^{\infty} \mathbb{E}[\mathrm{Cov}(g(X_t, Y_t), g(X_{t+k}, Y_{t+k}) \,|\, \boldsymbol{X})].$$

We note that statement (*i*) establishes a *conditional* Central Limit Theorem in which the observations $\boldsymbol{X}$ are considered to be fixed. In Section 3.4 we show that an unconditional version can be obtained if the conditional means are asymptotically normally distributed. Statement (*ii*) establishes a standard $\sqrt{n}$ rate of convergence, provided that $\sigma^2 > 0$:

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \big(g(X_t, Y_t) - \mathbb{E}[g(X_t, Y_t) \,|\, \boldsymbol{X}]\big) \,\Big|\, \boldsymbol{X} \quad \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Interestingly, in this case the asymptotic variance σ^2 does not depend on the particular realization of $\boldsymbol{X}$. In Algorithm 1 below we show how σ^2 can be estimated in practice.

Proofs. Before proving Theorem 2, we establish three technical lemmas. The proofs can be found in the extended version of this paper. The first lemma relates the mixing and contraction coefficients $\phi(\cdot)$ and $\gamma(\cdot)$ introduced in (1) and (3). With a slight abuse of notation, if $\boldsymbol{\pi} = (\pi_{ij})_{i,j\in\mathcal{Z}}$ is a stochastic matrix, we write $\gamma(\boldsymbol{\pi})$ for the contraction coefficient of the Markov kernel induced by $\boldsymbol{\pi}(i,C) := \sum_{j\in C}\pi_{ij}$.

Lemma 1. *For any stochastic matrix $\boldsymbol{\pi} = (\pi_{ij})$, we have the inequality $1-\gamma(\boldsymbol{\pi}) \geq \phi(\boldsymbol{\pi})$.*

Lemma 2. *Let z_1, z_2, ... be a real-valued sequence. Furthermore, let $\kappa > 0$ and suppose that the limit of $\frac{1}{n}\sum_{t=1}^n |z_t|^\kappa$ is well-defined and finite. Then the sequence $(m_n)_{n\in\mathbb{N}}$ with $m_n := \max_{1\leq t\leq n}|z_t|$ satisfies*

$$\lim_{n\to\infty}\frac{m_n^\kappa}{n} \quad = \quad 0.$$

Lemma 3. *Suppose that Assumption* (A2)-(A3) *are satisfied. Then the following statement holds true $\mathbb{P}$-almost surely:*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^n \mathrm{Var}(g(X_t,Y_t)\,|\,\boldsymbol{X}) \quad > \quad 0.$$

Proof of Theorem 2. (i) We need to show that condition (5) is satisfied where, in our case:

- $\mathrm{Var}(g_t(Z_t))$ corresponds to $\mathrm{Var}(g(X_t,Y_t)\,|\,\boldsymbol{X})$,
- c_n corresponds to $\max_{1\leq t\leq n}|g(X_t)|$,
- γ_n corresponds to $\max_{1<t\leq n}\gamma(\boldsymbol{P}_t(\boldsymbol{X}))$ with the matrices $\boldsymbol{P}_t(\boldsymbol{X})$ defined in Proposition 1.

According to Lemma 3, the inverse of the sum of variances is of order n^{-1}. Hence it remains to show that

$$\lim_{n\to\infty}\frac{c_n^2(1-\gamma_n)^{-3}}{n} \quad = \quad 0. \tag{7}$$

Using the result from Lemma 1, we obtain

$$(1-\gamma_n)^{-3} \quad \leq \quad \max_{1<t\leq n}\phi(\boldsymbol{M}(X_t))^{-3}.$$

With the lower bound (2) for $\phi(\cdot)$ and the mapping F defined in Assumption (A3), we arrive at

$$\phi(\boldsymbol{M}(X_t))^{-3} \quad \leq \quad \exp(F(X_t))^3.$$

Now let $p, q > 0$ be such that Assumption (A4) is satisfied. Since $\boldsymbol{X}$ is ergodic, the limits of the sequences $\frac{1}{n}\sum_{t=1}^n|g(X_t)|^{2p}$ and $\frac{1}{n}\sum_{t=1}^n\exp(F(X_t))^{3q}$ are well-defined and finite $\mathbb{P}$-almost surely. Hence, according to Lemma 2,

$$\lim_{n\to\infty}\frac{c_n^2}{n^{\frac{1}{p}}} = 0 \quad \text{and} \quad \lim_{n\to\infty}\frac{(1-\gamma_n)^{-3}}{n^{\frac{1}{q}}} = 0.$$

Multiplying the two sequences and using the fact that $\frac{1}{p}+\frac{1}{q}=1$, we arrive at (7).

(ii) We note that (A4) implies $\mathbb{E}[|g(X_t,Y_t)|^2] < \infty$. Hence, as an immediate consequence of Theorem 2 in (Sinn and Chen, 2012), σ^2 exists and is finite. It remains to show that $\frac{1}{n}\sigma_n^2(\boldsymbol{X})$ actually converges to σ^2. The details of the proof can be found in the extended version of this paper. □

3.3 Discussion

Let us discuss special cases, extensions and possible generalizations of the Central Limit Theorem.

Bounded features. If the feature functions of the Conditional Markov Chain are bounded (i.e., there exists a constant $u < \infty$ such that $|\boldsymbol{f}(x,i,j)| < u$ for all $x \in \mathcal{X}$ and $i,j \in \mathcal{Y}$), then a sufficient condition for the results in Theorem 2 (i)-(ii) is that Assumption (A2)-(A3) hold, and $\mathbb{E}[|g(X_t)|^{2+\epsilon}]$ for some $\epsilon > 0$.

Extension to functions of multiple observations and hidden states. It is straight-forward to extend the results to functions g depending on more than one observation and latent state. Similarly, without further technical difficulties it is possible to extend the result to Conditional Markov Chains of order $k > 1$.

Extension to vector-valued functions. To establish a multi-dimensional version of the Central Limit Theorem for vector-valued functions $\boldsymbol{g}$, we use the Cramér-Wold device (see, e.g., Lehmann (1999)). Without loss of generality, we may assume that the components of $\boldsymbol{g}$ have the following form: $g^{(i)}(x,y) = g^{(i)}(x)\mathbf{1}(y=i)$ for $x \in \mathcal{X}$, $y \in \mathcal{Y}$. We need to show that, for every non-zero vector $\boldsymbol{w}$, the partial sums $\sum_{t=1}^n \boldsymbol{w}^T\boldsymbol{g}(X_t,Y_t)$ are asymptotically normally distributed. The crucial part is to establish an equivalent result to Lemma 3. It is not difficult to see that such a result can be obtained if and only if

$$\mathbb{P}(\mathrm{Var}(\boldsymbol{w}^T\boldsymbol{g}(X_t,Y_t)\,|\,\boldsymbol{X}) > 0) \quad > \quad 0.$$

Hence, the individual components of $\boldsymbol{g}$ must satisfy $\mathbb{P}(\mathrm{Var}(g^{(i)}(X_t,Y_t)\,|\,\boldsymbol{X}) > 0) > 0$, and there must not be any linear dependence among them.

3.4 The Unconditional Case

While Theorem 2 establishes asymptotic normality for the partial sums $\sum_{t=1}^n g(X_t,Y_t)$ *conditional* on $\boldsymbol{X}$, it does not make any assertions about the limit distribution in the unconditional case. Statement (ii) shows that, under additional regularity assumptions, the asymptotic variance in the conditional case is constant and hence independent from $\boldsymbol{X}$. The conditional mean, however, is a random variable which in general does depend on $\boldsymbol{X}$. Hence, the unconditional limit distribution can be regarded as an infinite mixture of normal distributions with constant variance and random means. If the means are asymptotically normally distributed, then the mixture itself is converging to a normal distribution. This result is stated in the following theorem.

Theorem 3. *Suppose that Assumption* (A1) - (A5) *hold, and the conditional means* $\mathbb{E}[g(X_t, Y_t) \,|\, \boldsymbol{X}]$ *satisfy a Central Limit Theorem:*

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \big(\mathbb{E}[g(X_t, Y_t) \,|\, \boldsymbol{X}] - \mathbb{E}[g(X_t, Y_t)] \big) \xrightarrow{d} \mathcal{N}(0, \tau^2)$$

with the asymptotic variance τ^2 *given by*

$$\tau^2 = \sum_{k=-\infty}^{\infty} \mathrm{Cov}(\mathbb{E}[g(X_t, Y_t)|\boldsymbol{X}], \, \mathbb{E}[g(X_{t+k}, Y_{t+k})|\boldsymbol{X}])$$

and $0 < \tau^2 < \infty$. *If additionally the asymptotic variance* σ^2 *defined in Theorem 2* (*ii*) *is strictly positive, then we obtain*

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \big(g(X_t, Y_t) - \mathbb{E}[g(X_t, Y_t)] \big) \xrightarrow{d} \mathcal{N}(0, \sigma^2 + \tau^2)$$

where, according to the law of total variance,

$$\sigma^2 + \tau^2 \;\; = \;\; \sum_{k=-\infty}^{\infty} \mathrm{Cov}(g(X_t, Y_t), g(X_{t+k}, Y_{t+k})).$$

We leave it as an open problem to state conditions on $\boldsymbol{X}$ under which the conditional means are asymptotically normally distributed.

4 ESTIMATION OF THE ASYMPTOTIC VARIANCE

In this section we present an algorithm for estimating the asymptotic variance σ^2 in Theorem 2. The key idea is the following: Suppose we are given samples $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ and $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)$ of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, and g is a function such that Assumption (A1)-(A5) are satisfied. In order to estimate σ^2, we *resample* sequences of latent states conditional on $\boldsymbol{X}_n$ using a finite Conditional Markov Chain. See Algorithm 1 for a detailed outline.

Note that, alternatively, the conditional variance of $\sum_{t=1}^{n} g(X_t, Y_t)$ given $\boldsymbol{X}_n$ can be computed using the exact formulas in (Sutton and McCallum, 2006). The following theorem shows that the estimates $\hat{\sigma}^2_{n,M}$ obtained by Algorithm 1 are strongly consistent. The proof can be found in the extended version of this paper.

Algorithm 1 Estimation of the asymptotic variance

```
Input: Realizations $\boldsymbol{x}_n = (x_1, \ldots, x_n)$ and $\boldsymbol{y}_n = (y_1, \ldots, y_n)$ of the samples $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ and $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)$; feature functions $\boldsymbol{f}$; model weights $\boldsymbol{\lambda}$; real-valued function $g$; number of Monte Carlo replications $M$.
for $m = 1, \ldots, M$ do
    Simulate $\boldsymbol{Y}_n^{(m)} = (Y_1^{(m)}, \ldots, Y_n^{(m)})$ conditional on $\boldsymbol{X}_n = \boldsymbol{x}_n$ using a finite Conditional Markov Chain with feature functions $\boldsymbol{f}$ and model weights $\boldsymbol{\lambda}$ (Sutton and McCallum, 2006).
    Compute the sample statistic
        $g_m \;=\; \sum_{t=1}^{n} g(x_t, Y_t^{(m)})$
end for
Compute $\bar{g}_M = \frac{1}{M} \sum_{m=1}^{M} g_m$ and
    $\hat{\sigma}^2_{n,M} \;=\; \frac{1}{n(M-1)} \sum_{m=1}^{M} (g_m - \bar{g}_M)^2.$
Output: Estimate $\hat{\sigma}^2_{n,M}$ of the asymptotic variance $\sigma^2$ in Theorem 3.
```

Theorem 4. *Suppose that Assumption* (A1)-(A5) *hold. Then the following result holds* $\mathbb{P}$*-almost surely:*

$$\lim_{n \to \infty} \lim_{M \to \infty} \hat{\sigma}^2_{n,M} \;\; = \;\; \sigma^2.$$

In most practical situations, the feature functions $\boldsymbol{f}$ and model weights $\boldsymbol{\lambda}$ in Algorithm 1 are unknown and need to be replaced by empirical estimates. In our experiments in Section 5, we use the true feature functions, and estimate the model weights by maximizing the conditional likelihood of $\boldsymbol{Y}_n$ given $\boldsymbol{X}_n$ (Sha and Pereira, 2003). The results suggest that the estimates $\hat{\sigma}^2_{n,M}$ are still consistent in this case. A rigorous proof is left as an open problem for future research.

5 Experiments

In this section, we illustrate the theoretical findings from the previous sections using synthetically generated environmental data. In particular, we demonstrate how to construct hypothesis tests for Conditional Markov Chains. Important applications are, e.g., testing for model misspeficiations, signficance of feature functions, or dependencies between observations and latent states.

Data-Generating Model. In all our experiments, $\boldsymbol{X}$ is an autoregressive process, $X_t = \phi X_{t-1} + \epsilon_t$, with the autoregressive coefficient $\phi = \frac{1}{2}$ and independent

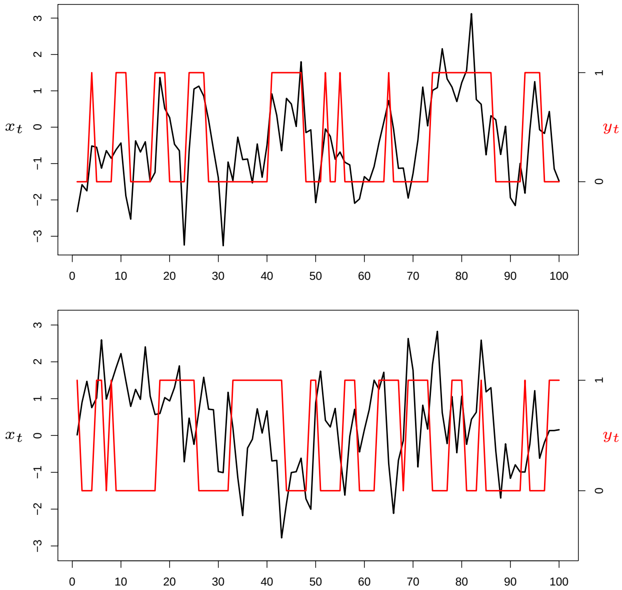

Figure 1: Realizations of a Conditional Markov Chain with model weights $\boldsymbol{\lambda} = (1,1)^T$ (upper plot) and $\boldsymbol{\lambda} = (0,1)^T$ (lower plot). In the first case, there is a positive correlation between the observations X_t and the events $Y_t = 1$; in the second case the observations and latent states are mutually independent.

standard normal innovations $(\epsilon_t)_{t \in \mathbb{Z}}$. The process of latent states $\boldsymbol{Y}$ has two different labels, $\mathcal{Y} = \{0, 1\}$. The distribution of $\boldsymbol{Y}$ conditional on $\boldsymbol{X}$ is induced by the feature functions $\boldsymbol{f} = (f_1, f_2)^T$ given by

$$\begin{aligned} f_1(x_t, y_{t-1}, y_t) &= x_t \mathbf{1}(y_t = 1), \\ f_2(x_t, y_{t-1}, y_t) &= \mathbf{1}(y_{t-1} = y_t), \end{aligned}$$

and the model weights $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$. In order to simulate samples $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ from an infinite Conditional Markov Chain, we generate sequences of observations and latent states of length $n + 2m$ using a finite Conditional Markov Chains, and then discard the first and the last m values. In our experiments, we found that discarding $m = 50$ values is sufficient; for more information on the effect of these "burn-in" periods, see (Sinn and Poupart, 2011b).

Figure 1 shows two examples: In the upper graph, we choose the model weights $\boldsymbol{\lambda} = (1,1)^T$. As can be seen, there is a positive correlation between the observations X_t and the events $Y_t = 1$. Furthermore, the latent states tend to persist, i.e., there is a high probability that $Y_t = Y_{t-1}$. In the lower graph, the model weights are $\boldsymbol{\lambda} = (0,1)^T$. In this case, the probability of the event $Y_t = 1$ is $\frac{1}{2}$ independently from the observation at time t, and the probability that $Y_t = Y_{t-1}$ equals $e/(1+e) \approx 73.1\%$.

Time series as those in Figure 1 are commonly encountered in environmental studies. For example, think of $\boldsymbol{X}$ as (normalized) daily average temperatures, and $\boldsymbol{Y}$ as indicators of extreme ozone level concentrations. More generally, multivariate observations can be taken into account, e.g., comprising CO_2 emissions or cloud cover data. An important question in such studies is whether particular covariates actually have an effect on the latent states.

Asymptotic normality. First, let us consider the asymptotic distributions both in the conditional and unconditional case. In this experiment we choose the model weights $\boldsymbol{\lambda} = (1,1)^T$, and consider the partial sums $\sum_{t=1}^n g(X_t, Y_t)$ for $g(x,y) = x\,\mathbf{1}(y = 1)$ with the sample size $n = 1000$.

Figure 2 (a)-(c) show examples of conditional distributions, each obtained for a fixed sequence of observations by repeatedly simulating sequences of latent states. Here and in all the following experiments, we use 1000 Monte Carlo replications. The red lines display the fit of normal distributions. As can be seen, the distributions all look approximately normal. The standard deviations are very similar, however, there is considerable variation in the conditional mean. Figure 2 (d) shows that the distribution of the conditional mean itself is approximately normal. Finally, the unconditional distribution of the partial sums is displayed in Figure 2 (e). In accordance with Theorem 3, this distribution can be regarded as the infinite mixture of the conditional distributions which again is approximately normal.

Estimation of the asymptotic variance. Next, we apply Algorithm 1 for estimating the asymptotic variance σ^2 in Theorem 2. By simulations, we obtain the true value $\sigma^2 \approx 0.128$. Table 1 shows the mean and the standard deviation of the estimates $\hat{\sigma}^2_{n,M}$ for $M = 10{,}000$. We report the results obtained using the true model weights $\boldsymbol{\lambda}$, and the maximum likelihood estimates. As can be seen, the estimates only have a small bias, and the standard deviation decreases as the sample size increases. Interestingly, the results for the true and the estimated model weights are practically identical, which supports our conjecture that Theorem 4 still holds if $\boldsymbol{\lambda}$ is replaced by strongly consistent estimates.

	True $\boldsymbol{\lambda}$		Estimated $\boldsymbol{\lambda}$	
n	100	1000	100	1000
Mean	0.129	0.128	0.129	0.128
Std.dev.	0.053	0.016	0.053	0.016

Table 1: Mean and standard deviation of the estimates $\hat{\sigma}^2_{n,M}$ for different sample sizes n and $M = 10{,}000$.

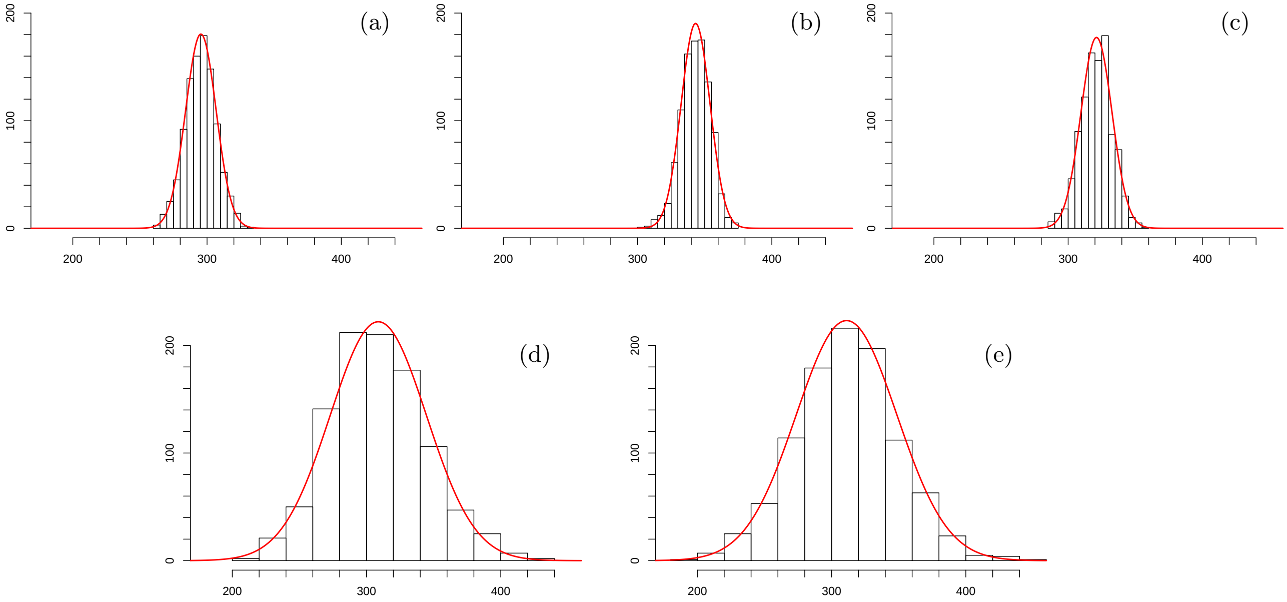

Figure 2: Illustration of the Central Limit Theorem for Conditional Markov Chains. (a)-(c): Conditional distributions, each obtained for a fixed sequence of observations. The red lines display the fit of normal distributions. (d): Distribution of the conditional means. (e) Unconditional distribution.

Hypothesis testing. Finally, let us show how the previous methodology can be used for hypothesis testing. Consider the null hypothesis $H_0 : \boldsymbol{\lambda} = (0,1)^T$. In order to construct a test at significance level $\alpha \in (0,1)$, we follow these steps: (S1) Compute the test statistic $T_n = \sum_{t=1}^{n} g(X_t, Y_t)$. (S2) Use Algorithm 1 to compute the estimate $\hat{\sigma}^2_{n,M}$ of σ^2. Analogously, compute an estimate $\hat{\mu}_{n,M}$ of the conditional mean. Use the model weights $\boldsymbol{\lambda} = (0,1)^T$ under H_0 in these computations. (S3) Compute the standardized test statistic

$$Z_n \quad = \quad \frac{1}{\sqrt{n}\hat{\sigma}_{n,M}} \big(T_n - \hat{\mu}_{n,M}\big).$$

(S4) Reject H_0 if $|Z_n|$ is larger than the $100(1-\frac{\alpha}{2})\%$ quantile of the standard normal distribution.

Table 2 shows the probability for rejecting H_0 when H_0 is true ($\boldsymbol{\lambda} = (0,1)^T$), and in a case ($\boldsymbol{\lambda} = (1,1)^T$) where H_0 is wrong. As can be seen, the test preserves the significance level, and achieves excellent power as the sample size increases.

	$\boldsymbol{\lambda} = (0,1)^T$	$\boldsymbol{\lambda} = (1,1)^T$
$n = 100$	0.048	0.205
$n = 1000$	0.045	0.967

Table 2: Probability for rejecting H_0 at signficance level $\alpha = 0.05$.

6 Conclusions

The present paper provides a rigorous treatment of Central Limit Theorems for real-valued functionals of Conditional Markov Chains. The main result is Theorem 2, which establishes asymptotic normality conditional on the sequence of observations. An unconditional version is obtained if the conditional means themselves are asymptotically normally distributed. Algorithm 1 shows how the asymptotic variance can be estimated by resampling the latent states conditional on a given sequence of observations. The experiments in Section 5 illustrate the Central Limit Theorems both in the conditional and unconditional case, and show the accuracy of the algorithm for the variance estimation. Moreover, it is demonstrated how the methodology can be used for hypothesis testing.

The paper opens interesting questions for future research. For example, it remains an open question when the conditional means (or higher conditional moments) are asymptotically normally distributed, which is a prerequisite for the unconditional Central Limit Theorem. Another problem is to prove that Algorithm 1 is still consistent if the true model weights are replaced by consistent estimates. One approach would be to bound the derivatives of conditional marginal distributions with respect to the model weights, which is an important problem in its own right.

References

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

R. Dobrushin (1956). Central limit theorems for nonstationary Markov chains I, II. *Theory of Probability and its Applications* 1, 65-80, 329-383.

L. Kontorovich, and K. Ramanan (2008). Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, Vol 36, No. 6, 2126-2158.

J. Lafferty, A. McCallum, and F. Pereira (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the 18th IEEE International Conference on Machine Learning (ICML).*

E.L. Lehmann (1999). *Elements of Large-Sample Theory.* New York, NY: Springer.

P.-M. Samson (2000). Concentration of measure inequalities for Markov chains and Φ-mixing processes. *The Annals of Probability*, Vol 28, No. 1, 416-461.

E. Seneta (2006). *Non-Negative Matrices and Markov Chains. Revised Edition.* New York, NY: Springer.

S. Sethuraman, and S.R.S. Varadhan (2005). A martingale proof of Dobrushin's theorem for nonhomogeneous Markov chains. *Electronic Journal of Probability*, Vol. 10, Article 36, 1221-1235.

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

M. Sinn, and B. Chen (2012). Mixing properties of conditional Markov chains with unbounded feature functions. In *Advances in Neural Information Processing Systems (NIPS).*

M. Sinn, and P. Poupart (2011a). Asymptotic theory for linear-chain conditional random fields. In *Proc. of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS).*

M. Sinn, and P. Poupart (2011b). Error bounds for online predictions of linear-chain conditional random fields. Application to activity recognition for users of rolling walkers. In *Proc. of the 10th International Conference on Machine Learning and Applications (ICMLA).*

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.

R. Xiang, and J. Neville (2011). Relational learning with one network: an asymptotic analysis. In *Proc. of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS).*